文章编号：1003-0530(2023)04-0571-17

语音情感识别综述

陶建华[1] 陈俊杰[2] 李永伟[3]
（1. 清华大学自动化系，北京 100084；2. 天津师范大学计算机与信息工程学院，天津 300382；
3. 中国科学院自动化研究所模式识别国家重点实验室，北京 100190）

摘 要： 语音情感识别是利用计算机建立语音信息载体与情感度量之间的关系，并赋予计算机识别、理解人类情感的能力，语音情感识别在人机交互中起着重要作用，是人工智能领域重要发展方向。本文从语音情感识别在国内外发展历史以及开展的一系列会议、期刊和竞赛入手，分别从6个方面对语音情感识别的研究现状进行了梳理与归纳：首先，针对情感表达从离散、维度模型进行了阐述；其次，针对现有的情感数据库进行了统计与总结；然后，回顾了近20年部分代表性语音情感识别发展历程，并分别阐述了基于人工设计的语音情感特征的情感识别技术和基于端到端的语音情感识别技术；在此基础之上，总结了近几年的语音情感识别性能，尤其是近两年在语音领域的重要会议和期刊上的语音情感识别相关工作；介绍了语音情感识别在驾驶、智能交互领域、医疗健康，安全等领域的应用；最后，总结与阐述了语音情感识别领域仍面临的挑战与未来发展方向。本文旨在对语音情感识别相关工作进行深入分析与总结，为语音情感识别相关研究者提供有价值的参考。

关键词： 语音情感识别；情感特征；情感分析；情感表达
中图分类号： TP37 **文献标识码：** A **DOI：** 10. 16798/j. issn. 1003-0530. 2023. 04. 001

引用格式： 陶建华，陈俊杰，李永伟. 语音情感识别综述[J]. 信号处理，2023，39(4)：571-587. DOI：10. 16798/j. issn. 1003-0530. 2023. 04. 001.

Reference format：TAO Jianhua，CHEN Junjie，LI Yongwei. Review on speech emotion recognition[J]. Journal of Signal Processing，2023，39(4)：571-587. DOI: 10. 16798/j. issn. 1003-0530. 2023. 04. 001.

Review on Speech Emotion Recognition

TAO Jianhua[1] CHEN Junjie[2] LI Yongwei[3]
（1. Department of Automation，Tsinghua University，Beijing 100084，China；2. College of Computer and Information Engineering，Tianjin Normal University，Tianjin 300382，China；3. National Laboratory of Pattern Recognition，Institute of Automation，Chinese Academy of Sciences，Beijing 100190，China）

Abstract: Speech emotion recognition is a kind of technology that uses computers to create the relationship between speech and emotion measurement，and provides computers with the ability to recognize and understand human emotions. Therefore，speech emotion recognition plays an important role in human-computer interaction and is a promising development direction in the field of artificial intelligence. Starting from the development history of speech emotion recognition and a series of conferences and competitions at home and abroad，this paper reviews the current research status of speech emotion recognition from six aspects. Firstly，the discrete and dimensional models for emotional representation are described. Secondly，the current commonly used speech emotion databases are summarized in detail. Thirdly，representative speech emotion recognition development history is reviewed in the past 20 years，and the speech emotion recognition technology based on hand-crafted speech emotion features and end-to-end framework are described，respectively. Then，the perfor-

收稿日期：2023-02-20；修回日期：2023-04-25
基金项目：国家自然科学基金(U21B2010，62201571)

mance of speech emotion recognition in recent years is summarized, especially the major conferences and journals in the speech signal field in the past two years. Then, the applications of speech emotion recognition in driving, intelligent interaction, medical health, safety and other fields are introduced. Finally, the challenges and trends in the field of speech emotion recognition are described from the three aspects, including speech emotion database, speech emotion features, and algorithms/models. This paper aims to analyze the related work of speech emotion recognition in detail and provides a valuable reference for researchers who are engaged in speech emotion recognition research filed.

Key words: speech emotion recognition; emotion feature; emotion analysis; emotion expression

1 引言

早在80年代,美国麻省理工学院(MIT)的Minsky教授就提出了“要让计算机具有情感能力”的想法[1]。1995年,MIT的Picard教授首次提出了“情感计算 Affective computing”,并将“情感计算”确立为计算机领域的新学科[2]。情感识别是利用计算机建立语音、文本、图像等信息载体与情感度量之间的关系,并赋予计算机识别与理解人类情感的能力[3]。

语音是人与人交流与沟通的最直接和最简便的方式之一。语音信号中不仅包含言语内容信息,而且包含着丰富的副语言信息(情感、年龄和性别等)[4]。大家常说的“言不尽意”就是针对语音中的副语言信息的一种解释,副语言信息中包含的情感信息不仅有助于言语内容的可懂度理解,而且有助于情感理解与交互[5]。语音情感识别已成为国家人工智能发展战略布局的重要组成部分。

近年来,国内外研究者对语音情感识别领域的关注度逐步提高,开展了一系列的会议和竞赛来推动该领域的发展。2003年,第一届中国情感计算与智能交互会议在北京召开。2005年,第一届国际Affective computing and intelligent interaction (ACII)会议在北京召开,推动了语音情感识别作为情感计算领域的重要分支得到了更广泛的关注和研究。2009年,情感计算的会刊IEEE Transactions on Affective Computing 创刊,目前影响因子13.99,属于计算机领域较高水平的国际期刊。2011年,第一届音/视频情感大赛(Audio/Visual Emotion Challenge and Workshop,简称AVEC 2011)在美国召开,促进了学术界和工业界之间的交流与合作。2013年发布了第一个国际情感相关的标准W3C Emotion Markup Language (EmotionML1.0),标志着情感识别走向了国际化和标准化。2016年,第一届多模态情感识别竞赛(MEC 2016)在成都召开。2018年,ACM多模态交互国际会议(ACM Int'1 Conf. on Multimodal Interaction, ICMI)包含音视频情感识别的子任务在美国召开,推动了多模态情感识别的发展。与此同时,国内也有相关会议和竞赛召开,2018年,亚洲情感计算学术会议(ACII Asia 2018)在北京召开,提高了语音情感识别在国内学术界的关注度。2019年,《人工智能 情感计算用户界面 框架》国标正式发布,为情感交互技术提供了统一的框架。2021年,多模态情感识别挑战赛 Multimodal Sentiment Analysis Challenge: Sentiment, Emotion, Physiological-Emotion, and Stress(MuSe2021)在成都召开,推动了国内对多模态情感识别的研究。2021年,由中国中文信息学会情感计算(筹)专委会牵头的第一届中国情感计算大会在北京召开。2022年,由之江实验室发起,多单位共同参与编写的《情感计算白皮书》面向全球正式发布,为情感计算相关研究人员提供了重要参考和指导。可以看出,从2011年开始,几乎每年都举行情感相关的会议和竞赛,情感计算已成为人工智能领域的研究热点。

对于语音情感识别相关领域的研究,已有30多年的历史。总体来讲,大体上可以分为两类:基于语音情感特征的语音情感识别方法和基于端到端的语音情感识别方法。传统的语音情感识别方法主要依赖于从语音信号中提取的声学特征,基于声学特征进行情感识别。因此,提取情感相关的声学特征是语音情感识别的重要研究部分。近年来,随着深度学习的发展,深度学习技术将原始语音信号进行非线性转换,形成数据深度表征,从而代替传统的声学特征,实现端到端的语音情感识别技术。

本文对语音情感识别的相关方面进行了综合性的介绍,包括语音的情感表示模型、情感语料库和近两年的情感识别方法。第二节主要介绍了不同的情感表示模型以及优缺点。第三节对主流的

一些情感语料库做了简要的信息概括。第四节阐述了近两年的语音情感识别方法，包括数据集、模态、特征、验证方法、正确率等方面。最后，描述了语音情感识别仍面临的挑战与展望。

2　情感表示模型

情感是人类智能的重要标志，是人类大脑的一种主观意识。人类的思维和决策随时随地会受到喜、怒、哀、乐等情感起伏变化的影响。人类的情感具有模糊性、主观性、复杂性、时变性等特点，目前对于情感的表示方法可以分为两大类：离散情感论和维度情感论。

2.1　离散情感

离散情感论是用离散状态来描述情感，如高兴、生气、悲伤、厌恶等。离散情感论又可以划分为基本情感类别和复杂情感类别，基本情感类别的定义要求是适用于不同文化的人群，复杂的情感类别是由基本情感类别相互组合从而形成更为复杂的情感。对于离散情感论而言，在情感计算领域已得到广泛认可，如美国心理学家 Dalgleish[6]提出了六类情感类别，包括生气、厌恶、恐惧、高兴、悲伤和惊讶。不同的学者对于基本情感定义并不相同，Orton 等人[7]总结了一些常见的基本情感类别，如表 1 所示。离散情感模型简单明了，容易理解，但只能对有限数量的情感状态进行定性描述，且情感状态之间常常出现混淆的状况。

2.2　维度情感

维度情感论则是用连续的维度空间来描述情感，将情感定义在了不同维度空间中的一个点，而不同情感之间的相似性和差异性由空间中点与点之间的距离和角度来表示。目前，常用的维度情感空间主要以二维情感空间和三维情感空间为主。Russell 等人[8]采用效价度-激活度（Valence-Arousal，VA）两个维度空间模型来表示情感，其中激活度表示个体的神经激活水平，效价度表示个体情感状态的积极或消极性。以 Valence-Arousal 情感维度空间为例，如图 1 所示。与离散情感模型相比，维度情感模型不仅可以对情感进行定性描述，而且可以对情感进行定量描述，反映更加细微的情感变化。因此，近年来，使用维度情感模型进行情感识别的研究呈上升趋势。

Russell 等人[9]采用愉悦度-激活度-支配度（Pleasure-Arousal-Dominance，PAD）三维度空间模型来描述情感，其中愉悦度和效价度与 Valence-Arousal 相似，增加的支配度则表示个体对他人的控制状态程度，表现在人际交往中的支配或顺从、采取行动或不采取行动的冲动、语速和音量的变化等方面。通过对情感状态的描述和理解，任何一个离散的情感类别都可以表示在情感维度空间中的一个点。

表 1　不同的离散基本情感理论[7]

Tab. 1　Different discrete emotion theory[7]

学者	情感类别
Amold	生气，厌恶，勇敢，沮丧，渴望，绝望，恐惧，讨厌，希望，爱，悲伤
Gray	愤怒，恐怖，焦虑，开心
Izard	生气，轻视，厌恶，悲痛，恐惧，内疚，有趣，开心，羞耻，惊喜
Ekman, Friesen, & Ellsworth	生气，厌恶，恐惧，高兴，悲伤，惊讶
Fridja	渴望，高兴，喜爱，惊喜，惊奇，懊悔
James	害怕，悲伤，喜爱，狂暴
Mowrer	疼痛，愉悦
Oatley & Johnson-Larid	愤怒，厌恶，焦虑，快乐，悲伤
Panksepp	期待，恐惧，愤怒，恐慌
Plutchik	接受，愤怒，期待，厌恶，喜悦，恐惧，悲伤，惊喜
Tomkins	愤怒，兴趣，轻蔑，厌恶，痛苦，恐惧，欢乐，羞耻
Watson	恐惧，热爱，愤怒
Weiner & Graham	高兴，伤心

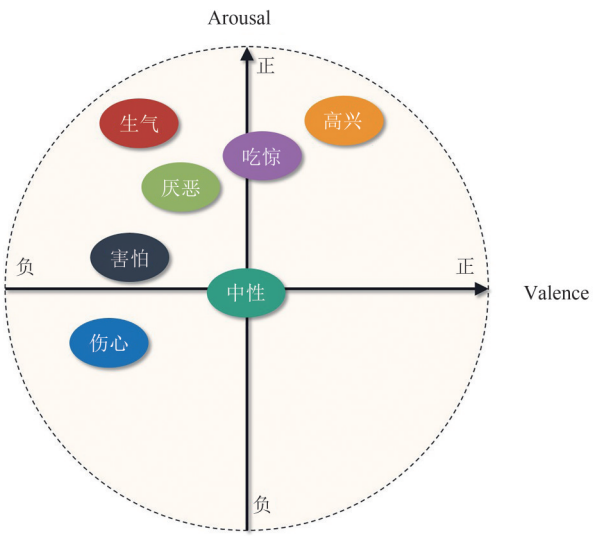

图 1 Valence-Arousal 维度情感空间

Fig. 1 Valence-Arousal emotion space

除两维和三维情感空间以外，还有多维情感空间。如：Fontaine 等人[10]采用四个维度来表达情感词含义的相似性和差异性。这些维度分别是愉悦度、激活度、支配度、和不可预测性（unpredictability）。该情感维度空间模型在 PAD 的基础上，增加了第四个维度不可预测度。它的特点是评价新奇性和不可预测性，常常表现为产生一些出乎意料的反应。此外，李海峰等人[11]还介绍了 Plutchik 抛物锥情感空间[12]，Schlosberg 倒圆锥三维情感空间[13]。

3 语音情感数据库

对于语音情感识别而言，语音情感数据库是关键因素之一。语音情感识别性能直接受数据库的质量影响，在流行的数据驱动的模型算法上表现尤为突出。因此，本文首先对语音情感数据库进行介绍。目前，关于语音情感数据库的构造并没有统一的规则，根据任务的需求构造不同的语音情感数据库。现有的情感语音库主要包括表演型、引导型、自发型。表演型情感数据库主要是在实验室环境下录制职业演员模拟不同的情感状态，职业演员的过度夸张的表情，使得易于情感识别。引导型情感数据库主要是在实验室环境下录制由工作人员激发的说话人的不同情绪状态，相比表演型语料库，其更加贴近真实的情绪反应。随着对自然场景下情感识别需求的增加，自发型语音情感数据库越来越受到重视，这类数据库主要包含真实环境下的说话人情感表达，虽然其更贴近真实场景，但也提升了情感识别的难度。本文总结了在语音情感计算领域，常用的一些离散和维度语音情感数据库，如表 2 所示。

3.1 表演型情感数据库

IEMOCAP[14]是由南加利福尼亚大学录制的音视频情感数据库，包含约 12 小时的视听数据。该数据库是由 10 名专业演员在有台词或即兴的对话场景下，引导出的情感数据。每一段对话被人为地切分成单句，每一句话至少由 3 个标注员进行离散情感（高兴、生气、中性、悲伤等）和 3 个维度（Valence、Activation、Dominance）的标注。

CASIA[15]是由中国科学院自动化研究所录制的语音情感语料库，共包括 4 个专业发音人，六种离散情感类别，总共 9600 句不同的发音。

CHEAVD[16]是由中国科学院自动化所构建的中文多模态情感语料库。语料库包含 140 分钟的电影、电视剧和脱口秀情感片段。238 位演讲者，从儿童到老人，保证了说话人的多样性。包括基本的六种情感。

EmoDB[17]是一个德语情感语料库，由柏林工业大学录制。10 名表演者（5 男 5 女）将 10 句德文（5 句短句 5 句长句）通过不同的情感表达出来，录音是在一个装有高质量录音设备的消声室中进行的。除了声音，还记录了电子声门图。演讲材料大约包括 800 个句子。

MSP-IMPROV[18]是一个行为视听情感数据库，探索自发二元即兴创作中的情感行为。场景经过精心设计，以激发现实情感。目前，该语料库含有来自六个二人组（12 名演员）的数据。参与者是来自艺术与人文学院的 UTD 学生，他们参加了戏剧课程，并有表演经验。

EMOVO[19]是第一个适用于意大利语的情感语料库。该语料库是在乌戈·博多尼基金会实验室用专业设备录制的，它由 6 个演员的声音组成，这些表演者用 14 句话模拟了 6 种情绪状态（厌恶、恐惧、愤怒、喜悦、惊讶、悲伤）和中性状态。

RAVDESS[20]是一个经过验证的情感语音和歌曲多模态数据集。24 名专业表演者用中性北美口音表述脚本，同时需要展现包括平静、快乐、悲伤、

表2　语音情感数据库

Tab. 2　Speech emotion database

数据库名称	语言	人数	句子数量	离散标签	维度标签	模态	时长	类型
IEMOCAP[14]	英语	10人	10039	中性、快乐、悲伤、愤怒、惊讶、恐惧、厌恶、沮丧、兴奋	效价度、唤醒度、支配度	音频 视频	12小时	表演 引导
CAISA[15]	汉语	4人	9600	生气、高兴、恐惧、悲伤、惊讶		音频		表演
CHEAVD[16]	汉语	238人	2322	中性、生气、高兴、悲伤、担心、焦虑、惊讶、厌恶		音频 视频	140分钟	表演
EmoDB[17]	德语	10人	800	愤怒、无聊、厌恶、焦虑/恐惧、快乐、悲伤、中性		音频 视频		表演
MSP-IMPROV[18]	英语	12人	240	生气、悲伤、快乐、中性、其他		音频 视频		表演
EMOVO[19]	意大利语	6人	588	厌恶、恐惧、愤怒、喜悦、惊讶、悲伤		音频	60分钟	表演
RAVDESS[20]	英语	24人	7356	平静、快乐、悲伤、愤怒、恐惧、惊讶、厌恶		音频 视频		表演
SEMAINE[21]	英语	20人	150		效价度、唤醒度、支配度、力量、强度	音频 视频	750分钟	引导
RECOLA[22]	法语	46人			效价度、唤醒度、支配度、一致性、参与度、表现力、融洽度	音频 视频 心电图 皮肤电信号	3.8小时	引导
eNTERFACE[23]	英语	43人	1166	愤怒、恐惧、惊讶、快乐、悲伤、厌恶		音频 视频		引导
CMU-MOSEI[24]	英语	1000多人	23453	快乐、悲伤、愤怒、恐惧、厌恶、惊讶	积极性、中性、消极性	音频 视频	65小时	自发
MSP-Podcast[25]	英语	100多人	62140	愤怒、快乐、悲伤、厌恶、惊讶、恐惧、蔑视、中立、其他	效价度、唤醒度、支配度	音频	100小时	自发

愤怒、恐惧、惊讶和厌恶在内的表情。歌曲则包含平静、快乐、悲伤、愤怒和恐惧的情绪。该语料库对7256条录音分别在情感有效性、强度和真实性方面进行了10次评级，由247名北美志愿者参与提供评分。

3.2　引导型情感数据库

SEMAINE[21]是一个情感对话语料库。在录制过程中，工作人员和体验者分别坐在单独的房间，通过屏幕和扬声器与对方交流，工作人员依次扮演四个唤起情绪反应的角色。为了实现高质量录制，该语料库采用五个高分辨率、高帧速率的摄像头和四个麦克风进行录制，共记录了20名参与者，总共100个对话记录和50个非对话记录，每个记录大约5分钟。所有记录的对话都被完全转录并注释为五个情感维度，部分注释为27个其他维度。

RECOLA[22]是一个法语多模态自发协作和情感互动语料库。在视频会议期间，参与者在完成一项需要协作的任务时被记录下来。连续同步记录不同模态的数据，即音频、视频、ECG和EDA。总共有46名参与者参加了测试，其中前5分钟的互动是为了方便注释。除了这些记录外，6名解说员还连续测量了情绪的两个维度：唤醒和效价，以及五个维度上的社会行为标签。

eNTERFACE[23]是一个视听情感数据集。该语料库的42位参与者来自14个不同的国家，在这些参与者中，有81%是男性，19%是女性。每个参与者被要求听六个连续的短篇故事，每个故事会引发特定的情感，两位专家会评估参与者的反应是否符合预期。

3.3　自发型情感数据库

CMU-MOSEI[24]是一个多模态情绪和情感识别数据集。在数据采集过程中，通过人脸检测分析来

自 YouTube 的视频，以确定帧中是否存在一个说话人，从而保证视频是独白，最后获得大约 5000 个视频。然后由 14 名专家在三个月内手动检查视频、音频和转录的质量。评估者还对每个视频进行了性别标注。该语料库拥有 65 个小时的标注视频，每个标注视频都有 6 种离散情感标签和 7 种维度标签。

MSP-Podcast[25]是一个自然主义的语音情感数据集。工作人员将播客的录制内容进行格式化和命名，然后这些内容将被自动分割和分析，排除了背景音乐、重叠语音等部分，以得到更加纯粹的包含情感信息的语音片段。在 1.7 版本中，该语料库包含了 62140 条语音，大约有 100 多个小时。在标注方面，该语料库使用基于属性的描述符（激活、支配和效价）和类别标签（愤怒、幸福、悲伤、厌恶、惊讶、恐惧、轻蔑、中立和其他）标注情感标签。

4 语音情感识别方法

语音情感识别本质上是模式识别。根据情感表示模型的不同，对语音情感特征进行分析与识别的任务。对于离散情感模型而言，其任务就是经典的分类问题，其任务是对不同情感类别进行分类。对于维度情感模型而言，其任务可理解为回归预测问题，其主要是预测情感在不同维度空间上的数值。图 2 表示语音情感识别基本框架，可以分为两类，一类是从情感数据库中提取声学特征，进入情感分类器（如：SVM，KNN，RNN，LSTM 等）；另一类是直接使用深度表征，直接进入情感分类器，进行情感分类或回归。目前，语音情感识别建模方法主要分为两类：基于声学特征提取与分析的语音情感识别建模方法和基于端到端的语音情感识别建模方法。本节主要首先介绍了近 20 年部分代表性语音情感识别发展历程，然后介绍了基于特征和端到端的语音情感识别方法，以及总结部分方法的情感识别性能。

4.1 部分代表性语音情感识别发展历程

图 3 总结了近 20 年部分代表性语音情感识别发展历程。自 20 世纪 60 年代起，隐马尔可夫模型 HMM（Hidden Markov Model）在语音识别上的成功应用使得它成为语音情感识别的首选方法之一。到 2005 年，Lin 等人[26]使用 HMM 和情感定制特征来分析时间动态和频谱特征，他们的工作取得了较好的分类结果，并加强了 Mel 子带能量和基本频率的时间动态是语音情感内容的重要指标的假设。1995 年，Vapnik 等人提出了支持向量机 SVM（Support Vector Machine），它的主要思想是使用核函数将原始输入映射到高维空间，在该空间中取得最优分类，SVM 在小样本情况下具有良好的表现。2010 年，Chavhan 等人[27]从音频文件中提取 MFCC 和 MEDC 特征，并采用 SVM 进行分类。他们在德语的柏林情感数据集上进行了实验，使用径向基函数和多项式核函数的结果分别为 93.75% 和 96.25%。

20 世纪 80 年代初，人工神经网络 ANN（Artificial Neural Network）技术开始兴起，以其思想衍生出的技术很快便应用于语音情感识别领域，并逐渐

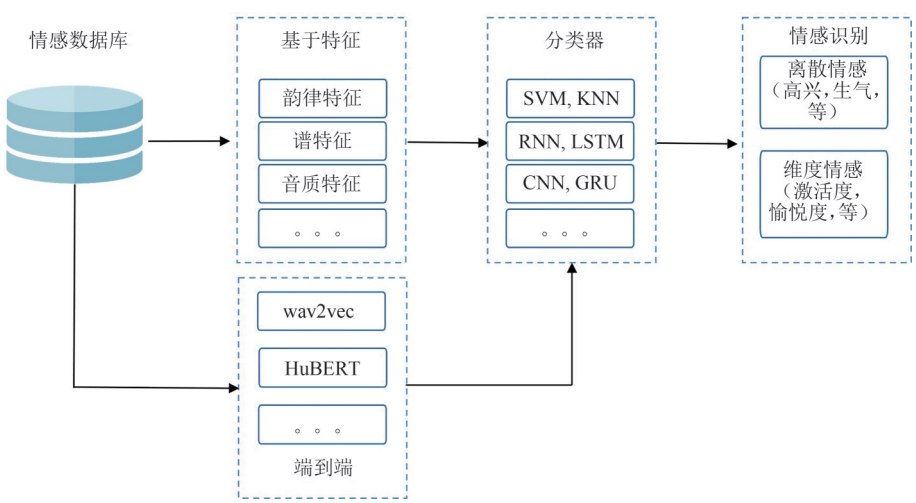

图 2 语音情感识别框架

Fig. 2 Speech emotion recognition framework

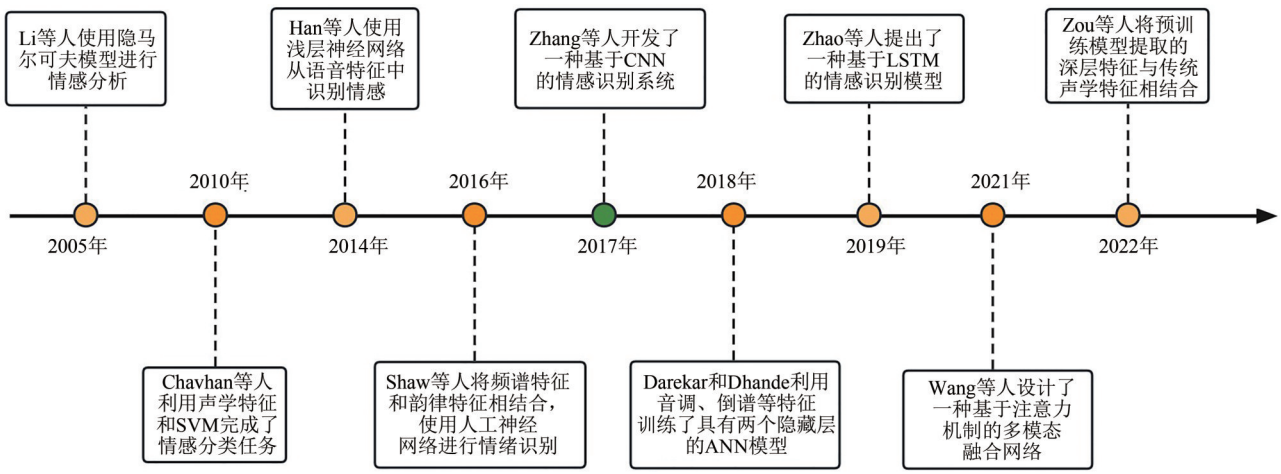

图3　部分代表性语音情感识别发展历程

Fig. 3　Representative speech emotion recognition development history

代替以往使用HMM和SVM的方法。相比传统的机器学习方法，神经网络能够进行并行计算，使得模型收敛速度大大提高，并且可以逼近数据与标签之间复杂的非线性关系，具有良好的信息综合能力。到2014年，Han等人[28]使用浅层单神经网络从话语特征中识别情感，他们的实验表明该方法的准确率比使用HMM和SVM的方法高5%-10%。2016年，Shaw等人[29]将频谱特征和韵律特征结合使用ANN识别快乐、愤怒、悲伤和中立四种情绪。2017年，Zhang等人[30]开发了一种基于深度卷积神经网络的情感识别系统。他们的实验证明该系统在EMODB数据集上获得了80%以上的准确率，比基线SVM高出约20%。2018年，Darekar和Dhande[31]提取NMF分析、音调分析和倒谱特征，然后将PCA应用于特征向量来降低维数，最后送入一个具有两个隐藏层的ANN进行训练。他们在RAVDESS数据集上应用了该模型，并表明他们比应用于RAVDESS的基线方法获得了10.85%的准确度提升。

1997年，Sepp Hochreiter和Jürgen Schmidhuber[32]提出了长短期记忆网络LSTM（Long Short Term Memory）。LSTM优化了RNN中存在的长期依赖问题，LSTM的门控结构使得它可以充分挖掘序列数据中的上下文相关信息，在时序数据预测方面具有良好的表现。2019年，Zhao等人[33]提出了一种基于两类网络块的框架，一类是学习局部特征的单层卷积块，另一类是学习全局特征的LSTM块，在语音情感识别任务得到了较好的性能。

虽然LSTM优化了RNN中上下文信息无依赖的问题，但它也存在着一些问题，例如数据在长距离上的依赖会被淡化，无法并行计算等，注意力机制很好地解决了这些问题，它能捕捉数据在全局上的信息联系，并且并行计算，解决了训练速度受序列数据长度影响的限制。2021年，Wang等人[34]采用共享权值的多模态Transformer捕捉模态间的依赖性，同样取得了很好的效果。最后，预训练模型在语音识别任务上的成功使得它越来越受语音情感识别研究的关注。2022年，Zou等人[35]利用Wav2vec预训练模型提取语音的深层特征，并与传统声学特征相结合进行情感识别。其消融实验证明加入深层特征后取得了更好的性能表现。

4.2　情感识别模型

4.2.1　基于语音声学特征的情感识别

基于语音声学特征的语音情感识别方法通常是从语音信号中提取一些人工设计的特征，将这些特征送入分类器完成识别任务。对于该语音情感识别方法而言，情感特征是语音情感识别中的重要环节，提取的情感特征直接影响最终的情感识别性能[11]。常用的语音情感特征主要分为三大类：韵律特征、谱特征和音质特征。韵律特征是指基频、音强、音长、音调、停顿、语速、时长等特征，它在语音情感识别领域已得到了广泛的认可。谱特征一般认为是反应发声运动和声道形状变化的特征，具体表现为频谱能量的分布（共振峰）、线性预测倒谱系数（Linear Prediction Cepstral Coefficients，LPCC）、

梅尔频率倒谱系数(Mel Frequency Cepstrum Coefficient, MFCC)等特征。音质特征是反应语音音质的特征,如喘息和哽咽等,具体表现为声门频率扰动(Jitter)、振幅扰动(Shimmer)等特征。Huang等人[36]分析了不同情感类别的语音信号特征差异,并对特征贡献度进行了排序,研究发现基频(F0)、音强和时长对情感种类的贡献度较大。Juslin等人[37]对这些特征在情感维度空间中进行了分析,结果表明停顿、音强和基频与arousal关联度较强,且F0和arousal值正相关。Mori等人[38]发现第一共振峰(F1)和第二共振峰(F2)与valence关联度较强,且F1和F2与valence值正相关。Grimm等人[39]采用基频、能量、MFCC等声学特征相结合的方式,采用support vector regression (SVR)作为回归分类器,对valence-arousal-dominance进行了回归预测,结果显示在arousal维度上的预测效果相对较好。陈逸灵等人[40]采用MFCC和语谱图相结合的方式,采用SVR作为回归预测模型,与单独用MFCC相比,维度语音情感识别性能提升明显。Tato等人[41]指出:传统的韵律特征、能量等特征可以很好地表征arousal维度,而valence维度则无法很好地表征。因此,为了更好地进行情感表征,有必要融合更多的特征对语音中的情感信息进行深入分析。

Schuller等人[42]提出了openSMILE工具集,它可以提取语音的韵律、音质、谱特征等相关高维统计特征,在情感识别领域得到了广泛应用,openSMILE特征提取方法也一直在被扩展与更新。如:384维特征参数集(Interspeech2009 emotion challenge, IS09)[43], 1582维特征参数集(Interspeech2010 emotion challenge, IS10)[44], 6373维特征参数集(the Interspeech 2013 computational parameters challenge, IS13_ComParE)[45], 88维特征参数集(the extended Geneva Minimalistic Acoustic Parameter Set, eGeMAPS)[46]等。Atmaja等人[47]采用88维的(eGeMAPS)特征集,作为Long Short-Term Memory (LSTM)网络的输入特征,深入讨论了不同的loss函数对维度情感识别性能的影响。韩文静等人[48]采用了1582维的(IS10)声学特征集,采用K-NearestNeighbor (KNN)和SVR进行了维度情感识别。Han[49]等人采用了6373维度的(IS13_ComParE)特征集,采用BiLSTM-RNN实现了情感识别。

日本北陆先端科学技术大学院大学研究团队Akagi等人[50]认为人类在感知语音中的情感信息不是直接来自声学特征的变化,而是由情感语义或形容词表达的组合描述传递的。因此,该研究团队构建了基于听觉感知的语音情感三层模型,底层是声学特征,中间层是一些形容词的组合,上层是情感类别或维度情感空间。Elbarougy等人[51-52]利用该感知三层模型,通过模糊推理系统将三层关联,构造了简单的语音情感识别模型,比直接将声学特征与情感映射的两层模型识别率更高。但由于中间层特征值是由被试感知语音后打分,所以三层模型的情感识别方法增加了人工成本。

4.2.2 基于端到端的语音情感识别

近年来,随着深度学习技术的发展,大量的端到端学习框架模型被广泛地应用于语音情感识别中。端到端的学习思路是利用深度神经网络直接从原始语音数据中自动学习到含有情感信息、二维空间信息和时序上下文信息的情感表征形式,无须人工特征提取,进而更好地适应情感识别任务。

Cai等人[53]提出了一种多任务学习框架,同时进行自动语音识别和语音情感分类任务。该方法主要采用基于wav2vec2.0的端到端的深度神经网络模型。Yeh等人[54]使用端到端的ASR来提取基于ASR的表征来进行语音情感识别,并且在预训练的ASR模型上设计了一种分解域适应方法来提高语音识别率和目标情感语料库的识别正确率。Trigeorgis等人[55]基于Convolutional Neural Networks (CNN)从原始语音信号中自动获得语音信号的情感表征,然后使用LSTM网络学习时序信息,在valence-arousal维度上的识别结果明显优于传统方法。Latif等人[56]采用并行CNN直接从原始语音信号中捕获情感表征信息,采用LSTM对CNN的输出特征捕获上下文关系,有效地提升了情感识别率。同样地,Tzirakis等人[57]采用CNN和LSTM相结合的方式,直接将原始语音信号输入CNN网络,以捕获语音情感特征表征,最后用LSTM捕获时序关系,有效地提升了语音情感识别的精度。Sarma等人[58]使用用于说话人识别任务的time-delay neural network (TDNN)网络代替CNN网络,结合LSTM网络,进一步提升了情感识别性能。Bakhshi等人[59]采用原始语音时域信号和频域信息作为深度Conv-RNN网络

的输入，有效提取了语音信号的情感表征，实现了端到端的维度情感识别。Zhang[60]等人在CNN层加入了注意力机制，可以从语音信号中学习到更有效的情感信息，从而提升了情感识别率。Li 等人[61]提出了一种使用端到端的多任务学习和自注意力机制的语音情感识别方法。该方法直接从语音谱图中提取情感标注，采用自注意力机制更好地关注语音中和情感有关的部分。此外，该方法还采用多任务学习技术，将性别分类作为辅助任务，提高语音情感识别率。Sun 等人[62]使用残差卷积神经网络从原始语音信号中捕获情感信息并进行分类。此外，该方法还加入了说话人性别信息以进一步提高识别率。

语音原始数据中含有丰富的信息，使用数学模型将其表示为语音声学特征可能会造成情感信息的遗漏，使用端到端的方式可以直接从原始信号中捕获更完整的情感信息。尽管端到端的语音情感识别方法取得了优异的成果。然而，深度学习模型提取的情感表征空间和人类可理解的空间存在着本质的区别与不可解释性[63]。

4.3　情感识别结果

在过去的二十年里，语音情感识别领域得到了广泛的研究。研究者探索了许多的声学特征及其提取技术。基于声学特征的分类算法已被证明可以有效地完成情绪识别的任务，基于声学特征的分类方法通常会选择某个语音情感语料库，然后从原始语音中提取合适的声学特征，并为这些特征的分类灵活地选择分类器。近年来，随着计算机硬件和软件的革新与发展，通用的预训练模型逐渐成为现实，使用通用的语音预训练模型在情感识别领域取得了不错的结果。

以下主要总结了近两年在Interspeech、ICASSP等语音会议和期刊上的部分代表性的离散和维度情感识别结果，表3和表4分别统计了基于语音声学特征和基于端到端的情感识别结果。从近两年的研究进展可以看出，很多语音情感识别的方法都已经在一些特定的数据集上取得了较好的结果，但仍有提升空间。

如表3和表4所示，基于语音声学特征和基于端到端的情感识别方法都含有单模态（仅语音）情感识别和多模态（语音、文本、视频等）情感识别。相比多模态情感识别，单模态情感识别仅使用语音模态作为输入且语音情感语料库体量相对较小，其更容易受说话人、语言、录音环境等非情感因素的影响。为了缓解语音情感信号在性别、年龄、语言和文化背景等不同领域之间差异过大的问题，Fan等人[78]提出了一种自适应领域感知表示学习方法，该方法利用领域知识来提取领域感知特征，使用注意力模型将领域知识嵌入到情感表示空间中；Zhu等人[67]提出了一种可以学习独立于语料库的情感编码网络结构。该网络由共享情感编码器、多个情感分类器和对抗语料库鉴别器组成，通过对抗性学习和多任务学习的训练，该方法可以在很大程度上减少不同语料库中语言、录音环境等因素对情感分类的负面影响；类似地，Lian等人[74]同样使用对抗性学习和多任务学习来学习独立于说话人的通用表征，两者均采用了多目标优化的多任务学习策略，但采用的评测方式不同，对性能亦有影响。此外，语音情感语料库通常是在录音室中使用专业设备录制，而真实环境下的语音信号可能会随录音环境变化而变化。因此，为了克服真实环境下噪声对情感识别性能的影响，Latif等人[79]融合了密集卷积网络（DenseNet）、长短期记忆网络（LSTM）和highway网络，以学习对噪声具有鲁棒性的表征；Chen等人[81]设计了一种关键稀疏Transformer网络，该方法采用预训练的Wav2vec模型和RoBERT模型提取语音模态和文本模态的深层特征用于情感识别，使用预训练模型提取的深层特征相比语音声学特征具有更好的泛化性，使用多种模态信息作为输入同样增强了模型的泛化性，因而该方法得到了比前者更好的性能表现。

相比单模态情感识别，多模态情感识别更关注于对模态融合的研究。Pan等人[75]和Wang等人[34]的研究验证了注意力机制作为多模态融合方式的有效性，前者使用openSMILE工具集从语音中提取出IS13-ComParE特征集，使用3D-CNN预训练模型提取视频中的肢体语言特征，使用Word2vec预训练模型将文本转化为特征向量，以语音特征为查询向量，视频特征与文本特征为键向量和值向量，使用注意力机制融合三模态的信息并输入长短期记忆网络LSTM，单模态特征也输入LSTM，单模态与多模态结果进行串联输入全连接层用于情感识别，在

表3 基于语音声学特征的情感识别结果统计

Tab. 3 Statistics of emotion recognition results based on acoustic features of speech

作者	数据库	模态	特征	分类器	使用数据	验证方法	UA /%	WA /%	CCC
Cai et al.[64]	IEMOCAP	Speech	log power-spectrum	CNN LSTM MAML	Angry, Happy, Sad, Neutral	5-fold cross validation	70.32	76.64	
Liu et al. [65]	IEMOCAP	Speech	MFCCs	CNN Bi-LSTM	Angry, Happy +(excited), Sad, Neutral	5-fold cross validation	78.30	79.52	
Peng et al.[66]	RECOLA SEWA	Speech	MMCG	LSTM	Arousal, Valence	5-fold cross validation			0.865 0.524 0.572 0.534
Zhu et al.[67]	EmoDB CREMA-D	Speech	Log-mel spectro-gram	ACRNN	Angry, Happy +(excited), Sad, Neutral, Disgust, Fear, Boredom	10-fold cross validation	77.20 73.40	77.90 79.60	
Latif et al.[68]	IEMOCAP MSP-IMPROV	Speech	Emobase2010	Mixup GAN	Angry, Happy +(excited), Sad, Neutral			46.60	
Liu et al. [69]	IEMOCAP	Speech Text	Mel-spectrogram MFCCs Glove	CNN Bi-LSTM Attention	Angry, Happy, Sad, Neutral	5-fold cross validation	70.08	72.39	
Chen et al. [70]	IEMOCAP	Speech Text	Log-mel spectrogram	CNN Bi-LSTM Attention ALBERT	Angry, Happy +(excited), Sad, Neutral	5-fold cross validation	72.05	71.06	
Li et al.[71]	IEMOCAP RAVDESS	Speech Text	Filterbank GloVe	Bi-LSTM	Angry, Happy, Sad, Neutral	5-fold cross validation	73.50 70.80	72.70 72.00	
Kumar et al.[72]	IEMOCAP	Speech Text	Mel-spectrogram MFCC Chroma	GRU Attention BERT	Angry, Happy +(excited), Sad, Neutral	5-fold cross validation	75.00	71.70	
Wang et al.[34]	IEMOCAP	Speech Text	MFCCs One-hot embedding	Transformer Attention	Angry, Happy, Sad, Neutral	5-fold cross validation	77.10	76.80	
Hou [73]	IEMOCAP MELD	Speech Text	LMBFs BERT	GRU Attention	Angry, Happy +(excited), Sad, Neutral	5-fold cross validation	77.60	75.60 64.90	
Lian et al.[74]	IEMOCAP	Speech Text	IS13-ComparE ELMo	GRU Multi-Head-Attention	Angry, Happy +(excited), Sad, Neutral	session 5 validation		82.68	
Pan et al.[75]	IEMOCAP	Speech Visual Text	IS13-ComparE Word2vec	LSTM Attention	Angry, Happy +(excited), Sad, Neutral, Frustrated			73.94	
Khare et al.[76]	CMU-MOSEI	Speech Visial Text	LFBE GloVe	Transformer Multi-Head-Attention	Angry, Happy, Sad, Disgust, Surprise, Fear			66.40	

IEMOCAP数据集上获得了73.94%的平均准确率；后者则采用共享权值的多模态Transformer捕捉模态间的依赖性，Transformer结构能够更好地学习时序信息，因此相比前者取得了2.86%的平均准确率提高。此外，现有的多模态融合方法一般是基于话语级别的特征融合，很大程度上忽略了细粒度级别上不同模态特征的动态交互。为了弥补这一缺陷，Shen等人[80]提出了一种多模态细粒度融合网络

表4　基于端到端的情感识别结果统计

Tab. 4　Statistics of emotion recognition results based on end-to-end

作者	数据库	模态	分类器	使用数据	验证方法	UA /%	WA /%	CCC
Yeh et al.[54]	LibriSpeech IEMOCAP	Speech	Bi-LSTM GRU Attention	Angry, Happy, Sad, Neutral	5-fold cross validation	64.40	63.10	
Feng et al.[77]	IEMOCAP	Speech	Bi-LSTM Attention	Angry, Happy+(excited), Sad, Neutral	5-fold cross validation	69.70	68.60	
Zou et al.[35]	IEMOCAP	Speech	Wav2vec AlexNet LSTM	Angry, Happy+(excited), Sad, Neutral	5-fold cross validation	71.05	69.80	
Fan et al.[78]	IEMOCAP	Speech	Attention	Angry, Happy, Sad, Neutral	10-fold cross validation	65.86	73.02	
Latif et al.[79]	IEMOCAP MSP-IMPROV	Speech	LSTM	Angry, Happy+(excited), Sad, Neutral	10-fold cross validation	64.10 56.20		
Cai et al.[53]	IEMOCAP	Speech	Wav2vec	Angry, Happy+(excited), Sad, Neutral	10-fold cross validation		78.15	
Sun et al.[62]	CASIA EMODB IEMOCAP	Speech	Residual-CNN	Angry, Happy, Sad, Neutral, Excited, Frustrated	10-fold cross validation	84.60 90.30 71.50		
Shen et al.[80]	IEMOCAP	Speech Text	GRU LSTM Attention	Angry, Happy, Sad, Neutral	5-fold cross validation	76.40	75.90	
Chen et al.[81]	IEMOCAP LSSED	Speech Text	Wav2vec Attention	Angry, Happy+(excited), Sad, Neutral	5-fold cross validation	75.30 55.70	74.30 65.10	
Li et al.[82]	IEMOCAP	Speech Text	Wav2vec BERT Attention	Angry, Happy+(excited), Sad, Neutral	5-fold cross validation	81.70	80.36	
Tzirakis et al. [57]	RECOLA	Speech, Video, ECG, EDA	CNN LSTM	Arousal, Valence				0.787 0.440

wise。该方法使用两个长短期记忆网络LSTM作为音频特征和文本特征的交互单元，以模拟音频特征和文本特征在单词级别上的动态交互，并获得了75.9%的加权准确率和76.4%的未加权准确率；Liu等人[69]提出了一种名为Gated Bidirectional Alignment Network（GBAN）的新模型。该模型由基于双向注意力的对齐网络和组门控融合层组成，对齐网络用于捕捉语音和文本之间的对齐信息，组门控融合层用于学习各模态对于最终预测结果的贡献；Li等人[71]设计了一种时间对齐的均值-最大值池化机制来捕捉话语中隐含的微妙和细粒度的情绪，此外，该方法还提出了一个跨模态激发模块，对跨模态嵌入进行样本的特定调整，并通过与其他模态对齐的潜在特征自适应地重新校准相应的值。细粒度多模态情感识别方法显式地对齐语音片段和文本单元，从而提高了情感识别的准确率，但在某些情况下，语音中的文本可能会发生漏识别或错误识别，使得对齐的结果不完整或不准确，进而对情感识别的性能造成影响。

此外，BERT等预训练模型在自然语言处理领域的成功也推动了语音通用预训练模型的发展，进而促进了语音情感识别领域的进步。语音预训练模型直接从原始语音信号中学习表征信息，通过在大规模的数据集上预训练，在体量相对较小的情感语料库上微调即可达到较好的效果。单模态情感识别方面，Zou等人[35]利用多层声学信息，包括预训

练模型 Wav2vec 提取的深层特征，MFCC 特征和声谱图，进行端到端的语音情感识别。在 IEMOCAP 数据集上获得了 69.8% 的平均准确率。在多模态情感识别方面，Li 等人[82]选择预训练模型 WavLM 和 BERT 分别将原始语音和文本编码为帧级和词级嵌入向量，利用 Transformer 结构进行模态融合用于情感分类。该方法充分利用了文本模态的信息，因此相比前者提升了 10.5% 的平均准确率。相比基于语音声学特征的情感识别方法，采用预训练模型的端到端的情感识别方法可以得到更具泛化性的深层表征，极大地缓解了语音情感语料库数据量小和说话人少的问题，因而取得了更好的性能表现。

最后，从统计情感识别结果来看，近两年的研究在实验数据集上已经取得了良好的成绩。如在常用的 IEMOCAP 数据集上，语音模态的情感识别率已经可以达到 65% 左右，语音融合语义信息的情感识别率达到 75% 左右。统计结果表明：(1)语音情感识别性能还有很大的提升空间。(2)结合多模态的语音情感识别有助于提高情感识别率，且正在呈上升趋势。(3)对于语音情感特征创新的研究较少，对于模型算法创新的研究较多。

5 语音情感识别应用

语音情感识别的研究涉及多个学科领域，包括计算机科学、心理学、情感学等。作为一项非接触式的情感识别技术，语音情感识别已被应用于多个领域。

5.1 交通安全领域

在交通安全领域，通过实时识别驾驶员的情感状态，帮助提高驾驶员的安全性和舒适性，从而降低交通事故发生的风险。以下是一些具体的应用场景：

驾驶员疲劳检测：通过对驾驶员的声音特征进行语音情感分析，判断驾驶员是否存在疲劳或困倦的情况。当驾驶员存在疲劳状态时，系统可以发出警报提醒驾驶员进行休息，进而保障驾驶员的安全[83]。

驾驶员情绪检测：语音情感识别系统实时检测驾驶员的情绪状态，当驾驶员的情绪处于不稳定状态时，车辆可以根据情绪状态进行相应的调整，例如调整车内音乐、灯光等，从而提高驾驶员的舒适感和安全性[84]。

5.2 智能交互领域

在智能交互领域，人性化的人机交互体验始终是智能交互的前进目标之一。语音交互作为一种非接触式的交互方式不仅方便而且快捷，以下是一些具体的应用场景：

语音助手：拥有情感识别技术的语音助手可以了解用户的情绪状态，从而更好地响应用户的需求和指令。例如，当用户出现焦虑或生气情绪时，语音助手可以通过更友好和亲切的回应来舒缓用户情绪[85]。

语音客服：语音情感识别技术帮助语音客服分析客户情绪状态，从而提供更加舒适的客户服务。当用户在与语音客服交互时出现不满或不耐烦的情况时，语音客服可以通过更加温和的回答来缓解用户情绪[86]。

5.3 医疗健康领域

在医疗健康领域，患者的情绪状态对诊断和治疗的效果有着很大的影响。通过语音情感识别技术，可以更好地了解患者的情绪状态，为医疗团队提供更全面和准确的诊断和治疗支持。以下是一些具体的应用场景：

心理诊断和治疗：在进行心理评估时，语音情感识别系统可以协助医生识别患者的情绪状态，判断患者是否存在抑郁、焦虑等心理障碍，并进行相应的治疗[87]。

老年人护理：当老年人在生活中出现情绪困扰或不适时，护理人员通过情感识别系统显式地了解老年人的情绪，及时采取相应的护理措施，缓解老年人的情绪和疼痛[88]。

5.4 信息安全领域

语音情感识别技术在信息安全领域的应用主要是通过识别说话人的情绪状态，进而识别出可能存在的欺诈、诈骗、间谍活动等威胁。在测谎的应用上，语音情感识别技术主要是通过分析声音中的语调、语速、声音质量、说话人的表现和情感等多个方面的特征，来分析说话人是否在说谎。在实际应用中，其他测谎技术可以结合语音情感识别技术，进一步提高测谎的准确性和可靠性[89]。

6　挑战与展望

随着计算机技术的发展，人们对于计算机的需求已不是仅仅满足功能需求，更需要照顾到人的情感需求，深度学习技术更是助推了这一需要的实现。在过去的十年里，语音情感识别的研究已经成为人机交互和语音处理领域的一项重要工作。情感识别涉及多个学科领域，特别是应用心理学和计算机科学方面，在对情感理解上，计算机与人有着巨大的差异。近年来，语音情感识别技术已经得到了巨大的提升，但仍然存在着一些值得研究的问题，如情感语料库，情感声学特征，情感识别模型等方面。

6.1　情感语料库

现有的深度学习方法需要大量数据进行训练，然而，与语音识别语料库相比，目前的语音情感数据集的语料总量较少，说话人较少，这常常会使得深度学习模型容易拟合于训练数据，降低其鲁棒性。另一方面，现有的多数语音情感语料库通常是在录音室内录制的，缺乏真实环境下的情感语料库，带有噪声或混响等复杂环境下的语音信息会极大地改变其声学特征，样本中的噪声会降低情感识别的性能。然而构建情绪定义准确的情感语料库仍然是一大挑战，受文化、性格、性别和年龄等因素的影响，人们对情感的定义缺乏共识。在对语音数据进行情感标注时，标注人员受主观感受影响较大，特别是使用维度情感标注法。因此，构建生态性较好的语音情感语料库将是未来语音情感识别领域的发展方向之一。

6.2　情感声学特征

目前，基于声学特征的语音情感识别方法常常使用从语音中提取的韵律、音质、谱等特征完成情感识别任务，探索寻找最能区分情感差别的情感声学特征仍然是研究者值得关注的问题。一方面，目前并没有找到统一的可以有效表达情感的特征，对于不同的语料库具有区分性的情感特征不同，导致情感特征没有普适性。另一方面，在现实场景中，情感不仅能通过语音表达，也能通过表情信息、肢体动作信息来表达。因此寻找具有鲁棒性、普适性的语音情感特征，并且和多模态融合的语音情感识别将是未来语音情感识别发展的必然趋势。

6.3　情感识别模型

目前，语音情感识别主要依靠数据驱动的深度学习模型。然而，深度学习模型提取的情感表征空间和人类可理解的空间存在着本质的区别与不可解释性，这些深度学习模型的不可解释性为建立安全可信的语音情感识别任务带来了挑战。另外，情感识别涉及感知、认知、心理等多个学科的知识，如何将这些相关知识融合到深度模型算法中，是语音情感识别技术的一大挑战。因此，构造知识驱动和数据驱动相结合的语音情感识别将是语音情感识别发展的另一方向。

7　结论

语音情感识别已经成为了人工智能领域的研究热点。从传统的声学特征到现在的深度特征，从传统机器学习到深度神经网络，从离散情感识别到维度情感识别等多方面得到了一定的发展，并被应用到了诸多领域。

本文对语音情感识别方法进行了系统性的梳理与分析。对情感表达模型包括离散、维度模型进行了阐述；其次，针对现有的语音情感数据库进行了总结；然后整理了基于语音情感特征和基于端到端的语音情感识别方法，在此基础上，介绍了近几年语音情感识别结果。最后介绍了语音情感识别的应用，以及面临的挑战与展望。

参考文献

[1] MINSKY M. Society of Mind [M]. New York: Simon & Schuster, 1988.

[2] PICARD R W. Affective Computing [M]. Massachusetts: The MIT Press, 2000.

[3] 张颖，罗森林. 情感建模与情感识别 [J]. 计算机工程与应用，2003，39(33)：98-102.
ZHANG Ying, LUO Senlin. Recognizing and expressing affect [J]. Computer Engineering and Applications, 2003, 39(33): 98-102. (in Chinese)

[4] 张会云，黄鹤鸣，李伟，等. 语音情感识别研究综述 [J]. 计算机仿真，2021，38(8)：7-17.
ZHANG Huiyun, HUANG Heming, LI Wei, et al. An overview of speech emotion recognition [J]. Computer Simulation, 2021, 38(8): 7-17. (in Chinese)

[5] 孙晓虎，李洪均. 语音情感识别综述 [J]. 计算机工程与应用，2020，56(11)：1-9.

SUN Xiaohu, LI Hongjun. Overview of speech emotion recognition [J]. Computer Engineering and Applications, 2020, 56(11): 1-9. (in Chinese)

[6] DALGLEISH T, POWER M. Handbook of Cognition and Emotion [M]. Chichester: Wiley, 1999.

[7] ORTONY A, TURNER T J. What's basic about basic emotions? [J]. Psychological Review, 1990, 97(3): 315-331.

[8] RUSSELL J A. A circumplex model of affect [J]. Journal of Personality and Social Psychology, 1980, 39(6): 1161-1178.

[9] RUSSELL J A, MEHRABIAN A. Evidence for a three-factor theory of emotions [J]. Journal of Research in Personality, 1977, 11(3): 273-294.

[10] FONTAINE J R J, SCHERER K R, ROESCH E B, et al. The world of emotions is not two-dimensional [J]. Psychological Science, 2007, 18(12): 1050-1057.

[11] 李海峰, 陈婧, 马琳, 等. 维度语音情感识别研究综述 [J]. 软件学报, 2020, 31(8): 2465-2491.
LI Haifeng, CHEN Jing, MA Lin, et al. Dimensional speech emotion recognition review [J]. Journal of Software, 2020, 31(8): 2465-2491. (in Chinese)

[12] PLUTCHIK R. Emotions: A general psychoevolutionary theory [J]. Approaches to Emotion, 1984, 1984(197-219): 2-4.

[13] SCHLOSBERG H. Three dimensions of emotion [J]. Psychological Review, 1954, 61(2): 81.

[14] BUSSO C, BULUT M, LEE Chichun, et al. IEMOCAP: interactive emotional dyadic motion capture database [J]. Language Resources and Evaluation, 2008, 42(4): 335-359.

[15] ZHANG J, JIA H. Design of speech corpus for mandarin text to speech [C]//The Blizzard Challenge 2008 Workshop, 2008.

[16] LI Ya, TAO Jianhua, CHAO Linlin, et al. CHEAVD: a Chinese natural emotional audio-visual database [J]. Journal of Ambient Intelligence and Humanized Computing, 2017, 8(6): 913-924.

[17] BURKHARDT F, PAESCHKE A, ROLFES M, et al. A database of German emotional speech [C]//Interspeech 2005, ISCA: ISCA, 2005: 1517-1520.

[18] BUSSO C, PARTHASARATHY S, BURMANIA A, et al. MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception [J]. IEEE Transactions on Affective Computing, 2017, 8(1): 67-80.

[19] COSTANTINI G, IADEROLA I, PAOLONI A, et al. EMOVO corpus: an Italian emotional speech database [C]//International Conference on Language Resources and Evaluation (LREC 2014), European Language Resources Association (ELRA), 2014: 3501-3504.

[20] LIVINGSTONE S R, RUSSO F A. The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English [J]. PLoS One, 2018, 13(5): e0196391.

[21] MCKEOWN G, VALSTAR M F, COWIE R, et al. The SEMAINE corpus of emotionally coloured character interactions [C]//2010 IEEE International Conference on Multimedia and Expo, Singapore, IEEE, 2010: 1079-1084.

[22] RINGEVAL F, SONDEREGGER A, SAUER J, et al. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions [C]//2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, Shanghai, China, IEEE, 2013: 1-8.

[23] MARTIN O, KOTSIA I, MACQ B, et al. The eNTERFACE′ 05 audio-visual emotion database [C]//22nd International Conference on Data Engineering Workshops (ICDEW′06). Atlanta, GA, USA. IEEE, 2006: 8.

[24] BAGHER ZADEH A, LIANG P P, PORIA S, et al. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph [C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia. Stroudsburg, PA, USA: Association for Computational Linguistics, 2018: 2236-2246.

[25] LOTFIAN R, BUSSO C. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings [J]. IEEE Transactions on Affective Computing, 2019, 10(4): 471-483.

[26] LIN Yilin, WEI Gang. Speech emotion recognition based on HMM and SVM [C]//In Proceedings of the 2005 International Conference on Machine Learning and Cybernetics. Guangzhou, China, 18-21 August 2005, 8: 4898-4901.

[27] CHAVHAN Y, DHORE M L, YESAWARE P. Speech emotion recognition using support vector machines [J]. Int. J. Comput. Appl., 2010, 1: 86-91.

[28] HAN K, YU D, TASHEV I. Speech emotion recognition using deep neural network and extreme learning machine [C]//In Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association. Singapore, September 2014: 14-18.

[29] SHAW A, VARDJAN R.K, SAXENA S. Emotion recognition and classification in speech using artificial neural networks [J]. Int. J. Comput. Appl., 2016, 145: 5-9.

[30] ZHANG S, ZHANG S, HUANG T, et al. Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching [J]. IEEE Transactions on Multimedia, 2017, 20(6): 1576-1590.

[31] DAREKAR R V, DHANDE A P. Emotion recognition from Marathi speech database using adaptive artificial neural network [J]. Biologically Inspired Cognitive Architectures, 2018, 23: 35-42.

[32] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780.

[33] ZHAO J, MAO X, CHEN L. Speech emotion recognition using deep 1D & 2D CNN LSTM networks [J]. Biomedical Signal Processing and Control, 2019, 47: 312-323.

[34] WANG Yuhua, SHEN Guang, XU Yuezhu, et al. Learning mutual correlation in multimodal transformer for speech emotion recognition [C]//Interspeech 2021. ISCA: ISCA, 2021: 4518-4522.

[35] ZOU H, SI Y, CHEN C, et al. Speech emotion recognition with co-attention based multi-level acoustic information [C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022: 7367-7371.

[36] HUANG Jian, TAO Jianhua, LIU Bin, et al. Learning utterance-level representations with label smoothing for speech emotion recognition [C]//Interspeech 2020. ISCA: ISCA, 2020: 4079-4083.

[37] JUSLIN P N, LAUKKA P. Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion [J]. Emotion (Washington, D C). 2001, 1(4): 381-412.

[38] MORI H, KASUYA H. Voice source and vocal tract variations as cues to emotional states perceived from expressive conversational speech [C]//Interspeech 2007. ISCA: ISCA, 2007: 102-105.

[39] GRIMM M, KROSCHEL K, NARAYANAN S. Support vector regression for automatic recognition of spontaneous emotions in speech [C]//2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP '07. Honolulu, HI, USA. IEEE, 2007: IV-1085.

[40] 陈逸灵，程艳芬，陈先桥，等. PAD三维情感空间中的语音情感识别 [J]. 哈尔滨工业大学学报，2018，50(11)：160-166.
CHEN Yiling, CHENG Yanfen, CHEN Xianqiao, et al. Speech emotion estimation in PAD 3D emotion space [J]. Journal of Harbin Institute of Technology, 2018, 50(11): 160-166. (in Chinese)

[41] TATO R, SANTOS R, KOMPE R, et al. Emotional space improves emotion recognition [C]//7th International Conference on Spoken Language Processing (ICSLP 2002). ISCA: ISCA, 2002.

[42] EYBEN F, WÖLLMER M, SCHULLER B. Opensmile: the Munich versatile and fast open-source audio feature extractor [C]//MM '10: Proceedings of the 18th ACM International Conference on Multimedia, 2010:1459-1462.

[43] SCHULLER B, STEIDL S, BATLINER A. The INTERSPEECH 2009 emotion challenge [C]//Interspeech 2009. ISCA: ISCA, 2009: 312-315.

[44] SCHULLER B, STEIDL S, BATLINER A, et al. The INTERSPEECH 2010 paralinguistic challenge [C]//Interspeech 2010. ISCA: ISCA, 2010: 2794-2797.

[45] SCHULLER B, STEIDL S, BATLINER A, et al. The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism [C]//Interspeech 2013. ISCA: ISCA, 2013.

[46] EYBEN F, SCHERER K R, SCHULLER B W, et al. The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing [J]. IEEE Transactions on Affective Computing, 2016, 7(2): 190-202.

[47] ATMAJA B T, AKAGI M. Evaluation of error- and correlation-based loss functions for multitask learning dimensional speech emotion recognition [J]. Journal of Physics: Conference Series, 2021, 1896(1): 012004.

[48] 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 [J]. 信号处理，2011，27(11)：1658-1663.
HAN Wenjing, LI Haifeng, MA Lin. Considering relative order of emotional degree in dimensional speech emotion recognition [J]. Signal Processing, 2011, 27(11): 1658-1663. (in Chinese)

[49] HAN Jing, ZHANG Zixing, SCHMITT M, et al. From hard to soft: towards more human-like emotion recognition by modelling the perception uncertainty [C]//MM '17: Proceedings of the 25th ACM International Conference on Multimedia, 2017: 890-897.

[50] HUANG Chunfang, AKAGI M. A three-layered model for expressive speech perception [J]. Speech Communication, 2008, 50(10): 810-828.

[51] ELBAROUGY R, AKAGI M. Speech emotion recognition

system based on a dimensional approach using a three-layered model [C]//Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference. Hollywood, CA, USA. IEEE, 2012: 1-9.

[52] ELBAROUGY R, AKAGI M. Cross-lingual speech emotion recognition system based on a three-layer model for human perception [C]//2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. Kaohsiung, Taiwan, China. IEEE, 2012: 1-10.

[53] CAI Xingyu, YUAN Jiahong, ZHENG Renjie, et al. Speech emotion recognition with multi-task learning [C]//Interspeech 2021. ISCA: ISCA, 2021: 4508-4512.

[54] YEH S L, LIN Yunshao, LEE Chichun. Speech representation learning for emotion recognition using end-to-end ASR with factorized adaptation [C]//Interspeech 2020. ISCA: ISCA, 2020: 536-540.

[55] TRIGEORGIS G, RINGEVAL F, BRUECKNER R, et al. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network [C]// 2016 IEEE International Conference on Acoustics, Speech and Signal Processing. Shanghai, China. IEEE, 2016: 5200-5204.

[56] LATIF S, RANA R, KHALIFA S, et al. Direct modelling of speech emotion from raw speech [C]//Interspeech 2019. ISCA: ISCA, 2019: 3920-3924.

[57] TZIRAKIS P, ZHANG Jiehao, SCHULLER B W. End-to-end speech emotion recognition using deep neural networks [C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing. Calgary, AB, Canada. IEEE, 2018: 5089-5093.

[58] SARMA M, GHAHREMANI P, POVEY D, et al. Emotion identification from raw speech signals using DNNs [C]//Interspeech 2018. ISCA: ISCA, 2018: 3097-3101.

[59] BAKHSHI A, WONG A S W, CHALUP S. End-to-end speech emotion recognition based on time and frequency information using deep neural networks [C]//ECAI 2020. IOS Press, 2020: 969-975.

[60] ZHANG Zixing, WU Bingwen, SCHULLER B. Attention-augmented end-to-end multi-task learning for emotion prediction from speech [C]//ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Brighton, UK. IEEE, 2019: 6705-6709.

[61] LI Yuanchao, ZHAO Tianyu, KAWAHARA T. Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning [C]//Interspeech 2019. ISCA: ISCA, 2019: 2803-2807.

[62] SUN Tingwei. End-to-end speech emotion recognition with gender information[J]. IEEE Access, 2020, 8: 152423-152438.

[63] 张钹, 朱军, 苏航. 迈向第三代人工智能 [J]. 中国科学: 信息科学, 2020, 50 (9): 1281-1302.
ZHANG Ba, ZHU Jun, SU Hang. Toward the third generation of artificial intelligence [J]. Scientia Sinica Informationis, 2020, 50(9): 1281-1302. (in Chinese)

[64] CAI Ruichu, GUO Kaibin, XU Boyan, et al. Meta multi-task learning for speech emotion recognition [C]//Interspeech 2020. ISCA: ISCA, 2020: 3336-3340.

[65] LIU Jiawang, WANG Haoxiang. A speech emotion recognition framework for better discrimination of confusions [C]//Interspeech 2021. ISCA: ISCA, 2021: 4483-4487.

[66] PENG Zhichao, DANG Jianwu, UNOKI M, et al. Multi-resolution modulation-filtered cochleagram feature for LSTM-based dimensional emotion recognition from speech [J]. Neural Networks, 2021, 140: 261-273.

[67] ZHU Zhi, SATO Y. Reconciliation of multiple corpora for speech emotion recognition by multiple classifiers with an adversarial corpus discriminator [C]//Interspeech 2020. ISCA: ISCA, 2020: 2342-2346.

[68] LATIF S, ASIM M, RANA R, et al. Augmenting generative adversarial networks for speech emotion recognition [C]//Interspeech 2020. ISCA: ISCA, 2020: 521-525.

[69] LIU Pengfei, LI Kun, MENG H. Group gated fusion on attention-based bidirectional alignment for multimodal emotion recognition [C]//Interspeech 2020. ISCA: ISCA, 2020: 379-383.

[70] CHEN Ming, ZHAO Xudong. A multi-scale fusion framework for bimodal speech emotion recognition [C]//Interspeech 2020. ISCA: ISCA, 2020: 374-378.

[71] LI Hang, DING Wenbiao, WU Zhongqin, et al. Learning fine-grained cross modality excitement for speech emotion recognition [C]//Interspeech 2021. ISCA: ISCA, 2021: 3375-3379.

[72] KUMAR P, KAUSHIK V, RAMAN B. Towards the explainability of multimodal speech emotion recognition [C]//Interspeech 2021. ISCA: ISCA, 2021: 1748-1752.

[73] HOU Mixiao, ZHANG Zheng, LU Guangming. Multimodal emotion recognition with self-guided modality calibration [C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing. Singapore, Singapore. IEEE, 2022: 4688-4692.

[74] LIAN Zheng, TAO Jianhua, LIU Bin, et al. Context-dependent domain adversarial neural network for multimodal emotion recognition [C]//Interspeech 2020. ISCA: ISCA, 2020: 394-398.

[75] PAN Zexu, LUO Zhaojie, YANG Jichen, et al. Multi-modal attention for speech emotion recognition [C]//Interspeech 2020. ISCA: ISCA, 2020: 364-368.

[76] KHARE A, PARTHASARATHY S, SUNDARAM S. Multi-modal embeddings using multi-task learning for emotion recognition [C]//Interspeech 2020. ISCA: ISCA, 2020: 384-388.

[77] FENG Han, UENO S, KAWAHARA T. End-to-end speech emotion recognition combined with acoustic-to-word ASR model [C]//Interspeech 2020. ISCA: ISCA, 2020: 501-505.

[78] FAN Weiquan, XU Xiangmin, XING Xiaofen, et al. Adaptive domain-aware representation learning for speech emotion recognition [C]//Interspeech 2020. ISCA: ISCA, 2020: 4089-4093.

[79] LATIF S, RANA R, KHALIFA S, et al. Deep architecture enhancing robustness to noise, adversarial attacks, and cross-corpus setting for speech emotion recognition [C]//Interspeech 2020. ISCA: ISCA, 2020: 2327-2331.

[80] SHEN Guang, LAI Riwei, CHEN Rui, et al. WISE: word-level interaction-based multimodal fusion for speech emotion recognition [C]//Interspeech 2020. ISCA: ISCA, 2020: 369-373.

[81] CHEN Weidong, XING Xiaofeng, XU Xiangmin, et al. Key-sparse transformer for multimodal speech emotion recognition [C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing. Singapore, Singapore. IEEE, 2022: 6897-6901.

[82] LI Jinchao, WANG Shuai, CHAO Yang, et al. Context-aware Multimodal Fusion for Emotion Recognition [C]// Proc. Interspeech, 2022: 2013-2017.

[83] DUTTA K, SARMA K K. Multiple feature extraction for RNN-based Assamese speech recognition for speech to text conversion application [C]//2012 International Conference on Communications, Devices and Intelligent Systems (CODIS). IEEE, 2012: 600-603.

[84] PRAVENA D, GOVIND D. Significance of incorporating excitation source parameters for improved emotion recognition from speech and electroglottographic signals [J]. International Journal of Speech Technology, 2017, 20 (4): 787-797.

[85] SWAIN M, ROUTRAY A, KABISATPATHY P. Databases, features and classifiers for speech emotion recognition: a review [J]. International Journal of Speech Technology, 2018, 21: 93-120.

[86] HSIEH Y H, CHEN S C. A decision support system for service recovery in affective computing: an experimental investigation [J]. Knowledge and Information Systems, 2020, 62: 2225-2256.

[87] SCHELINSKI S, KRIEGSTEIN K. The relation between vocal pitch and vocal emotion recognition abilities in people with autism spectrum disorder and typical development [J]. Journal of Autism and Developmental Disorders, 2019, 49: 68-82.

[88] PARIS M, MAHAJAN Y, KIM J, et al. Emotional speech processing deficits in bipolar disorder: The role of mismatch negativity and P3a [J]. Journal of Affective Disorders, 2018, 234: 261-269.

[89] FAN Xiaohe, ZHAO Heming, CHEN Xueqin, et al. Deceptive speech detection based on sparse representation [C]//2016 IEEE 12th International Colloquium on Signal Processing & Its Applications (CSPA). IEEE, 2016: 7-11.

作者简介

陶建华(通讯作者)　男,1972年生,江苏人。清华大学自动化系教授。主要研究方向为多源信息处理、情感计算、语音信息处理等。

E-mail: jhtao@tsinghua. edu. cn

陈俊杰　男,1998年生,江苏人。天津师范大学研究生。主要研究方向为语音情感分析与识别、情感计算等。

E-mail: 2110090009@stu. tjnu. edu. cn

李永伟　男,1988年生,内蒙古人。中国科学院自动化研究所助理研究员。主要研究方向为语音情感分析与识别、情感计算、语音信号处理等。

E-mail: yongwei. li@nlpr. ia. ac. cn