

基于粗神经网络的语音情感识别

曾光菊^{1,2}

(1. 电子科技大学电子工程学院, 成都 610054; 2. 四川理工学院理学院, 四川 自贡 643000)

摘要:语音情感识别是从语音信号中提取一些有效的声学特征,然后利用智能计算或者识别的方法对话者的情感状态进行识别。介绍了国内外在该领域中关于语音情感数据库、特征提取、识别方法的研究现状。基于对该领域现状的了解,发现特征提取对识别率有着非常大的影响。录制了1050句语音,每句语音提取了30个特征,从而形成了一个1050×30的数据库。提出了用粗糙集理论中的信息一致性对数据库中的30个特征进行化简,最后得到了12个特征。用神经网络中的BP网络对话者的情感状态进行识别,最高识别率达到了84%。从实验结果发现不同的情感用不同的方法识别结果更好。

关键词:语音情感识别;情感分类;特征提取;粗糙集;BP网络

中图分类号:TP391

文献标识码:A

引言

通过人类的语音信号,我们既能得到其中包含的语义信息同时也能感知语义以外的情感信息。情感信息是语音信号的重要组成部分。斯坦福大学的Reeves和Nass的研究发现表明,在人机交互中需要解决的问题实际上与人和人交流中的重要因素是一致的,最关键的都是“情感智能”的能力。为了让计算机能获得像人一样的“情感智能”,首先计算机就得像人一样能识别他人的情感状态。语音情感识别是从语音信号中提取部分特征让计算机自动地识别出话者的情感信息(情感状态、情感类别等)。

目前大量的研究^[1-5]都是从语音信号中提取出一些声学特征,然后在将特征化简提取出一些更有效的特征,再用模式识别中的一些方法尝试去识别语音信号中的情感。正确识别话者的情感,目前的研究主要是将语音情感分成大众较为熟悉的六种情感状态:高兴、愤怒、惊讶、害怕、悲伤、平静。

1 语音情感识别的研究现状

1.1 语音情感分类

语音情感分类在心理学或者工程研究中都没有一

个统一的分类标准。下面是最近研究者做语音情感识别中所用提及的语音情感分类。

1996年,Dellaert提出以基音频率相关信息为主要特征的分类方法,并考虑了恐惧、愤怒、悲伤和高兴四种情感状态。

1999年,Joy Nicholson采用ANN作为分类器,考虑了八种情感达到识别率50%。

2000年,ASSESS提出的系统能识别恐惧、愤怒、悲伤和高兴四种情感类别。

2001年,Albino Nogueiras等考虑了MPEG-4标准中的六种情感,用HMM模型进行识别,达到了80%的识别率。

2003年,Oh-wook Kwon等采用SVM和HMM对高兴、愤怒、悲伤、平静进行分

类,达到了分类精度为70.1%。

国内的许多学者^[1]将情感分成欢快、愤怒、恐惧、悲伤这种四情感。这样的情感分类好处是增大了情感的粒度,容易区分辨别,能有效提高实验的准确度,实验表明单纯通过语音数据分析,情感识别率能达到80%以上。当然这四种情感模型中,某些情感就不能很好的归结到上述四种情感中。另有一些学者提出了连续情感

收稿日期:2011-06-25

基金项目:四川理工院校内项目(2009XJKYL005)

作者简介:曾光菊(1970-),女,四川富顺人,讲师,研究生,主要从事信号与信息处理、语音情感识别等方面的研究。

维问题^[2]。即将情感用三维空间来描述或表示,它的三个维度分别命为:唤醒度、愉悦度、控制度。用这三个基本维度建立一个情感空间,该空间是连续的。但目前只有“激励”(镇静/激动)和“诱力”(负面/正面)所建立的二维情感空间得到了广泛的认可。

由以上的研究可看出,尽管有许多研究者都在做关于情感识别的工作,但是在情感分类的工作上基本没有一个统一的分类标准。这样大家的研究结果之间没有可比性,这也增加了语音情感识别研究的难度。心理学和认知学理论家已经为情感状态讨论了小数量的一组类别。在这两个领域里研究者给“基本情感”下了多种定义。可能 Paul Ekman 给出的定义最详尽^[3]。他将基本情感,和对应有明显的通用面部表情的情感,以及其他 8 种属性连续起来。用这个标准,Ekman 提出了 6 种基本情感:恐惧、愤怒、悲伤、高兴、厌恶和惊奇。这 6 种基本情感在心理学和精神病学领域得到了广泛的认同和应用。本文也借鉴了 6 种基本情感作为情感识别研究的起点。

1.2 语音情感数据库

语音情感数据库同语音情感分类一样在目前的研究中没有一个公开的数据库平台是资源共享的。大部分的研究者若要开展有关语音情感识别的相关研究,都得自组织团队录制语音,然后从语音信号中通过一些软件(主要是 cool edit 和 praat)得到一些语音声学特征的数据。得到的这些数据可以制作成数据库,也可以做好标记后直接保存在 excel 文件中。但是也有几个数据库是比较有代表性的见表 1。

表 1 代表性的数据库

来源国家	语种	样本数	语音类型
英国 Queen's 大学	英语	200	剪辑视频文件
日本 Meikai 大学	俄语	3660	朗读式
中科院自动化研究所	汉语	9600	朗读式
台湾大同大学	汉语	839	讲述式

1.3 语音情感特征提取

目前关于语音情感识别的研究几乎都是从语音信号中提取相关的声学特征,然后再进行识别。语音情感识别的研究过程如图 1 所示:

从上图中可以了解到是否有效的特征提取将直接影响到后面的识别结果。大量的研究显示^[3,4,6]韵律特征和音质特征与语音中的情感识别有很大的相关性。韵律特征中用得较为广泛的是:音强、音长、音高、重音、声调和语调等以及这些特征的衍生参数。音质特征主

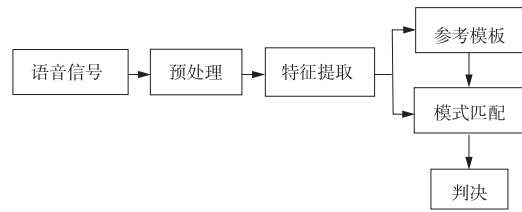


图 1 语音情感识别系统框图

要包括音色和语谱方面的特征。它们反映发音时声门波形状的变化,其影响因素有肌肉张力、声道中央压力以及声道长度张力。主要的音质类特征有:振幅、共振峰、以及 Mel 频率倒谱系数特征、LPC 系数、基音、能量、频率等及其相关的衍生参数。有研究证实^[3,7]仅用韵律特征或者音质特征中的某一类进行语音情感识别效果甚微。若是将二者结合起来效果更佳。这二者并不是相互孤立的,相互结合才能表达情感。

1.4 语音情感识别方法

语音情感识别是上世纪末兴起的一个新的研究领域。在语音情感识别诞生之前,在模式识别领域里已经有很多关于生物特征识别的研究如:语音识别、指纹识别、车牌识别等等这些都是已经研究成功并已经或者是正在走向市场化了。有很多研究者利用模式识别领域里上述的生物特征识别方法对语音情感进行识别或者是将这些算法进行一些改进。主要的一些方法有:基于人工神经网络(ANN)的识别方法、基于支持向量机(SVM)的识别方法、基于高斯混合模型(GMM)的识别方法、基于主成分分析(PCA)的识别方法、基于隐马尔可夫(HMM)的识别方法、线性判别分类器(LDC)等等。这些识别方法都是从语音信号中提取一些特征然后设计一个识别模型进行模式匹配再输出决策最后得到识别的结果。

2 基于粗神经网络的语音情感识别

本研究按照文献^[7]中介绍的录制语音数据库的准则自制建立了一个包含 1050 个样本的数据库。该数据库的录音样本来自中国科学院 2005 年建立的汉语情感语料库^[8]。本研究邀请了 5 名 20 岁左右的在校大学生来参加录音(2 女 3 男)。部分录音样本见表 3。

表 2 部分录音样本语句

语句	语句	语句
就是下雨也去	孙英开飞机	四万四块四
十头峡河牛	小宝逮老鼠	银行拥有保安
他们支持中国	国华来完成	大岳去种菜
我马上拿来	台球很有意思	苏联代表世界

表3 部分录音样本语句

情感状态	语句
高兴	今天天气真好啊
愤怒	你给我滚出去
惊讶	这真的是你吗? 一点都不像啊
难过	为什么这次我又不及格, 我不想活了
害怕	不要杀我啊

表2中所示的录音样本其语义均为中性,用高兴、愤怒、悲伤、害怕、惊讶、平静六种情感都可以表达。在录音时,这5名人员分别以上述的六种情感朗读一次。表3中的语句其语义自身与情感相关,在本次录音中对于这部分语句则是用5名录音人员都以对应的情感朗读了一次。本次录音共获得了1750句录音,但是经过专家听取录音后,去掉了其中600句情感状态模糊不清的录音。最终得到了1050句情感状态较为清晰的语音。用cool edit2.0录音,然后再用cool edit2.0和praat2.0这两种语音处理软件从每句语音信号中通过分帧(每帧语音信号取10ms)、加窗(hamming)等预处理过程后得到了每句语音信号的声学特征数据。在本研究中的原始声学特征包含韵律类和音质类共30个。本文所涉及的数据库为 1050×30 构成。该数据库中每条语音信号对应的声学特征及其数据如表4所示。

2.1 粗糙集下的情感语音声学特征化简

在上述数据库的基础上,建立一个用数据表表达的语音情感识别知识系统 S ,令语音情感识别论域 U 中研究的对象数目为 m ,语音情感识别条件特征属性 $C = \{c_1, c_2, \dots, c_n\}$,令 x 是论域 U 中研究的对象,则 x 可以表达为:

$x_i = \{c_{i1}, c_{i2}, \dots, c_{in}\} \quad i = 1, 2, \dots, m$. $c_{i1}, c_{i2}, \dots, c_{in}$ 是 x 的 n 个特征属性的属性值。

根据粗糙集理论^[11],基于条件特征属性 C 不可分辨关系的等价族表示为 $U|C$,基于决策特征属性 D 不可分辨关系的等价族表示为 $U|D$ 。

定义1 设 $S = (U, A, V, f)$ 是一个语音情感识别系统, $A = C \cup D$ 是属性集, C 和 D 分别是条件属性集和决策属性,基于条件属性 C 描述决策属性 D 表达的信息一致性表达为:

$$Q_C = \frac{\text{card}(U|C)}{\text{card}(U)} \quad (1)$$

这里 $\text{card}(U|C)$ 表示 U 基于条件特征属性 C 不可分辨关系的等价族的数目, $\text{card}(U)$ 表示有限论域 U 中可分辨的集合的基数。

定理1^[12] 设 $S = (U, A, V, f)$ 是一个语音情感识别系统, $A = C \cup D$ 是属性集, C 和 D 分别是条件属性集

表4 声学特征及其数据

序号	特征	数据
1	最小采样值(smpl)	-2156
2	最大采样值(smpl)	2013
3	振幅峰值(dB)	-23.18
4	直流偏移(%)	-0.353
5	RMS激励最小值(dB)	-69.18
6	RMS激励最大值(dB)	-30.83
7	RMS激励平均化r(dB)	-41.96
8	RMS激励总计(dB)	-39.14
9	最高分贝(dB)	-50.68
10	最高分贝对应的频率(Hz)	15.62
11	最低分贝(dB)	-111
12	最低分贝对应的频率(Hz)	7921
13	分析对象起点时间(s)	0
14	分析对象末点时间(s)	2.016
15	分析对象时长(s)	2.016
16	分析对象基频点个数(frames)	126
17	分析对象基频(Hz)	163.73
18	最高强度(dB)	61.35
19	最高强度对应的时间(s)	0.88
20	最低强度(dB)	19.83
21	最低强度对应的时间(s)	0.416
22	共振峰时间(s)	1.008
23	第一共振峰频率(Hz)	505.534
24	第一共振峰带宽(Hz)	258.151
25	第二共振峰频率(Hz)	1451.589
26	第二共振峰带宽(Hz)	329.891
27	第三共振峰频率(Hz)	2963.094
28	第三共振峰带宽(Hz)	115.087
29	第四共振峰频率(Hz)	3508.097
30	第四共振峰带宽(Hz)	212.051

和决策属性,如果存在:

$$Q_C = 1 \quad (2)$$

这里 Q_C 表示基于条件特征属性 C 描述决策属性 D 表达的信息一致性,该决策系统是一个信息一致的协调系统,否则是一个不协调系统。如果存在:

$$Q_C = Q_{C-R_i} = 1 \quad (3)$$

称 $R_i, R_i \in C$ 为 C 中相对于 D 可省略的,否则 R_i 为 C 中相对于 D 不可省略的。

这里 Q_C 表示基于条件特征属性 C 描述决策属性 D 表达的信息一致性, Q_{C-R_i} 表示基于 R_i 为 C 中相对于 D 可省略后的条件特征属性描述决策属性 D 表达的信息一致性。

语音情感识别系统化简实现:

输入:给定一个语音情感识别 $S = (U, A, V, f)$, $A = C \cup D$ 是属性集, C 和 D 分别是条件属性集和决策属性。

输出:简化的特征集

(1) 通过定义的欧氏距离聚类方法进行属性的属

性值的离散归一化处理,构成一个协调的数据表。在聚类产生一种新的划分时,直到协调度大于或等于原始数据的协调度就聚类产生一种新的划分。定义协调度:

$$L_d = \sum \text{card}(C_-(X)) / \text{card}(U) \quad (4)$$

这里 $\text{card}(\cdot)$ 表示括号内集合元素的个数, $C_-(X)$ 表示集合 X 的下近似。

(2) 基于信息一致性知识约简的方法,利用式(1)进行条件属性的约简。

(3) 对可约简的条件属性中属性值的集合的基数大的条件属性进行删除,保留不可约简的条件属性和可约简的条件属性中属性值的集合的基数小的属性组合,利用式(1、2、3 及 4),得到一个能提供最大数据信息覆盖率的语音情感识别最简条件属性集见表 5。

表 5 化简后的声学特征

特征	特征
振幅峰值	最高分贝
RMS 激励最小值	最高分贝对应的频率
RMS 激励最大值	最低分贝
RMS 激励平均化	最低分贝对应的频率
最高强度	最低强度
第一共振峰带宽	第二共振峰带宽

2.2 BP 神经网络下的语音情感识别

BP 网络是一种具有三层或三层以上的多层神经网络,每一层都由若干个神经元组成,如图 2 所示,它的左、右各层之间各个神经元实现全连接,即左层的每个神经元与右层的每个神经元都有连接。

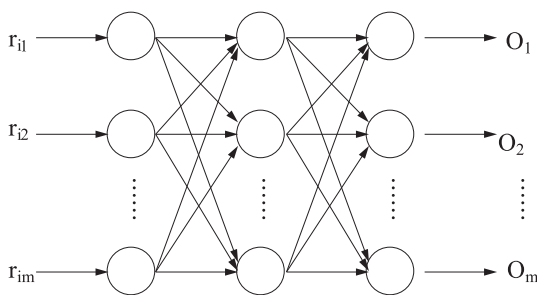


图 2 多层 BP 神经网络结构

本研究的输入层是表 5 中用粗糙集化简后的 12 个声学特征。输入层有 12 个节点,对应这 12 个声学特征,隐含层为 12 个节点,输出层为 6 个节点,分别对应高兴、愤怒、悲伤、害怕、惊讶、平静这六种情感。神经网络中各层间的激活函数选用的是双曲正切 S 型函数,一反向传播算法进行训练,学习速率为 0.001,训练误差为 0.000 1,最大迭代步数为 500。

2.3 实验结果

数据库中的 1/2 数据用以训练网络,1/2 数据用以测试网络。本研究中测试了 525 个样本,用 matlab 中的神经网络数据包进行仿真得到识别结果见表 6。

表 6 识别结果

	高兴	愤怒	悲伤	害怕	惊讶	平静
样本数	86	100	80	90	89	80
识别数	50	54	46	66	45	67
识别率	58%	54%	58%	73%	51%	84%

从识别的结果可以看出,本次试验的识别率只有害怕和安静这两类情感的识别率较高,其余的情感识别率都较低,而且在试验中发现高兴和愤怒不容易区分,容易误判成对方。导致识别率偏低的原因如下:

(1) 录音人员表达情感的强度不够,由于环境、心境等因素的影响朗读人员在表达高兴和惊讶的情感时没有区分开。

(2) 本次录音有部分语音中的噪声较大,在语音信号的预处理过程中去噪的效果偏低导致后面特征的数据由误差。

(3) 处理语音信号中声学特征的数据时,尽管都是用软件处理,但是其中也有手动的过程,这样难免有一些误差在里面,直接导致了数据的人为误差性。

(4) 设计 BP 网络时用的 S 型正切函数,但是其中的算法设计上还是存在一些问题。

在以后的研究中,可以剪辑影视文件中的语音信号,因为专业演员对情感的表达更具有真实性。通过本研究发现 BP 网络比较容易识别害怕和安静。在以后的研究尝试对不同的情感用不同的识别方法。

参考文献:

- [1] 赵力,将春辉.语音信号中的情感特征分析和识别的研究[J].电子学报,2004,30(3):423-429.
- [2] Yuan J,Shen L,Chen F.The Acoustic Realization of Anger,Fear,Joy and Sadness in Chinese[A].in International Conference on Spoken Language Processing 2002,Denver,Colorado,USA.
- [3] 余华,黄程韦,赵力.基于粒子群优化神经网络的语音情感识别[J].数据采集与处理,2011,26(1):57-62.
- [4] 石瑛,胡学钢.基于神经网络的语音情感识别[J].计算机工程与应用,2008,44(24):191-193.
- [5] Kleinginna P R,Kleinginna A M.A categorized list of emotion definitions with suggestions for a consensual definition[J].Motivation and Emotion,1981:345-379.

- [6] Jianxia C. A summary about emotional speech recognition[C]//1st Chinese Conference on Affective Computing and Intelligent Interaction. Beijing: [s. n.], 2003.
- [7] 谢波, 陈岭, 陈根才, 等. 普通话语音情感识别的特征选择技术[J]. 浙江大学学报, 2007, 41(11): 1816-1822.
- [8] 汉语情感语料库 <http://www.chineseldc.org>.
- [9] 曾黄麟. 智能计算[M]. 重庆: 重庆大学出版社, 2004.
- [10] Zeng Huanglin, Lan Hengyou, Zeng Xiaohui. Redundant Data Processing Based on Rough-Fuzzy Approach, Rough Sets and Knowledge Technology. RSKT, Chongqing, China, July 24-26, 2006[C]. 156-16.
- [11] Ekman P. An argument for basic emotions[J]. Cognition and Emotion, 1992, 6(3/4): 169-200.
- [12] Ekman. Are there basic emotions? [J]. Psychological Review, 1992, 99(3): 550-553.

Speech Emotion Recognition Based on Rough Set and ANN

ZENG Guang-ju^{1,2}

- (1. School of Electronic Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China;
2. School of Science, Sichuan University of Science & Engineering, Zigong 643000, China)

Abstract: Speech emotion recognition is about extracting effect acoustic features from speech signals and recognizing emotion state of human by using of intelligent computation. The domestic related research of emotion speech database, features extraction and recognition ways are studied. Learning from these related researches, the features extraction was found to have important affections on the speech emotion recognition. 1050 sentences was recorded and 30 features extracted from every sentence and then formed to a database of 1050×30 . The information consistence of rough set is applied to simplify 30 features of database to 12 features. Then artificial neural network is used to recognize emotion state of 525 sentences, it attains to the highest recognition rate of 84%. The results shows that using different ways to recognize different emotion has better effects.

Key words: speech emotion recognition; emotion classification; features extraction; rough set; BP network