软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2014,25(1):37−50 [doi: 10.13328/j.cnki.jos.004497]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

语音情感识别研究进展综述*

韩文静[1], 李海峰[1], 阮华斌[2], 马 琳[1]

[1](哈尔滨工业大学 计算机科学与技术学院,黑龙江 哈尔滨 150001)

[2](清华大学 计算机科学与技术系,北京 100084)

通讯作者: 韩文静, E-mail: hanwenjing07@gmail.com

摘 要: 对语音情感识别的研究现状和进展进行了归纳和总结,对未来语音情感识别技术发展趋势进行了展望.从5个角度逐步展开进行归纳总结,即情感描述模型、具有代表性的情感语音库、语音情感特征提取、语音情感识别算法研究和语音情感识别技术应用,旨在尽可能全面地对语音情感识别技术进行细致的介绍与分析,为相关研究人员提供有价值的学术参考;最后,立足于研究现状的分析与把握,对当前语音情感识别领域所面临的挑战与发展趋势进行了展望.侧重于对语音情感识别研究的主流方法和前沿进展进行概括、比较和分析.

关键词: 人机交互;情感计算;情感描述模型;情感语音库;情感声学特征;语音情感识别

中图法分类号: TP391 **文献标识码**: A

中文引用格式: 韩文静,李海峰,阮华斌,马琳.语音情感识别研究进展综述.软件学报,2014,25(1):37−50. http://www.jos.org.cn/1000-9825/4497.htm

英文引用格式: Han WJ, Li HF, Ruan HB, Ma L. Review on speech emotion recognition. Ruan Jian Xue Bao/Journal of Software, 2014,25(1):37−50 (in Chinese). http://www.jos.org.cn/1000-9825/4497.htm

Review on Speech Emotion Recognition

HAN Wen-Jing[1], LI Hai-Feng[1], RUAN Hua-Bin[2], MA Lin[1]

[1](School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

[2](Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

Corresponding author: HAN Wen-Jing, E-mail: hanwenjing07@gmail.com

Abstract: This paper surveys the state of the art of speech emotion recognition (SER), and presents an outlook on the trend of future SER technology. First, the survey summarizes and analyzes SER in detail from five perspectives, including emotion representation models, representative emotional speech corpora, emotion-related acoustic features extraction, SER methods and applications. Then, based on the survey, the challenges faced by current SER research are concluded. This paper aims to take a deep insight into the mainstream methods and recent progress in this field, and presents detailed comparison and analysis between these methods.

Key words: human-computer interaction; affective computing; emotion representation model; emotional speech corpora; emotion-related acoustic feature; speech emotion recognition

人类之所以能够通过聆听语音捕捉对方情感状态的变化,是因为人脑具备了感知和理解语音信号中的能够反映说话人情感状态的信息(如特殊的语气词、语调的变化等)的能力.自动语音情感识别则是计算机对人类上述情感感知和理解过程的模拟,它的任务就是从采集到的语音信号中提取表达情感的声学特征,并找出这些声学特征与人类情感的映射关系.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,是实现自然

* 基金项目: 国家自然科学基金(61171186, 61271345); 语言语音教育部微软重点实验室开放基金(HIT.KLOF.2011XXX); 中央高校基本科研业务费专项资金(HIT.NSRIF.2012047)

收稿时间: 2013-05-08; 定稿时间: 2013-09-02; jos 在线出版时间: 2013-11-01

CNKI 网络优先出版: 2013-11-01 13:49, http://www.cnki.net/kcms/detail/11.2560.TP.20131101.1349.001.html

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.

语音情感识别研究的开展距今已有 30 余年的历史,在此期间,它得到了世界范围内相关研究者们的广泛关注,也取得了一些令人瞩目的成绩,但同时也面临着诸多问题的考验与挑战.本文将立足于语音情感识别研究领域的已有成果,对领域内的研究进展进行总结,并对未来的技术发展趋势加以展望.

一般说来,语音情感识别系统主要由 3 部分组成:语音信号采集、情感特征提取和情感识别,系统框图如图 1 所示.语音信号采集模块通过语音传感器(例如,麦克风等语音录制设备)获得语音信号,并传递到下一个情感特征提取模块对语音信号中与话者情感关联紧密的声学参数进行提取,最后送入情感识别模块完成情感的判断.需要特别指出的是,一个完整的语音情感识别系统除了要完善上述 3 部分以外,还离不开两项前期工作的支持:(1) 情感空间的描述;(2) 情感语料库的建立.情感空间的描述有多重标准,例如离散情感标签、激励-评价-控制空间和情感轮等,不同的标准决定了不同的情感识别方式,会对情感语料的收集标注、识别算法的选择都产生影响.情感语料库更是语音情感识别研究的基础,负责向识别系统提供训练和测试用语料数据.国内外相关研究根据研究者的出发点不同会各有侧重,但归根结底都可以涵盖到上述 5 个关键模块之中.

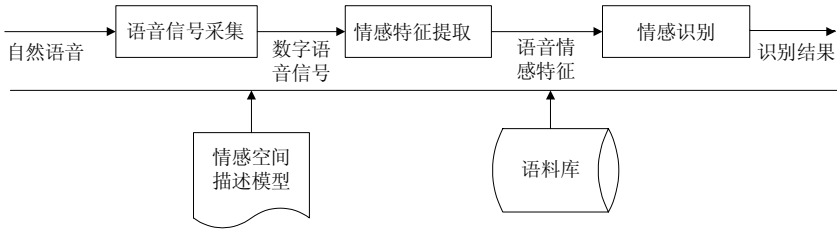

Fig.1 Framework of a standard speech emotion recognition system

图 1 语音情感识别系统框图

因此,本文将首先对语音情感识别接近 40 年的发展历程进行简要的回顾,然后从情感描述模型、情感语音数据库、语音情感相关声学特征提取、语音情感识别算法、语音情感识别技术应用这 5 个角度对当前的语音情感识别技术主流方法和前沿进展进行系统的总结和分析,最后给出技术挑战与展望.

1 语音情感识别历史回顾

最早的真正意义上的语音情感识别相关研究出现在 20 世纪 80 年代中期,它们开创了使用声学统计特征进行情感分类的先河[1,2].紧接着,随着 1985 年 Minsky 教授“让计算机具有情感能力”观点的提出,以及人工智能领域的研究者们对情感智能重要性认识的日益加深,越来越多的科研机构开始了语音情感识别研究的探索.

在 20 世纪 80 年代末至 90 年代初期,麻省理工学院多媒体实验室构造了一个“情感编辑器”对外界各种情感信号进行采集,综合使用人体的生理信号、面部表情信号、语音信号来初步识别各种情感,并让机器对各种情感做出适当的简单反应[3];1999 年,Moriyama 提出语音和情感之间的线性关联模型,并据此在电子商务系统中建造出能够识别用户情感的图像采集系统语音界面,实现了语音情感在电子商务中的初步应用[4].整体而言,语音情感识别研究在该时期仍旧处于初级阶段,语音情感识别的研究主要侧重于情感的声学特征分析这一方面,作为研究对象的情感语音样本也多表现为规模小、自然度低、语义简单等特点,虽然有相当数量的有价值的研究成果相继发表,但是并没有形成一套被广泛认可的、系统的理论和研究方法.

进入 21 世纪以来,随着计算机多媒体信息处理技术等研究领域的出现以及人工智能领域的快速发展,语音情感识别研究被赋予了更多的迫切要求,发展步伐逐步加快.2000 年,在爱尔兰召开的 ISCA Workshop on Speech and Emotion 国际会议第 1 次把致力于情感和语音研究的学者聚集在一起.近年来,先后又有若干以包括语音情感计算在内的情感计算为主题的会议和期刊被创立,并得到了世界范围内的注目,其中较为著名的有:始于 2005 年的 Affective Computing and Intelligent Interaction 双年会,始于 2009 年的 INTERSPEECH Emotion Challenge 年度竞赛,创刊于 2010 年的《IEEE Transactions on Affective Computing》期刊以及始于 2011 年的 International

Audio/ Visual Emotion Challenge and Workshop(AVEC)年度竞赛等.同时,越来越多国家的大学或科研机构涉足到语音情感识别研究的工作中来,著名的有:贝尔法斯特女王大学 Cowie 和 Douglas-Cowie 领导的情感语音小组;麻省理工大学 Picard 领导的媒体研究实验室;慕尼黑工业大学 Schuller 负责的人机语音交互小组;南加州大学 Narayanan 负责的语音情感组;日内瓦大学 Soberer 领导的情绪研究实验室;布鲁塞尔自由大学 Canamero 领导的情绪机器人研究小组等.国内对语音情感识别研究的关注起始于 21 世纪初,经过近 10 年的发展,目前已有越来越多的科研单位加入该领域的研究,著名的有东南大学无线电工程系、清华大学人机交互与媒体集成研究所、模式识别国家重点实验室、浙江大学人工智能研究所和中国科学院语言研究所等.

近 10 余年来,语音情感识别研究工作在情感描述模型的引入、情感语音库的构建、情感特征分析等领域的各个方面都得到了发展.Cowie 等人[5]开发的 FEELTRACE 情感标注系统为语音情感数据的标注提供了标准化工具.Grimm 等人[6,7]将三维情感描述模型(activation-evaluation-power space)引入到自发语音情感识别的研究中,并将维度情感识别问题建模为标准的回归预测问题.Grimm 的工作为维度语音情感识别研究的发展争取到更多的关注,激发了维度语音情感识别的热潮[7−11].慕尼黑工业大学的 Eyben 等人[12]开发了面向语音情感特征提取的开放式工具包 openSMILE,实现了包括能量、基频、时长、Mel 倒谱系数等在内的常用语音情感特征的批量自动提取,并逐渐得到广泛认可[13,14].McKeown 等人[15]以科研项目为依托,创建了一个以科学研究为目的的大型多媒体情感数据库 SEMAINE,并提供了情感数据的维度标注结果,为语音情感识别的研究和发展提供了公开的、丰富的、高质量的自然情感语料.正是这些研究成果的不断涌现,为构建语音情感识别标准化平台做出了里程碑式的贡献.

2 两类主流情感描述模型

情感描述方式大致可分为离散和维度两种形式.

前者将情感描述为离散的、形容词标签的形式,如高兴、愤怒等,在人们的日常交流过程中被广泛使用,同时还被普遍运用于早期的情感相关研究中.丰富的语言标签描述了大量的情感状态,那么,其中哪些情感状态的研究价值更具有普遍性呢?这个问题可以归结为对基本情感类别的确定.一般认为,那些能够跨越不同人类文化,甚至能够为人类和具有社会性的哺乳动物所共有的情感类别为基本情感.表 1[16]列举了不同学者对基本情感的定义和划分,其中,美国心理学家 Ekman 提出的 6 大基本情感(又称为 big six)在当今情感相关研究领域的使用较为广泛[17].

Table 1 Various definitions of emotion from different researchers[16]

表 1 不同学者对基本情感的定义[16]

学者	基本情感
Arnold	Anger, aversion, courage, dejection, desire, despair, dear, hate, hope, love, sadness
Ekman, Friesen, Ellsworth	Anger, disgust, fear, joy, sadness, surprise
Fridja	Desire, happiness, interest, surprise, wonder, sorrow
Gray	Desire, happiness, interest, surprise, wonder, sorrow
Izard	Anger, contempt, disgust, distress, fear, guilt, interest, joy, shame, surprise
James	Fear, grief, love, rage
McDougall	Fear, disgust, elation, fear, subjection, tender-emotion, wonder
Mower	Pain, pleasure
Oatley, Johnson-Laird	Anger, disgust, anxiety, happiness, sadness
Panksepp	Anger, disgust, anxiety, happiness, sadness
Plutchik	Acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise
Tomkins	Anger, interest, contempt, disgust, distress, fear, joy, shame, surprise
Watson	Fear, love rage
Weiner, Graham	Happiness, sadness

后者则将情感状态描述为多维情感空间中的点.这里的情感空间实际上是一个笛卡尔空间,空间的每一维对应着情感的一个心理学属性(例如,表示情感激烈程度的激活度属性以及表明情感正负面程度的效价属性).理论上,该空间的情感描述能力能够涵盖所有的情感状态.换句话说,任意的、现实中存在的情感状态都可以在情感空间中找到相应的映射点,并且各维坐标值的数值大小反映了情感状态在相应维度上所表现出来的强弱

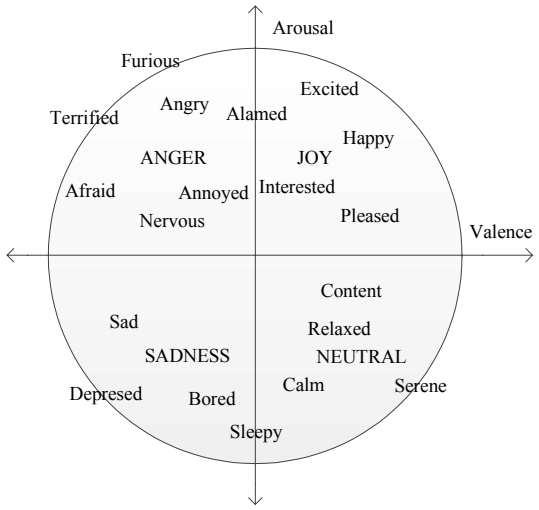

Fig.2 Arousal-Valence emotional space

图 2 激活度-效价情感空间

程度.由于维度情感模型使用连续的实数值来刻画情感,因此在有些文献中又被称作连续情感描述模型[18].一些既简单又能被广泛使用的维度情感描述模型有二维的激活度-效价空间理论(arousal-valence space)、三维的激励-评估-控制空间理论(valence-activation-dominance space)[19]和情感轮理论(emotion wheel)[18]等.其中,激活度-效价空间理论如图 2 所示[18]:垂直轴是激活度维,是对情感激烈程度的描述;水平轴是效价维,是对情感正负面程度的评价.情感状态的日常语音标签和该坐标空间可以进行相互转化,通过对情感状态语言描述的理解和估计,就可以找到它在情感空间中的映射位置.

两种表达模型各有千秋:从模型复杂度而言,离散描述模型较为简洁、易懂,有利于相关研究工作的着手和开展,而维度模型却要面对定性情感状态到定量空间坐标之间如何相互转换的问题;从情感描述能力的角度而言,离散情感模型的情感描述能力则显示出较大的局限性,多数情况下,它只能刻画单一的、有限种类的情感类型,然而人们在日常生活中所体验的情感却是微妙而多变的,甚至是复杂而模糊的(例如,人们在受到惊吓时所表现出来的情感不仅有吃惊,往往还包含害怕甚至恐惧的成分;又比如,人们对愉悦的表达可以呈现出若干的程度,可以从喜上眉梢,到眉飞色舞,再到手舞足蹈),可以说,离散描述方式和自发情感的描述之间还存在着较大的障碍,然而维度情感模型从多侧面、连续的角度进行情感的描述,很好地化解了自发情感的描述问题,并且以精确的数值很大程度上回避了离散情感标签的模糊性问题.最后,我们以表格的形式对两个情感描述模型之间的区别进行了直观的总结和展示,见表 2.

Table 2 Comparison of two emotional representation models

表 2 两种情感描述模型的区别

考察点	离散情感描述模型	离散情感描述模型
情感描述方式	形容词标签	笛卡尔空间中的坐标点
情感描述能力	有限的几个情感类别	任意情感类别
被应用到语音情感识别领域的时期	1980s	2000s
优点	简洁、易懂、容易着手	无限的情感描述能力
缺点	单一、有限的情感描述能力 无法满足对自发情感的描述	将主观情感量化为客观实数值的过程 是一个繁重且难以保证质量的过程

3 具有代表性的情感语音数据库

语音情感识别研究的开展离不开情感语音数据库的支撑.情感语音库的质量高低,直接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,按照激发情感的类型可分为表演型、引导型、自然型这 3 个类别;按照应用目标可分为识别型和合成型两个类别;按照语种不同可分为英语、德语、汉语等.不同于一般文献中的分类方法,本文将依据情感描述模型的不同,将数据语料资源划分为离散情感数据库和维度情感数据库两个分支,二者的区别在于情感标注形式的不同,前者以离散的语言标签(如高兴、悲伤等)作为情感标注,而后者则以连续的实数坐标值表示情感.

由此,我们称以语言标签进行标注的情感语料库为离散情感语料库,而以情感空间坐标值进行标注的语料库为维度情感语料库.目前,就国内外整个研究领域而言,以离散情感语料库居多,而维度情感语料库还有待丰富.本文将依照上述两个分支对当前国内外颇具代表性的情感语音库进行简要综述.它们虽然没有涵盖领域内大部分的语音资源,但都是经过精挑细选的、语料质量较高、影响较为广泛的情感语音库.若需了解更多的情

感语料库情况,可以参考文献[20–22]中的相关内容.

3.1 离散情感数据库

一个离散情感数据库一般包括有限的几类基本情感类型,并且希望每类情感的演绎都能达到单一、浓重、易辨识的标准,然而这恰恰是生活化的自然语音难以满足的.因此,目前的离散情感数据库多属于表演型或者引导型,或者二者的融合.例如,下面列举的代表性数据库中只有 FAU AIBO 属于自然型.

3.1.1 Belfast 英语情感数据库

Belfast 情感数据库[5,23]由 Queen 大学的 Cowie 和 Cowie 录制,由 40 位录音人(18 岁~69 岁,20 男 20 女)对 5 个段落进行演绎得到.每个段落包含 7~8 个句子,且具有某种特定的情感倾向,分别为生气/anger、悲伤/sadness、高兴/happiness、恐惧/fear、中性/neutral.

3.1.2 柏林 EMO-DB 德语情感语音库

DMO-DB[24]是由柏林工业大学录制的德语情感语音库,由 10 位演员(5 男 5 女)对 10 个语句(5 长 5 短)进行 7 种情感(中性/nertral、生气/anger、害怕/fear、高兴/joy、悲伤/sadness、厌恶/disgust、无聊/boredom)的模拟得到,共包含 800 句语料,采样率 48kHz(后压缩到 16kHz),16bit 量化.语料文本的选取遵从语义中性、无情感倾向的原则,且为日常口语化风格,无过多的书面语修饰.语音的录制在专业录音室中完成,要求演员在演绎某个特定情感前通过回忆自身真实经历或体验进行情绪的酝酿,来增强情绪的真实感.经过 20 个参与者(10 男 10 女)的听辨实验,得到 84.3%的听辨识别率.

3.1.3 FAU AIBO 儿童德语情感语音库

FAU AIBO[25]录制了 51 名儿童(10 岁~13 岁,21 男 30 女)在与索尼公司生产的电子宠物 AIBO 游戏过程中的自然语音,并且只保留了情感信息明显的语料,总时长为 9.2 小时(不包括停顿),包括 48 401 个单词.语音通过一个高质量的无线耳麦进行收集,并由 DAT-recorder 录制,48kHz 采样(而后压缩到 16kHz),16bit 量化.为了记录真实情感的语音,工作人员让孩子们相信 AIBO 能够对他们的口头命令加以反应和执行,而实际上,AIBO 则是由工作人员暗中人为操控的.标注工作由 5 名语言学专业的大学生共同完成,并通过投票方式决定最终标注结果,标注共涵盖包括 joyful,irritated,angry,neutral 等在内的 11 个情感标签.该数据库中的 18 216 个单词被选定为 INTERSPEECH 2009 年情感识别竞赛用数据库[26].

3.1.4 CASIA 汉语情感语料库

该数据库(http://www.chineseldc.org/resource_info.php?rid=76)由中国科学院自动化研究所录制,由 4 位录音人(2 男 2 女)在纯净录音环境下(信噪比约为 35db)分别在 5 类不同情感下(高兴、悲哀、生气、惊吓、中性)对 500 句文本进行的演绎得到,16kHz 采样,16bit 量化.经过听辨筛选,最终保留其中 9 600 句.

3.1.5 ACCorpus 系列汉语情感数据库

该系列情感数据库(http://hcsi.cs.tsinghua.edu.cn/accenter/fruit/database.html)由清华大学和中国科学院心理研究所合作录制,包含 5 个相关子库:1) ACCorpus_MM 多模态、多通道的情感数据库;2) ACCorpus_SR 情感语音识别数据库;3) ACCorpus_SA 汉语普通话情感分析数据库;4) ACCorpus_FV 人脸表情视频数据库;5) ACCorpus_FI 人脸表情图像数据库.其中,ACCorpus_SR 子库共由 50 位录音人(25 男 25 女)对 5 类情感(中性、高兴、生气、恐惧和悲伤)演绎得到,16kHz 采样,16bit 量化.每个发音者的数据均包含语音情感段落和语音情感命令两种类型.

3.2 维度情感数据库

对维度情感语音数据库的建立而言,由于维度情感描述模型的使用,使得数据的采集不再受情感类别的制约,理论上,蕴含任意情感信息的自然语音都可以被收纳到数据库中.然而,接下来的维度情感标注工作却显得并不轻松.目前而言,维度情感的标注工作一般都是基于打分制进行的(例如著名的情感标注工具 FEELTRACE[5]),即要求标注者在各个情感维度上对语音中的情感程度进行听辨,并赋以合适的分值.然而看似简单的打分工作,实际上却伴随了标注者们“将主观情感直接量化为客观实数值”的思考过程,尤其是当数据量

变得庞大时,相应的标注工作也会变得枯燥、劳累、令人难以忍受.近些年来,随着研究者们对维度情感识别领域的关注,尤其是维度情感识别竞赛(例如,2012 年 Continuous AVEC 2012[14])的开展,一些公开的维度情感数据库逐渐被发布出来.

3.2.1 VAM 数据库

VAM 数据库[27]是一个以科学研究为目的的无偿数据库,通过对一个德语电视谈话节目“Vera am Mittag”的现场录制得到,语音和视频被同时保存,因此,数据库包含语料库、视频库、表情库这 3 个部分.谈话内容均为无脚本限制、无情绪引导的纯自然交流.以 VAM-audio 库为例,该子库包含来自 47 位节目嘉宾的录音数据 947 句,wav 格式,16kHz 采样,16bit 量化.所有数据以句子为单位进行保存(1 018 句),标注在 Valence,Activation 和 Dominance 这 3 个情感维度上进行,标注值处于−1~1 之间.标注工作由多个标注者共同完成,最终的情感值是相关标注者的平均值.VAM-audio 是一个应用较为广泛的情感语料库,在本文的后续研究中也会加以使用.

3.2.2 Semaine 数据库

Semaine[15]数据库是一个面向自然人机交互和人工智能研究的数据库,可供科研人员无偿使用(http://semaine-db.eu/).数据录制在人机交互的场景下进行,20 个用户(22 岁~60 岁,8 男 12 女)被要求与性格迥异的 4 个机器角色进行交谈(实际上,机器角色由工作人员扮演).这 4 个角色分别是:1) 温和而智慧的 Prudence;2) 快乐而外向的 Poppy;3) 怒气冲冲的 Spike 和 4) 悲伤而抑郁的 Obadiah.录音过程在专业配置录音室内进行,同时有 5 个高分辨率、高帧频摄像机和 4 个麦克风进行数据的收集,其中,音频属性为 48kHz 采样,24bit 量化,数据时长在 7 小时左右.标注工作由多个参与者借助标注工具 FEELTRACE[5]在 Valence,Activation,Power,Expectation 和 Intensity 这 5 个情感维度上进行.该数据库中的部分数据被用于 AVEC 2012 的竞赛数据库[14].

3.3 语音情感特征提取

当前,用于语音情感识别的声学特征大致可归纳为韵律学特征、基于谱的相关特征和音质特征这 3 种类型.这些特征常常以帧为单位进行提取,却以全局特征统计值的形式参与情感的识别.全局统计的单位一般是听觉上独立的语句或者单词,常用的统计指标有极值、极值范围、方差等.

3.3.1 韵律学特征

韵律是指语音中凌驾于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,却决定着一句话是否听起来自然顺耳、抑扬顿挫.韵律学特征又被称为“超音段特征”或“超语言学特征”,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,使用非常普遍[28−31],其中最为常用的韵律特征有时长(duration)、基频(pitch)、能量(energy)等.

Luengo 等人[31]在一个 Basque 情感语音数据的基础上进行了一系列的韵律特征分析研究,他们首先为每个情感语句提取能量和基频曲线和对数曲线,然后继续为各条曲线计算相应的一阶差分和二阶差分曲线,最后统计出每条曲线的最大值、最小值、均值、方差、变化范围、偏斜度(skewness)、峰度(kurtosis),从而获得了 84 个特征组成的韵律特征集.经过特征选择与分析,最后共有基频均值、能量均值、基频方差、基频对数的斜交、基频对数的动态范围和能量对数的动态范围这 6 维特征被认为具有最佳的情感区分能力.Origlia 等人[32]使用基频和能量相关的最大值、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接近 60%的识别率.Seppänen 等人[33]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.Iliou 等人[30]和 Wang 等人[34]则分别将基频、能量、时长的韵律特征用于德语的说话人不相关的情感识别和汉语普通话情感的说话人相关的情感识别,分别得到了 51%和 88%的识别率.

除此之外,学者们还针对韵律特征与特定情感类型之间的关联上展开了研究[3,19,35−38],这些研究工作进一步验证了韵律特征区分情感的性能,但也出现了一些不甚一致的结论.例如,Murray 等人认为,较快的语速与愤怒的情感相关;而 Oster 等人却在文献[35]中给出了相反的结论.再者,学者们还发现:韵律特征区的情感区分能力是十分有限的.例如,愤怒、害怕、高兴和惊奇的基频特征具有相似的表现[3,36].

3.3.2 基于谱的相关特征

基于谱的相关特征被认为是声道(vocal tract)形状变化和发声运动(articulator movement)之间相关性的体现[39],已在包括语音识别、话者识别等在内的语音信号处理领域有着成功的运用[40−42].Nwe 等人[43]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.近年来,有越来越多的研究者们将谱相关特征运用到语音情感的识别中来[43−47],并起到了改善系统识别性能的作用,相关谱特征的情感区分能力是不可忽视的.在语音情感识别任务中使用的线性谱特征(linear-based spectral feature)一般有:LPC(linear predictor coefficient)[36],OSALPC(one-sided autocorrelation linear predictor coefficient)[48],LFPC(log-frequency power coefficient)[43]等;倒谱特征(cepstral-based spectral feature)一般有:LPCC(linear predictor cepstral coefficient),OSALPCC(cepstral-based OSALPC)[44],MFCC(mel-frequency cepstral coefficient)等.

目前,对线性谱特征和倒谱特征情感区分能力高低的判定似乎并无定论.Bou-Ghazale[44]对倒谱特征和线性谱特征在压力语音检测(detecting speech under stress)任务中的性能表现进行了研究,研究发现,倒谱特征 OSALPCC, LPCC 和 MFCC 的区分能力明显优于线性谱特征 LPC 和 OSALPC.然而,Nwe 等人[43]却得出了相反的结论.具体地,HMM 被用作分类器对包括生气、厌恶、恐惧、愉悦、悲伤和惊奇在内的 6 类情感进行话者相关的识别,结果表明,LFPC 取得了 77.1%的识别率,而 LPCC 和 MFCC 的识别率分别为 56.1%和 59.0%.

3.3.3 声音质量特征

声音质量是人们赋予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[49].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、难以抑制的情形之下[19].语音情感的听辨实验中,声音质量的变化被听辨者们一致认定为与语音情感的表达有着密切的关系[49].在语音情感识别研究中,用于衡量声音质量的声学特征一般有:共振峰频率及其带宽(format frequency and bandwidth)、频率微扰和振幅微扰(jitter and shimmer)[50]、声门参数(glottal parameter)等.

Lugger 等人[51−53]在一系列工作中提取第 1 和第 4 共振峰频率和相应的带宽作为声音质量特征,连同基频等韵律特征一起用于话者不相关的语音情感识别.Li 等人[54]提取了频率微扰和振幅微扰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)被作为识别器.与仅使用 MFCC 的基线性能 65.5%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%由 MFCC、频率微扰和振幅微扰的共同组合获得.相对前两类特征而言,声门参数的应用相应较少.一般地,人类的发声机制被建模为气流冲过声门,再通过声道进行滤波,继而输出的过程.从信号处理的角度来看,语音信号可视为声门激励信号和声道冲激响应的卷积.因此,提取声门参数的关键任务就是要去除语音中声道滤波的影响,从而获得声门激励相关的信号.然而,不论是激励信号还是声道滤波器的各项参数,我们都无从得知,仍需进一步估算,声门参数的提取可参见文献[49].

另外,Sun 等人在文献[55]中对声门参数和基频、能量等韵律特征在情感识别中发挥的作用进行了比较和探讨.

3.3.4 融合特征

上述 3 种特征分别从不同侧面对语音情感信息进行表达,自然会想到使用它们的融合用于语音情感的识别,从而达到提高系统识别性能的目的.目前,使用融合特征进行语音情感识别研究是本领域的主流方法[56−60].例如:Sanchez 等人[56]将基频、能量、共振峰、谱倾斜(spectral tilt)的 90 维全局统计特征用于 WCGS 数据库中沮丧情绪的检测;Schuller 等人[57]将过零率、能量、基频、声音质量、谐波噪声比、0~15 阶 MFCC 等特征的 5 967 维相关统计量用于 eNTERFACE[61]、柏林情感语料库 EMO-DB[23]以及合成语料库的交叉数据库情感识别研究;Malandrakis 等人[59]使用基频、强度、对数能量、过零率、频谱重心(spectral centroid)、频谱通量(spectral flux)、MFCC、PLPC(perceptual linear prediction coefficient)等特征的统计值用于电影维度情感的跟踪等.

3.3.5 基于 i-vector 的特征

i-vecotr 在近些年来的说话人识别领域有着广泛的应用,是一项将高维高斯混合模型(Gaussian mixture models,简称 GMM)超向量空间映射到低维总变异空间的技术,然而在语音情感识别领域的应用还较为新颖.文献[62]提出使用串联结构的情感 i-vector 特征用于语音情感的识别,他们首先使用 openSMILE 提取 1 584 维的声学特征,并使用这些特征为自然情感状态的语音训练得到一个通用模型(universal background model),然后在该通用模型的基础上为每类情感状态生成各自的 GMM,继而得到每类情感状态的 GMM 超向量用于 i-vector 的生成.最后,对应于各个情感状态的 i-vector 被串连在一起作为支持向量机的输入,用于 angry,happy,neutral,sad 这 4 类语音情感的识别,取得了优于原始 1 584 维声学特征的识别性能.

3.4 语音情感识别算法研究进展

寻找合适的识别算法,是本领域研究者们一直以来为之不懈努力的一个目标.整体而言,依据情感描述模型的不同,当今语音情感识别系统所采用的识别算法可以分为如下两类.

3.4.1 离散语音情感分类器

本文将基于离散情感描述模型的语音情感识别研究称作离散语音情感识别,它们一般被建模为标准的模式分类问题,即使用标准的模式分类器进行情感的识别[7].常用于语音情感识别领域的分类器,线性的有:Naïve Bayes Classifier,Linear ANN(artificial neural network),Linear SVM(support vector machine)等;非线性的有:Decision Trees,*k*-NN(*k*-nearest neighbor algorithm),Non-linear ANN,Non-linear SVM,GMM (Gaussian mixture model),HMM (hidden Markov model)以及稀疏表示分类器等.

如上所示,已有不少模式分类器被语音情感识别研究者们所尝试.其中,使用最为广泛的有 HMM[43,63,64], GMM[65−67],ANN[68−70]和 SVM [69,71].

Nwe 等人[43]使用基于 HMM 的识别器用于 6 类情感的识别.具体地,LFPC,MFCC 和 LPCC 被用作情感特征,为每个话者的每类情感构建一个四状态、全连接的 HMM,一个缅甸语语料库和一个汉语普通话语料库被分别用于 HMM 的训练和测试,系统最优性能分别可达到 78.5%和 75.5%.Lee 等人[64]分别以情感类别和音素类别为单位建立 HMM 模型,并在说话人不相关的情形下对模型性能进行测试.实验结果表明,基于音素类别的 HMM 模型具有更优的表现.

GMM 是一种用于密度估计的概率模型[72],可以被看作是只包含一个状态的连续 HMM 模型[73].文献[65]中,GMM 分类器被用于对面向婴儿的(infant-directed)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[74]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%.Tang 等人[66]针对语音情感识别构造了一种使用 Boosting 算法进行类条件分布估计的 GMM 模型,并称其为 Boosted-GMM,与传统的使用 EM(expectation maximization)方法进行分布估计的 EM-GMM 相比,Boosted-GMM 表现出更优的性能.

MLP(multi-layer perceptron)是语音情感识别中应用最为广泛的一种人工神经网络模型,这与 MLP 完善的工具包支撑和成熟的学习算法有着很大的关系.Nichoson 等人[68]基于 MLP 建立了一个 OCON(one class in one neural network)网络模型,对包括 joy,teasing,fear,sadness,disgust,anger,surprise 和 neutral 在内的 8 种情感进行识别.该 OCON 模型由 8 个 4 层 MLP 网络和一个决策逻辑控制构成:每个子网络对应于一种情感类型的识别,并在输出层唯一的神经元处输出某测试语句属于某种情感的概率预测值,模型最终会将待识别语句的情感分配为具有最大输出值的子网络所对应的情感.文中使用的数据库是自行录制的,有 100 位话者参与录制.实验过程中,挑选其中 30 位话者的语料用于模型的训练,剩余 70 位话者的语料用于性能测试.实验结果表明,该模型的最优识别率为 52.87%.Petrushin 等人[69]对普通 MLP 和 Bagging-MLP 在语音情感识别中的性能进行了比较.Bagging 是一种用于为一个分类器生成多个版本,从而合并为一个性能更为强大的聚合分类器的策略.实验结果表明,Bagging-MLP 的性能与普通 MLP 网络相比提高了 5.0%.

SVM 分类器的关键在于核函数的运用,它负责将原始特征以非线性的方式映射到高维空间中,从而提高数据的可分性.SVM 在语音情感识别领域有着广泛的应用,这里以文献[75]为例进行说明.文中共有 3 种策略被用

来构建基于二分类 SVM 的多分类模型:前两种策略中都首先为每类情感构建一个二分类的 SVM,不同的是,第 1 种策略将待识别语句分配给距离其余情感距离最远的情感类型,而第 2 种策略则将各个二分类 SVM 的输出作为一个 3 层 MLP 网络的输入,通过进一步的计算做出最终的分配决定;第 3 种策略被称为多层次的分类模型(hierarchical classification model),各个 SVM 子分类器按照树形结构进行排列,从根节点开始由粗略到细致地实现情感的逐步划分,在叶节点处给出最终识别结果.实验结果表明:在 FERMUS III 数据库[75]之上,3 种策略的识别率分别为 76.12%,75.45%和 81.29%,第 3 种策略表现最优.

而稀疏表示分类器则是近年来随着压缩感知技术的兴起发展而来的一项分类技术.在文献[76]中,该分类器首先采用稀疏分解的方法,用训练样本对测试样本进行最稀疏表示,即把训练样本看作是一组基,通过求解 1-范数最小化的方法得到测试样本的最稀疏表示系数,最后用测试样本与稀疏表示后的残差来进行分类.在柏林 EMO-DB 德语情感语音库上进行 7 类情感状态的识别时,取得了相比线性判别分类器、*k*-NN、ANN、SVM 更好的识别性能.

3.4.2 维度语音情感预测器

本文将基于维度情感描述模型的语音情感识别研究称为维度语音情感识别,它的出现与传统的离散语音情感识别相比较为新兴,但也已得到领域内研究者们越来越多的关注[7−11,77].该研究一般被建模为标准的回归预测问题,即使用回归预测算法对情感属性值进行估计,在当前的维度语音情感识别领域使用较多的预测算法有:Linear Regression,*k*-NN,ANN,SVR(support vector regression)等.其中,SVR 因为性能稳定、训练时间短等优点应用得最为广泛.例如,Grimm 等人[7]在 VAM 数据库[27]上对基于规则的逻辑分类器(rule-based fuzzy logic classifier)、*k*-NN 和 SVR 在包括 Valence,Activation 和 Dominance 在内的三维情感属性上的预测能力进行比较,结果表明,SVR 的预测能力更胜一筹.我们可以看出:相比离散情感分类器的繁荣发展,维度情感预测算法的研究较为薄弱,更多针对情感识别任务的高性能算法仍有待进一步加以开发.

3.5 语音情感识别技术应用

语音情感识别在众多具有自然人机交互需求的领域内有着广泛的应用,例如:可以用于对电话服务中心(call center)用户紧急程度的分拣,从而提高服务质量[29].具体地,可通过及时发现负面情绪较为激烈的用户,并将他们的电话及时转接给人工客服,达到优化用户体验的目的;用于对汽车驾驶者的精神状态进行监控,从而在驾驶员疲劳的时候加以提醒[78,79],从而避免交通事故的发生;用于对远程网络课堂(E-learning)用户在学习过程中的情感状态进行监控,从而及时调整授课重点或者进度[80,81];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[82];用于辅助、指导自闭症儿童对情感理解和表达能力的学习[83]等.这些技术应用从算法实现要求上可分为实时类和性能类:实时类包括电话服务中心用户紧急程度分拣、驾驶员疲劳检测、E-learning 学员情感监控等,这类应用的特点为对识别速度要求很高,但相对而言对识别准确性具有一定程度的容忍性;而性能类则对算法的识别效果有着较高的要求,例如抑郁症患者情绪跟踪等,因为此时的识别结果关系到医生对患者病情的判断以及治疗方案的定制,那么为了获得较高的识别性能,此时的技术实现可在一定程度上做出识别速度的妥协.

3.6 结束语

本文在充分调研和深入分析的基础上对当今的语音情感识别领域研究进展进行了综述,其中重点介绍了语音情感识别研究中的几个关键问题,包括情感描述模型选取、情感语音数据库建立、语音情感相关声学特征提取、语音情感识别算法建模、语音情感识别技术应用等.可以说,自从该领域在 20 世纪末期被创立以来,在世界范围内的研究者们数十年的不懈努力下,语音情感识别研究取得了令人欢欣鼓舞的进步与发展.然而,鉴于“语音情感”其本身自有的复杂性,该领域仍旧面临着若干值得深入探索的问题.这里,我们基于大量的调研和近几年来的研究经验提出一些值得进一步挖掘的研究点,希望对本领域的其他研究者有所启发.

3.6.1 情感语料问题

一个丰富、优质的情感语音数据库是开展语音情感计算研究的必要基础,可以为研究工作提供可靠的训练

和测试用数据.然而,由于情感本身的复杂性,使得情感语音数据的采集和整理工作非常困难,进而导致了高质量的情感语料难以获取.尤其是相比于语音识别领域的大规模自然语音数据库以及音乐计算领域的海量歌曲数据库,现已公布的情感语料数据堪称稀少.对离散情感语音数据库而言,如何同时满足语料的自然度和情感的纯净度是其面临的最大挑战.虽然经历了数十年的发展和积累,也不断有数据库被录制和发布,但是,为研究者们所认可的高质量数据库却为数不多.对维度情感语音数据库的建立而言,困难不在于语料的获取,而在于语料的整理和情感的标注.为了将语料中的情感量化为精确的实数值,标注者担负了繁重的听辨和打分工作,并且标注结果的好坏、正误也难以评判.当前,已有的维度情感语音数据库资源仍较为稀少.面对语料资源的上述现状,应该如何对现有资源进行补充和丰富?能否通过技术手段对训练语料的选择进行系统的指引和帮助?都是研究者们亟待解决的实际问题.

3.6.2 情感与声学特征之间的关联问题

语音情感识别的最终目标是人脑的识别水平.从情感语音信号的形成开始,计算机与人脑的情感识别机制的最初差异就是情感相关声学特征的提取以及情感与声学特征之间的关联方式的确定.因此,如果计算机不能准确地或者以尽可能接近人类的方式对情感语音进行声学特征提取并加以正确的关联和映射,就会使得计算机语音情感识别系统被建立于一个偏离实际的基础之上,从而导致其后的识别机制与人脑处理机制间的差距越来越大,无法达到期望的效果.然而,目前并没有一个相当于语音识别领域中的 Mel 倒谱系数同样地位的情感声学特征被提出.一般情况下,研究者们使用包括韵律学、声音质量、频谱在内的多种相关声学特征的合集作为语音情感特征的代表.因此,如何从现有的声学特征中选择区分能力最优的特征子集、如何探究与情感表达关联更加密切的新特征都是当前领域内十分重要的研究课题.并且一般认为,基于语句时长的全局特征与情感状态之间的关联最为紧密,因为它可以在一定程度上削弱文本差异对声学特征的干扰.然而,这种所谓的干扰削弱,却是以减弱部分表征情感状态的声学特征的细节效用为代价的.从该角度来看,如何界定情感声学特征的最优提取时长,抑或是对不同时长的声学特征进行融合,也都是不容忽略的研究课题.

3.6.3 语音情感识别的建模问题

构建合理、高效的语音情感识别模型是语音情感识别研究的重中之重,它负责对大量的训练语料进行学习,从中挖掘由各种声学特征通往对应情感状态的映射通路,从而实现对测试语料情感状态的正确判断与识别.理想的语音情感识别模型应该是对人脑语音情感处理机制的模拟和重建,然而,由于人脑情感处理机制的复杂性以及目前的认知科学水平,当前领域内构建的识别模型仍停留在功能模拟的水平,与机制模拟的目标还存在一定的差距.例如,离散情感识别任务一般被建模为普通的模式分类器,而维度情感识别任务一般被建模为标准的回归预测问题.那么,如何在现有的认知科学水平之上,以尽可能贴近人脑情感处理机制的方式来构建语音情感识别模型,是一项艰巨却有着重大意义的任务.

3.6.4 语音情感识别技术的推广问题

伴随着人机语音交互技术的不断发展,越来越多的语音交互技术从实验室环境进入了商业应用,并对人们的生活方式产生着影响.例如,midomi 哼唱检索网络使用的分布式旋律比对技术(http://www.midomi.com)、苹果公司的 Siri 语音搜索软件使用的分布式语音识别及合成技术等.然而,鉴于语音情感识别的新兴性,目前并没有成熟的相关应用问世.利用互联网平台推广语音情感识别技术的应用,对于加快人机交互的情感智能化进程有着非常重要的实际意义,应当给予足够的重视.

References:

[1] van Bezooijen R, Otto SA, Heenan TA. Recognition of vocal expressions of emotion: A three-nation study to identify universal characteristics. Journal of Cross-Cultural Psychology, 1983,14(4):387−406. [doi: 10.1177/0022002183014004001]

[2] Tolkmitt FJ, Scherer KR. Effect of experimentally induced stress on vocal parameters. Journal of Experimental Psychology Human Perception Performance, 1986,12(3):302−313. [doi: 10.1037/0096-1523.12.3.302]

[3] Cahn JE. The generation of affect in synthesized speech. Journal of the American Voice Input/Output Society, 1990,8:1−19.

[4] Moriyama T, Ozawa S. Emotion recognition and synthesis system on speech. In: Proc. of the 1999 IEEE Int'l Conf. on Multimedia Computing and Systems (ICMCS). Florence: IEEE Computer Society, 1999. 840−844. [doi: 10.1109/MMCS.1999.779310]

[5] Cowie R, Douglas-Cowie E, Savvidou S, McMahon E, Sawey M, Schroder M. Feeltrace: An instrument for recording perceived emotion in real time. In: Proc. of the 2000 ISCA Workshop on Speech and Emotion: A Conceptual Frame Work for Research. Belfast: ISCA, 2000. 19−24.

[6] Grimm M, Kroschel K. Evaluation of natural emotions using self assessment manikins. In: Proc. of the 2005 ASRU. Cancun, 2005. 381−385. [doi: 10.1109/ASRU.2005.1566530]

[7] Grimm M, Kroschel K, Narayanan S. Support vector regression for automatic recognition of spontaneous emotions in speech. In: Proc. of the 2007 IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing (ICASSP). IEEE Computer Society, 2007. 1085−1088. [doi: 10.1109/ICASSP.2007.367262]

[8] Eyben F, Wollmer M, Graves A, Schuller B, Douglas-Cowie E, Cowie R. On-Line emotion recognition in a 3-D activation-valence-time continuum using acoustic and linguistic cues. Journal on Multimodal User Interfaces, 2010,3(1-2):7−19. [doi: 10.1007/s12193-009-0032-6]

[9] Giannakopoulos T, Pikrakis A, Theodoridis S. A dimensional approach to emotion recognition of speech from movies. In: Proc. of the 2009 IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing (ICASSP). Taibe: IEEE Computer Society, 2009. 65−68. [doi: 10.1109/ICASSP.2009.4959521]

[10] Wu DR, Parsons TD, Mower E, Narayanan S. Speech emotion estimation in 3d space. In: Proc. of the 2010 IEEE Int'l Conf. on Multimedia and Expo (ICME). Singapore: IEEE Computer Society, 2010. 737−742. [doi: 10.1109/ICME.2010.5583101]

[11] Karadogan SG, Larsen J. Combining semantic and acoustic features for valence and arousal recognition in speech. In: Proc. of the 2012 Int'l Workshop on Cognitive Information Processing (CIP). IEEE Computer Society, 2012. 1−6. [doi: 10.1109/CIP.2012.6232924]

[12] Eyben F, Wollmer M, Schuller B. openSMILE—The Munich versatile and fast open-source audio feature extractor. In: Proc. of the 2010 ACM Multimedia. Firenze, 2010. 1459−1462. [doi: 10.1145/1873951.1874246]

[13] Schuller B, Valstar M, Eyben F, McKeown G, Cowie R, Pantic M. AVEC 2011—The first international audio/visual emotion challenge. In: Proc. of the 2011 Affective Computing and Intelligent Interaction, SER. Lecture Notes in Computer Science, Memphis: Berlin, Heidelberg: Springer-Verlag, 2011. 415−424. [doi: 10.1007/978-3-642-24571-8_53]

[14] Schuller B, Valstar M, Eyben F, Cowie R, Pantic M. Avec 2012 the continuous audio/visual emotion challenge. In: Proc. of the 2012 Int'l Audio/Visual Emotion Challenge and Workshop (AVEC), Grand Challenge and Satellite of ACM ICMI 2012. Santa Monica: ACM Press, 2012. [doi: 10.1145/2388676.2388758]

[15] McKeown G, Valstar MF, Cowie R, Pantic M. The semaine corpus of emotionally coloured character interactions. In: Proc. of the 2010 IEEE Int'l Conf. on Multimedia and Expo (ICME). Singapore: IEEE Computer Society, 2010. 1079−1084. [doi: 10.1109/ICME.2010.5583006]

[16] Ortony A, Turner TJ. What's basic about basic emotions. Psychological Review, 1990,97(3):315−331. [doi: 10.1037/0033-295X.97.3.315]

[17] Ekman P, Power MJ. Handbook of Cognition and Emotion. Sussex: John Wiley & Sons, 1999.

[18] Xie B. Research on key issues of Mandarin speech emotion recognition [Ph.D. Thesis]. Hangzhou: Zhejiang University, 2006 (in Chinese with English abstract).

[19] Cowie R, Douglas-Cowie E, Tsapatsoulis N, Votsis G, Kollias S, Fellenz W, Taylor JG. Emotion recognition in human-computer interaction. In: Proc. of the IEEE Signal Processing Magazine. 2001. 32−80. http://www.signalprocessingsociety.org/

[20] Ververidis D, Kotropoulos C. Emotional speech recognition: Resources, features, and methods. In: Proc. of the Speech Communication. 2006. 1162−1181. [doi: 10.1016/j.specom.2006.04.003]

[21] Ayadi ME, Kamel MS, Karray F. Survey on speech emotion recognition: Features, classification schemes, databases. Pattern Recognition, 2011,44(3):572−587. [doi: 10.1016/j.patcog.2010.09.020]

[22] Ververidis D, Kotropoulos C. A state of the art review on emotional speech databases. In: Proc. of the 2003 Richmedia Conf. Lausanne. Switzerland, 2003. 109−119.

[23] McGilloway S, Cowie R, Douglas-Cowie E, Gielen S, Westerdijk M, Stroeve S. Approaching automatic recognition of emotion from voice: A rough benchmark. In: Proc. of the 2000 ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research. Belfast: ISCA, 2000. 207−212.

[24] Burkhardt F, Paeschke A, Rolfes M, Sendlmeier W, Weiss B. A database of german emotional speech. In: Proc. of the 2005 INTERSPEECH. Lisbon: ISCA, 2005. 1517−1520.

[25] Steidl S. Automatic classification of emotion-related user states in spontaneous children's speech [Ph.D. Thesis]. Erlangen: University at Erlangen Nurberg, 2009.

[26] Schuller B, Steidl S, Batliner A. The INTERSPEECH 2009 emotion challenge. In: Proc. of the 2009 INTERSPEECH. Brighton: ISCA, 2009. 312−315.

[27] Grimm M, Kroschel K, Narayanan S. The vera am mittag german audiovisual emotional speech database. In: Proc. of the 2008 IEEE Int'l Conf. on Multimedia and Expo (ICME). Hannover: IEEE Computer Society, 2008. 865−868. [doi: 10.1109/ICME.2008.4607572]

[28] Schroder M, Cowie R. Issues in emotion-oriented computing—Towards a shared understanding. In: Proc. of the 2006 Workshop on Emotion and Computing. Hannover: IEEE Computer Society, 2006. 865−868.

[29] Lee CM, Narayanan SS. Toward detecting emotions in spoken dialogs. IEEE Trans. on Speech and Audio Processing, 2005,13(2): 293−303. [doi: 10.1109/TSA.2004.838534]

[30] Iliou T, Anagnostopoulos CN. Statistical evaluation of speech features for emotion recognition. In: Proc. of the 2009 Int'l Conf. on Digital Telecommunications. Colmar: IEEE Computer Society, 2009. 121−126. [doi: 10.1109/ICDT.2009.30]

[31] Luengo I, Navas E, Hernaez I, Sanchez J. Automatic emotion recognition using prosodic parameters. In: Proc. of the 2005 INTERSPEECH. Lisbon: ISCA, 2005. 493−496.

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

[33] Seppanen T, Vayrynen E, Toivanen J. Prosody-Based classification of emotions in spoken finnish. In: Proc. of the 2003 European Conf. on Speech Communication and Technology (EUROSPEECH). Geneval: ISCA, 2003. 717−720.

[34] Wang Y, Du SF, Zhan YZ. Adaptive and optimal classification of speech emotion recognition. In: Proc. of the 2008 Int'l Conf. on Natural Computation. Ji'nan: IEEE Computer Society, 2008. 407−411. [doi: 10.1109/ICNC.2008.713]

[35] Oster A-M, Risberg A. The Identification of Mood of a Speaker by Hearing Impaired Listeners. Stockholm, 1986. 79−90. http://www.speech.kth.s/qpsr

[36] Rabiner LR, Schafer RW. Digital Processing of Speech Signal. London: Prentice Hall, 1978.

[37] Borchert M, Düsterhöft A. Emotion in speech—Experiments with prosody and quality features in speech for use in categorical and dimensional emotion recognition environments. In: Proc. of the 2005 IEEE Int'l Conf. on Natural Language Processing and Knowledge Engineering. IEEE Computer Society, 2005. 147−151. [doi: 10.1109/NLPKE.2005.1598724]

[38] Tao JH, Kang YG, Li AJ. Prosody conversion from natural speech to emotional speech. IEEE Trans. on Audio, Speech, and Language Processing, 2006,14(4):1145−1154. [doi: 10.1109/TASL.2006.876113]

[39] Benesty J, Sondhi MM, Huang Y. Springer Handbook of Speech Processing. Berlin: Springer-Verlag, 2008. [doi: 10.1007/978-3-540-49127-9]

[40] O'Shaughnessy D. Invited paper: Automatic speech recognition: History, methods and challenges. Pattern Recognition, 2008,41 (10):2965−2979. [doi: 10.1016/j.patcog.2008.05.008]

[41] Wang LB, Minami K, Yamamoto K, Nakagawa S. Speaker identification by combining MFCC and phase information in noisy environments. In: Proc. of the 2010 IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing (ICASSP). Dallas: IEEE Computer Society, 2010. 4502−4505. [doi: 10.1109/ICASSP.2010.5495586]

[42] Nakagawa S, Wang LB, Ohtsuka S. Speaker identification and verification by combining mfcc and phase information. IEEE Trans. on Audio, Speech, and Language Processing, 2012,20(4):1085−1095. [doi: 10.1109/TASL.2011.2172422]

[43] Nwe TL, Foo SW, De Silva LC. Speech emotion recognition using hidden Markov models. Speech Communication, 2003,41(4): 603−623. [doi: 10.1016/S0167-6393(03)00099-2]

[44] Bou-Ghazale SE, Hansen JHL. A comparative study of traditional and newly proposed features for recognition of speech under stress. IEEE Trans. on Speech and Audio Processing, 2000,8(4):429−442. [doi: 10.1109/89.848224]

[45] Bitouk D, Verma R, Nenkova A. Class-Level spectral features for emotion recognition. Speech Communication, 2010,52(7-8): 613−625. [doi: 10.1016/j.specom.2010.02.010]

[46] Chauhan R, Yadav J, Koolagudi SG, Rao KS. Text independent emotion recognition using spectral features. In: Proc. of the 2011 Int'l Conf. on Contemporary Computing. Berlin, Heidelberg: Springer-Verlag, 2011. 359−370. [doi: 10.1007/978-3-642-22606-9_37]

[47] Wu SQ, Falk TH, Chan WY. Automatic speech emotion recognition using modulation spectral features. Speech Communication, 2011,53(5):768−785. [doi: 10.1016/j.specom.2010.08.013]

[48] Hernando J, Nadeu C. Linear prediction of the one-sided autocorrelation sequence for noisy speech recognition. IEEE Trans. on Speech and Audio Processing, 1997,5(1):80−84. [doi: 10.1109/89.554273]

[49] Gobl C, Chasaide AN. The role of voice quality in communicating emotion, mood and attitude. Speech Communication, 2003, 40(1-2):189−212. [doi: 10.1016/S0167-6393(02)00082-1]

[50] Gelfer MP, Fendel DM. Comparison of jitter, shimmer, and signal-to-noise ratio from directly digitized versus taped voice samples. Journal of Voice, 1995,9(4):378−382. [doi: 10.1016/S0892-1997(05)80199-7]

[51] Lugger M, Yang B. The relevance of voice quality features in speaker independent emotion recognition. In: Proc. of the 2007 IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing (ICASSP). Honolulu: IEEE Computer Society, 2007. 17−20. [doi: 10.1109/ICASSP.2007.367152]

[52] Lugger M, Yang B. Psychological motivated multi-stage emotion classification exploiting voice quality features. In: Proc. of the Speech Recognition. 2008.

[53] Lugger M, Janoir ME, Yang B. Combining classifiers with diverse feature sets for robust speaker independent emotion recognition. In: Proc. of the 2009 European Signal Processing Conf. Glagow: EURASIP, 2009. 1225−1229.

[54] Li X, Tao JD, Johnson MT, Soltis J, Savage A, Leong KM, Newman JD. Stress and emotion classification using jitter and shimmer features. In: Proc. of the 2007 IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing (ICASSP). Honolulu: IEEE Computer Society, 2007. 1081−1084. [doi: 10.1109/ICASSP.2007.367261]

[55] Sun R, Moore E, Torres JF. Investigating glottal parameters for differentiating emotional categories with similar prosodics. In: Proc. of the 2009 IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing (ICASSP). Taibe: IEEE Computer Society, 2009. 4509−4512. [doi: 10.1109/ICASSP.2009.4960632]

[56] Sanchez MH, Vergyri D, Ferrer L, Richey C, Garcia P, Knoth B, Jarrold W. Using prosodic and spectral features in detecting depression in elderly males. In: Proc. of the 2011 INTERSPEECH. Florence: ISCA, 2011. 3001−3004.

[57] Schuller B, Burkhardt F. Learning with synthesized speech for automatic emotion recognition. In: Proc. of the 2010 IEEE Int'l Conf. on Acoustics Speech and Signal Processing (ICASSP). Dallas: IEEE Computer Society, 2010. 5150−5153. [doi: 10.1109/ICASSP.2010.5495017]

[58] Espinosa HP, García CA, Pineda LV. Features selection for primitives estimation on emotional speech. In: Proc. of the 2010 IEEE Int'l Conf. on Acoustics Speech and Signal Processing (ICASSP). Dallas: IEEE Computer Society, 2010. 5138−5141. [doi: 10.1109/ICASSP.2010.5495031]

[59] Malandrakis N, Potamianos A, Evangelopoulos G, Zlatintsi A. A supervised approach to movie emotion tracking. In: Proc. of the 2011 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP). Prague: IEEE Computer Society, 2011. 2376−2379. [doi: 10.1109/ICASSP.2011.5946961]

[60] Lee CC, Mower E, Busso C, Lee S, Narayanan S. Emotion recognition using a hierarchical binary decision tree approach. Speech Communication, 2011,53(9-10):1162−1171. [doi: 10.1016/j.specom.2011.06.004]

[61] Martin O, Kotsia I, Macq B, Pitas I. The enterface 2005 audio-visual emotion database. In: Proc. of the 2006 Int'l Conf. on Data Engineering Workshops. Washington: IEEE Computer Society, 2006. 8.

[62] Xia R, Liu Y. Using i-vector space model for emotion recognition. In: Proc. of the INTERSPEECH 2012. Portland: ISCA, 2012. 2230−2233.

[63] Schuller B, Rigoll G, Lang M. Hidden Markov model-based speech emotion recognition. In: Proc. of the 2003 IEEE Int'l Conf. on Acoustics, Speech, Signal Processing. Hong Kong: IEEE Computer Society, 2003. II-1.

[64] Lee CM, Yildirim S, Bulut M, Kazemzadeh A, Busso C, Deng Z, Lee S, Narayanan S. Emotion recognition based on phoneme classes. In: Proc. of the 2004 INTERSPEECH. Jeju Island: ISCA, 2004. 889−892.

[65] Breazeal C, Aryananda L. Recognition of affective communicative intent in robot-directed speech. Autonomous Robots, 2002,12(1): 83−104. [doi: 10.1023/A:1013215010749]

[66] Tang H, Chu SM, Hasegawa-Johnson M, Huang TS. Emotion recognition from speech via boosted gaussian mixture models. In: Proc. of the 2009 IEEE Int'l Conf. on Ultimedia and Expo (ICME). Piscataway: IEEE Press, 2009. 294−297. [doi: 10.1109/ICME.2009.5202493]

[67] Yuan G, Lim TS, Juan WK, Ringo HMH, Li Q. A GMM based 2-stage architecture for multi-subject emotion recognition using physiological responses. In: Proc. of the 2010 Augmented Human Int'l Conf. New York: ACM Press, 2010. 3:1−3:6. [doi: 10.1145/1785455.1785458]

[68] Nicholson J, Takahashi K, Nakatsu R. Emotion recognition in speech using neural networks. Neural Computing & Applications, 2000,9(4):290−296. [doi: 10.1007/s005210070006]

[69] Petrushin VA. Emotion recognition in speech signal: Experimental study, development, and application. In: Proc. of the 2000 Int'l Conf. on Spoken Language Processing. 2000. 222−225.

[70] Bhatti MW, Wang YJ, Guan L. A neural network approach for human emotion recognition in speech. In: Proc. of the 2004 Int'l Symp. on Circuits and Systems. Vancouver, 2004. 181−184. [doi: 10.1109/ISCAS.2004.1329238]

[71] Hassan A, Damper RI. Multi-Class and hierarchical svms for emotion recognition. In: Proc. of the 2010 INTERSPEECH. Chiba: ISCA, 2010. 2354−2357.

[72] Vlassis N, Likas A. A greedy em algorithm for gaussian mixture learning. Neural Processing Letters, 2002,15(1):77−87. [doi: 10.1023/A:1013844811137]

[73] Reynolds DA, Rose RC. Robust text-independent speaker identification using gaussian mixture speaker models. IEEE Trans. on Speech and Audio Processing, 1995,3(1):72−83. [doi: 10.1109/89.365379]

[74] Vlassis N, Likas A. A kurtosis-based dynamic approach to Gaussian mixture modeling. IEEE Trans. on Systems, Man and Cybernetics, Part A: Systems and Humans, 1999,29(4):393−399. [doi: 10.1109/3468.769758]

[75] Schuller B. Towards intuitive speech interaction by the integration of emotional aspects. In: Proc. of the 2002 IEEE Int'l Conf. on Systems, Man and Cybernetics. Hammamet: IEEE Computer Society, 2002. 6. [doi: 10.1109/ICSMC.2002.1175635]

[76] Zhao XM, Zhang SQ. Robustness speech emotion recognition methods based on compressed sensing. China Patent, CN103021406A, 2013-04-03 (in Chinese).

[77] Gunes H, Schuller B, Pantic M, Cowie R. Support vector regression for automatic recognition of spontaneous emotions in speech. In: Proc. of the Int'l Workshop on EmoSPACE, Held in Conjunction with the 9th Int'l IEEE Conf. on Face and Gesture Recognition 2011 (FG 2011). Santa Barbara: IEEE Computer Society, 2011. 827−834.

[78] Schuller B, Rigoll G, Lang M. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In: Proc. of the 2004 IEEE Int'l Conf. on Acoustics, Speech, Signal Processing (ICASSP). Montreal: IEEE Computer Society, 2004. 577−580. [doi: 10.1109/ICASSP.2004.1326051]

[79] Boril H, Sadjadi SO, Kleinschmidt T, Hansen JHL. Analysis and detection of cognitive load and frustration in drivers' speech. In: Kobayashi T, Hirose K, Nakamura S, eds. Proc. of the 2010 INTERSPEECH. Chiba: ISCA, 2010. 502−505.

[80] Chen K, Yue GX, Yu F, Shen Y, Zhu AQ. Research on speech emotion recognition system in e-learning. In: Proc. of the 2007 Int'l Conf. on Computational Science, SER. LNCS 4489, Berlin, Heidelberg: Springer-Verlag, 2007. 555−558. [doi: 10.1007/978-3-540-72588-6_91]

[81] Wang WS, Wu JB. Emotion recognition based on CSO&SVM in e-learning. In: Proc. of the 2011 7th Int'l Conf. on Natural Computation (ICNC). Beijing: IEEE Computer Society, 2011. 566−570. [doi: 10.1109/ICNC.2011.6022071]

[82] France DJ, Shiavi RG, Silverman S, Silverman M, Wilkes MD. Acoustical properties of speech as indicators of depression and suicidal risk. IEEE Trans. on Biomedical Engineering, 2000,47(7):829−837.

[83] Marchi E, Schuller B, Batliner A, Fridenzon S, Tal S, Golan O. Emotion in the speech of children with autism spectrum conditions: Prosody and everything else. In: Proc. of the 2012 Workshop on Child, Computer and Interaction (WOCCI 2012), Satellite Event of INTERSPEECH. Portland: ISCA, 2012.

附中文参考文献:

[18] 谢波.普通话语音情感识别关键技术研究[博士学位论文].杭州:浙江大学,2006.

[76] 赵小明,张石清.基于压缩感知的鲁棒性语音情感识别方法.中国专利,CN103021406A,2013-04-03.

韩文静(1983－),女,河南正阳人,博士,CCF 学生会员,主要研究领域为语音情感识别,自然人机交互.

E-mail: hanwenjing07@gmail.com

阮华斌(1983－),男,博士,主要研究领域为智能信息处理,算法优化与加速,可重构计算.

E-mail: ruanhuabin@mail.tsinghua.edu.cn

李海峰(1969－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为语音情感处理,自然人机交互,智能信息处理,人工脑与认知科学.

E-mail: lihaifeng@hit.edu.cn

马琳(1967－),女,博士,副教授,CCF 会员,主要研究领域为模式识别,智能信息处理,图像处理,生物信息学.

E-mail: malin_li@hit.edu.cn