软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, [doi: 10.13328/j.cnki.jos.006078]

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

维度语音情感识别研究进展*

李海峰[1,2], 陈 婧[1], 马 琳[1,2], 薄洪健[2], 徐 聪[1], 李洪伟[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](深圳航天科技创新研究院,广东 深圳,518057)

通讯作者：李海峰,E-mail: lihaifeng@hit.edu.cn

摘 要： 情感识别是多学科交叉的研究方向,涉及认知科学、心理学、信号处理、模式识别、人工智能等领域的研究热点,目的是使机器理解人类情感状态,进而实现自然人机交互.本文首先从心理学及认知学角度介绍了语音情感认知研究进展,详细介绍了情感的认知理论、维度理论、脑机制以及基于情感理论的计算模型,旨在为语音情感识别提供科学的情感理论模型.然后,从人工智能角度系统地总结了目前维度情感识别的研究现状和发展,包括语音维度情感数据库、特征提取、识别算法等技术要点.最后,分析了维度情感识别技术目前面临的挑战以及可能的解决思路,对未来研究方向进行了展望.

关键字： 情感维度理论；语音情感认知；情感计算模型；语音情感特征提取；维度情感识别算法

中图法分类号: TP391.42

中文引用格式: 李海峰,陈婧,马琳,薄洪健,徐聪,李洪伟.维度语音情感识别研究进展.软件学报,2020,31. http://www.jos.org.cn/1000-9825/6078.htm

英文引用格式: Li HF, Chen J, Ma L, Bo HJ, Xu C, Li HW. Review of Speech Dimensional Emotion Recognition. Ruan Jian Xue Bao/Journal of Software, 2020,31 (in Chinese). http://www.jos.org.cn/1000-9825/6078.htm

Review of Dimensional Emotion Recognition From Speech

LI Hai-Feng[1,2], CHEN Jing[1], MA Lin[1,2], BO Hong-Jian[2], XU Cong[1], LI Hong-Wei[1]

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

[2](Shenzhen Academy of Aerospace Technology, Shenzhen 518057, China)

Abstract: Emotion recognition is an interdisciplinary research field which relates to cognitive science, psychology, signal processing, pattern recognition, artificial intelligence and so on, aiming at helping computer understand human emotion state to realize natural human-computer interaction. In this survey, the psychological theory of emotion is firstly introduced as the theoretical basis for the emotion model used in emotion recognition, including appraisal theory, dimensional models of emotion, brain mechanisms and computing models. Then, the advanced technologies of dimensional emotion recognition from the artificial intelligence perspective, such as the speech emotion corpora, feature extraction, classification, are presented in detail. Finally, the challenges of dimensional emotion recognition are discussed and the workable solutions and future research directions are proposed.

Key words: dimensional emotion model; cognitive theory of speech emotions; affective computing model; emotion-related feature extraction from speech ; speech emotion recognition algorithms

* 基金项目: 国家重点研发计划项目(2018YFC0806800),国家自然科学基金项目(61671187),深圳市基础研究项目(JCYJ20180507183608379),深圳市创新环境建设计划重点实验室项目(ZDSYS201707311437102),语言语音教育部-微软重点实验室开放基金资助项目(HIT.KLOF.20160xx),应用基础研究项目(CJN13J004)

Foundation item: National Key Research and Development Program of China (2018YFC0806800), National Natural Science Foundation of China (61671187), Shenzhen Foundational Research Funding (JCYJ20180507183608379), Shenzhen Key Laboratory of Innovation Environment Project (ZDSYS201707311437102), Open Funding of MOE-Microsoft Key Laboratory of Natural Language Processing and Speech (HIT.KLOF.20160xx), Applied Basic Research Programs(CJN13J004).

收稿时间: 2019-06-30; 修改时间: 2019-09-24; 采用时间: 2020-05-04; jos 在线出版时间: 2020-05-26

1 研究背景

情感是人类智能的重要组成部分,使计算机拥有情感,像人一样识别和表达情感仍是一个亟需解决的问题.Picard 提出情感计算的概念,开辟了计算机科学的新领域.目前情感识别的研究主要集中在语音情感识别、基于人脸的情感识别、文本情感识别、肢体行为情感识别.语音是人类交流情感和思想的最自然、最有效的方式之一[1],是人类生存和社会活动极其重要的信息传递和情感表达交流的方式.语音是人的发音器官发出的具有一定社会意义的声音,是表示语言的声音符号,不仅承载了语义信息而且包含与情感相关的声学信息,如音高、响度、韵律、音色等[2].语音的情感信息包含在声学参数随时间的变化中,如基频、能量、频谱、语调变化等[3-5].与基于人脸的情感识别相比,语音信号具有时序性,承载丰富的上下文信息.与文本相比,语音可以通过声学属性改变情感强度.肢体行为情感交互涉及较多的心理学范畴, 表达情感时存在较大的模糊不确定性,在特征提取与情感分类方面仍面临较大困难,应用较少.

语音情感识别研究已有 30 余年的历史,吸引世界范围内的研究单位、学者们的重点研究.如美国 MIT 多媒体实验室以 Picard 教授带领的情感计算研究组(https://affect.media.mit.edu/),研究方向包括多维信号建模,计算机视觉及模式识别,机器学习,人机交互和情感计算等.Picard 的《Affective Computing》开创了计算机科学和人工智能学科的新分支-“情感计算”;德国奥格斯堡大学 Björn Schuller 团队,长期致力于人工智能、音频识别、情感计算、机器学习的相关算法和研究领域,其开发的 OpenSMILE 情感特征提取工具被广泛应用;微软 Microsoft 研究院研究员利用 CNN, RNN,LSTM 等多种深度学习方法检测语音信号中的情感信息;南加州大学的 Jonathan Gratch 教授的研究方向主要包括虚拟机器人以及情感计算模型,以及研究认知与情感的关系,SAIL(Signal Analysis and Interpretation Laboratory)实验室研究以人类交流为核心的信号及信息处理技术,包括行为信号处理、情感计算、多模态信号处理、计算多媒体智能、计算语音科学等;卡内基梅隆大学的人机交互研究(https://hcii.cmu.edu/research/audio-emotion-recognition),将提出的两阶段分层语音情感识别方法(two-stage hierarchical classification approach)应用于中风康复治疗虚拟教练中,建议患者是否该休息,是否进行不同的锻炼;Virginia Affective Neuroscience Laboratory 研究设计情感的神经科学机制研究、行为学研究、情感健康研究,旨在为人类情感研究提供基础的理论研究,利用 EEG 脑电图分析、fMRI 成像技术研究人类大脑对情感的处理机制,为推动情感识别、情感计算等的发展提供认知理论支撑及指导.瑞士情感中心(Swiss Center for Affective Sciences)是一个跨学科研究中心,研究重点为情感或情感科学,涉及到认知神经科学、心理学、语言学、情感计算领域.除此之外,日本北陆先端科学技术大学院大学、新加坡南洋理工大学、新加坡国立大学、新加坡资讯通信研究院、爱尔兰都柏林圣三一学院、英国格拉斯哥大学、德国帕绍大学、加拿大滑铁卢大学、美国德克萨斯州大学等国际众多院校或机构致力于情感智能相关领域的研究.

国内也有越来越多的科研单位加入该领域的研究,如中科院自动化研究所主要研究听觉模式的分析和理解,包括情感交互技术等;清华大学多年从事语音信号处理方面的研究,开发的“汉语文语转换系统 Sonic”在文本分析、韵律模型、合成语音的自然度方面有重要突破;东南大学从事语音信号处理、情感信息处理等研究,在汉语连续语音韵律特征、F0 的生成模型、声调处理、语音信号中的情感信息处理等方面取得了一些有价值的研究成果;天津大学在语音识别、对话、言语认知脑机理、言语理解、情感计算等领域的研究成果也均处于领先地位;哈尔滨工业大学在语音情感识别、情感大脑认知领域进行深入研究等;浙江大学与阿里巴巴建立前沿技术联合研究中心,在人工智能、情感计算及跨媒体分析等领域取得很好成果,并联合发布‘懂情感’人工智能系统 Aliwood,可以为视频所配音乐建立情感模型.除此之外,北京邮电大学、电子科技大学、大连理工大学、华南理工大学、中国科学技术大学、山东大学、西北大学、南京邮电大学、太原理工大学等都在语音情感识别或多模态情感识别领域做出重要贡献.

近几年来,随着研究者对人工智能领域的关注,越来越多的会议与竞赛也进一步推动了情感识别研究的发展.语音识别领域顶会 INTERSPEECH 和 ICASSP 每年都有语音情感识别的议题,2016 年举办了第六届音/视频情感大赛(the Audio/Visual Emotion Challenge and Workshop ,AVEC 2016)[6],2017 年召开第一届国际情感计算与情感识别大会(1st Int. Workshop on Affective Computing and Emotion Recognition, ACER 2-17),会议议题涵

盖了情感计算的方方面面.2018 年 ACM 多模态交互国际会议(ACM International Conference on Multimodal Interaction ,ICMI)中的 Emotion Recognition in the Wild(EmotiW)竞赛[7]包括音视频情感识别子任务.国内也召开了该领域相关会议,2016 年全国模式识别学术会议的特殊议题即为第一届多模态情感识别竞赛(MEC 2016)[8],该竞赛包括音频情感识别、表情识别和音视频融合的情感识别三个子任务,选用 CHEAVD(CASIA Chinese Emotional Audio-Visual Database)作为数据库,国内外共 43 个团队参加,爱奇艺媒体智能组通过迁移学习的方法在 8 类音频情感识别任务中取得最高识别率 44.22%,会议针对情感语料库建立、情感识别方法及应用展开深入讨论,促进了整个领域的发展.2017 年开展了第二届多模态情感识别竞赛(MEC2017)[9],目标是提高真实环境下的情感识别性能,数据库采用 CHEAVD 的扩展版 2.0,促进了汉语多模态情感识别的研究.2018 年 5 月首届亚洲情感计算学术会议(ACII Asia 2018)在中科院自动化所召开, 围绕情感计算与智能交互进行探讨:情感认知、情感识别与生成、情感交互界面与系统、情感表达评价、情感对话系统、情感代理与机器人等, 是首个聚焦跨学科情感计算的亚洲论坛.

2018 年中国科协发布了 12 个领域 60 个重大科学问题和工程技术难题,其中信息科技领域的‘人与机器的情感交互’位列其中,‘无情感不智能’已成为众多研究者的共识.如何赋予机器人“情商”,使之具有情感处理能力就成为了服务机器人领域当前亟待突破的方向.目前美国、日本、德国、中国等纷纷开展了情感机器人的研究,而识别情感则是实现情感交互的第一步.

语音情感识别的研究涉及诸多学科,例如神经科学、心理学、认知科学、计算机科学等.情感理论是研究语音情感识别的基础,人类情感极其复杂,心理学领域已产生众多情感理论来解释人类情感[10, 11].目前,基于语音的情感识别技术常用的情感理论模型有两种:一种是离散情感模型,定义几种“基本情感”,其他情感由“基本情感”不同程度修改和组合[12].该模型虽然简洁但对情感的描述能力有限,很难准确地描述自发情感.另一种是维度情感模型,把情感看作是逐渐的、平滑的转变,不同的情感可以映射到高维空间上的一点[13].近年来该领域的研究也明显地呈现出了由离散情感模型发展到维度情感模型的总体趋势[14-16].本文将首先从情感的心理学研究基础展开,介绍情感的评估理论与维度情感模型;在语音情感的认知学研究进展方面,将综述包括语音情感的大脑处理机制、情感计算模型以及脑启发的情感识别算法;在语音信号分析方面,将着重介绍语音维度情感识别技术,包括语音音频信号预处理方法、特征提取方法以及情感预测算法,语音情感识别技术实现所需要用到的算法实现工具,最后分析了该领域存在的问题,并提出今后研究的关键问题.

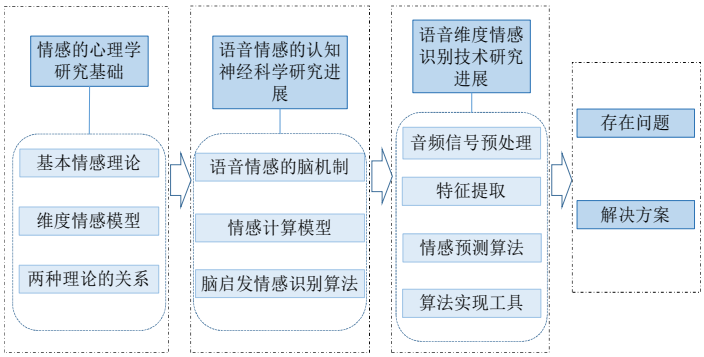

Fig.1 The survey framework of Speech Dimensional Emotion Recognition

图 1 语音维度情感识别研究综述框架

2 情感的心理学研究基础

2.1 基本情感理论

基本情感理论认为情感具有原型模式,既存在数种基本情感类型.该理论将情感分为基本情感(basic/primary/fundamental emotions)和次级情感(non-basic/secondary emotions).基本情感固化在人类神经自主系统之中,每类基本情感对应一个独特的、专门的神经通路,能以特定的方式推动对他人和情境做出反应,如语言声调、面部表情、身体姿态等.次级情感是根据情感的调色板理论[17],由基本情感混合而成.这些情感的表达方式具有跨文化差异,其表达方式由社会化过程所决定.Izard 把次级情感分为 3 类:第一类是由 2～3 种基本情感混合组成;第二种是基本情感与内驱力的混合;第三种为基本情绪与认知的组合.

基本情感的定义往往利用情感评估模型.情感是在比较个人需求与外部要求过程中诱发的,反映个人与环境的关系,可按照一套标准来描述或评估,这套标准叫做评估变量(例如 likelihood, desirability, unexpectedness, controllability, urgency, future expectancy)、检查项或评价维度.

1)Scherer 成分处理模型

1984 年,日内瓦的瑞士情感科学研究中心的心理学教授 Scherer 提出情感成分处理模型[18](component process model),将情感定义为产生认知活动(Cognitive component)、调控过程(Peripheral efference component)、行为动机(Motivational component)、行为表达(Motor expression component)以及个人情感状态(Subjective feeling component)的过程,情感表达是情感过程的成分表达,通过评价结果进行模式化.Scherer[19]在后续研究中,指出当人类接触到事件后,会产生简单、原始的动机趋力,可通过含义评估(implication appraisal)检验事件的起因与可能带来的影响、应对评估(coping appraisal)检验自己控制该事件的能力有多少,或是当无法控制它时有多少调整的空间和标准显著度评价(normative significance appraisal)评估上述处理结果与自我道德规范标准或社会道德规范标准之间的一致性)对该动机趋力进行评估与调整.

2)OCC 情感模型

在评估理论中最有影响力的是 1990 年 Ortony、Clore 和 Collins 提出的 OCC 模型[20].OCC 情感模型是早期对人类情感研究提出的最完整的离散认知情感论模型之一,也是第一个以计算机实现为目的发展起来的模型.OCC 模型定义了 22 类基本情感种类的形成规则以及三个层级(事件 events、智能体 Agents、目标 object),通过以下五个步骤实现从最初事件的分类到产生个性行为的完整系统:1) 对事件、行为或目标进行分类;2) 量化受到影响的情感的强度;3) 新产生情感与已存在情感的相互作用;4) 将情感状态映射到某种情感表达;5) 对情感状态进行表达.

3)Roseman 评价理论

1996 年,美国罗格斯大学心理学教授 Roseman[21]提出了具体的事件评价因素和执行计算的框架结构,通过它们的相互作用来推断所合成的情感.评价因素分为意外、动机、情境、可能性、控制度、事件引发原因及问题类型,其中,动机与控制度是评估情感的最重要两个因素,如当情境与主体的目标不一致时,常诱发消极情感,例如生气或者后悔.他根据这七种评价因素给出 17 种基本情绪,其中积极情感(动机一致)包括希望、高兴、安慰、喜欢、自豪,消极情感包括生气、轻视、恐惧、悲伤、悲痛、厌恶、挫折、遗憾、内疚、羞愧,某些情感,如欲望、惊讶可根据事件引发原因决定积极情感或消极情感.Roseman 所提出的基于事件评价的情感模型,形成了一个较为完善的理论体系.

目前研究者们对基本情感尚未达成共识,大部分观点认为存在六种基本情感:恐惧、高兴、愤怒、厌恶、悲伤和惊奇,Ortony 和 Turner 将这些观点整理如下表 1[12].

基本情感理论借助情感评估模型以不同的方式解释情感是如何产生以及演变的,社会心理学研究者利用这种理论解释和预测人对事件的反应机制以及情绪模式.评估模型主要用于情感建模与合成,如文献[22, 23]利用 OCC 模型合成情感,且在机器人情感研究中广泛应用,设计不同个性的情感机器人[24-26].评估模型基于离散情感描述模型,可表达的情感类别有限,且有些情感类别非常相似,以至于环境很难触发这些情感[27].

Table 1 Basic emotion theories[12]

表 1 研究者们提出的基本情感理论[12]

研究者	基本情感
Arnold	Anger(生气), aversion(厌恶), courage(勇敢), dejection(沮丧), desire(渴望), despair(绝望), fear(恐惧), hate(讨厌), hope(希望), love(爱), sadness(悲伤)
Ekman, Friesen,&Ellsworth	anger, disgust(厌恶), fear, joy(高兴), sadness(悲伤), surprise(惊讶)
Fridja	desire, happiness(开心), interest(喜爱), surprise, wonder(惊奇), sorrow(懊悔)
Gray	Rage(愤怒) and terror(恐怖), anxiety(焦虑), joy
Izard	anger, contempt(轻视), disgust, distress(悲痛), fear, guilt(内疚), interest, joy, shame(羞耻), surprise
Plutchik	acceptance(容忍), anger, anticipation(期望), disgust, joy, fear, sadness, surprise
Oatley & Johnson-Laird	anger, disgust, anxiety, happiness, sadness
Panksepp	expectancy(期望), fear, rage, panic(恐慌)
Tomkins	anger, interest, contempt, disgust, distress, fear, joy, shame, surprise

2.2 维度情感模型(Dimensional emotion model)

任何情感发生时,在某一属性或特性上可以有不同的幅值,情感维度就是对情感某种属性的度量,维度具有极性.情感维度理论认为情感状态不是独立存在的,多个维度构成了人类情感空间,不同情感之间是平滑过渡的,利用维度空间中的距离可以表示不同情感的差异度与相似度.迄今为止,研究者提出的维度划分方法多种多样,并没有统一的标准评测哪种维度划分方法更好.典型的维度理论有:

1、Wundt 的情感三度说

Wundt 在 1863 年提出情感的维度理论[28],认为情感由愉悦(pleasure)-不愉悦(displeasure)、激动(excitement)-平静(inhibition)和紧张(tension)-松弛(relaxation)三个维度组成,每一种特定感情都是这三个维度以不同方式的独特组合.在一个特定的时间里作用于意识的感情总和被称之为总体感情(total feeling),它是同时存在的不同性质的器官感受的总和,它们结合起来,形成一个具有确定性质和强度的感情特征的组合体.从感情与观念的关系来看,感情可以看作是伴随观念形成的一种过程,某一时刻的情感在三维情感空间中表示为一个独立的点,当对具体事件作出反应时,情感可以表示成一条轨迹,一般情况下,轨迹的起始和重点都位于原点(如图 2 所示).

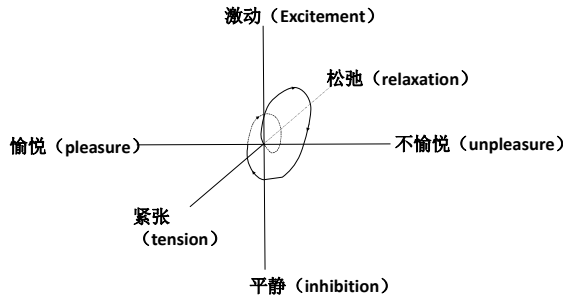

Fig2 The three principal axes of emotion space[28]

图 2 Wundt 理论中的情感轨迹[28]

2、Schlosberg 倒圆锥三维情感空间

Schlosberg[29](1954)对 Wundt 理论中的激动-平静维度进行了进一步研究,提出了激活度的概念,并通过对面部表情的情感分类研究,提出了由愉悦度、注意度、激活水平三个维度构成的倒立圆锥形情感空间模型,圆锥切面的长轴代表了情感的愉悦度变化,短轴代表了情感的注意度变化,垂直于椭圆面的轴表示激活度强度变化(如图 3 所示).Schlosberg 提出,与愉悦情感相比,不愉悦的情感具有更高的激活度.

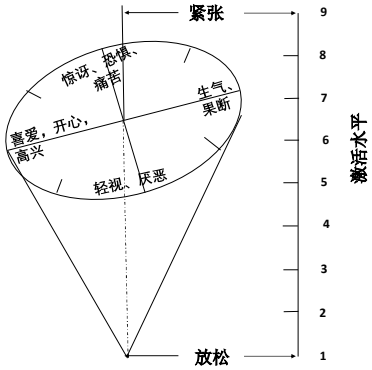

Fig3 Schlosberg's three dimensional emotion model[29]

图 3 Schlosberg 提出的倒圆锥三维情感空间模型[29]

3、PAD 情感空间模型

Russell & Mehrabian[30]于 1977 年利用回归分析的方法研究愤怒(anger)和焦虑(anxiety)情感,发现愤怒和焦虑都具有高激活度和低愉悦度,但两者的优势度(dominance)明显不同,愤怒具有控制倾向,焦虑具有服从倾向.结合先前的研究,他们提出了 PAD 维度模型.该模型简洁且相对完善,通过 SAM(self assessment manikin)量表可以快速测定个体的情感状态,因此被人工智能领域广泛认可.PAD 模型由三个维度组成:

1)P 代表情感的愉悦度维度(Pleasure-Displeasure):表征情绪状态的正负性,已通过脑成像研究证实了愉悦度维度.

2)A 代表情感的唤醒度/激活度维度(Arousal-Nonarousal):表示情绪生理激活水平和警觉性.

3)D 代表情感的优势度维度(Dominance-Submissiveness):该维度反映在相对动机的比较中,表示情绪对他人和外界环境的控制力和影响力.

4、Plutchik 抛物锥情感空间模型

Robert Plutchik 于 1984 年提出 8 种基本的“两极”情感:高兴—悲伤;愤怒—恐惧;厌恶—信任;惊奇—期望[31].类似于三维颜色表达空间,利用强度、相似性和两极性三个维度来描述情绪模型,基本情感可以表达为不同的强度,基本情绪相互混合演化出次级情感.Plutchik 采用倒锥体来描述情绪三个维度之间的关系.上述 8 种基本情绪组成了椎体的截面(如图 4 所示),相邻位置的情绪相似,对角位置的情绪对立,锥体自下而上表明情绪强度由弱到强.该模型的优点在于清晰地界定情绪,并将情绪的相似性与对立性很形象地表达.Plutchik 的情感结构理论与 Schlosberg 的情感模型相似,都将激活度与颜色强度进行对比,但 Schlosberg 提出的锥形情感空间未提出基本情感,而是从理论上推导出三个维度.

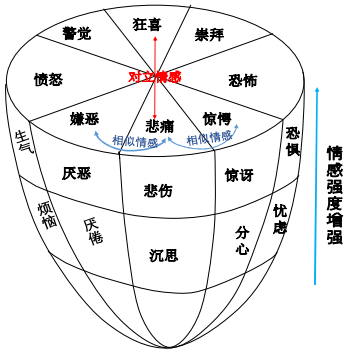

Fig.4 Plutchik's three-dimensional structural model of emotions[31]

图 4 Plutchik 提出的情感三维结构模型[31]

5、Russell 的愉快度和强度环形模型[32]

Russell 他的后续研究表明 Schlosberg 所提出的注意-拒绝和激活度是很难区分的.于是他于 1980 进一步研究了情感的环状模式,提出了二维情感描述模型：愉悦度和强度.

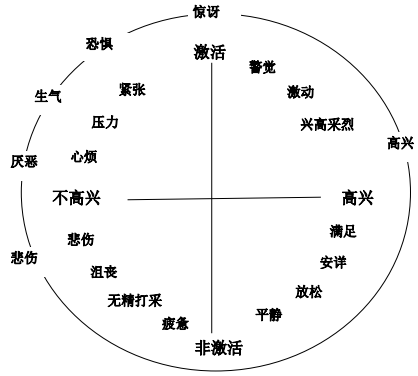

Fig5 Russell's circumplex model of core affect

图 5 Russell 提出的二维情感环形模型[32]

6、情感的高维空间模型

由于情感空间维度的数量没有定论,所以部分学者根据自己的研究提出了高维空间模型.1974 年 Krench[33]利用强度、紧张水平、复杂度和快乐度四维模型来评定人体所处的情感状态；1991 年,Izard[34]提出的四维度分别是愉悦度、紧张度、激动度和确信度,并编制了情感维度量表(DRS、DES)对情感体验的评定比较准确.Frijda 也根据自己的研究提出六维情感模型,分别是愉悦度、激活度、兴趣度、惊奇度、复杂度、社会评价.

2.3 离散情感描述模型与维度情感描述模型的关系

尽管情感层次理论与维度空间理论分别利用不同的方法描述情感,但是两者之间并不是对立的,而是可以相互转换的.维度理论利用欧氏空间描述情感,坐标轴的不同取值组合表示一种特定的情感状态,但基本情感可以通过一定方式映射到情感空间中.Mehrabian[35]利用个性(personality)代表长期的情感,采用开放性(openness)、尽责性(conscientiousness)、外向型(extraversion)、亲和性(agreeableness)和情绪稳定性(neuroticism)五大特质来分析个性,并研究了五个特质与 PAD 空间模型的内在关系,提出了利用五个特质预测 PAD 值的方法.基于 Mehrabian 的理论,Gebhard[36]将 OCC 理论中的基本情感映射到三维 PAD 情感维度模型,见表 2.Becker-Asano[37]根据情感的动态理论,提出了将基本情感向 PAD 模型映射的方法.

李海峰,韩文静[38]在对语音情感识别综述中对比了离散情感描述模型与维度情感描述模型的优缺点,离散描述模型虽然较为简洁,但只能刻画有限种类的情感类型,其情感描述能力显示出较大局限性.维度模型很好地化解了这一问题,利用维度空间精确地量化情感,减小情感标签的模糊性,具有无限的情感描述能力,更利于自发情感的描述,近年来受到越来越多的关注.

Tabel2 Mapping of OCC emotions into PAD space[36]

表 2 OCC 基本情感与 PAD 维度空间的映射[36]

情感	P	A	D	情绪象限
羡慕(Admiration)	0.5	0.3	-0.2	+P+A-D 依赖的(Dependent)
感谢的(Gratitude)	0.4	0.2	-0.3	
喜欢(Liking)	0.4	0.16	-0.24	
希望(Hope)	0.2	0.2	-0.1	
生气(Anger)	-0.51	0.59	0.25	-P+A+D 敌意的(Hostile)
不喜欢(Disliking)	-0.4	0.2	0.1	
厌恶(Hate)	-0.6	0.6	0.3	
失望(Disappointment)	-0.3	0.1	-0.4	-P+A-D 焦虑(Anxious)
恐惧(Fear)	-0.64	0.6	-0.43	
同情(Remorse)	-0.3	0.1	-0.6	
羞愧(Shame)	-0.3	0.1	-0.6	
悲痛(Distress)	-0.4	-0.2	-0.5	-P-A-D 无聊的(Bored)
遗憾(Pity)	-0.4	-0.2	-0.5	
FearsConfirmed	-0.5	-0.3	-0.7	
怨恨(Resentment)	-0.2	-0.3	-0.2	
满意(Gratification)	0.6	0.5	0.4	+P+A+D 兴高采烈(Exuberant)
高兴(Happy For)	0.4	0.2	0.2	
快乐(Joy)	0.4	0.2	0.1	
热爱(Love)	0.3	0.1	0.2	
自豪(Pride)	0.4	0.3	0.3	
安慰(Relief)	0.2	-0.3	0.4	+P-A+D 放松的(Relaxed)
满足(Satisfaction)	0.3	-0.2	0.4	
耻辱(Reproach)	-0.3	-0.1	0.4	-P-A+D 轻蔑的(Disdainful)
心满意足(Gloating)	0.3	-0.3	-0.1	+P-A-D 温顺的(Docile)

3 语音情感的认知神经科学研究进展

3.1 情感的神经机制

情感产生的脑机理研究经历了一个较长的过程,受到神经解剖学、神经生理与认知心理学等相关科学发展的影响.思想家和科学家对情绪奥秘的探讨可以追溯到古代的臆测和神秘主义.直到文艺复兴以后,如霍布斯(Hobbes)、洛克(Locke)、笛卡儿(Descartes)等带有唯物主义色彩的哲学家才把知觉、思维、知识、情绪等和神经与脑的活动联系了起来.1872 年,达尔文(Darwin)在《人和动物的表情》一书里论述了情绪的生物学基础,强调了环境对情绪行为的作用,形成了情绪生理心理学的雏形.其后的詹姆斯(James)提出了最早的情绪生理—心理学理论,为探讨情绪的性质指出了一条必由之路.James-Lang 理论(1885 年),即情绪外周理论,强调情绪的产生是植物神经系统活动的产物.1912 年,Mills 首次提出了情感的大脑右半球假说,右脑更多地决定了人的空间感、抽象思维、音乐感与艺术性.1931 年,Cannon 提出情绪的丘脑学说,认为丘脑对情绪调节起着重要作用.随后,Papez 提出了 Papez 环路理论,认为下丘脑是情绪表达中心,边缘系统是情绪体验部位.但当时这一回路并没得到科学研究证实.Maclean 于 1952 年提出情绪脑的概念,划分了较为精细的情绪相关脑区网络,得到研究者的

广泛认同.

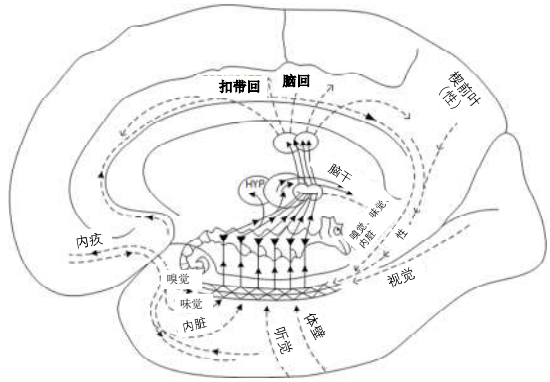

Fig 6 MacLean's limbic system theory of functional neuroanatomy of emotion[39]

图 6 MacLean 提出的边缘系统理论[39]

上世纪 60 年代随着情绪生理—心理学的发展,形成了诸多情绪理论学派：阿诺德(Arnold)的评价—兴奋论[40],认为情绪的发生决定于对感觉刺激的评估,而皮质兴奋是情绪行为的基础.普里布拉姆(Pribram)的"不协调"论[41],把大脑高级中枢实现的认识活动与情绪联系起来.20 世纪中叶的信息革命,导致了认知心理学的建立,把人脑理解为一个信息加工系统,形成了情绪的信息加工论.拉扎勒斯(Lazarus)的认知—评价理论[42],从心理学的角度填充了信息加工过程的心理内容,着重于外界刺激与行为反应之间的认知评价环节,丰富了脑内信息加工的内容.LeDoux[43]根据神经生理学上的研究提出,边缘系统对听觉刺激引起的情感响应起着至关重要的作用.边缘系统负责处理情感刺激,主要包括四部分：感觉皮层、丘脑、眼眶额叶皮层、杏仁体[44, 45].随着脑成像技术的发展,研究者对情感的大脑活动的研究也越来越精确.2004 年,Florin 利用 fMRI 对不同唤醒度、效价度情感刺激下前额叶皮层活动进行研究,实验结果说明前额叶皮层(PFC)左侧对效价度积极的情感反应更活跃,背外侧 PFC 对唤醒度更加敏感.2005 年,LeDoux 与 Phelps[46]研究了动物模型及人类行为中杏仁体对情感处理的作用.2008 年,Mathersul[47]研究了脑电信号 EEG 的 alpha 波段与悲伤生气情感的关系.2014 年,康奈尔大学神经学家 Adam Anderson[48]研究眼窝前额皮层的精细神经活动模式,发现虽然情感是个人的和主观的,但是人的大脑会把它们转换成一个标准的代码,这个代码客观地代表着不同感官、情况甚至人的情感.2018 年,Kirkby[49]等利用半慢性颅内脑电图(iEEG)记录边缘系统的多位点,并周期性的评估被试的情绪,研究情绪和焦虑的神经编码,并揭示一个生物指标,有助于诊断和治疗情绪和焦虑障碍.

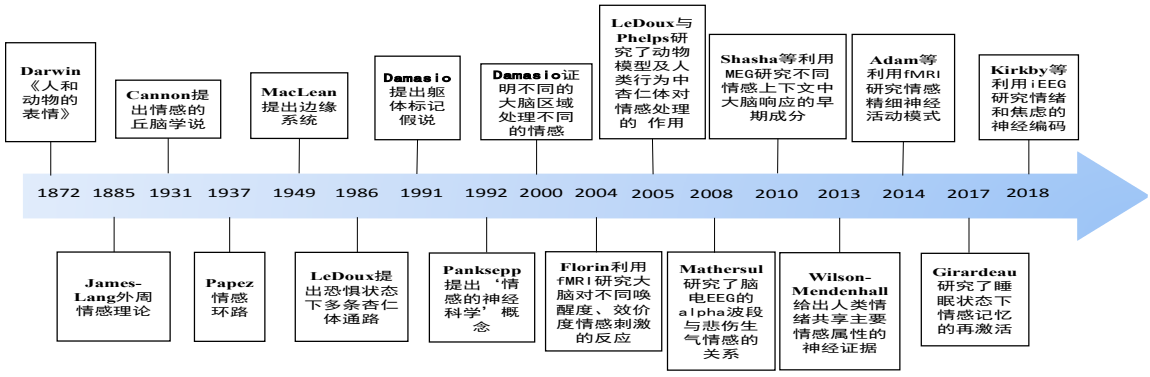

Fig 7 The timeline of historical milestones in understanding the emotional brain

图 7 情感大脑研究的重要里程碑工作

近年来功能性磁共振成像 fMRI(functional magnetic resonance imaging)技术与脑电图 EEG

(Electroencephalography)技术为人类情绪的中枢神经机制研究提供了大量的研究证据,初步揭示了人类情绪管理过程中大脑的区域功能和神经机制(图 8):1、情绪感知:枕叶加工视觉信息,顶叶进行躯体感觉整合和空间视觉整合,颞叶进行听觉性言语功能处理,岛叶接受来自内脏和躯体状态改变的感知信号;2、认知评价:眶额皮层、腹内侧前额皮层对情绪信息进行高级再加工,完成对情绪刺激动机意义的评价;3、主观调整:前部扣带回负责情绪加工中的冲突监控;杏仁核通过与海马系统的相互作用,可以使情绪性事件的陈述性记忆变得更加巩固;4、自主活动:颞上回与社会性情绪相关,完成对精细感觉的加工;后扣带皮层与评断道德价值有关;5、外显行为:脑干和下丘脑调节情绪活动中的躯体与自主反应,实现人类的情感行为表达.

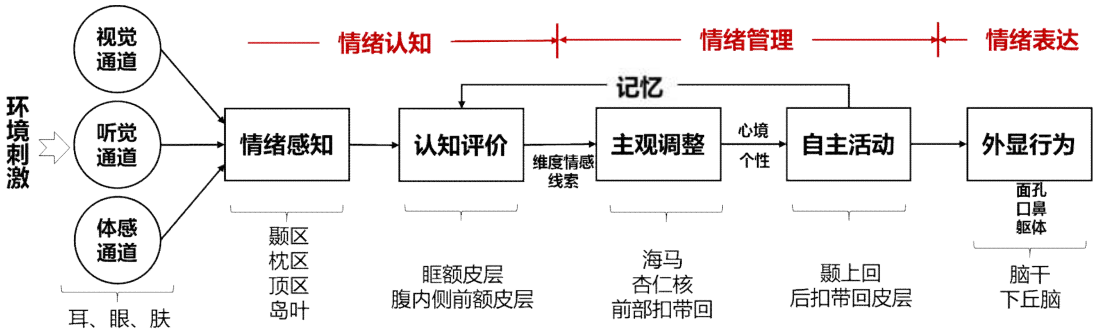

Fig 8 Emotional experience

图 8 人类情绪管理系统示意图

在情绪神经机制研究方面,Lindquist[50]对比了两种情感加工脑机制的研究方法,一种方法是 Locationist 方法,该方法假设离散的情感类别是由其对应的不同脑区产生,例如,恐惧对应于杏仁核(amygdala)的激活,厌恶对应于脑岛区(insula) 的激活,生气对应于眶额叶皮层(orbitofrontal cortex, OFC)的激活,悲伤对应于前扣带皮层(anterior cingulate cortex, ACC)的激活.另外一种方法是心理学建构论方法(psychological constructionist approach),该方法假设情感状态是由大脑功能网络的相互作用形成,杏仁核、脑岛、腹内侧眶额皮层、前扣带皮层、丘脑都参与多个主要情感的形成.Lindquist 等人通过对大量人类情感的神经影像学文献的总结,认为更多地证据与构建论一致,不同的大脑区域相互作用共同参与情感的体验与感知.

更具体地,大脑如何处理语音情感也是听觉语言处理研究的一个热门课题.语义信息以及韵律线索对语音情感的理解起着重要作用.有研究表明大脑右半球负责处理情感韵律信息[51-54],但实验的任务类型或者被试默读复述也可能引起双边激活模式.Ross[55, 56]的偏侧性假设认为无论情感激活度如何,大脑右半球在处理情感语音时更具有优越性.与之相比,激活度假设[57]认为大脑左半球对积极情感具有主导性控制,右半球主要控制消极情感.由于韵律信息随着声学参数变化,如基频 f0,强度以及时长等,Zatorre[58]提出右半脑负责基频信息的感知,左半脑处理强度以及时长信息.文献[59-62]利用 fMRI 技术研究语音情感表达时脑区的激活程度.Kotz[63]研究发现具体的语音情感表达由大脑的额叶-岛盖-颞叶(fronto-operculo-temporal)区进行编码,颞叶区负责副语言声学处理,额叶区进行情感评估,左侧颞叶-小脑(temporo-cerebellar)区负责时序处理,右侧额下回(inferior frontal)区分不同的情感表达.文献[64]研究发现通过情境上下文的学习,通过语义与非语言获得情感意图的途径一致.语境学习假设认为情感状态基于个人对该情感以往的学习经验,情感系统由预先定义的概念进行评估,然后根据经验进行精细处理.

3.2 情感计算模型

情感相关的认知神经科学的研究促进了情感计算模型的发展,产生了一系列能实现情感计算的系统.目前,较多的情感计算模型是基于情感认知理论,Elliott 实现了一个基于 OCC 模型的情感推理机(Affective Reasoner)系统[65],每一种情绪都由一组不同的认知导出条件通过推理得出.Reilly 和 Bates 实现了一个可以及时更新情绪状态的 EM 系统[66].Gratch 和 Marsella 将认知过程引入情感的研究,提出了一种能解释情感动态变化过程的 EMA[67]系统.MIT 人工智能实验室 Velasquez 提出了一种新的情感更新规则,由此开发了一个能够控制

各类情感现象的动态变化的 Cathexis 模型[68].

ALMA 多层次情感模型[36]利用 OCC Model 测量短期情感、PAD 情感量表中期情感(mood)以及五大人格特质来衡量长期情感状态,该模型对情感进行了更完整的定义,可以更自然地实现不同情感的语言或非语言的情感表达.Becker-Asano 提出了 WASABI[37]情感计算模型,该模型融合了基于维度情感理论的情感动态更新规则以及 OCC 情感评估理论,与其他基于 OCC 理论的计算模型相比,该模型建立了更加完整的反馈机制.Marsella[69]将情感计算模型总结如下图:

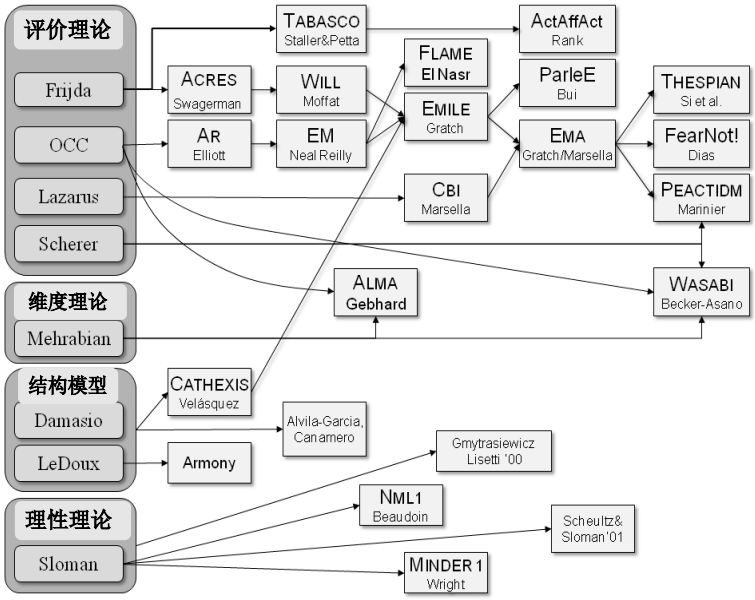

Fig 9 A history of computational models of emotion[69]

图 9 情感计算模型发展史[69]

3.3 类脑语音情感识别算法

听觉通路从听觉信息的感知、说话人识别、语音感知到言语生成分为不同的等级[70],语音进入左右耳蜗,耳蜗相当于一个滤波器组,将声音以时频谱的形式呈现,并以相应的神经电信号方式传递至低位脑干,低位脑干负责预处理、缩放和归一化,之后信号进入下丘脑、上丘脑和丘脑区,丘脑负责控制注意力,并产生信号传递至边缘系统和主要的听觉皮层.最后,经边缘系统和听觉皮层处理的信号再经过特定的通路进行语音识别、言语生成、说话人识别和音乐感知等[70].

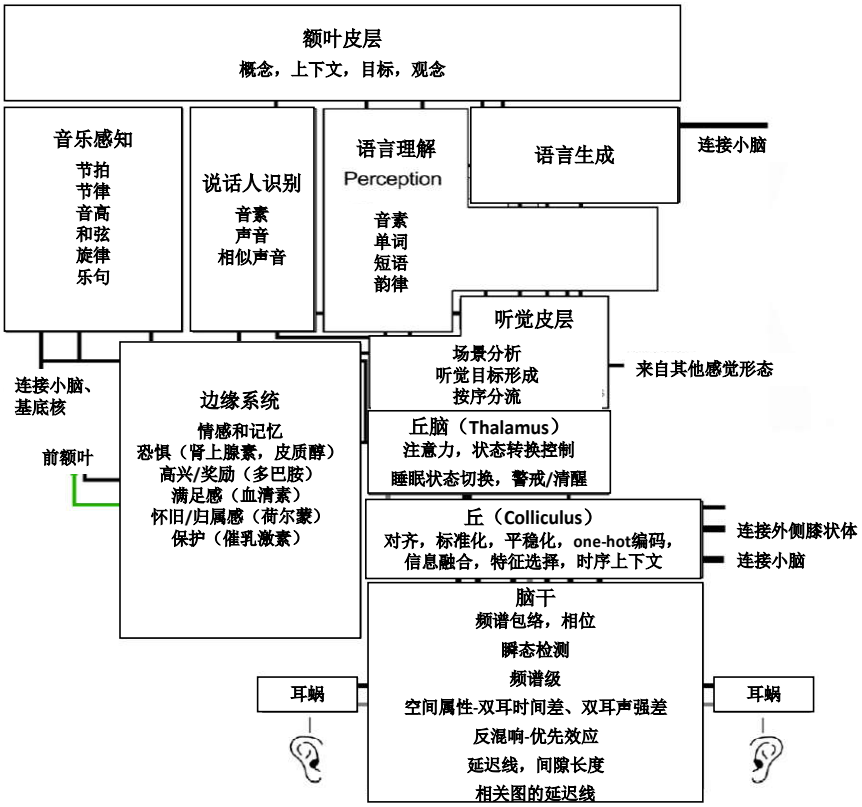

Fig 10 Block diagram of the Human Auditory Pathway[70]

图 10 人类听觉通路功能模型[70]

根据大脑边缘系统的结构,Moren 和 Balkenius 提出了大脑情感学习模型(Brain Emotional Learning model, BEL model)[45],对边缘系统四个部分之间的情感学习机制进行数学建模,采用一种基于奖励信号的强化学习方法调节模型参数,并通过实验证明 BEL 模型的输出对奖励信号有明显依赖性.该模型在混沌时序预测领域取得广泛应用[71-73],与神经网络模型相比,具有结构简单、计算复杂度低等优点；但是关于奖励信号的设定方法目前没有统一的规定.随后,出现了一系列优化 BEL 模型参数的研究,如 Lotfi[74]设计了竞争型 BEL 模型并采用遗传算法优化其参数,增强了其处理高维多分类数据的能力；Lucas[75]在 BEL 模型的基础上,利用感知输入与情感线索的行为产生机制,提出了 BELBIC 智能控制器,并将该控制器用于非线性系统中,验证了其具有很好的控制能力、抗干扰能力和系统鲁棒性.Parsapoor[76]利用模糊推理系统(Fuzzy Inference System)对 BEL 模型的杏仁体和眶额叶皮层模块进行优化得到 BELFIS 模型.S Motamed 等人[77]利用自适应神经模糊推理系统(Adaptive Neuro-Fuzzy Inference System,ANFIS)和多层感知器(Multilayer Perception, MLP)对 BEL 模型进行改进用于语音情感识别,并在 Berlin 语音情感数据库上进行实验,与 SVM、KNN、BEL、BELFIS、BELBLA 模型的实验结果进行了对比,提出的算法取得更高的识别率.

借鉴人类情绪机制的类脑情感计算研究已经开始,在人脑这个“巨象”上,研究工作者面临着如何深入解读大脑功能和揭示这个开放的复杂巨系统运行机制的挑战.

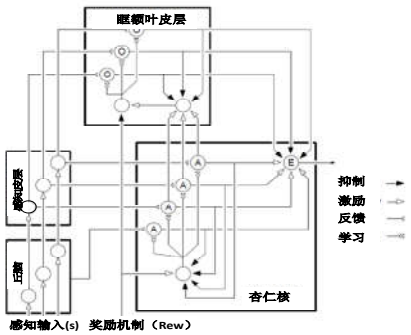

Fig 11 BEL model proposed by Moren and Balkenius[45]

图 11 Moren 和 Balkenius 提出的 BEL 模型[45]

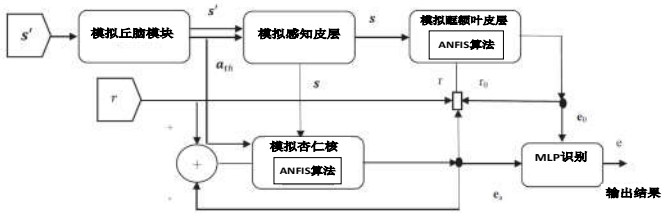

Fig 12 An optimized model of brain emotional learning (BEL) that merges the Adaptive Neuro-Fuzzy Inference System (ANFIS) and Multilayer Perceptron (MLP) for speech emotion recognition[77].

图 12 基于 ANFIS 和 MLP 改进的 BEL 模型用于语音情感识别[77]

4 语音维度情感识别技术研究进展

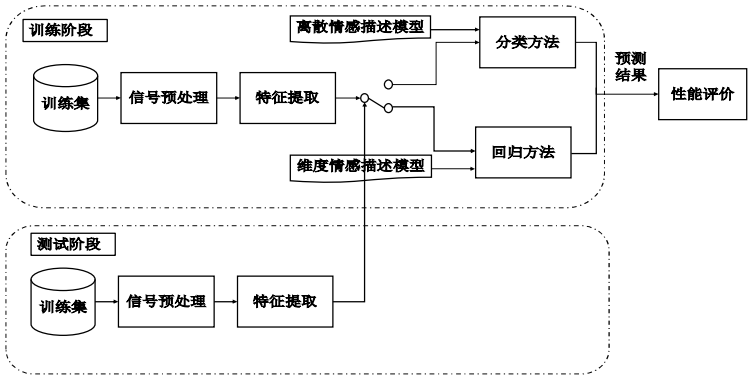

Fig 13 Framework of a speech emotion recognition system

图 13 语音情感识别系统框架图

语音情感识别系统是经典的模式识别系统,包括系统训练阶段和测试阶段.对于采集的语音信号均先进行预处理后,根据情感空间描述模型的不同,进行特征分析与识别任务技术设计.对于离散情感描述模型,语音情感识别任务可视为多分类问题,为样本预测离散型类别标签；对于维度情感模型,其任务可视为回归预测问题,为样本预测连续输出量的问题.分类问题与回归问题采用的建模方法以及性能评价指标不同.分类模型经常为输入样本预测得到与每一类别对应的像概率一样的连续值,这些概率可以被解释为样本属于每个类别的似然度或者置信度,预测到的概率可以通过选择概率最高的来转换成类别标签.回归预测问题预测的是情感在不同维度上的连续数值,其性能可以用预测结果中的错误来评价.在特定条件下,分类问题和回归问题是可

以相互转换的.如 Grimm 等在离散情感识别任务中,首先将提取的全局统计特征利用模糊逻辑系统(fuzzy logic system)映射到连续三维情感空间,再利用 KNN 识别为离散的七类情感[78].虽然 DNN 技术的广泛使用使得大量工作不需要进行数据预处理,但语音信号有着低信噪比的特殊性,众多学者对语音信号的预处理方法进行了大量研究.因此,在本文中,将就预处理技术、特征提取技术及分类器设计等方面进行综述.

4.1 语音情感特征提取

特征提取与处理是语音情感识别中重要的部分,特征集直接影响识别器的识别能力和鲁棒性.特征提取的目的是从语音信号中,提取一方面能表征不同识别单元的声学差异,另一方面有能表征相同识别单元不同样本之间的声学相似性的信息.

语音情感信息通过语义和非语义两种形式传递.语义信息以一定的语言规则(语法、修辞等)传递说话者的情感,非语义语音情感信息包括两种形式,情绪韵律(emotional prosody)[79]和非语言发声(non-linguistic vocalizations)[80, 81].

4.1.1 声学特征

人们可通过感知语音中的声学线索,从中提取出所携载的情感倾向.声学特征是独立于语言内容而传递的情感信息,不受文化差异的影响,对于不同语种的情感数据库均可通过提取声学特征进行情感识别[82-86].声学特征可分为LLDs特征(Low-level descriptors)和统计特征(functions),其中LLDs特征常常以帧为单位进行提取,可以从韵律特征、谱特征、音质特征对语音情感信息进行表达.统计特征一般是将LLD特征在独立的语句或单词上进行统计,包括极值、方差、峰度、偏斜度等.

4.1.1.1 LLDs 特征

(1)韵律特征

韵律特征被认为与发音单元(音节、单词、短语、句子)相关联的声学特征,又被称为“超音段特征”,在情感识别中应用非常广泛[87-89],主要包括时间特性、基频、能量等,被认为与情感的感知具有明显的关系.文献[90]得出韵律特征与唤醒度相关,音质特征与愉悦度相关的结论.Pereira[91]等人分析了语音韵律特征与情感维度的相关性,数据结果表明基音等韵律参数与维度空间中的唤醒度对应,一般认为音质参数与维度空间中的效价度对应[92].

近来研究者提出了一些新的韵律特征.Arias[93]利用函数型数据分析(FDA)建立中性参照模型,计算基音频率的主成分分析(PCA)映射矩阵作为每条语音的特征.具有高激活度的语音情感信号,其能量多集中在高频成分,低激活度的情感语音信号的基频较低[94].Sant’Ana[95]提出赫斯特指数(Hurst exponent)用于说话人识别,L.Zao[96]进一步提出pH时频声源特征与情感的愉悦度相关,取得了较MFCC、TEO-CB-Auto-Env特征集更高的识别率.Mencattini[97]提出了基于CQT的频域幅值包络特征,并结合能量、小波近似分量和细节分量、过零率、共振峰、TEO等特征,共520维特征用于维度情感识别.

(2)音质特征

音质特征描述声门属性, 语音的音质特征主要指的是具有不同情感状态的说话人发音方式上的区别.Scherer的情感成分处理模型提到音质特征影响情感的变化.Tato[98]等人探讨了情感维度对语音识别的贡献,研究发现音质类特征对于区分唤醒维接近而效价维远离的情感(生气和喜悦)有较好的效果.

M Borchert[92]将共振峰、不同频带的频谱能力分布、谐波噪声比、频率微扰和振幅微扰在内的音质特征用于效价度预测,将韵律学特征用于激活度预测,实验结果表明音质特征更适用于区分唤醒度相同、效价度不同的情感.I Idris[99]利用音质特征集、韵律学特征集以及二者混合特征集,选用多层感知器网络分别在柏林情感数据库上进行情感识别,平均识别率分别是59.63%、64.67%和75.51%.M Kachele[100]将谱特征、韵律学特征和音质特征用于表达语音的长时信息,并利用改进的前向选择/后向剔除算法进行特征选择,在公开的柏林情感数据库上进行测试,平均识别率为88.97%.

(3)谱特征

谱特征通常用来表示发声器官的物理特征,是信号的短时表示,一般认为在很短时间内(10～30ms)相对平

稳,可以通过某时刻附近一段短语音信号得到一个频谱.频谱表示频率与能量的关系,有助于更好地观察音素.常见的频谱图主要有线性振幅谱、对数振幅谱、自功率谱.谱特征主要有线性预测系数(Linear Predictor Coefficients,LPC),线谱对参数(Line Spectrum Pair, LSP),单边自相关线性预测系数(One-sided Autocorrelation Linear Predictor Coefficients,OSALPC)等.频谱图中的共振峰携带了声音的辨识属性,利用倒谱可以提取包络信息,得到共振峰用于识别.常见的倒谱特征有感知线性预测倒谱系数(Perceptual Linear Predictive Cepstral Coefficients,PLP),线性预测倒谱系数(Linear Predictor Cepstral Coefficients,LPCC),单边自相关线性预测倒谱系数(One-side Autocorrelation Linear Predictor Cepstral Coefficients,OSALPCC).考虑到人耳听觉系统响应不同频率信号的灵敏度不同,将线性频谱映射到基于听觉感知的Mel非线性频谱中,再进行倒谱转换,得到Mel倒谱系数(Mel Frequency Cepstrum Coefficients,MFCC).MFCC已广泛应用于语音识别、情感识别领域.

另外,最近研究者们也提出了一些新的谱特征.Huang[101]提出一种基于小波包的自适应滤波器组构建方法(Wavelet Packet Cepstral Coefficients, WPCC),对 MFCC 有很好的扩展作用,而且可以利用2D的小波包进行图像处理,适用于语音视觉多模态情感识别系统.M Ziolko[102]提出了Fourier-Wavelet特征提取方法,首先对语音信号进行小波变换,然后再进行傅里叶变换.Inshirah Idris[103]提出两种谱特征优化方法,一种方法是基于离散谱特征的优化,一种是融合谱特征,利用这两种优化方法得到的特征集合进行情感识别,识别率较优化前分别提高2%和4%.Espinosa[104]等在VAM数据集上测试了韵律学特征集合、音质特征集、谱特征集对PAD维度空间识别率的影响.KunxiaWang等[105]提出新颖的傅里叶参数模型组合傅里叶参数及其一阶、二阶差分用于语音情感识别,并利用提出的特征与MFCC结合提供了说话人独立的语音情感识别.Sayan Ghosh[106]等从语音信号及声门流量信号中提取频谱图,利用堆叠的自编码方法进行频谱图编码,最后利用RNN进行四类情感识别,采用基于声门流量信号的表征学习模型与基于效价度和唤醒度分类训练的迁移模型来提高RNN训练效率,实验结果显示表征模型与迁移模型的加入可以提高1.17%的识别率.

4.1.1.2 统计特征

进行语音情感识别时,帧特征往往不直接作为网络输入进行学习,而是利用这些特征的一些统计值进行神经网络训练.表3给出了常用的统计特征.Schuller等[107]在一个AVIC(Audiovisual Interest Corpus)语料库上分别利用帧特征和全局统计特征进行语音对话兴趣识别,他们首先提取了包括基频、能量、MFCC、共振峰、频率微扰、振幅微扰、谐噪比等37维LLD特征曲线,然后统计出每条曲线的最大值、最小值、均值、方差、峰度、偏斜度等共19维全局特征统计值,最后分别利用MI-SVM(Multi-Instance learning-SVM)和 SVM 对LLD特征和统计特征进行兴趣识别,定量对比其识别准确,实验结果表明基于统计特征的识别结果比帧特征的识别结果更加准确.

情境上下文对情感的识别具有关键性作用,长时统计特征在区分高激活度和低激活度情感语音的效果较好,但是对激活度相同情感的区分能力较弱,如很难区分具有相同激活度的生气和欢乐情感语音.具有时序信息的帧特征在区别效价度不同的情感语音[108].

目前已有少量文献尝试选取不同窗长来提高情感识别率,但存在的文献没有统一的答案.Origlia[109]认为目前特征提取方法是基于整个语音信号,没有考虑语音内容的变化,这与韵律研究的理论基础是矛盾的.并以此提出一种基于音节的特征提取办法,同时考虑音节核,可以减少信息的处理量.Sethu[110]认为帧特征和全局统计特征不足以全面的表征情感的时序信息,因此提出以段为单位的特征提取,可通过基音频率和前三个共振峰的轮廓进行提取,将该特征与短时帧特征和全局统计特征融合可以提高情感识别率.李海峰[111]等人使用"语段特征"用于识别,并给出了各类情感状态对应的"最佳识别段长",构建了全局控制Elman神经网络用于将全局统计特征与基于语段的时序特征相融合.随后该团队又提出了一种基于不同时间单元的多粒度特征提取方法,以及可以融合多粒度特征的基于认知机理的回馈神经网络(Cognition-Inspired Recurrent Neural Network, CIRNN)[112].该网络既突出了情感的时序性,也保留了全局特性对情感识别的作用,实现多层级信息融合.Deng J[113]等利用Bag-of-Audio-Words(BoAW)算法代替传统的统计特征,该方法针对LLDs特征利用k均值聚类方法或随机采样方法生成编码本(codebook),再利用多重赋值量化技术(multi-assignment quantisation)将每帧语音信号提取的LLDs

特征分配到相应的编码本得到直方图,将直方图归一化后作为特征用于识别.

Table 3 Low-level descriptors and functionals

表3 常见LLDs特征以及统计特征

特征Low-level descriptors (LLDs)	统计函数(Functionals)
基频(Fundamental frequency),能量(energy),强度(intensity), 谐波噪声比(harmonics-to-noise ratio, HNR), 语速(speech rate),Mel倒频谱系数(Mel frequency cepstral coefficients ,MFCCs),共振峰振幅(formant amplitude),共振峰带宽(formant bandwidth),共振峰频率(formant frequency),线性预测倒谱系数(Linear Predictor Cepstral Coefficients,LPCC),线谱对参数(Line Spectrum Pair, LSP),谱斜率(spectral tilt),振幅比(normalized amplitude quotient)	极值(Extreme values), 最大值(maximum), 最小值(minimum),平均值(means),标准差(standard deviation), 方差(variance), 峰度(kurtosis), 偏斜度(skewness), 百分数(percentiles),百分比范围(percentile ranges), 四分位数(quartiles), 中心(centroids), 偏离量(offset), 斜率(slope), 均方误差(mean squared error), 时长(time/durations)

4.1.2 语音信号中的语义信息

语音信号中传递的语义信息对于情感识别具有一定的作用,有些特定的词汇可以表达相应的情感倾向.Lee C M[114]等将声学特征、句法、语篇信息相结合用于情感识别,引入情感显著性的信息理论来表达语言层面的情感信息,对电话服务中心数据的实验结果表明融合特征可以有效地提高情感识别率. Schuller[115]提出一种新的方法将声学特征与语义信息融合用于情感识别.首先,提取声学特征利用分类器进行识别,然后利用置信网络根据语义上下文进行识别,最后利用Neural Net将两次识别结果进行决策融合.Wu[116]等(2011)将语义标签识别结果与声学韵律信息融合来提高语音情感识别结果.语义标签来自知网汉语知识库(Chinese knowledge base HowNet)用于自动提取情感关联规则(Emotion Association Rules ,EARs).

4.2 语音维度情感预测器

情感识别通过获取人类情感信息,识别人类的情感,提高机器与人之间自然交互能力.根据情感描述模型的不同,语音情感识别系统采用的识别算法亦不同.维度语音情感识别问题可建模为回归预测问题.常见的回归预测算法包括线性回归(Linear Regression)、k-NN、ANN、PLS、SVR,当前新兴的深度神经网络如LSTM、RNN等.

偏最小二乘法(PLS)[117, 118]结合了主成分分析PCA和典型相关分析CCA的思想,适用于特征集较大并且存在多重共线性的预测建模问题. Mencattini[97]将7类离散情感投影到二维情感空间描述模型(效价度-激活度)中,采用偏最小二乘法回归(PLSR)模型在印度语音数据库EMOVO上对男性、女性发音语料进行情感预测,平均判决系数分别为0.89和0.72.

SVR是支持向量在函数回归领域的应用[119-121] ,M. Grimm[122]等在VAM数据库上利用SVR在效价度、激活度和控制度三维情感属性上进行情感预测,其性能优于k-NN、基于规则的逻辑分类器(Rule-based Fuzzy Logic Classifier).Giannakopoulos[123]等利用效价度－唤醒度的二维情感空间描述情感状态,并使用 k 近邻算法(k-NN)对电影剪辑语句的情感坐标值进行估计；Kanluan[124]等在VAM数据库上进行多模态情感识别,提取韵律学特征、谱特征等声学特征以及基于二维离散余弦变换的面部图片特征,利用SVR分别进行语音情感识别和面部情感识别,再利用决策级融合方法将两种模态预测结果进行权重线性融合,预测结果较语音情感识别提高12.3%.

LSTM网络使用特殊的神经元在长时间范围内存储并传递信息,适合于处理和预测时间序列中长时间延迟的信号,因此该网络可以记忆情感随时间的变化信息.利用长短时记忆循环网络(LSTM-RNN)进行维度情感识别取得了比传统方法更好地效果.Martin Wöllmer[125]采用AVEC2011 (Audio/ Visual Emotion Challenge 2011)[126]情感竞赛提供的声学特征结合面部运动特征,在SEMAINE情感数据库上进行音视频情感维度识别,实验结果表明,与其他参赛者提供的情感识别模型相比,基于LSTM网络的平均识别效果最好.Ringeval[15]等将LSTM-RNN用于音频、视频、生理信号的多模态维度情感识别,该网络可以动态地利用长时间的上下文信息,同时避免RNN网络的梯度消失问题.文中比较了不同窗长对各模态情感识别结果的影响,以及特征级融合与决策级融合方法

的识别效果,研究结果表明,效价度的情感识别比激活度需要更长的窗长,决策级融合取得更好地识别效果.在RECOLA数据库上,该模型在激活度和效价度上的一致相关系数分别可达到0.804和0.528.Chao[127]等利用时间池对输入特征进行时间建模,并引入ε不敏感损失函数改进LSTM-RNN模型,使其对标注噪声具有更好的鲁棒性,该模型在RECOLA数据库上对效价度和唤醒度进行情感识别都取得了更好地效果,但在唤醒度上存在过拟合现象.

国内也越来越多学者提出新颖的语音维度情感识别方法,陈逸灵[128]等利用MFCC特征结合语谱图中提取时间点火序列特征、点火位置信息特征三种特征分别用于语音情感识别,并将识别结果与PAD (Pleasure、Arousal、Dominance) 维度进行相关性分析得到特征的权重系数,加权融合后获得情感语音的最终 PAD 值.李海峰[129]等通过构建对情感程度相对顺序敏感的Dim-SER系统,提出顺序敏感的神经网络算法,实验结果表明,该网络性能较常用的k近邻算法和支持向量回归算法相比有了提升.

目前上述基于单一数据的语音情感识别性能已经取得了很大的提升.然而,在很多实际应用情境下,系统必须考虑文化、语言、种族、个体、年龄等影响下数据的情感分类.从大脑工作神经机制来讲,不同种族、文化等人群对情感的反应具有一致生理生化基础,康奈尔大学神经学家Adam Anderson的一项研究表明人的大脑会使用一种标准的代码,来说出同样的情感语言[48],人的大脑会对从愉悦到不愉悦、好到坏的感觉,产生一种特殊的代码,读起来就像一个‘神经价表’,在这个价表中,一组神经元在一个方向倾斜等同于积极情绪,其他方向的倾斜则等同于消极情绪.虽然存在一致性的生理基础,但是文化对于个体的态度、行为、语言或非语言的反应都有着潜移默化的影响,这些差异影响了人类跨文化情感表达、感知与理解.有情感心理学研究表明文化背景对于个体如何利用面部和声音线索从多感官刺激中有意识地评估情感含义有着重要的影响[130-132].Elfenbein 和Ambady[133]发现同种族或者区域的人群具有比较一致的情感表达和识别方式,情感识别会更加精准一些.上述心理学及认知学研究表明,从大脑脑区的精细神经活动模式角度看,情感感知存在着相似性,但是文化背景、语言、个体差异又影响着情感的感知.在共同的信息加工机制下进行跨文化、跨种族等语音情感识别有了理论基础.Peng [134]提出一种迁移线性子空间学习(transfor linear subspace learning, TLSL)网络框架进行跨库语音情感识别,在学习的投影子空间中提取鲁棒的跨库特征表征,其优势是解决了当前大多数迁移学习只专注于寻找最可能迁移的特征的缺陷,通过结合实验,证明TLSL用于跨库语音情感识别是有效的.Hesam[135]等利用基于自动语言检测的模型可以提高多语言情感识别的准确率.在三种语言(德语、罗曼语系、汉藏语)的六个数据库上进行测评,将情感分别在效价度与唤醒度上进行分类,实验结果说明,尽管语音情感识别更多的依赖于声学特征,但其语言学信息可以提供话者文化背景相关的有用信息.通过识别话者的语言作为先验知识,基于该知识的学习模型可以提高情感识别系统的性能.Kaya[136]等采用融合了线性说话人归一化、能量归一化、特征向量归一化的级联归一化方法以减少跨库以及不同说话人差异带来的影响,并利用极限学习机(Extreme Learning Machines,ELM)在跨语系的五种语言情感数据库上测试该归一化方法的有效性.Silvia与Bjorn Schuller[137]于2015年情感计算国际会议ACII上对跨语言声学情感识别做了综述及前景展望.

5 维度语音情感识别研究的相关资源

5.1 语音维度情感数据库

情感数据库是语音情感识别的先决条件,提供训练与测试用语音样本,数据库的质量直接影响情感识别率以及研究结果的可靠性.目前,语音情感识别领域以离散情感数据库居多,如 Belfast 情感数据库、EMO-DB 德语情感数据库、FAU AIBO 儿童德语情感数据库、CASIA 汉语情感语料库、ACCorpus 汉语情感数据库等,维度情感语料库有待进一步丰富.下文首先介绍维度语音数据库的建立与标注方法,然后介绍一些代表性的维度情感数据库.

5.1.1 情感数据库的建立

根据语料的情感自然度程度的不同,情感语音数据库的建立方法主要有三种:

(1)自然情感语料

从现实生活中采集真实的自然语料,进一步通过人工筛选与标注的方法获得可用语料.这类情感语料具有最高的自然度,可以认为是真实意义上的情感语料.这种语料在使用前必须进行分类标注,由于分类的标准不统一,并且有些情感人类自身也难以区分,因此,这类情感语料具有一定局限性.

(2)模拟情感语料

由专业或善于表达情感的人进行情感模仿录制语料.这种有目的性录制的特定情感语料,具有更好的区分性.但这种语料的情感自然度取决于录音者的模仿能力,有时情感成分被夸大而不能体现真实的情感.

(3)诱导情感语料

利用情景短片或者角色扮演的方式营造相应的环境氛围,从而诱导录音者产生特定情感后录音.利用该方法获得的语料接近真实情感,但由于环境诱发刺激效果很难评测,导致较难判断诱发的情感是否强烈.

5.1.2 情感数据库的标注

语音情感数据库的标注是一个困难但又极为重要的工作,数据标注的质量对基于语音的情感研究有着重要的意义.实现较为精确的语音情感标注通常需要三个方面: 音字转写(transcription) 、注解(annotation) 、标注(labelling)[138].音字转写是将音频中的语言信息以文字的形式转写标注,即将语音转化为文字; 注解是在转写基础上进一步的标注韵律信息、语速、音量/调变化等副语言特征; 标注是对语句进行情感状态的标记.目前转写与注解已经有一些较为成熟的工具和软件,如 Anvil、EX-MARaLDA、Partitur Editor、Praat 等,这些软件各有优势.情感标注(labelling)工具可以方便地实现对语音情感的连续性变化的跟踪(此节以维度情感的标注方法为主).Cowie [139]等人开发了实时的效价度-唤醒度二维情感标注工具 Feeltrace,可用于动态情绪的标注与分析,标注者根据自己感知的情感实时的利用鼠标拖动圆形光标到合适位置即可实现标注.Emocards 量表根据 Russell 的情感环状理论,用环状布局的 16 张卡通表情图片描述情感,在愉悦度和紧张度两个维度上测量情感[140].Bradley[141]等人依据 PAD 情感空间模型提出 SAM 量表,以图形化的方式从愉悦度、唤醒度和优势度由弱到强进行 9 级评分,每个维度由逐渐变化的小人图片代表.SAM 量表已经被证实可以有效地评定被试的情感感觉[142].Broekens[143]开发了在线情感测量工具 AffectButton,仅包含一个按钮,按钮表面是一张动态变化的卡通脸部图片,鼠标的(x,y)坐标映射到 PAD 三维空间模型中,表情图片随鼠标的移动而改变.AffectButton 比 SAM 更加形象、简便,一个按键可以反馈三维信息.ANNEMO[144]是基于网页的音视频维度情感标注工具,可同时显示音视频与标注界面,可进行时间连续的标记.Ikannotate[145]工具将上述三方面融合,可以实现转写、注解、标注、以及标注的不确定.

标注时须有一定的规则,包括标注的一致性、连贯性、标注符号的易记性,但同时还需要遵循的一条原则是允许标注的不确定性和差异性存在,即允许不同的标注者对同一条语音中的情感、重音、声调等有不同的理解,避免向用户提供错误信息.

5.1.3 具有代表性的维度情感数据库

近些年来随着研究者们对维度情感识别领域的关注,一些公开的以科学研究为目的的维度情感数据库逐渐被发布.尽管完整的语音情感数据库应包括转写、注解、情感标注,但目前维度语音情感数据库的标注往往只包含对整句或段的情感标注,因此构建公认的有效、全面、优质的语音情感数据库是语音情感计算研究的重中之重.

VAM 数据库(Vera am Mittag Database)现场录制了 12 个小时的德语电视谈话节目[146],谈话内容均为无脚本限制、无情绪引导的纯自然交流,该库是一个包含视频库、语音库、表情库的多模态情感数据库,视频库(VAM-Video)包括 104 个说话人的 1421 个视频,语音库与表情库是从该视频库中分离获得,其中语音库又分为两部分,一部分为非常明显的情感表达,包括 19 个不同说话人的 499 个语句,由 17 个听者在 Valence、Activation、Dominance 三个维度利用 SAM 进行标注,可用于维度语音情感识别研究; 另一部分包括 28 位说话人的 519 个语句,由 6 位听者进行标注.表情库包括 20 位说话者的 1867 幅表情图片,涵盖高兴、生气、悲伤、厌恶、恐惧、惊讶的六类情感以及中性情感,可用于表情识别研究.

Semaine 数据库是一个音视频情感数据库[147],数据录制了用户与性格迥异的四个机器角色的交谈对话,

在三种情景下录制,一种是 Solid SAL(Sensitive Artificial Listener),该情境下操作者扮演了 SAL 的角色,录制了用户与角色的 95 段交互,190 个视频片段;第二种是半自动 SAL(Semi-Automatic SAL),该情景需要操作者选择一系列日常用语,该语句已提前被表演者以与某种性格匹配的声音录制,再以图形界面交互方式展现给用户,总共录制了 1410 分钟用户与机器角色的视频数据.第三种是自动 SAL(Automatic SAL),该情境下角色表达的语句及非言语表达完全由 SEMAINE 系统自动的生成,该系统同时检测用户的情感变化并由摄像头记录下来,用户与角色交互视频共计 1266 分钟.对话由多个参与者借助标注工具 Feeltrace 在 Activation, Valence, Power, Anticipation/Expectation 和 Intensity 这五个情感维度上进行标注.该数据库中的部分数据被用作 AVEC2012 的竞赛数据库[148].

Recola 数据库是一个多模态法语情感数据库[144],包括音频、视频、生理数据(ECG 和 EDA).该数据库录制了 9.5 小时的视频会议,连续同步的记录了 46 名参与者自然交流.6 名法语助理通过 ANNEMO 标注工具在 Arousal, Valence 维度上进行标注,最终 34 名参与者同意共享数据,数据时长共计 7 小时,其中,包括 27 名参与者 5.5 小时的生理数据.

USC IEMOCAP(Interactive Emotional Dyadic Motion Capture)数据库[149]是一个英语情感数据库,包括 10 个说话人参与的相互交流的音视频,将总计 12 小时的音视频数据进行分割成 10039 段语句,既包括有情感脚本的情感表演也包括即兴情感表达场景.每个语句由 3 名标注者进行离散情感标注,包括高兴、生气、悲伤、中性、挫败感五类情感,标注者也可根据理解标注为其他情感类别,2 名标注者在 Arousal、Valence、Dominance 三个维度进行维度空间标注,每个维度标注的范围为[1,5],标注间隔为 0.5,可用于离散或维度情感识别.

5.2 语音情感特征提取工具

目前已有公开的程序或工具箱广泛应用于语音信号的处理、标注、频谱分析、特征提取等,例如 PRAAT(http://www.fon.hum.uva.nl/praat)可实现对语音信号的采集、分析、标注、合成、统计分析等功能;OpenSMILE(http://audeering.com/research/opensmile/)软件对于音频处理的特征提取是一款很有效的工具,是一种以命令行形式运行的而不是图形界面的操作软件,通过配置config文件对音频进行特征提取;pyAudioAnalysis(An Open-Source Python Library for Audio Signal Analysis https://github.com/tyiannak/pyAudioAnalysis/wiki/2.-General)是Python下的一个音频处理工具包,可用于音频特征提取;Librosa(https://librosa.github.io/)也是基于python的工具包,可以提取各种语音特征,window和Linux均可;HTK Speech Recognition Toolkit(http://htk.eng.cam.ac.uk/)是基于C语言的特征提取工具包,代码成熟稳定,目前支持GPU,windows和Linux环境均可;Kaldi ASR(http://kaldi-asr.org/)是一个语音识别工具包,开发效率高,Linux使用方便.

5.3 识别算法工具

开源的深度学习神经网络正步入成熟,目前有许多框架具备为语音情感识别提供先进的机器学习的能力.例如,TensorFlow(https://www.tensorflow.org/)是谷歌发布的开源工具,编程接口支持 Python 和 C++, 还可在谷歌云和亚马孙云中运行.TensorFlow 支持细粒度的网格层,而且允许用户在无需用低级语言实现的情况下构建新的复杂的层类型,子图执行操作允许开发者在图的任意边缘引入和检索任意数据的结果.Caffe(http://caffe.berkeleyvision.org/)是自 2013 年底以来第一款主流的工业级深度学习工具包,具有优秀的卷积模型,是计算机视觉界最流行的工具包之一.CNTK(https://github.com/Microsoft/CNTK/wiki)是微软最初面向语音识别的框架,支持 RNN 和 CNN 类型的网络模型,从而在处理图像、手写字体和语音识别问题上,它是很好的选择.MXNet(http://mxnet.io/)是一个全功能、可编程和可扩展的深度学习框架,它支持深度学习架构,如卷积神经网络(CNN)、循环神经网络(RNN)和其包含的长短时间记忆网络(LTSM),为图像、手写文字和语音的识别和预测以及自然语言处理提供了出色的工具.PyTorch(http://pytorch.org/)是一种 Python 优先的深度学习框架,特点是快速成形、代码可读和支持最广泛的深度学习模型.Theano(http://deeplearning.net/software/theano/) 开创了将符号图用于神经网络编程的趋势,但缺乏分布式

应用程序管理框架,只支持一种编程开发语言.

6 目前存在的问题及未来发展方向

6.1 计算模型缺乏脑科学、心理学等学科研究成果的指导

现有的语音情感识别是基于计算机科学进行研究的,利用机器学习的算法进行训练与识别,但情感是人类极其复杂的心理状态,研究人类大脑的情感处理机制将尤为重要.目前情感识别的算法太简单,缺乏心理学对情感研究成果的指导,如何更全面地建立情感的描述模型,不同情感之间是否有关联,例如Ekman等[150]惊讶情感是对一件意料之外的事件的反应,这种情感往往容易会跟随在高兴或者恐惧情感之后,Davidson[151]认为对惊讶情绪的识别需要考虑情境上下文.Banse[152]等研究发现生气或者恐惧情绪的语音在声学特征上具有明显区分性,也很少受到文化差异的影响,更容易进行识别.

除此之外,目前的情感识别框架缺乏人类大脑的复杂机制和工作模式的指导,与认知功能之间的交互与协同较少.随着认知科学的快速发展,科学家越来越多的了解人类大脑复杂的信息处理机制,将这些成果与机器学习算法结合,将有助于突破目前情感识别研究的瓶颈,实现真正的人工智能.

6.2 语音情感数据标注困难

语音情感类数据在收集与标注上存在的困难,导致当下用于研究的数据规模较小,种类较为贫乏.在上下文语境未知的情况下,标注变得更加困难,公认的有效、全面、优质的语音情感数据库是语音情感计算研究的基础.目前高质量的情感语料库很少,而且缺乏大规模跨语言的公认语料库,研究者们利用不同的数据库进行情感识别,导致识别结果难以进行比较评价.目前用于情感标注的都是自我评价(self-report)方法,如SAM量表等.研究者们可制定情感数据库标注的相关国标以明确详细的标注规则和方法；借助数据标注公司、情感心理学专家的帮助,建立拥有完整情感标注信息的优质语音情感数据库.

6.3 情感特征与语音情感之间存在鸿沟

与离散情感识别类似,进行维度情感识别的首要工作是特征提取,决定了回归预测器准确率的高低.目前大多数特征是基于语音的声学特征,这些声学特征能否有效地表征情感并没有详细的论证.情感特征的提取需要考虑两方面问题,首先,所提取的声学特征与情感之间是否存在鸿沟,能否有效地区分情感,实现类内的特征距离较小,类间的特征距离较大；其次,情境上下文对情感的识别具有关键性作用,需选取合适的时间粒度来提高情感识别率.

解决上述问题,探索特征与情感类别之间映射关系,提出对情感具有区分度的新特征将是非常有价值的研究方向.同时,探索人类大脑对情感的处理机制,结合心理学、认知学研究成果,研究语音的各个层面(语素、词素、句法、语篇)对情感识别的影响.在此基础上,提取不同粒度上的特征,提高语音情感识别率.

6.4 用于维度情感识别的机器学习策略有待提高

语音识别的快速发展得益于人工神经网络的支持,特别是近年来深度神经网络的发展,使语音识别性能进一步提升.研究者们往往借鉴语音识别中使用的神经网络模型进行情感识别,但是情感是较语言更高层次的表达,需要包含更多信息,甚至推理、记忆、决策能力.因此,目前用于情感识别的网络模型需要基于认知理论进一步改进,探索人类情感处理机制,并对认知模型进行实用化实现,提出相应的机器学习方法,进一步建立类脑多尺度神经网络计算模型以及类脑人工智能算法,将是突破语音情感识别研究瓶颈的有效策略.

7 结束语

语音情感识别是使机器实现自然地人机交互的重要方面,不仅对推动信号处理、计算机、人工智能、人机交互、控制、认知等学科发展具有重要的学术意义,而且具有重要的经济价值和社会意义,如具有社交能力的情感机器人、情绪检测与监控、呼叫中心情绪考核等.基于情感的维度空间描述模型较传统的离散情感模型,可以

更精确地描述情感,减小情感标签的模糊性,具有无限的情感描述能力.基于维度情感模型的语音情感识别系统也日益受到越来越多的关注.相关研究人员已在语音情感认知、语音维度情感数据库、情感相关的语音特征提取、以及识别算法方面取得长足的进步,本文也主要针对这四个方面详细介绍了基于维度情感描述模型的语音情感识别进展,填补了目前语音维度情感识别综述的空缺,同时,提出了该技术当前仍面临的一系列挑战,如进一步探究人脑对语音情感认知规律,提出表征情感的语音特征,利用人脑情感认知机制指导识别算法的改进等.

References

[1] Crystal, D. Non-segmental phonology in language acquisition: A review of the issues. Lingua, 1973, 32(1-2): 1-45.

[2] Liebenthal, E., D.A. Silbersweig, and E. Stern. The language, tone and prosody of emotions: Neural substrates and dynamics of spoken-word emotion perception. Frontiers in neuroscience, 2016, 10: 506.

[3] Murray, I.R. and J.L. Arnott. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. The Journal of the Acoustical Society of America, 1993, 93(2): 1097-1108.

[4] Williams, C.E. and K.N. Stevens. Emotions and speech: Some acoustical correlates. The Journal of the Acoustical Society of America, 1972, 52(4B): 1238-1250.

[5] Murray, I.R. and J.L. Arnott. Synthesizing emotions in speech: Is it time to get excited? In: *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*. IEEE, 1996.

[6] Valstar, M., J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In: *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016.

[7] Dhall, A., A. Kaur, R. Goecke, and T. Gedeon. Emotiw 2018: Audio-video, student engagement and group-level affect prediction. In: *Proceedings of the 2018 on International Conference on Multimodal Interaction*. ACM, 2018.

[8] Li, Y., J. Tao, B. Schuller, S. Shan, D. Jiang, and J. Jia. Mec 2016: The multimodal emotion recognition challenge of ccpr 2016. In: *Chinese Conference on Pattern Recognition*. Springer, 2016.

[9] Li, Y., J. Tao, B. Schuller, S. Shan, D. Jiang, and J. Jia. Mec 2017: Multimodal emotion recognition challenge. In: *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*. IEEE, 2018.

[10] Christianson, S.-A. The handbook of emotion and memory: Research and theory: Psychology Press, 2014.

[11] Wager, T., L.F. Barrett, E. Bliss-Moreau, K. Lindquist, S. Duncan, H. Kober, J. Joseph, M. Davidson, and J. Mize. The handbook of emotion. Lewis, M, 2008: 249-271.

[12] Ortony, A. and T.J. Turner. What's basic about basic emotions? Psychological review, 1990, 97(3): 315.

[13] Gunes, H., B. Schuller, M. Pantic, and R. Cowie. Emotion representation, analysis and synthesis in continuous space: A survey. In: *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*. IEEE, 2011.

[14] Chen, S. and Q. Jin. Multi-modal dimensional emotion recognition using recurrent neural networks. In: *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2015.

[15] Ringeval, F., F. Eyben, E. Kroupi, A. Yuce, J.-P. Thiran, T. Ebrahimi, D. Lalanne, and B. Schuller. Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. Pattern Recognition Letters, 2015, 66: 22-30.

[16] Fontaine, J. The dimensional, basic, and componential emotion approaches to meaning in psychological emotion research. In: *Components of emotional meaning: A sourcebook*.Oxford University Press, 2013. p. 31-45.

[17] Cowie, R. and R.R. Cornelius. Describing the emotional states that are expressed in speech. Speech communication, 2003, 40(1-2): 5-32.

[18] Scherer, K.R. On the nature and function of emotion: A component process approach. Approaches to emotion, 1984, 2293: 317.

[19] Scherer, K.R. Vocal communication of emotion: A review of research paradigms. Speech communication, 2003, 40(1-2): 227-256.

[20] Ortony, A., G.L. Clore, and A. Collins. The cognitive structure of emotions: Cambridge university press, 1990.

[21] Roseman, I.J. Appraisal determinants of emotions: Constructing a more accurate and comprehensive theory. Cognition & Emotion, 1996, 10(3): 241-278.

[22] Soleimani, A. and Z. Kobti. Toward a fuzzy approach for emotion generation dynamics based on occ emotion model. IAENG International Journal of Computer Science, 2014, 41(1): 48-61.

[23] Olgun, Z.N., Y. Chae, and C. Kim. A system to generate robot emotional reaction for robot-human communication. In: *2018 15th International Conference on Ubiquitous Robots (UR)*. IEEE, 2018.

[24] Masuyama, N., C.K. Loo, and M. Seera. Personality affected robotic emotional model with associative memory for human-robot interaction. Neurocomputing, 2018, 272: 213-225.

[25] Cavallo, F., F. Semeraro, L. Fiorini, G. Magyar, P. Sinčák, and P. Dario. Emotion modelling for social robotics applications: A review. Journal of Bionic Engineering, 2018, 15(2): 185-203.

[26] Rincon, J.A., A. Costa, P. Novais, V. Julian, and C. Carrascosa. A new emotional robot assistant that facilitates human interaction and persuasion. Knowledge and Information Systems, 2018: 1-21.

[27] Bartneck, C., M.J. Lyons, and M. Saerbeck. The relationship between emotion models and artificial intelligence. arXiv preprint arXiv:1706.09554, 2017.

[28] Wundt, W. Vorlesungen über menschen-und thierseele: Leop. Voß, 1863.

[29] Schlosberg, H. Three dimensions of emotion. Psychological review, 1954, 61(2): 81.

[30] Russell, J.A. and A. Mehrabian. Evidence for a three-factor theory of emotions. Journal of research in Personality, 1977, 11(3): 273-294.

[31] Plutchik, R. Emotions: A general psychoevolutionary theory. Approaches to emotion, 1984, 1984: 197-219.

[32] Russell, J.A. A circumplex model of affect. Journal of personality and social psychology, 1980, 39(6): 1161.

[33] Krech, D., R.S. Crutchfield, and N. Livson. Elements of psychology: Alfred A. Knopf, 1974.

[34] Izard, C.E. The psychology of emotions: Springer Science & Business Media, 1991.

[35] Mehrabian, A. Analysis of the big‐five personality factors in terms of the pad temperament model. Australian journal of Psychology, 1996, 48(2): 86-92.

[36] Gebhard, P. Alma: A layered model of affect. In: *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*. ACM, 2005.

[37] Becker-Asano, C. Wasabi: Affect simulation for agents with believable interactivity: IOS Press, 2008.

[38] WJ, H., L. HF, R. HB, and M. L. Review on speech emotion recognition. Ruan Jian Xue Bao/Journal of Software, 2014, 25(1): 37-50 (in Chinese).

[39] MacLean, P.D. Psychosomatic disease and the" visceral brain"; recent developments bearing on the papez theory of emotion. Psychosomatic medicine, 1949.

[40] Arnold, M.B. Emotion and personality. 1960.

[41] Pribram, K.H. Feelings as monitors. In: *Feelings and emotions. The Loyola Symposium*. Academic Press New York, 1970.

[42] Lazarus, R.S. and S. Folkman. Stress, appraisal, and coping: Springer publishing company, 1984.

[43] LeDoux, J. and J.R. Bemporad. The emotional brain. Journal of the American Academy of Psychoanalysis, 1997, 25(3): 525-528.

[44] Moren, J. Emotion and learning-a computational model of the amygdala. 2002.

[45] Morén, J. and C. Balkenius. A computational model of emotional learning in the amygdala. From animals to animats, 2000, 6: 115-124.

[46] Phelps, E.A. and J.E. LeDoux. Contributions of the amygdala to emotion processing: From animal models to human behavior. Neuron, 2005, 48(2): 175-187.

[47] Mathersul, D., L.M. Williams, P.J. Hopkinson, and A.H. Kemp. Investigating models of affect: Relationships among eeg alpha asymmetry, depression, and anxiety. Emotion, 2008, 8(4): 560.

[48] Chikazoe, J., D.H. Lee, N. Kriegeskorte, and A.K. Anderson. Population coding of affect across stimuli, modalities and individuals. Nature neuroscience, 2014, 17(8): 1114.

[49] Kirkby, L.A., F.J. Luongo, M.B. Lee, M. Nahum, T.M. Van Vleet, V.R. Rao, H.E. Dawes, E.F. Chang, and V.S. Sohal. An amygdala-hippocampus subnetwork that encodes variation in human mood. Cell, 2018, 175(6): 1688-1700. e14.

[50] Lindquist, K.A., T.D. Wager, H. Kober, E. Bliss-Moreau, and L.F. Barrett. The brain basis of emotion: A meta-analytic review. Behavioral and brain sciences, 2012, 35(3): 121-143.

[51] Buchanan, T.W., K. Lutz, S. Mirzazade, K. Specht, N.J. Shah, K. Zilles, and L. Jäncke. Recognition of emotional prosody and verbal components of spoken language: An fmri study. Cognitive Brain Research, 2000, 9(3): 227-238.

[52] George, M.S., P.I. Parekh, N. Rosinsky, T.A. Ketter, T.A. Kimbrell, K.M. Heilman, P. Herscovitch, and R.M. Post. Understanding emotional prosody activates right hemisphere regions. Archives of neurology, 1996, 53(7): 665-670.

[53] Paulmann, Silke, and S.A. Kotz. Temporal interaction of emotional prosody and emotional semantics: Evidence from erps. International Conference on Speech Prosody, 2006.

[54] Pihan, H., E. Altenmüller, and H. Ackermann. The cortical processing of perceived emotion: A dc-potential study on affective speech prosody. Neuroreport, 1997, 8(3): 623-627.

[55] Ross, E.D., R.D. Thompson, and J. Yenkosky. Lateralization of affective prosody in brain and the callosal integration of hemispheric language functions. Brain and language, 1997, 56(1): 27-54.

[56] Ross, E.D., J.A. Edmondson, G.B. Seibert, and R.W. Homan. Acoustic analysis of affective prosody during right-sided wada test: A within-subjects verification of the right hemisphere's role in language. Brain and Language, 1988, 33(1): 128-145.

[57] Davidson, R.J., H. Abercrombie, J.B. Nitschke, and K. Putnam. Regional brain function, emotion and disorders of emotion. Current opinion in neurobiology, 1999, 9(2): 228-234.

[58] Zatorre, R.J., P. Belin, and V.B. Penhune. Structure and function of auditory cortex: Music and speech. Trends in cognitive sciences, 2002, 6(1): 37-46.

[59] Ethofer, T., D. Van De Ville, K. Scherer, and P. Vuilleumier. Decoding of emotional information in voice-sensitive cortices. Current Biology, 2009, 19(12): 1028-1033.

[60] Wildgruber, D., A. Riecker, I. Hertrich, M. Erb, W. Grodd, T. Ethofer, and H. Ackermann. Identification of emotional intonation evaluated by fmri. Neuroimage, 2005, 24(4): 1233-1241.

[61] Wildgruber, D., H. Ackermann, B. Kreifelts, and T. Ethofer. Cerebral processing of linguistic and emotional prosody: Fmri studies. Progress in brain research, 2006, 156: 249-268.

[62] Grandjean, D., D. Sander, G. Pourtois, S. Schwartz, M.L. Seghier, K.R. Scherer, and P. Vuilleumier. The voices of wrath: Brain responses to angry prosody in meaningless speech. Nature neuroscience, 2005, 8(2): 145.

[63] Kotz, S.A., C. Kalberlah, J. Bahlmann, A.D. Friederici, and J.D. Haynes. Predicting vocal emotion expressions from the human brain. Human Brain Mapping, 2013, 34(8): 1971-1981.

[64] Fritsch, N. and L. Kuchinke. Acquired affective associations induce emotion effects in word recognition: An erp study. Brain and language, 2013, 124(1): 75-83.

[65] Elliot, C. The affective reasoner: A process model of emotions in a multi-agent system. 1992. Northwestern University Institute for the Learning Sciences: Northwestern, IL, 48.

[66] Reilly, W.S. Believable social and emotional agents. In.: Carnegie-Mellon Univ Pittsburgh pa Dept of Computer Science, 1996.

[67] Gratch, J. and S. Marsella. Evaluating the modeling and use of emotion in virtual humans. In: *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems-Volume 1*. IEEE Computer Society, 2004.

[68] Velásquez, J.D. and P. Maes. Cathexis: A computational model of emotions. In: *Proceedings of the first international conference on Autonomous agents*. ACM, 1997.

[69] Marsella, S., J. Gratch, and P. Petta. Computational models of emotion. A Blueprint for Affective Computing-A sourcebook and manual, 2010, 11(1): 21-46.

[70] Watts, L. Reverse-engineering the human auditory pathway. In: *Advances in computational intelligence*.Springer, 2012. p. 47-59.

[71] Abdi, J., B. Moshiri, B. Abdulhai, and A.K. Sedigh. Forecasting of short-term traffic-flow based on improved neurofuzzy models via emotional temporal difference learning algorithm. Engineering Applications of Artificial Intelligence, 2012, 25(5): 1022-1042.

[72] Falahiazar, A., S. Setayeshi, and Y. Sharafi. Computational model of social intelligence based on emotional learning in the amygdala. Journal of mathematics and computer Science, 2015, 14: 77-86.

[73] Milad, H.S., U. Farooq, M.E. El-Hawary, and M.U. Asad. Neo-fuzzy integrated adaptive decayed brain emotional learning network for online time series prediction. IEEE Access, 2017, 5: 1037-1049.

[74] Lotfi, E., O. Khazaei, and F. Khazaei. Competitive brain emotional learning. Neural Processing Letters, 2018, 47(2): 745-764.

[75] Lucas, C., D. Shahmirzadi, and N. Sheikholeslami. Introducing belbic: Brain emotional learning based intelligent controller. Intelligent Automation & Soft Computing, 2004, 10(1): 11-21.

[76] Parsapoor, M. and U. Bilstrup. Brain emotional learning based fuzzy inference system (belfis) for solar activity forecasting. In: *Tools with Artificial Intelligence (ICTAI), 2012 IEEE 24th International Conference on*. IEEE, 2012.

[77] Motamed, S., S. Setayeshi, and A. Rabiee. Speech emotion recognition based on a modified brain emotional learning model. Biologically inspired cognitive architectures, 2017, 19: 32-38.

[78] Grimm, M., K. Kroschel, E. Mower, and S. Narayanan. Primitives-based evaluation and estimation of emotions in speech. Speech Communication, 2007, 49(10-11): 787-800.

[79] Hammerschmidt, K. and U. Jürgens. Acoustical correlates of affective prosody. Journal of Voice, 2007, 21(5): 531-540.

[80] Laukka, P., H.A. Elfenbein, N. Söder, H. Nordström, J. Althoff, F.K.e. Iraki, T. Rockstuhl, and N.S. Thingujam. Cross-cultural decoding of positive and negative non-linguistic emotion vocalizations. Frontiers in Psychology, 2013, 4: 353.

[81] Sauter, D.A., F. Eisner, P. Ekman, and S.K. Scott. Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations: Correction. 2015.

[82] Tickle, A. English and japanese speakers' emotion vocalisation and recognition: A comparison highlighting vowel quality. In: *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*. 2000.

[83] Yang, L.-c. and N. Campbell. Linking form to meaning: The expression and recognition of emotions through prosody. In: *4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis*. 2001.

[84] Thompson, W.F. and L.-L. Balkwill. Decoding speech prosody in five languages. Semiotica, 2006, 2006(158): 407-424.

[85] Pell, M.D., L. Monetta, S. Paulmann, and S.A. Kotz. Recognizing emotions in a foreign language. Journal of Nonverbal Behavior, 2009, 33(2): 107-120.

[86] Bryant, G. and H.C. Barrett. Vocal emotion recognition across disparate cultures. Journal of Cognition and Culture, 2008, 8(1-2): 135-148.

[87] Émond, C., L. Ménard, M. Laforest, F. Bimbot, C. Cerisara, C. Fougeron, G. Gravier, and L. Lamel. Perceived prosodic correlates of smiled speech in spontaneous data. In: *INTERSPEECH*. 2013.

[88] Wang, Y.T., J. Han, X.Q. Jiang, J. Zou, and H. Zhao. Study of speech emotion recognition based on prosodic parameters and facial expression features. In: *Applied Mechanics and Materials*. Trans Tech Publ, 2013.

[89] Rao, K.S., S.G. Koolagudi, and R.R. Vempada. Emotion recognition from speech using global and local prosodic features. International journal of speech technology, 2013, 16(2): 143-160.

[90] Pao, T.-L., Y.-T. Chen, J.-H. Yeh, and W.-Y. Liao. Detecting emotions in mandarin speech. International Journal of Computational Linguistics & Chinese Language Processing, Volume 10, Number 3, September 2005: Special Issue on Selected Papers from ROCLING XVI, 2005, 10(3): 347-362.

[91] Pereira, C. Dimensions of emotional meaning in speech. In: *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*. 2000.

[92] Borchert, M. and A. Dusterhoft. Emotions in speech-experiments with prosody and quality features in speech for use in categorical and dimensional emotion recognition environments. In: *2005 International Conference on Natural Language Processing and Knowledge Engineering*. IEEE, 2005.

[93] Arias, J.P., C. Busso, and N.B. Yoma. Shape-based modeling of the fundamental frequency contour for emotion detection in speech. Computer Speech & Language, 2014, 28(1): 278-294.

[94] Cowie, R., E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor. Emotion recognition in human-computer

interaction. IEEE Signal processing magazine, 2001, 18(1): 32-80.

[95] Sant'Ana, R., R. Coelho, and A. Alcaim. Text-independent speaker recognition based on the hurst parameter and the multidimensional fractional brownian motion model. IEEE Transactions on Audio, Speech, and Language Processing, 2006, 14(3): 931-940.

[96] Zao, L., D. Cavalcante, and R. Coelho. Time-frequency feature and ams-gmm mask for acoustic emotion classification. IEEE signal processing letters, 2014, 21(5): 620-624.

[97] Mencattini, A., E. Martinelli, G. Costantini, M. Todisco, B. Basile, M. Bozzali, and C. Di Natale. Speech emotion recognition using amplitude modulation parameters and a combined feature selection procedure. Knowledge-Based Systems, 2014, 63: 68-81.

[98] Tato, R., R. Santos, R. Kompe, and J.M. Pardo. Emotional space improves emotion recognition. In: *Seventh International Conference on Spoken Language Processing*. 2002.

[99] Idris, I. and M.S.H. Salam. Emotion detection with hybrid voice quality and prosodic features using neural network. In: *2014 4th World Congress on Information and Communication Technologies (WICT 2014)*. IEEE, 2014.

[100] Kächele, M., D. Zharkov, S. Meudt, and F. Schwenker. Prosodic, spectral and voice quality feature selection using a long-term stopping criterion for audio-based emotion recognition. In: *2014 22nd International Conference on Pattern Recognition*. IEEE, 2014.

[101] Huang, Y., G. Zhang, Y. Li, and A. Wu. Improved emotion recognition with novel task-oriented wavelet packet features. In: *International Conference on Intelligent Computing*. Springer, 2014.

[102] Ziółko, M., P. Jaciów, and M. Igras. Combination of fourier and wavelet transformations for detection of speech emotions. In: *2014 7th International Conference on Human System Interactions (HSI)*. IEEE, 2014.

[103] Idris, I. and M.S. Salam. Improved speech emotion classification from spectral coefficient optimization. In: *Advances in machine learning and signal processing*.Springer, 2016. p. 247-257.

[104] Espinosa, H.P., C.A.R. García, and L.V. Pineda. Features selection for primitives estimation on emotional speech. In: *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010.

[105] Wang, K., N. An, B.N. Li, Y. Zhang, and L. Li. Speech emotion recognition using fourier parameters. IEEE Transactions on Affective Computing, 2015, 6(1): 69-75.

[106] Ghosh, S., E. Laksana, L.-P. Morency, and S. Scherer. Representation learning for speech emotion recognition. In: *Interspeech*. 2016.

[107] Schuller, B. and G. Rigoll. Recognising interest in conversational speech-comparing bag of frames and supra-segmental features. In: *Proc. Interspeech 2009, Brighton, UK*. 2009.

[108] El Ayadi, M., M.S. Kamel, and F. Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition, 2011, 44(3): 572-587.

[109] Origlia, A., F. Cutugno, and V. Galatà. Continuous emotion recognition with phonetic syllables. Speech Communication, 2014, 57: 155-169.

[110] Sethu, V., E. Ambikairajah, and J. Epps. On the use of speech parameter contours for emotion recognition. EURASIP Journal on Audio, Speech, and Music Processing, 2013, 2013(1): 19.

[111] WJ, H., L. HF, and H. JQ. Speech emotion recognition with combined short and long term features. J Tsinghua Univ (Sci & Tech), 2008, 48(1): 708-714 (In Chinese).

[112] J, C., L. HF, M. L, C. X, and C. XM. Multi-granularity feature fusion for dimensional speech emotion recognition. Journal of Signal Processing, 2017, 33(3): 374-382 (in Chinese).

[113] Deng, J., N. Cummins, J. Han, X. Xu, Z. Ren, V. Pandit, Z. Zhang, and B. Schuller. The university of passau open emotion recognition system for the multimodal emotion challenge. In: *Chinese Conference on Pattern Recognition*. Springer, 2016.

[114] Lee, C.M. and S.S. Narayanan. Toward detecting emotions in spoken dialogs. IEEE transactions on speech and audio processing, 2005, 13(2): 293-303.

[115] Schuller, B., G. Rigoll, and M. Lang. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In: *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2004.

[116] Wu, C.-H. and W.-B. Liang. Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. IEEE Transactions on Affective Computing, 2011, 2(1): 10-21.

[117] Wold, S., M. Sjöström, and L. Eriksson. Pls-regression: A basic tool of chemometrics. Chemometrics and intelligent laboratory systems, 2001, 58(2): 109-130.

[118] Vinzi, V.E., L. Trinchera, and S. Amato. Pls path modeling: From foundations to recent developments and open issues for model assessment and improvement. In: *Handbook of partial least squares*.Springer, 2010. p. 47-82.

[119] Vapnik, V. The nature of statistical learning theory: Springer science & business media, 2013.

[120] Campbell, C. An introduction to kernel methods. Studies in Fuzziness and Soft Computing, 2001, 66: 155-192.

[121] Smola, A.J. and B. Schölkopf. A tutorial on support vector regression. Statistics and computing, 2004, 14(3): 199-222.

[122] Grimm, M., K. Kroschel, and S. Narayanan. Support vector regression for automatic recognition of spontaneous emotions in speech. In: *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*. IEEE, 2007.

[123] Giannakopoulos, T., A. Pikrakis, and S. Theodoridis. A dimensional approach to emotion recognition of speech from movies. In: *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009.

[124] Kanluan, I., M. Grimm, and K. Kroschel. Audio-visual emotion recognition using an emotion space concept. In: *2008 16th European Signal Processing Conference*. IEEE, 2008.

[125] Wöllmer, M., M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll. Lstm-modeling of continuous emotions in an audiovisual affect recognition framework. Image and Vision Computing, 2013, 31(2): 153-163.

[126] Schuller, B., M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic. Avec 2011–the first international audio/visual emotion challenge. In: *International Conference on Affective Computing and Intelligent Interaction*. Springer, 2011.

[127] Chao, L., J. Tao, M. Yang, Y. Li, and Z. Wen. Long short term memory recurrent neural network based multimodal dimensional emotion recognition. In: *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2015.

[128] Chen YL, Cheng YF, Chen XQ, Wang HX, and L. C. Speech emotion estimation in pad 3d emotion space. Journal of Harbin institute of technology, 2018, 50(11): 160-166 (in Chinese).

[129] WJ, H., L. HF, and M. L. Considering relative order of emtional degree in dimensional speech emotion recognition. Signal Processing, 2011, 27(11): 1658-1663 (In Chinese).

[130] Tanaka, A., A. Koizumi, H. Imai, S. Hiramatsu, E. Hiramoto, and B. de Gelder. I feel your voice: Cultural differences in the multisensory perception of emotion. Psychological science, 2010, 21(9): 1259-1262.

[131] Liu, P., S. Rigoulot, and M.D. Pell. Culture modulates the brain response to human expressions of emotion: Electrophysiological evidence. Neuropsychologia, 2015, 67: 1-13.

[132] Liu, P., S. Rigoulot, and M.D. Pell. Cultural differences in on-line sensitivity to emotional voices: Comparing east and west. Frontiers in human neuroscience, 2015, 9: 311.

[133] Elfenbein, H.A. and N. Ambady. On the universality and cultural specificity of emotion recognition: A meta-analysis. Psychological bulletin, 2002, 128(2): 203.

[134] Song, P. Transfer linear subspace learning for cross-corpus speech emotion recognition. IEEE Transactions on Affective Computing, 2017(1): 1-1.

[135] Sagha, H., P. Matejka, M. Gavryukova, F. Povolný, E. Marchi, and B.W. Schuller. Enhancing multilingual recognition of emotion in speech by language identification. In: *Interspeech*. 2016.

[136] Kaya, H. and A.A. Karpov. Efficient and effective strategies for cross-corpus acoustic emotion recognition. Neurocomputing, 2018, 275: 1028-1034.

[137] Feraru, S.M. and D. Schuller. Cross-language acoustic emotion recognition: An overview and some tendencies. In: *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2015.

[138] Böck, R., I. Siegert, M. Haase, J. Lange, and A. Wendemuth. Ikannotate–a tool for labelling, transcription, and annotation of emotionally

coloured speech. In: *International Conference on Affective Computing and Intelligent Interaction*. Springer, 2011.

[139] Cowie, R., E. Douglas-Cowie, S. Savvidou*, E. McMahon, M. Sawey, and M. Schröder. 'Feeltrace': An instrument for recording perceived emotion in real time. In: *ISCA tutorial and research workshop (ITRW) on speech and emotion*. 2000.

[140] Zenk, R., M. Franz, and H. Bubb. Emocard–an approach to bring more emotion in the comfort concept. SAE International Journal of Passenger Cars-Mechanical Systems, 2008, 1(2008-01-0890): 775-782.

[141] Bradley, M.M. and P.J. Lang. Measuring emotion: The self-assessment manikin and the semantic differential. Journal of behavior therapy and experimental psychiatry, 1994, 25(1): 49-59.

[142] Lang, P.J. International affective picture system (iaps): Affective ratings of pictures and instruction manual. Technical report, 2005.

[143] Broekens, J. and W.-P. Brinkman. Affectbutton: A method for reliable and valid affective self-report. International Journal of Human-Computer Studies, 2013, 71(6): 641-667.

[144] Ringeval, F., A. Sonderegger, J. Sauer, and D. Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In: *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 2013.

[145] Siegert, I. and A. Wendemuth. Ikannotate2–a tool supporting annotation of emotions in audio-visual data. Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2017, 2017: 17-24.

[146] Grimm, M., K. Kroschel, and S. Narayanan. The vera am mittag german audio-visual emotional speech database. In: *2008 IEEE international conference on multimedia and expo*. IEEE, 2008.

[147] McKeown, G., M.F. Valstar, R. Cowie, and M. Pantic. The semaine corpus of emotionally coloured character interactions. In: *2010 IEEE International Conference on Multimedia and Expo*. IEEE, 2010.

[148] Schuller, B., M. Valster, F. Eyben, R. Cowie, and M. Pantic. Avec 2012: The continuous audio/visual emotion challenge. In: *Proceedings of the 14th ACM international conference on Multimodal interaction*. ACM, 2012.

[149] Busso, C., M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, and S.S. Narayanan. Iemocap: Interactive emotional dyadic motion capture database. Language resources and evaluation, 2008, 42(4): 335.

[150] Ekman, P. and W.V. Friesen. Measuring facial movement. Environmental psychology and nonverbal behavior, 1976, 1(1): 56-75.

[151] Davidson, R.J. Affective style, psychopathology, and resilience: Brain mechanisms and plasticity. American Psychologist, 2000, 55(11): 1196.

[152] Banse, R. and K.R. Scherer. Acoustic profiles in vocal emotion expression. Journal of personality and social psychology, 1996, 70(3): 614.

附中文参考文献:

[38] 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

[111] 韩文静,李海峰,韩纪庆. 基于长短时特征融合的语音情感识别方法. 清华大学学报: 自然科学版, 2008. 48(1):708-714.

[112] 陈婧,李海峰,马琳,陈肖,陈晓敏.多粒度特征融合的维度语音情感识别方法. 信号处理, 2017. 33(3): 374-382.

[128] 陈逸灵,程艳芬,陈先桥,王红霞,李超. PAD 三维情感空间中的语音情感识别. 哈尔滨工业大学学报, 2018. 50(11):160-166.

[129] 韩文静,李海峰,马琳. 考虑情感程度相对顺序的维度语音情感识别. 信号处理, 2011. 27(11): 1658-1663.