软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software [doi: 10.13328/j.cnki.jos.007232]

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

离散语音情感识别研究进展*

郭丽丽[1], 王龙标[2,3], 党建武[2,3,4], 丁世飞[1]

[1](中国矿业大学 计算机科学与技术学院, 江苏 徐州 221116)

[2](天津大学 智能与计算学部, 天津 300350)

[3](天津市认知计算与应用重点实验室 (天津大学), 天津 300350)

[4](Japan Advanced Institute of Science and Technology, Ishikawa 9231292, Japan)

通信作者: 王龙标, E-mail: longbiao_wang@tju.edu.cn; 丁世飞, E-mail: dingsf@cumt.edu.cn

摘　要: 语音情感识别是情感计算的重要组成部分, 在人机交互中占据重要的地位. 准确地识别说话人的情感信息, 有助于机器更好地理解用户的意图, 进而提供良好的交互性以提升用户的体验. 以离散语音情感为对象, 对语音情感识别的理论和方法进行综述. 首先在全面回顾情感识别发展历程的同时, 提出一个语音情感识别综述框架. 其次, 介绍情感描述方法以及常用的情感语料库, 旨在为语音情感识别提供基础支撑. 然后, 概述语音情感识别过程, 主要包括特征提取和识别模型, 重点归纳总结传统分类模型、经典深度模型、其他先进模型, 并介绍常用的评价指标, 同时基于评价指标对模型进行总结. 最后, 探讨语音情感识别领域所面临的挑战, 并对未来的发展趋势进行展望.

关键词: 语音情感识别; 声学特征; 相位信息; 分类模型; 深度学习

中图法分类号: TP18

中文引用格式: 郭丽丽, 王龙标, 党建武, 丁世飞. 离散语音情感识别研究进展. 软件学报. http://www.jos.org.cn/1000-9825/7232.htm

英文引用格式: Guo LL, Wang LB, Dang JW, Ding SF. Research Progress of Discrete Speech Emotion Recognition. Ruan Jian Xue Bao/Journal of Software (in Chinese). http://www.jos.org.cn/1000-9825/7232.htm

Research Progress of Discrete Speech Emotion Recognition

GUO Li-Li[1], WANG Long-Biao[2,3], DANG Jian-Wu[2,3,4], DING Shi-Fei[1]

[1](School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China)

[2](College of Intelligence and Computing, Tianjin University, Tianjin 300350, China)

[3](Tianjin Key Laboratory of Cognitive Computing and Application (Tianjin University), Tianjin 300350, China)

[4](Japan Advanced Institute of Science and Technology, Ishikawa 9231292, Japan)

Abstract: Speech emotion recognition is an important part of affective computing and plays an important role in human-computer interaction. Accurately distinguishing emotions helps machines understand users' intentions and provide better interactivity to enhance user experience. This study reviews the theories and methods of speech emotion recognition focusing on discrete speech emotions. Firstly, the study reviews the development of emotion recognition and presents an architecture of speech emotion recognition to summarize research progress. Secondly, emotion representation models and commonly used corpora are introduced to provide basic support for speech emotion recognition. Then, the process of speech emotion recognition is outlined, including feature extraction and recognition models, with a focus on traditional classification models, classical deep models, and other advanced models. Meanwhile, commonly used evaluation indicators are introduced and applied to provide a summary of models. Finally, the study discusses the challenges in speech emotion recognition and suggests possible directions for future research.

Key words: speech emotion recognition (SER); acoustic feature; phase information; classification model; deep learning

* 基金项目: 国家自然科学基金 (62276265, 62176182, 62276185); 中央高校基本科研业务费专项资金 (2022QN1096)
收稿时间: 2021-12-23; 修改时间: 2022-08-24, 2023-07-20, 2023-10-26, 2024-01-05, 2024-04-15; 采用时间: 2024-06-05; jos 在线出版时间: 2024-09-25

情感是人类生存和进化过程中形成意图表达的方式之一, 在人际交往过程中起着极其重要的作用. 人类的智能除了体现在正常的逻辑推理和理性思维能力之外, 还有情感生成和感知能力. 情感理解是人类智能的重要体现, 甚至在理性行为和决策中起到至关重要的作用. 情感状态的任何细微改变都有可能对主观创造性以及问题解决产生重要的影响. 情感可以帮助人们更好地理解说话者的意图, 从而促进交流沟通. 既然情感能帮助人类更好地理解彼此, 那么我们希望计算机同样也具备情感能力[1].

语音作为人类最基本、有效的沟通方式, 其中蕴含着丰富的情感信息[2]. 目前许多的人机交互产品, 如苹果的 Siri、微软的 Cortana、谷歌的 Google Assistant 等, 都是以语音交互为主. 由此, 大量的情感识别研究基于语音信号开展. 语音情感识别 (speech emotion recognition, SER) 在人机交互中应用十分广泛, 对人们的工作和生活带来巨大的影响. 例如, 在智能客服中, 通过实时检测客户的情绪变化, 并给予适当回应来提高服务质量[3], 当检测到用户情绪焦虑、激烈不满时, 可以给予安慰并及时转为人工服务, 达到优化用户体验的目的. 在教育上, 通过检测课堂的互动信息, 实时分析学生的表现, 并进行情绪分析, 以得到学生兴趣度和知识接受度的实时反馈, 进而辅助课题教学[2]. 在医疗上, 医生可以通过情感监测系统跟踪精神疾病、抑郁症等患者的情感变化, 从而给疾病诊断和治疗提供一种参考[4]. 在驾驶方面, 通过对驾驶员语速、音量、清晰度等因素的分析, 可以实时检测其情绪状况, 当驾驶员呈现疲劳或情绪激动的时候给以提醒, 尽可能地避免交通事故的出现[5]. 此外, 语音情感识别还在刑事侦查、情感陪护、虚拟现实、信息检索、安全监测等多方面发挥作用, 拥有十分广阔的应用场景. 可见, 语音情感识别有助于推动新型人机交互环境的开发. 总之, 语音情感识别已经成为一个非常重要的研究课题, 无论从学术研究价值, 还是从实际应用前景的角度来看, 语音情感识别均有重大的研究价值.

语音情感识别经历了 30 多年的发展, 逐渐受到了国内外研究学者的广泛关注. 经典的语音情感识别框架首先是提取声学特征; 然后将这些特征输入到分类模型中完成情感识别[6]. 早期的语音情感识别研究都是基于启发式特征 (如基频、能量、梅尔频率倒谱系数等) 展开, 其中常用的识别算法包括隐马尔可夫模型 (hidden Markov model, HMM)[7]、高斯混合模型 (Gaussian mixture model, GMM)[8]、支持向量机 (support vector machine, SVM)[9] 等. 此外, 还有一些基于决策树 (decision tree, DT), K 近邻 (K-nearest neighbor, KNN), K 均值 (K-means), 朴素贝叶斯 (naive Bayes) 的方法. 近年来, 随着深度学习的发展与进步, 提出了一些基于深度学习的模型, 如深度神经网络 (deep neural network, DNN)、循环神经网络 (recurrent neural network, RNN)、长短时记忆网络 (long short-term memory, LSTM) 等. 由于人对语音情感识别的认知是有局限性的, 单纯利用启发式特征很难提取到丰富全面的特征. 近几年基于振幅时频特征的语音情感识别研究发展迅速, 并提出了一些经典模型. 目前, 利用卷积神经网络 (convolutional neural network, CNN) 从振幅时频特征中提取深度声学特征也成为语音情感识别领域最常用的方法之一[10].

语音情感识别通常包含语音信号输入、情感特征提取和识别 3 个主要部分. 此外, 语音情感识别的完整过程还需要另外两个基础工作的支撑, 即情感空间描述和情感数据库. 虽然已经有一些综述论文对语音情感识别的相关内容进行了总结[11−13], 但早期的综述对语音情感特征总结的不够全面, 缺失相位相关特征的介绍. 此外, 在总结识别模型的时候没有包括最新的进展, 缺乏对经典深度模型和其先进模型的总结. 为了给对语音情感识别感兴趣的研究者提供一个综合的指导, 本文对离散语音情感识别进行了综述, 主要贡献如下.

• 我们对语音情感识别进行了全面、广泛的概述, 并提出一个综述框架如图 1 所示, 以概括语音情感识别领域的研究进展. 它为理解语音情感识别提供了坚实的基础, 并可作为设计和使用不同的情感识别方法的指南.

• 和以往语音情感识别综述论文不同, 本综述介绍了相位相关特征, 并对利用振幅与相位信息互补的语音情感识别模型进行了总结.

• 为了更好地理解现有的 SER 分类模型, 本文不仅介绍了传统的 SER 分类模型和经典的深度模型, 还总结概括了近些年提出的先进 SER 模型, 如基于相位信息、Transformer、图神经网络 (graph convolutional network, GCN) 的方法.

• 经过对识别方法进行全面的分析, 我们确定了基于语音信号的情感识别中仍然有待解决的关键挑战, 并展望了未来发展的 5 个方向.

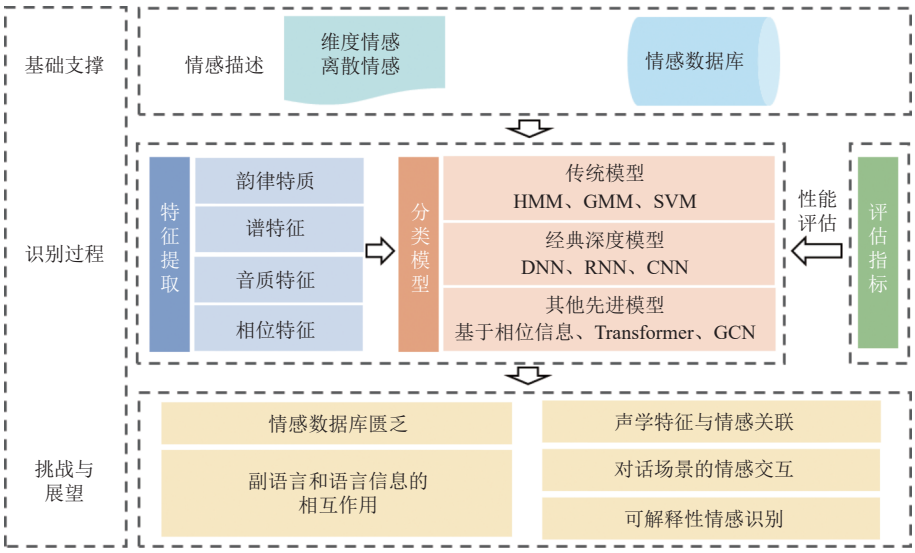

图 1　语音情感识别综述框架

本文第 1 节介绍两种主流的情感描述方法, 并介绍常用的情感语料库. 第 2 节概述常用的声学特征, 包括韵律特征、谱特征和音质特征, 还介绍相位特征. 第 3 节总结归纳语音情感识别中的传统分类模型、经典深度模型和其他先进模型, 同时介绍常用的评价指标, 并对各种模型进行总结. 最后对语音情感识别所面临的挑战和发展趋势进行总结与展望.

1　情感分类与表示及其情感数据库

对情感的定义和描述是实现语音情感识别系统的前提. 本节首先对情感描述方法进行概述; 然后, 总结介绍情感相关的语料库.

1.1　情感描述方法

关于情感状态的描述, 研究者提出了许多不同的方式. 在 20 世纪提出了 90 多个情感定义. 然而, 关于情感的定义还没有达成共识, 它仍然是心理学上一个悬而未决的问题. 情感是一种复杂的心理状态, 由个人经历、生理、行为和交流反应等几部分组成[13]. 基于这些定义, 有两种常见的情感描述方法被提出: 离散情感描述和维度情感描述.

离散情感理论又包括基本情感和复合情感. 基本情感模型把情感描述为几类独立的情感类别, 如: 高兴、生气、愤怒等. 在多年的发展中, 对于情感类别的定义也是大不相同. 1990 年, 美国神经学科学家 Oitony 等人[14]对基本型类型进行了总结. 但究竟情感可以分为哪些类别, 至今还存有争议. 目前, 研究者普遍承认的是美国心理学家 Ekman 提出的 6 大基本情感[15], 包括恐惧 (fear)、生气 (anger)、讨厌 (disgust)、高兴 (joy)、惊奇 (surprise) 和悲伤 (sadness) 这 6 种. 然而, 在实际应用中, 中性 (neutral) 类别也经常出现, 于是构成了常见的 7 种基本情感. 另外, 还有由几种基本情感一起组成的复合情感, 也被称作情感调色板理论[16]. 大多数的语音情感识别系统基于基本的情感类别开展. 在日常生活中, 情感类别的标签是直观的, 人们通常使用离散情感模型来定义他们所观察到的情感. 但这些离散的情感类别并不能定义日常交流中观察到的一些复杂的情绪状态.

和离散情感不同, 在维度情感模型中情感被看成是由一个多维的维度空间来描述的, 情感之间是连续变化的. 维度空间里的维度表示情感的属性, 如效价 (valence), 唤醒度 (arousal), 期望度 (expectation), 优势度 (dominance) 等. 在这些维度的表示下, 情感被量化成多维空间里的一个坐标点, 而坐标点的值体现了情感在该维度上的强度. 在维度模型中, 情感不是彼此独立的, 是可以相互类推的. 目前, 被大众认可并广泛使用的维度空间是“唤醒度-效

价”二维情感空间表示[17]. 其中, 唤醒度用于描述说话人情感的激烈程度, 效价维度则用于评价说话人情感的正负性程度. 由于维度情感描述可以实现对情感连续性变化的追踪, 所以这种表示方法也受到研究者的关注.

离散情感和维度情感都有各自的研究价值, 看似对立实则是可以相互转化, 研究者需要根据具体的任务选择合理的情感描述方式. 离散情感模型主要反映的是人类的基本情感类型, 区分较为清晰, 具有天然的可解释性[11], 更加符合人类的直觉和常识, 与人类的语言和语义接轨. 而维度情感是把人类主观情感量化为客观数值的过程, 不够直观, 可能需要特殊的训练来标记[18]. 不仅繁琐而且难以确保质量, 并且匮乏一致认可的评价标准. 此外, 有些情感的纬度表示相似, 如生气和害怕. 从模型复杂度而言, 离散情感模型较为直观、简洁、易懂, 有利于开展相关的研究工作. 因此, 目前的情感识别研究中, 基于离散情感描述的方法占据主流地位[19], 本文以离散情感为主要研究对象进行综述.

1.2 情感数据库

合适的情感语音数据库是进行语音情感识别的必要先决条件[20], 数据库的质量和类型决定了情感识别系统的可靠性和用途. 因此, 应该仔细设计和收集数据库[21]. 由于情感数据本身的多样性, 录制和采集并没有统一的标准. 收集数据库的目标和方法因系统开发的动机而异. 依据情感语音产生方式的不同, 语音情感数据库可以分为3类: 表演型情感数据库, 诱导型情感数据库和自然型情感数据库.

表1简要地总结了每种数据库类型的优缺点.

表1 不同类型数据库在语音情感识别中的特性

数据库类型	优点	缺点
表演型	标准、情感覆盖范围广、结果易于比较	数据库建设成本大、缺乏语境、缺乏自然的情感
诱导型	更贴近自然、有上下文信息	上下文是人为的、并非所有情感都可用、数据不平衡
自然型	自然的表达和真实情感、数据库建设成本低	版权和隐私问题、背景噪音的存在、话语中多种情感同时发生和话语本身的重叠

表演型数据是由经验丰富和训练有素的演员在隔音工作室中录制的, 演员要用不同的情感去演绎给定的句子. 与其他类型相比, 这种数据库可操作性比较强, 录制也相对容易, 被认为是获得各种语音情感数据库的最简单的方法. 通过演绎, 可以得到所有的情感类别, 且音频质量高, 避免了麦克风距离、编解码器效应、噪声和混响数据等记录问题. 这是标准且最为常见的数据库. 一般来说, 模拟的情感往往比真实情绪更具有表现力[22], 当训练和测试多是模型情感的时候, 识别精度会很高. 但是, 研究人员指出, 表演的情感不能反映现实生活中的真实情感, 甚至可能被夸大, 从而离人类真实情感有明显差异, 这会降低对真实情感的识别率.

诱导型数据是将说话人置于能够激发各种情绪的模拟环境中来录制的[23]. 这些数据有时是通过受试者与计算机进行语言交互来记录, 而计算机的语言反应则由人类在受试者不知情的情况下控制. 通过与受试者对话产生不同的语境状态, 从而诱发受试者主动表达出不同的情感类别. 虽然这可能不会产生所有类别的情绪, 但它们很接近人类自然环境下的真实情绪, 录制起来也比较容易. 与表演数据库相比, 这些数据库可能更自然, 包含了上下文信息.

与以上两种情感数据不同, 自然情感数据库是从现实生活中录制的自然语料. 数据的情感是自然且真实的, 对真实世界的情感建模很有帮助. 这些数据来源于谈话节目、呼叫中心对话、公共场所录音等, 录制成本比较低. 但是, 通常这种数据是在说话人毫不知情的情况下录制的, 便会造成伦理和法律的问题, 因此这种自然情感数据的获取也是有难度的. 在情感标注过程中, 对情感的判别会出现分歧, 不同的标注者可能会给一句话赋予不同的标签, 这个问题在自然数据库中更为突出. 另外, 自然对话包含不受控制的噪音、重叠的谈话和多种情绪同时发生等情况, 在识别方面存在困难[24].

数据库的准确性还取决于情感的数量和类型. 例如, 悲伤和无聊、愤怒和高兴的声学特征非常相似, 分类器很容易将他们混淆. 由于不同的数据库包含不同数量和类型的情感, 因此无法对他们的结果进行直接比较. 目前, 依靠自发、表演、诱发的机制已经建立了许多的情感数据库, 同时根据第1.1节介绍的情感描述方法可以分为离散数据和维度数据. 不过, 语音情感识别领域以离散情感数据库居多, 表2总结了一些重要的离散情感数据库.

表 2 常用情感数据库总结

数据库名称	类型	语种	描述	情感	模态
Emo-DB[25]	表演	德语	德国柏林工业大学录制, 10个说话人 (5男5女) 对10个句子进行演绎, 包含800句语料	愤怒、无聊、厌恶、高兴、恐惧、悲伤、中性	音频
CASIA[26]	表演	普通话	中国科学院自动化研究所录制, 由4个说话人 (2男2女) 对500句文本进行演绎, 最终保留9 600句	高兴、悲伤、愤怒、恐惧、中性、惊喜	音频
eNTERFACE[27]	诱导	英语	42个受试 (34男8女) 听6个连续的短篇小说, 并根据每个短篇小说引发的特定的情感作出反应, 共计1 116段视频序列	愤怒、讨厌、恐惧、高兴、悲伤、惊喜	音频/视频
IEMOCAP[28]	表演/自然	英语	南加利福尼亚大学录制, 由10个说话人 (5男5女) 进行5个session的会话, 包括即兴自发和演绎两种类型, 共约12 h	广泛使用的是愤怒、高兴、悲伤、中性4类情感	音频/视频
CASS[29]	自然	普通话	收集汉语口语, 包含语音转录的自然语音, 7个说话人 (2男3女), 共计6 h音频	愤怒、恐惧、高兴、悲伤、惊喜、中性	音频
FAU AIBO[30]	自然	德语	通过51个小孩与索尼公司的AIBO机器狗进行自然交互, 进行情感数据的采集, 共计9 h音频	快乐、愤怒、中性等11种情感	音频
Belfast[31]	表演	英语	由Queen大学对40名受试者 (年龄在18–69岁之间, 20男20女) 对5个段落进行演绎	生气、悲伤、高兴、恐惧、中性	音频
CHEAVD[32]	自然	普通话	包含了取自电影、电视剧和脱口秀的140 min的情感话语, 共计约2 322个样本记录	除常见的基本情感, 还标注了骄傲、窘迫等一些非典型情感	音频/视频
JAVED[33]	表演	日语	数据从脚本对话和独白中收集, 由24名男性进行演绎, 共包含100 min数据	高兴、愤怒、悲伤、中性和满足	音频/视频
HEU Emotion[34]	半自然	普通话	该数据库分为两部分, 第1部分收集自Tumblr、Google和gypsy, 第2部分收集自电影、电视剧和综艺节目	愤怒、无聊、失望、困惑、恐惧、厌恶、高兴、中性、悲伤和惊讶等10种情绪	音频/视频
EMOVO[35]	表演	意大利语	第1个意大利情感语料库, 由6名 (3男3女) 演绎而来, 含有588条语音样本	厌恶、喜悦、恐惧、惊讶、悲伤、喜悦和中性	音频

2 语音情感特征

特征抽取是情感识别的重要组成部分, 抽取的特征质量会对情感识别的最终结果产生直接影响. 因此, 如何筛选和提取出足够有效的情感特征是情感识别的一个挑战性问题. 为了探究声学信号中哪些特征可以有效地表征说话人的情感状态, 研究学者们结合语音语言学、生理学、心理学等多个学科基础进行了充分和深入研究[36]. 人类在识别情感的时候, 会从声音的大小、高低、音色等各方面的特性去感知. 比如, 当说话人生气时, 语调升高、语速加快; 悲伤时, 语调低沉、语速缓慢. 听者可以轻易地通过这些特征去感受说话人的情绪. 研究者基于此设计了一些启发式特征, 可以反映人类感知情感的知识. 当前, 用于语音情感识别的声学特征主要分为 3 类: 韵律特征、谱特征和音质特征. 此外, 近几年, 研究者开始探究相位相关特征的作用, 并已验证相位信息具有情感判别能力. 表 3 对不同种类的语音特征进行了汇总, 详细介绍在后面小节中展开.

表 3 语音特征总结

特征种类	具体特征
韵律特征	时长 (duration)、基频 (F0)、能量 (energy)
谱特征	对数频率功率系数 (LFPC)、线性预测倒谱系数 (LPCC)、梅尔频率倒谱系数 (MFCC)、伽马通滤波器倒谱系数 (GFCC)
音质特征	声门参数 (glottal parameter)、频率微扰 (jitter)、振幅微扰 (shimmer)、谐波噪声比 (HNR) 等
相位特征	群延迟 (GD)、改进群延迟倒谱系数 (MGDCC)、相对相位 (RP)、动态相对相位 (DRP) 等

2.1 韵律特征

韵律特征 (prosodic feature) 也称作“超音段特征”或“超语言学特征”, 是指语音中凌驾于语义符号之上的音高、音长、快慢和轻重等可以被人类感知的特征, 是语音情感表达的重要形式之一[11]. 它的存在影响着一句话的语速、停顿、音调等主观感受, 这些特征携带有重要的情感信息. 最广泛使用的韵律特征包括时长 (duration)、基频 (fundamental frequency, F0)、能量 (energy) 等相关的特征. 其中, F0 是基本频率, 由声带的振动产生的. 在话语过程中基频的变化会产生基频轮廓, 其统计值可以用作声学特征. 基频产生了说话的节奏和音调, 在很大程度上反映了不同情感状态. 例如, 当音调较高时表达的情感可能是高兴或者愤怒; 而当音调较低时可能表示的是悲伤. 能量被称为音量或强度, 反映了声学信号的幅度随时间的变化. 研究表明愤怒、快乐或惊讶等高唤醒情绪会能量增加, 而厌恶和悲伤会能量降低[37]. 持续时间是指在演讲中形成元音、单词以及类似结构的时间. 语音速度、沉默区持续时间、清音区和清音区持续时间等是使用最为广泛的时长相关特征.

关于韵律特征有许多不同方面的研究, 如 Frick 等人[38]研究了韵律特征与情感状态的关系. Busso 等人[39]分析了各种 F0 轮廓的统计数据, 从而找到情感突出的时刻. Origlia 等人[40]使用基频和能量相关的最大、最小、平均、标准差等统计计算组成了一个 31 维的韵律特征集, 在一个包含有多语种情感语料库上取得近 60% 的识别率. Seppänen 等人[41]使用基频、能量、时长相关的 43 维全局韵律特征进行芬兰语的情感识别, 在说话人独立的情况下取得了 60% 的识别率. 文献 [42] 分别将基频、能量、时长等韵律特征用于德语语料的说话人独立的情感识别和汉语普通话的说话人不独立的情感识别, 分别得到了 51% 和 88% 的情感识别率. Luengo 等人[43]在情感语音上进行了一系列的韵律特征分析研究. 经过特征选择与分析, 最后结合基频均值、能量均值、基频方差、基频对数的斜交、基频对数的动态范围和能量对数的动态范围具有良好的情感判别能力. 研究表明, 当使用韵律特征时, 语音情感识别系统可以得到与人类判断类似或者更好的表现[44]. 韵律特征已经得到了研究者们的广泛认可, 并且许多研究都证实了其在语音情感识别中的作用[42]. 但韵律特征的情感判别能力存在一定的限制, 如愤怒与高兴的基频特征十分相似, 不利于情感区分.

2.2 谱特征

谱特征 (spectral feature) 是声道 (vocal tract) 形状变化和发声运动相关性的体现. 当一个人发出声音时, 会被声道的形状过滤, 发出的声音是由声道形状决定的. 因此, 如果能精确地模拟这个形状就可以得到声道和所产生的声音的精确表示. 研究人员发现声道的特征在频率域得到了很好的表现[45]. 人们利用傅里叶变换 (Fourier transform, FT) 将时域信号转化为频域信号, 进而得到谱特征. 谱相关的特征通常可以分为线性谱特征和倒谱特征, 线性谱特征有线性预测系数 (linear prediction coefficient, LPC)、对数频率功率系数 (log-frequency power coefficient, LFPC) 等; 倒谱特征有线性预测倒谱系数 (linear prediction cepstral coefficient, LPCC)、梅尔频率倒谱系数 (Mel-frequency cepstral coefficient, MFCC)[46]等. 其中, MFCC 是最常用的频谱特征, 它代表了声学信号的短期功率谱. 提取 MFCC 的时候, 首先对信号进行预加重、分帧、加窗处理, 然后利用短时傅里叶变换 (short-time Fourier transform, STFT) 将每一帧信号转换成频域特征. 随后, 利用梅尔滤波器组计算子带能量, 并计算这些子带的对数. 最后, 利用离散余弦变换 (discrete cosine transform, DCT) 计算 MFCC 系数阶数.

除此之外, 研究者们对谱特征也开展了其他深层次的探索. Kim 等人[47]提出一种新的说话人独立的特征: 频谱平坦度与频谱中心的比值特征 (the ratio of a spectral flatness measure to a spectral center, RSS), 并在语音情感识别中取得了一定成效. 调制频谱特征 (modulation spectral feature, MSF) 是 Wu 等人提出的[48], MSF 特征利用声学滤波器组和调制滤波器组对声学信号进行处理, 从而提取到时频信息. Bitouk 等人[49]提出一套新的谱特征, 他们对 3 种感兴趣的音素类型, 即重音、非重音元音和辅音进行 MFCC 统计, 得到了更高的准确度. 此外, 将这些特征与韵律特征结合起来也可以提高准确性. 越来越多的研究者们将谱特征应用到语音情感的识别中, 并起到了改善情感识别性能的作用[50]. Sato 等人[51]使用多模板 MFCC 聚类来标记每一帧, 进而采用分段 MFCC 特征进行语音情感识别. 与使用 KNN 的韵律算法和使用 HMM 的传统 MFCC 算法相比取得了更好的性能. 研究者探究了新的傅里叶参数模型[52], 将傅里叶参数及其一阶、二阶差与 MFCC 结合用于说话人独立的语音情感识别. 然而谱特征

通常以频谱图的形式表示, 对于长时间的信号, 频谱图的维度非常高, 导致计算和处理的复杂性增加. 此外, 谱特征对于时域特征的变化很敏感, 音调、速度等信息可能无法很好地反映.

2.3 音质特征

音质特征 (voice quality feature) 指与发音方式有关的声学特征, 反应说话人的个性化表现, 可以作为衡量语音清晰度、是否容易辨识的主观评价指标. 当人在情绪激动、无法控制的时候, 说话时往往会表现出哽咽、颤音、喘息等现象, 这些都会影响声音质量[17]. 可见, 情感信息也会在音质特征中流露, 并且研究者通过情感的听辨实验证实了这种关系[53]. 常用的音质特征有声门参数 (glottal parameter), 频率微扰 (jitter), 振幅微扰 (shimmer), 共振峰频率及其带宽 (format frequency and bandwidth), 谐波噪声比 (harmonics to noise ratio, HNR) 等. 其中, 频率微扰描述连续振动周期之间基础频率的变化, 主要体现噪声的程度; 而振幅微扰描述的是振幅的变化, 主要体现嘶哑声的程度; HNR 是对噪声相对水平的测量, 指谐波能量与噪声能量的比值.

在识别语音情感的过程中, 研究者通常会将音质特征与其他类型的声学特征结合在一起使用, 例如, Li 等人[54]将频率微扰和振幅微扰作为音质特征, 并与谱特征 MFCC 结合使用, 取得了更高的识别率; Zhang 等人[55]将频率微扰、振幅微扰、HNR 等音质特征与韵律特征联合使用, 与只使用韵律特征相比识别率提高了 10%; Kacheel 等人[56]将音质特征、韵律特征和谱特征结合用于语音情感识别, 在公开的柏林情感数据库上达到了 88.97% 的平均识别率. Lugger 等人[57]提取第 1 和第 4 共振峰频率及其相应的带宽作为音质特征, 并与基频等韵律特征联合用于说话人独立的语音情感识别. Sun 等人[58]比较和探讨了声门参数与韵律特征 (如基频、能量等) 在情感识别中发挥的作用. Borchert 等人[59]利用共振峰、HNR、振幅微扰等音质特征结合韵律特征在说话人独立的语音情感识别中取得了 70% 的平均识别率. 音质特征是一种主观评价指标, 无法完全客观地描述一个语音的声音质量. 此外, 不同的音质特征可能会给出不同的评价结果, 缺乏一致性.

以上这些声学特征通常以帧为单位提取, 得到低级描述 (low level descriptor, LLD). 然而, 在实际使用过程中, 人们通常采用这几种特征的统计计算, 如均值、最大值、最小值、标准差等来获得全局特征. 与局部韵律特征相比, 将局部韵律特征与全局韵律特征相结合, 其表现略有提高. 研究者开发出一个模块化的、灵活的特征提取工具箱 openSMILE[60], 用来提取常用的声学特征集合. 可用于语音情感识别的特征集有 IS09_emotion[61], IS10_paraling[62], emobase2010[62], IS13_ComParE[63]等. 振幅时频特征是声学时序信息的可视化表达, 包含丰富的信息, 近些年逐渐被用在语音情感识别的研究中.

2.4 相位特征

语音信号在经过时频变换后为复函数, 其信息包含振幅和相位两部分. 相位信息也是声学特征的重要一部分. 为了更直观地分析振幅和相位的关联, 振幅决定了幅值大小, 相位决定了角度, 只有两种信息同时考虑才能获取更精准细化的特征.

有关相位信息的研究经历了一个长期的探索[64], 在语音处理领域相位信息受到越来越多研究者的关注. 语音信号在经过傅里叶变换后, 相位的计算公式如下:

$$\theta(\omega,t)=\tan^{-1}\left(\frac{X_I(\omega,t)}{X_R(\omega,t)}\right) \tag{1}$$

其中, $\theta(\omega,t)$ 表示在频率 ω 、时刻 t 的相位谱, $X(\omega,t)$ 表示语音信号的谱图, R 和 I 分别表示实部和虚部. 早期的时候, 由于相位缠绕等问题容易造成难以对其进行有效建模[65], 因此相位相关的特征在许多应用中被忽略.

为了解决上述问题, 一些新的相位相关的特征被提出. 其中, 群延迟 (group delay, GD) 是常用的相位特征之一[66], 其定义为信号的傅里叶变换相位的负导数, 可以实现对相位信息的简单操作. Hegde 等人[67]提出了改进的群延迟 (modified group delay, MGD), 取得了更优的效果. 但相位信息 $\theta(\omega,t)$ 会随着输入语音信号的裁剪位置而变化, 即便是在相同的频率. 为了克服这个问题, Nakagawa 等人提出了一种相位归一化方法, 称为相对相位 (relative phase, RP), 直接从信号的傅里叶变换中提取[68]. RP 选定一个基础频率并让其相位值保持不变, 然后以此估计其他频率的相位, 具体计算如下:

$$\tilde{\theta}(\omega,t)=\theta(\omega,t)+\frac{\omega}{\omega_b}(-\theta(\omega_b,t)) \tag{2}$$

其中, $\tilde{\theta}$ 代表 RP, ω_b 表示选定的基础频率. 相关研究表明 RP 在语音识别[68]、说话人识别[69,70]和语音增强[71]等任务上起到一定的作用. Guo 等人[72]探究了相位信息对于区分情感的作用, 将改进群延迟倒谱系数 (MGDCC) 和 RP 相位相关特征用于语音情感识别, 结果表明与振幅特征相比, 振幅和相位信息联合使用能显著的提升情感识别率. 此外, 为了进一步缓解传统相位特征对帧裁剪位置依赖的问题, 研究者对 RP 进行了改进, 提出了动态相对相位 (dynamic relative phase, DRP)[73], 并用于语音情感识别任务, 取得了比 RP 更高的识别率. 但相位特征受到噪声和信号失真的影响较大; 相位特征通常基于信号的周期性来表示, 对于频率变化较大的信号, 其可靠性受到限制. 频率的变化可能会导致相位特征的改变, 进而影响其准确性和稳定性.

综上, 在只考虑振幅信息时, 韵律特征和谱特征在语音情感识别中更常用. 作为声学特征中的一部分, 相位特征中包含的情感信息有限, 不适合单独用来识别情感, 需要联合振幅特征以形成更完整的情感表达空间. 在实际应用中, 将各种不同的特征融合在一起能获得更好的识别效果[73].

3 语音情感识别模型

本节首先介绍了几种传统的语音情感识别模型, 其次详细讨论了经典的深度模型, 然后对近些年提出的先进深度模型进行了总结概括. 最后, 总结了常用的评价指标.

3.1 传统模型

传统的语音情感识别算法主要包括隐马尔可夫模型 (HMM)、高斯混合模型 (GMM)、支持向量机 (SVM) 等.

3.1.1 隐马尔可夫模型

隐马尔可夫模型 (HMM) 是基于概率统计的模型, 由 Markov 链演变而来, 依赖于 Markov 属性, 即系统在 t 时刻的当前状态只依赖于之前在 $t-1$ 时刻的状态. HMM 能够较好地体现出声学信号的局部平稳性, 早期被广泛应用于语音识别领域, 并成为该领域的主要技术. 随着语音信号的不断深入研究, HMM 已成功地推广到语音情感识别领域.

HMM 的一个优点是它能够对情绪的时间动态进行建模[74]. Nogueiras 等人[44]使用 HMM 对基频和能量特征及其轮廓进行建模, 取得了与人为主观评测相似的结果. 使用对数频率功率系数 (LFPC)、MFCC 等作为声学特征, HMM 作为分类器, 取得的准确率比人高. 虽然, HMM 在语音信号处理领域取得了不错的成绩, 但其需要进行大量的统计训练来选取模型的系数, 对数据库的规模需求较大, 并且计算资源需求大.

3.1.2 高斯混合模型

高斯混合模型 (GMM) 是一种概率方法, 相当于只包含一个状态的连续 HMM 的特例. 混合模型的思想是根据几个成分的组合对数据进行建模, 其中每个成分都有一个简单的参数形式, 如高斯分布. GMM 假设每个数据点属于一个组件, 并尝试分别推断每个组件的分布. GMM 训练后得到的数据形态和原始随机数据十分相似, 也就是说 GMM 能够对任何连续形态的数据分布进行表征, 而声学信号就符合连续的分布.

Schuller 等人[7]设计了两种情感识别模型对 HMM 和 GMM 进行了比较. GMM 使用基频和能量轮廓的全局统计特征, 而 HMM 使用的是没有进行统计计算的 LLD 特征, 最终结果表明 GMM 结果明显优于 HMM. Ververidis 等人[8]使用 GMM 模型对音高、能量和共振峰参数这些声学特征进行情感分类, 取得了不错的结果. 以上两种模型均采用的全局统计特征, Neiberg 等人[75]提取每一帧 MFCC 为特征, 然后采用 GMM 作分类器, 结果表明在帧级别上使用 GMM 是一种可行的语音情感识别方法[76]. 然而, 当情感特征的维度太多时, GMM 就不起作用了, 并且它不能模拟情感特征之间的非线性关系[12].

3.1.3 支持向量机

支持向量机 (SVM) 是一种为线性可分模式寻找最优超平面的、应用广泛的广义判别分类器. SVM 分类器背后的假设是减少训练误差以及经验风险 (测试) 误差, 即给定训练数据, 在两类数据点之间找到具有最大边界的超

平面. 如果这些模式不是线性可分的, SVM 利用核函数将原始特征空间映射到一个新的高维空间进行分类. 与 HMM 和 GMM 相比, SVM 具有训练算法的全局最优性, 且在非线性、小样本以及高维的识别问题上可以呈现出较好的优势, 已经被普遍应用到语音情感识别中.

Schuller 等人[9]使用多层支持向量机 (ML-SVM) 进行语音情感识别, 取得了比常规 SVM, GMM 更好的结果. Seehapoch 等人[77]利用基频、能量、MFCC 等特征, 并训练不同的 SVM 来组合特征, 在德语数据库上均表现良好. 目前, SVM 已经成为语音情感识别领域常用的分类器[78]. 但是, 核函数的选择是 SVM 的一个关键问题. 事实上, 没有系统的方法来选择核函数, 因此, 变换后的特征的区分度是不能保证的.

3.2 经典深度模型

尽管上述传统语音情感识别模型在很长一段时间里占据情感识别领域的主导地位, 但均存在情感表示能力较弱的问题. 于是, 基于深度学习的模型开始进入研究者的视野. “深度”这个词来自于隐藏层的数量, 因为它可以达到数百层. 与传统的分类器相比, 基于深度学习的情感识别模型表现出了良好的性能, 同时对信号处理的要求也大大降低[12], 因此研究的重点转向了深度学习算法. 比较常用的深度学习模型有深度神经网络 (DNN)、循环神经网络 (RNN) 和卷积神经网络 (CNN).

3.2.1 深度神经网络

随着深度学习的不断发展进步, 基于深度学习的语音情感识别模型被相继提出, 其中就包括深度神经网络 (DNN). DNN 是一种包含多个隐藏层的神经网络, 其网络层有输入层、输出层和隐藏层这 3 种. DNN 能够从原始特征中学习高级表征, 进而完成识别[79], 已经在许多任务中表现出优势[80]. Han 等人[81]提出了一种基于 DNN 和极限学习机 (extreme learning machine, ELM) 的模型 DNN-ELM. 将 DNN 输出的概率分布输入到分类器 ELM 中识别情感. 该模型取得了比 HMM 更高的识别率. Wang 等人[82]又对该模型进行了改进, 作者输出 DNN 最后一个隐藏层的激活值代替原来的输出每种情感类别的概率值, 进一步提高了语音情感识别的精度.

3.2.2 循环神经网络

与 DNN 不同, 循环神经网络 (RNN) 是一种专门处理时序数据的神经网络. 通过使用内部存储器, RNN 可以记住接收到的输入数据, 并对接下来发生的事情做出预测. RNN 已成成功应用于语音信号处理、自然语言处理等领域. 然而, RNN 具有较短的时间记忆, 很难学习到序列之间的长期依赖关系. 为了解决这个问题, 提出了长短时记忆网络 (long short-term memory, LSTM)[83]. LSTM 可以看作是一种特殊的循环神经网络.

目前, RNN 相关的模型已经在语音情感识别中广泛应用. Eyben 等人[84]将 LSTM 与 RNN 结合, 提出了一个混合在线情感识别系统. Lee 等人[85]提出了一种新的混合模型 RNN-ELM, 取得了比 DNN-ELM 更好的识别结果. 考虑到 LSTM 是对单一方向序列的历史信息进行编码, 没有考虑有助于信息理解的未来信息. 因此, 研究者开始采用双向的 LSTM 网络 (BLSTM) 对声学特征进行建模. BLSTM 的主要思想是利用正向 LSTM 和反向 LSTM 提取上下文隐藏信息形成最终输出[86]. BLSTM 可以很好地利用上下文信息, 这对于学习语音情感表征也是很重要的. 此外, Hsiao 等人[87]提出将注意机制整合到 BLSTM 中进行语音情感识别. 通过引入注意机制, 使系统学习如何将注意力集中在输入信号中更具鲁棒性或信息性的片段上.

3.2.3 卷积神经网络

卷积神经网络 (CNN) 是受到猫的视觉感受野 (receptive field) 启发而提出的一种神经网络, 最早应用在图像领域[88], 并且成为该领域的标志性方法. 近年来, CNN 在语音信号处理领域中的应用越来越多. Huang 等人[89]提出了混合的 CNN-SVM 模型, 将 CNN 从时频特征中自动提取的声学特征输入到 SVM 中完成情感的识别. SVM 作为一种静态分类器在应用到语音领域时存在很多局限性, 例如不能有效地利用情感动态信息和上下文信息. 后来提出了 CNN 和 LSTM 结合的模型[10], 将 CNN 提取的特征传输给可以有效利用时序信息的 LSTM. 目前, 基于振幅时频特征的 CNN-BLSTM 框架已经成为语音情感识别任务中最常用的方法之一. 这类方法虽然能够学习到情感的综合表征, 但人类识别情感的一些感性知识没有被有效利用, 很难突显一些显著情感特征 (如 F0) 的作用. 为了研究基于感性知识的启发式特征对基于时频特征的综合情感表征学习的作用, 研究者提出了融合启发式特征和

振幅时频特征的语音情感识别算法[90]. 该方法既保证了特征的丰富性, 又可以突出 F0 等关键声学特征的作用, 从而提升了情感表征区分度.

尽管上述基于 CNN 的方法可以学习到情感的综合表征, 但研究表明语音情感信息可能同时嵌入在时间域和频率域[91]. 为了利用时域/频域信息, 一些方法被提出. Li 等人[92]提出一种基于注意力池化的表征学习方法, 利用 CNN 从振幅时频特征中分别学习时间维度和频率维度的特征. Liu 等人[93]提出时频卷积神经网络 (time-frequency CNN, TFCNN), 利用 CNN 分别提取时间相关和频率相关的表征. 然而这两种方法都没有明确地对时间/频率轴上特征的关联或相对重要性进行建模. 此外, 没有考虑卷积中通道域的影响, 而每个通道的特征图重要性不同[94]. 为此, Guo 等人[95]提出基于时间-频率-通道 (spectro-temporal-channel, STC) 注意力的深度表征学习模型, 旨在通过 STC 注意力模块来增强情感表示能力. 目前, 基于 CNN 模型从振幅时频特征中学习高层情感表征已经成为语音情感识别中最常用的方法之一.

3.3 先进 SER 模型

上述方法都是从振幅信息角度研究语音情感识别而忽略了相位信息. 振幅和相位是语音声学特征的两个基本要素, 忽略相位信息的情感表征可能会有缺失. 基于此考虑, 研究者从声学特征的完整性出发, 探究情感语音中相位特征的准确抽取和有效使用. 此外, 一些基于 Transformer、图卷积网络 (GCN) 的先进模型也相继被提出.

3.3.1 利用振幅与相位互补的 SER 模型

声学特征的完备描述是由振幅信息和相位信息共同承担的. 传统的语音情感识别方法中, 研究者大多关注于振幅声学特征, 而忽略了相位信息. 忽略相位信息的声学特征不够完整, 从而造成情感表征不够细化具体. 但是如何准确抽取相位特征并将其应用于情感识别以获取更为完整的声学表征是语音情感识别所面临的又一挑战, 具有重要的研究价值.

研究者相继提出了一些能够利用振幅和相位信息的语音情感识别模型. Deng 等人[96]利用相位相关的特征进行耳语语音情感识别. 他们将 MFCC 与基于群延迟 (GD) 的特征结合, 然后利用 SVM 进行分类, 取得了优于 MFCC 特征的结果. 但仅利用浅层模型很难提取到一些深层的情感表征, 从而影响情感识别的精度. Guo 等人[97]提出基于振幅时频特征和改进的群延迟 (MGD) 特征的深度语音情感识别模型, 利用卷积神经网络 (CNN) 对振幅和相位信息进行建模. 考虑到群延迟特征 (GD) 相关的特征中仍包含一些振幅信息[66,67], 不利于定量分析相位对语音情感识别的影响. 于是, 研究者通过相对相位 (RP) 特征探究了相位信息对于语音情感识别的作用[72]. 随后, 研究者通过对相位特征进行定量分析表明相位中也包含可以用来区分情感的信息, 并且提出了一种动态相对相位 (DRP) 特征提取方法以解决相对相位 (RP) 存在难以确定基础基频等问题, 从而进一步缓解传统相位对于帧裁剪位置的依赖问题[73]. 此外, 文中还提出了两种基于振幅和相位信息的融合机制以提取到较为完整声学特征. 一种是单通道模型 (SCM), 即首先将振幅特征和相位特征进行拼接, 然后输入到模型中提取互补特征; 另一种是多通道注意力模型 (MCMA), 即分别利用模型从振幅和相位中提取特征, 然后通过注意力层将两种特征进行融合. Prabhakar 等人[98]提出一种基于 MFCC 和修正群延迟 (modified group delay function, MODGD) 的多通道 CNN-BLSTM 框架, 然后将学习到振幅和相位表征结合后传递到 SVM 进行分类. 此外, 文中利用深度典型相关分析 (deep canonical correlation analysis, DCCA) 使振幅信息和相位信息之间的相关性最大化, 从而提高情感识别性能. 上述研究表明相位信息对语音情感识别具有一定的作用, 可以对振幅信息进行某种程度的互补.

3.3.2 基于 Transformer 的 SER 模型

Transformer 是由谷歌在 2017 年提出的 Seq2Seq 模型[91], 基于注意力机制构建. Transformer 不同于传统的 CNN 和 RNN, 其网络结构由自注意和前馈神经网络组成. 与 RNN 相比, Transformer 中所有时间步的数据都是通过自注意力计算, 这样便使得整个过程可以并行计算. 与 CNN 相比, Transformer 将序列中任意两个位置之间的距离缩小为一个常量, 计算两个位置之间的关联所需的操作次数不随距离增长. Transformer 已经成功应用到许多领域, 如自然语言处理[99]、视频处理[100]等领域.

在语音情感识别领域, 一些基于 Transformer 的深度模型别相继被提出. Wang 等人[101]提出一种端到端语音情感识别框架, 其新颖之处在于在经典的语音情感识别模型顶部嵌入堆叠的 Transformer 层. Andayani 等人[102]建立了 LSTM-Transformer 模型, 从 MFCC 中学习情感的长期依赖关系, 同传统的 LSTM 相比, 极大地提升了识别性能. 但该模型仍然避免不了 LSTM 串行的问题, 计算效率有待进一步提升. Gumelar 等人[103]提出 Transformer-CNN 模型, 对语音特征做时间和空间的处理. 由于 Transformer 可以对不同时间序列的信息进行建模, 许多研究者将语音模态和其他模态信息结合开展多模态情感识别研究. Huang 等人[104]利用 Transformer 网络来探索跨语音和视觉模态的情感表征融合; Lian 等人[105]提出一种用于对话情感识别的学习框架 (CTNet), 其中利用基于 Transformer 的结构来建模声学模态和文本模态之间的交互; Chen 等人[106]提出一种只关注情感信息的关键稀疏 Transformer (KS-Transformer) 模型, 并级联交叉注意力模块实现语音和文本模态之间的深度交互. 此外, 还有一些联合语音、视频的三模态模型被提出. Tran 等人[107]提出了第 1 个基于音频和视频的预训练多模态 Transformer 模型, 旨在从面部和听觉行为的互动中捕获有用的情感信息. Wang 等人[108]提出了一种多模态 Transformer 增强的 SER 方法, 该方法采用混合融合策略, 结合特征级融合和模型级融合方法, 在模态内部和模态之间进行细粒度的信息交互. 设计了一个由 3 个交叉 Transformer 编码器组成的模型融合模块, 用于模态引导和信息融合. 虽然基于 Transformer 的框架在 SER 中展示出了效果, 但现有的工作并没有评估模型大小和预训练数据对下游性能的影响, 并且对泛化、鲁棒性、公平性和效率的关注有限. 于是, Wagner 等人[109]讨论了基于 Transformer 的模型在大量未标记数据上预训练的结果, 在最后一个 Transformer 层的隐藏状态上应用平均池化, 并通过隐藏层和最终输出层提供结果. Liu 等人[110]利用自监督学习增强语音特征的鲁棒性, 提出了一种由卷积层、Transformer 模块和双向长短期记忆 (BLSTM) 模块组成的特征融合模型 (Dual-TBNet). 综上, 基于 Transformer 的模型在情感表征学习以及多模态交互建模上表现良好. 虽然 Transformer 能够实现并行计算以及输入和输出之间的全局关系, 但在局部信息获取方面不如 RNN 和 CNN.

3.3.3 基于图卷积网络的 SER 模型

传统的语音情绪识别方法大多以 RNN 或 LSTM 为基础对时间序列进行模型, 但这类模型难以刻画子序列间的长距离依赖或全局依赖, 不利于挖掘深层的依赖关系. 图是一种紧凑、高效且可扩展的数据表示方式, 可以将不同序列看作图中的节点, 进而利用图神经网络进行建模. 虽然上文介绍的 Transformer 是图神经网络的特例, 但其不支持与任意图结构进行交互, 具有一定的局限性. 而图卷积网络 (GCN)[111]可以对节点之间的依赖关系进行建模, 在处理广义拓扑图结构上发挥着重要作用. GCN 能够深入挖掘不规则数据的特征和规律, 通过聚合每个节点的邻接节点特征而获得该节点的聚合特征表示.

目前, 图卷积网络 (GCN) 已经成功用于解决计算机视觉和自然语言处理的各种问题, 如动作识别[112]、目标跟踪[113]、文本分类[114]等. 受此启发, 研究者在情感识别领域开始探究 GCN 的作用. Shirian 等人[115]将一个语音信号转换为一个简单的图, 其中分割的语音段充当每个节点, 且每个节点只连接两个相邻点, 极大地简化了图上的卷积操作, 取得了比标准 GCN 更好的识别率. 为了能够对可变长度的句子进行建模, Liu 等人[116]提出 GraphSAGE 模型, 将语音情感识别问题转换为一个图分类问题. 将可变长度的句子转换为图, 以避免填充或切割. 在该方法中语音的每一帧表示为图中的节点, 从帧中提取的声学特征作为节点特种向量, 进而根据帧的时间关系连接节点. 以上方法都是对单句进行建模, 为了利用句子间的语境信息, Liu 等人[117]在对话语音情感识别模型中引入图卷积网络对句子之间的关联性进行建模, 从而学习到有效的上下文情感信息. Ghosal 等人[118]提出了 DialogGCN 模型, 将每段对话视为一个图, 其中每个句子与周边句子相连. 后续又有一些改进模型, 例如, Fu 等人[119]提出一种面向上下文对话和知识的图卷积网络 (ConSK-GCN) 的情感识别方法, 利用 GCN 进行上下文建模的基础上引入知识图谱对对话内容进行知识增强. 该模型将上下文建模、知识图谱和多模态 (文本和音频) 融入到 GCN 的学习中. Shen 等人[120]提出了用有向无环图编码句子的新想法, 以更好地模拟对话语句中的内在结构, 并设计了一个有向无环神经网络, 其识别性能超过了 DialogGCN. Chandola 等人[121]提出了一种基于 GCN 的语音对话情感识别模型 (SERC-GCN), 首先提取话语级语音信号的情感特征, 然后将这些特征形成对话图, 这些对话图被用来训练图卷积

网络来执行情感判别.

3.3.4 其他 SER 模型

用于识别情感的表征可能来自能量谱、频谱等多种物理属性, 这些属性可以作为多视图表征进行收集. 为了充分探讨多个情感表征之间潜在的交互关系, Hou 等人[122]提出了一种新的集体多视图关系网络 (CMRN), 利用多视图语音表征的内在特性来进行语音情感识别. 所提出的 CMRN 由 3 个子网络组成, 即特定视图注意网络、多视图共享注意网络和集体关系网络. 该方法可以综合利用多个表征共享的和特定视图的信息, 最终实现聚合多视图特征的异质信息, 进行准确的情感识别. Liu 等人[123]提出的注意力时间动态激活 (ATDA) 模型中也引入了多视图模块, 基于多个注意力视图和粒度进一步检测和加强情感相关的动态特征. 由于个体情感感知的差异, 多个注释者会产生不同的标签. Li 等人[124]提出基于跨类差异损失的响应残差网络用于多标签语音情感识别, 使网络能够自适应地学习所有话语中的情绪分布. 由于每个个体都有其独特的表达习惯, 所以个体差异会导致分布偏差, 跨个体情感表征学习面临挑战. Fan 等人[125]提出了一种个体标准化网络 (ISNet) 用于语音情感识别, 以缓解个体差异造成的个体间情感混淆问题. 此外, Lu 等人尽可能减少不同说话人之间语音样本特征分布的差异提出了一种新的域不变特征学习方法[126], 从多源无监督域适应的角度出发, 通过消除不同说话人引起的训练和测试数据之间的域转移来学习说话人不变的情感特征. 这类方法旨在消除说话人差异的影响, 不利于在一些追求个性化服务的人机交互中开展.

自监督预训练特征在自然语言处理 (NLP) 领域展现出了显著的效果, 在 SER 领域也有许多相关研究被提出. Morais 等人[127]介绍了一个基于上游+下游架构范式的模块化端到端 SER 系统, 该系统允许轻松使用/集成各种自监督功能. 通过对自监督语音表示模型进行微调, 可以在语音情感识别 (SER) 任务中取得良好的表现. 但在实际应用中部署时, 仍然需要使其适应嘈杂的目标环境. 因此, Leem 等人[128]提出了一个对比的师生学习框架来重新训练一个自监督的语音表示模型. 为了保留原始模型的知识, 最小化了原始 SER 模型的干净嵌入与重新训练模型的噪声嵌入之间的均方根误差. 为了获得目标噪声条件下的判别知识, 还选择相应的干净嵌入作为正样本, 选择其他具有不同情感标签的噪声嵌入作为负样本, 以最小化损失. 为了充分利用不同种类的声学特征, Li 等人[129]提出了基于多特征和参数优化的级联深度学习网络; Chen 等人[130]提出了一种基于连接注意机制的多尺度 SER 并行网络 (AMSNet), AMSNet 融合了细粒度的帧级手动特征和粗粒度的话语级深度特征, 同时根据语音信号的时空特征, 采用不同的语音情感特征提取模块, 丰富了特征, 提高了表征能力.

3.4 评价指标

由于本文主要以离散情感为核心展开综述, 所以本节主要介绍分类模型的常用评价指标. 一般地, 把精确率 (precision, *P*)、召回率 (recall, *R*) 及两者的调和值 *F*1 作为评价指标来衡量模型对每类情感的识别性能, 用准确率 (accuracy) 评估情感识别的整体性能. *F*1 是精确率和召回率的调和平均值, 值越大表示模型识别性能越好. *F*1 相当于对精确率和召回率进行了加权. 准确率 (accuracy) 是分类任务的一个常用指标, 分为加权准确率 (weighted accuracy, *WA*) 和非加权准确率 (unweighted accuracy, *UA*). *WA* 是指所有样本中预测正确的比例, 适用于平衡数据库, 不适用于不平衡数据库, 因为 *WA* 对样本数量较大的类别给予了更多的权重, 而给予小样本类别的权重较少. 而 *UA* 是先计算每一类别的正确率即召回率, 然后取平均值. *UA* 在数据不平衡的情况下可以对模型起到很好的评判作用.

下面具体介绍各个评价指标, 其中 *TP* 表示预测为正, 真实为正; *FP* 表示预测为正, 真实为负; *TN* 表示预测为负, 真实为负; *FN* 表示预测为负, 真实为正.

(1) 精确率

精确率也叫查准率, 指被正确预测为正的样本 (*TP*) 占所有预测为正样本的比例, 计算公式为:

$$P = \frac{TP}{TP + FP} \tag{3}$$

(2) 召回率

召回率也叫查全率, 指所有预测为正的样本在所有真实为正样本中的比例, 是一般意义上的正确率, 公式

如下:

$$R=\frac{TP}{TP+FN} \tag{4}$$

关于精确率和召回率的使用取决于具体的应用场景, 但当应用系统对查全率和查准率都要求较高的时候, 研究者需要用 $F1$ 来进行评测.

(3) $F1$ 值

$F1$ 是精确率和召回率的调和平均值, 值越大表示模型识别性能越好. $F1$ 相当于对精确率和召回率进行了加权, 具体计算公式如下:

$$F1=\frac{2\times P\times R}{P+R} \tag{5}$$

(4) 准确率

准确率 (accuracy) 是分类任务的一个常用指标, 可以用来衡量所有类别的整体识别效果, 又可以分为加权准确率 (WA) 和非加权准确率 (UA). WA 是指所有样本中预测正确的比例, 可以表示为:

$$WA=\frac{TP+TN}{TP+TN+FP+FN} \tag{6}$$

WA 适用于平衡数据库, 不适用于不平衡数据库, 因为 WA 对样本数量较大的类别给予了更多的权重, 而给予小样本类别的权重较少. 而 UA 是先计算每一类别的正确率即召回率, 然后取平均值. UA 在数据不平衡的情况下可以对模型起到很好的评判作用.

3.5 模型总结

结合 WA, UA, $F1$ 等评价指标, 表 4 对传统模型、经典深度模型、先进 SER 模型进行了总结对比.

表 4 用于语音情感识别的模型总结

模型类别	文献	模型	声学特征	数据库及情感类别	结果 (%)
传统机器学习模型	Nogueiras等人[44] (2001)	HMM	基频、能量特征及其轮廓	IESSDB: 愤怒、厌恶、恐惧、高兴、悲伤、惊喜和中性	*WA*: 82.52 *UA*: 82.56 *F*1: 82.44
	Schuller等人[7] (2003)	HMM/GMM	基频、能量的统计值/低级瞬时特征	德语录制: 愤怒、厌恶、惊喜、恐惧、高兴、悲伤和中性	*WA*: 86.80 *UA*: 86.80
	Ververidis等人[8] (2005)	GMM	音高、能量和共振峰	丹麦数据库DES: 愤怒、高兴、中性、悲伤、惊喜	*WA*: 55.6 *UA*: 55.20
	Albornoz等人[76] (2011)	GMM、HMM和MLP组合	韵律特征、MFCC	EmoDB: 愤怒、无聊、厌恶、高兴、恐惧、悲伤和中性	*WA*: 71.75 *UA*: 71.75 *F*1: 71.82
	Seehapoch等人[77] (2013)	SVM	基频、能量、MFCC等特征	EmoDB: 愤怒、无聊、厌恶、高兴、恐惧、悲伤和中性	*WA*: 89.80 *UA*: 89.15
经典深度模型	Han等人[81] (2014)	DNN-ELM	基频、MFCC	IEMOCAP (完整数据): 愤怒、高兴、悲伤和中性	*WA*: 54.3 *UA*: 48.2
	Hsiao等人[87] (2018)	BLSTM-attention	基频、过零率、MFCC、能量等的统计值	FAU AIBO: 愤怒、中性、积极、强调、放松	*UA*: 46.3
	Satt等人[10] (2017)	CNN-BLSTM	振幅时频特征	IEMOCAP (即兴数据): 愤怒、高兴、悲伤和中性	*WA*: 68.8
	Guo等人[90] (2018)	CNN-ELM	基频、能量、MFCC和振幅时频特征	EmoDB: 愤怒、无聊、厌恶、高兴、恐惧、悲伤和中性	*WA*: 92.45 *UA*: 91.97 *F*1: 92.5
	Li等人[92] (2018)	CNN_TF_Att.pooling	振幅时频特征	IEMOCAP (即兴数据): 愤怒、高兴、悲伤和中性	*WA*: 71.75 *UA*: 68.06
	Guo等人[95] (2023)	DSTCNet	振幅时频特征	IEMOCAP (完整数据): 愤怒、高兴、悲伤和中性	*WA*: 61.80 *UA*: 61.78 *F*1: 62.51

表 4 用于语音情感识别的模型总结 (续)

模型类别	文献	模型	声学特征	数据库及情感类别	结果 (%)
先进SER模型	Guo等人[73] (2022)	多通道模型MCMA	振幅时频特征和相位特征 (MGD, DRP)	EmoDB: 愤怒、无聊、厌恶、高兴、恐惧、悲伤和中性	*WA*: 94.02 *UA*: 93.66 *F*1: 94.19
	Chen等人[106] (2022)	KS-Transformer	Wav2vec (加文本特征)	EMOCAP (完整数据): 愤怒、高兴、悲伤和中性	*WA*: 74.3 *UA*: 75.3
	Liu等人[116] (2022)	GraphSAGE	振幅时频特征	EMOCAP (完整数据): 愤怒、高兴、悲伤和中性	*WA*: 65.43 *UA*: 66.40
	Liu等人[123] (2022)	ATDA	MFCC	IEMOCAP (即兴数据): 愤怒、高兴、悲伤和中性	*WA*: 76.2 *UA*: 75.4 *F*1: 75.8
	Liu等人[110] (2023)	Dual-TBNet	基频、能量等手动特征及其统计值	IEMOCAP (即兴数据): 愤怒、高兴、悲伤和中性	*UA*: 64.8 *F*1: 65.25
	Chen等人[130] (2023)	AMSNet	振幅时频特征	IEMOCAP (即兴数据): 愤怒、高兴、悲伤和中性	*WA*: 69.22 *UA*: 70.51
	Chandola等人[121] (2024)	SER-GCN	LLD	IEMOCAP (即兴数据): 愤怒、高兴、悲伤、中性和兴奋	*WA*: 66.85 *UA*: 67.55

从表 4 中可以看出, 传统机器学习模型使用的声学特征大都为启发式特征 (基频、能量、MFCC 等). 这类特征是基于人类识别情感的感性知识设计的, 可以直观地体现情感信息. 这个时期研究者主要关注如何设计提取不同的启发式特征. 我们还可以发现, 韵律特征的应用更为广泛. 主要原因是基频 (F0) 直接反映声音的音调, 能量 (energy) 则直观的体音量和声音强度, 这些因素都是和情感表达息息相关的. 随着深度学习的发展, 深度模型开始占据语音情感识别领域的主导地位. 针对启发式特征, 常采用 DNN/RNN 等模型进行高级表征的学习. 近些年, 声学特征开始从手动设计向机器自动学习发展. 其中振幅时频特征的应用更为广泛, 主要是因为时频特征是声学时序信息的可视化表达, 可以给出情感的综合表征, 包含更为丰富的情感信息. 此时, DNN 的表征学习能力便会受到限制, 无法挖掘谱图的局部信息. 在基于振幅时频特征建模的方法中 CNN 的应用更为普遍, 主要是因为 CNN 丰富的表示能力和权值共享等特性, 可以从谱图中捕获到有效的情感信息. 近些年, 声学特征的使用趋向于多样化. 除了传统的振幅相关特征, 相位特征也被有效利用. 这些特征的使用也不再仅局限于特定模型中. 综合运用适合的模型对不同特征建模, 然后提取融合情感表征有助于进一步完善情感表征. 此外, 针对不同类型的特征进行多视图建模也表现良好.

4 挑战与展望

尽管语音情感识别系统已经取得了许多进展, 但想要精准地识别情感并实现情感交互, 仍有一些挑战需要解决. 在本节中, 我们简要讨论这些挑战和未来潜在的研究方向.

4.1 情感数据库的匮乏

语音情感识别的基础就是要拥有优质、全面的情感数据库, 但由于情感本身是复杂的, 其采集录制过程都是有困难的, 进而导致可用于研究的高质量语音情感数据库规模较小[11]. 此外, 在情感标注过程中也存在一些问题. 在人工标注者对语音数据进行标注时, 讲话者实际要表达的情绪与人类注释者感知到的情绪可能存在差异、标注者之间也会存在判别分歧[131], 且在上下文语境未知的情况下, 标注会变得更加难以进行. 因此, 如何对情感数据库进行补充以及有效利用现有资源进行研究都是亟待解决的实际问题.

4.2 声学特征和情感之间的关联

研究声学特征与情感的关联性. 当前研究使用韵律特征、谱特征、音质特征、相位信息等声学特征进行语音情感识别, 虽然保证了声学特征的丰富性且识别性能得到了一定的保障, 但是缺乏对情感特征与情感类别的针对性关联分析[132], 没有详细讨论这些特征对情感的表达是否一定有效, 进而无法挑选出情感判别能力最优的特征集

合. Banse 等人[133]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. 因此, 开展情感识别任务特有的表征学习探究, 分析挖掘与情感类别关联更加密切的特征集合是语音情感识别领域十分重要的研究.

4.3 副语言信息和语言信息的相互作用

语音是通过语言调制的声学信号, 人类在语音中通常是通过副语言信息和语言信息来共同表达情感. 研究表明语音中的语言内容在表达某些负性情感的时候会起到重要的作用, 而对于中性语言内容则主要靠副语言信息来赋予不同的情感[134]. 可见, 语言信息和副语言信息在表达情感的时候具有相互补充的作用. 因此, 通过语音去识别情感的时候, 声学信号中承载的语言信息不能被忽略. 如何有效利用语言信息和副语言信息之间的相互作用, 获取更加互补的情感表征, 也是一个具有挑战性的难题.

4.4 对话场景的情感交互

研究面向对话的情感交互. 在对话系统中注入情感可以使会话主体更加人性化, 有利于人机交互[135]. 而目前的情感识别研究大都在情感识别专用的数据库展开, 没有和对话系统进行较好的结合[1]. 以后的情感识别研究应更多地结合真实应用场景展开, 做到将情感真正融入到人机交互系统, 实现自然和谐的人机情感交互, 从而为人类提供更人性化的服务. 因此, 研究如何结合对话系统进行人机情感交互, 也是未来具有挑战性的研究问题. 此外, 人的情感在交流过程中是一个动态变化的过程, 关于情感动态演绎的过程也值得关注.

4.5 可解释性情感识别

除上述挑战外, 用于识别语音情感的深度模型也有待继续探究. 近些年语音情感识别的发展主要依靠深度神经网络, 而深度神经网络就像一个黑盒, 缺乏可解释性[136]. 情感是高层次的表达, 识别过程需要推理、记忆、决策等能力. 因此, 构建适用于语音情感识别特定任务, 且解释能力强的深度神经网络将有助于促进语音情感识别的发展. 此外, 研究人脑的情感处理机制[36], 用于启发/指导构建语音情感识别模型或建立类脑语音情感识别智能算法都是十分具研究价值的, 将有助于突破语音情感识别研究瓶颈.

5 总 结

本文对语音情感识别进行了一个全面的研究综述, 旨在通过一个统一架构来概述语音情感识别的理论和方法基础. 本文首先介绍了语音情感识别需要的基础支撑 (即情感描述和情感数据库). 然后, 总结了常用的声学特征, 尤其是在语音情感识别综述中首次进行了相位相关特征的分析与总结. 随后, 详细讨论了语音情感识别分类模型, 并从传统模型、经典的深度模型和先进深度模型这 3 个方面, 进行了全面回顾、比较和总结. 其中着重总结了利用振幅和相位信息互补的语音情感识别模型和基于 Transformer、GCN 等的先进深度模型. 另外, 简要地叙述了常用的评价指标. 最后, 基于对当前语音情感识别的总结与分析指出了领域内亟待解决的难题与值得进一步研究的方向. 虽然本文不能涉及所有关于语音情感识别的研究, 但希望我们对最新情感特征和识别模型的总结能够给从事语音情感识别的其他研究者带来启发, 进而促进语音情感识别的未来发展.

References:

[1] Guo LL. Research on emotion recognition based on paralinguistic and linguistic information [Ph.D. Thesis]. Tianjin: Tianjin University, 2021 (in Chinese with English abstract). [doi: 10.27356/d.cnki.gtjdu.2021.000107]

[2] Bahreini K, Nadolski R, Westera W. Towards multimodal emotion recognition in e-learning environments. Interactive Learning Environments, 2016, 24(3): 590–605. [doi: 10.1080/10494820.2014.908927]

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

[4] France DJ, Shiavi RG, Silverman S, Silverman M, Wilkes M. Acoustical properties of speech as indicators of depression and suicidal risk. IEEE Trans. on Biomedical Engineering, 2000, 47(7): 829–837. [doi: 10.1109/10.846676]

[5] Bořil H, Sadjadi SO, Kleinschmidt T, Hansen JHL. Analysis and detection of cognitive load and frustration in drivers' speech. In: Proc. of the 2010 Annual Conference of the International Speech Communication Association. Makuhari: ISCA, 2010. 502–505.

[6] Stuhlsatz A, Meyer C, Eyben F, Zielke T, Meier G, Schuller B. Deep neural networks for acoustic emotion recognition: Raising the

benchmarks. In: Proc. of the 2011 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP). Prague: IEEE, 2011. 5688–5691. [doi: 10.1109/ICASSP.2011.5947651]

[7] Schuller B, Rigoll G, Lang M. Hidden Markov model-based speech emotion recognition. In: Proc. of the 2003 IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing. Hong Kong: IEEE, 2003. 1–4. [doi: 10.1109/ICASSP.2003.1202279]

[8] Ververidis D, Kotropoulos C. Emotional speech classification using Gaussian mixture models. In: Proc. of the 2005 IEEE Int'l Symp. on Circuits and Systems. Kobe: IEEE, 2005. 2871–2874. [doi: 10.1109/ISCAS.2005.1465226]

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

[10] Satt A, Rozenberg S, Hoory R. Efficient emotion recognition from speech using deep learning on spectrograms. In: Proc. of the 2017 Annual Conference of the International Speech Communication Association. Stockholm: ISCA, 2017. 1089–1093. [doi: 10.21437/Interspeech.2017-200]

[11] Han WJ, Li HF, Ruan HB, Ma L. Review on speech emotion recognition. Ruan Jian Xue Bao/Journal of Software, 2014, 25(1): 37–50 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4497.htm [doi: 10.13328/j.cnki.jos.004497]

[12] Shah Fahad M, Ranjan A, Yadav J, Deepak A. A survey of speech emotion recognition in natural environment. Digital Signal Processing, 2021, 110: 102951. [doi: 10.1016/j.dsp.2020.102951]

[13] Akçay MB, Oğuz K. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Communication, 2020, 116: 56–76. [doi: 10.1016/j.specom.2019.12.001]

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

[15] Ekman P. An argument for basic emotions. Cognition and Emotion, 1992, 6(3–4): 169–200. [doi: 10.1080/02699939208411068]

[16] Cowie R, Cornelius RR. Describing the emotional states that are expressed in speech. Speech Communication, 2003, 40(1–2): 5–32. [doi: 10.1016/S0167-6393(02)00071-7]

[17] Cowie R, Douglas-Cowie E, Tsapatsoulis N, Votsis G, Kollias S, Fellenz W, Taylor JG. Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine, 2001, 18(1): 32–80. [doi: 10.1109/79.911197]

[18] Zeng ZH, Pantic M, Roisman GI, Huang TS. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2008, 31(1): 39–58. [doi: 10.1109/TPAMI.2008.52]

[19] Song P, Zheng WM, Zhao L. Joint subspace learning and feature selection method for speech emotion recognition. Journal of Tsinghua University (Science and Technology), 2018, 58(4): 347–351 (in Chinese with English abstract). [doi: 10.16511/j.cnki.qhdxxb.2018.26.014]

[20] Ververidis D, Kotropoulos C. Emotional speech recognition: Resources, features, and methods. Speech Communication, 2006, 48(9): 1162–1181. [doi: 10.1016/j.specom.2006.04.003]

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

[22] Williams CE, Stevens KN. Emotions and speech: Some acoustical correlates. The Journal of the Acoustical Society of America, 1972, 52(4B): 1238–1250. [doi: 10.1121/1.1913238]

[23] Ringeval F, Sonderegger A, Sauer J, Lalanne D. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In: Proc. of the 10th IEEE Int'l Conf. and Workshops on Automatic Face and Gesture Recognition (FG). Shanghai: IEEE, 2013. 1–8. [doi: 10.1109/FG.2013.6553805]

[24] Cowie R, Douglas-Cowie E, Cox C. Beyond emotion archetypes: Databases for emotion modelling using neural networks. Neural networks, 2005, 18(4): 371–388. [doi: 10.1016/j.neunet.2005.03.002]

[25] Burkhardt F, Paeschke A, Rolfes M, Sendlmeier WF, Weiss B. A database of German emotional speech. In: Proc. of the 9th European Conf. on Speech Communication and Technology. Lisbon: ISCA, 2005. 1517–1520. [doi: 10.21437/Interspeech.2005-446]

[26] Tao J, Liu F, Zhang M, Jia H. Design of speech corpus for mandarin text to speech. In: Proc. of the Blizzard Challenge 2008 Workshop, 2008. 1–4.

[27] Martin O, Kotsia I, Macq B, Pitas I. The eNTERFACE'05 audio-visual emotion database. In: Proc. of the 22nd Int'l Conf. on Data Engineering Workshops (ICDEW). Atlanta: IEEE, 2006. 8. [doi: 10.1109/ICDEW.2006.145]

[28] Busso C, Bulut M, Lee CC, Kazemzadeh A, Mower E, Kim S, Chang JN, Lee S, Narayanan SS. IEMOCAP: Interactive emotional dyadic motion capture database. Language Resources and Evaluation, 2008, 42(4): 335–359. [doi: 10.1007/s10579-008-9076-6]

[29] Li AJ, Zheng F, Byrne W, Fung P, Kamm T, Liu Y, Song ZJ, Ruhi U, Venkataramani V, Chen XX. CASS: A phonetically transcribed

corpus of mandarin spontaneous speech. In: Proc. of the 6th Int'l Conf. on Spoken Language Processing. Beijing: ISCA, 2000. 485–488. [doi: 10.21437/ICSLP.2000-120]

[30] Batliner A, Steidl S, Nöth E. Releasing a thoroughly annotated and processed spontaneous emotional database: The FAU AIBO emotion corpus. In: Proc. of f a Satellite Workshop of LREC 2008 on Corpora for Research on Emotion and Affect. Marrakesh: LREC, 2008. 28–31.

[31] Cowie R, Douglas-Cowie E, Savvidou S, McMahon E. FEELTRACE: An instrument for recording perceived emotion in real time. In: Proc. of the 2000 ISCA Workshop on Speech and Emotion. Newcastle: ISCA, 2000. 19−24.

[32] Li Y, Tao JH, Chao LL, Bao W, Liu YZ. CHEAVD: A Chinese natural emotional audio-visual database. Journal of Ambient Intelligence and Humanized Computing, 2017, 8(6): 913–924. [doi: 10.1007/s12652-016-0406-z]

[33] Lubis N, Gomez R, Sakti S, Nakamuraz K, Yoshino K, Nakamura S, Nakadai K. Construction of Japanese audio-visual emotion database and its application in emotion recognition. In: Proc. of the 10th Int'l Conf. on Language Resources and Evaluation. Portorož: ACL, 2016. 2180–2184.

[34] Chen J, Wang CH, Wang KJ, Yin CQ, Zhao C, Xu T, Zhang XY, Huang ZQ, Liu MC, Yang T. HEU Emotion: A large-scale database for multimodal emotion recognition in the wild. Neural Computing and Applications, 2021, 33(14): 8669–8685. [doi: 10.1007/s00521-020-05616-w]

[35] Costantini G, Iaderola I, Paoloni A, Todisco M. EMOVO corpus: An Italian emotional speech database. In: Proc. of the 9th Int'l Conf. on Language Resources and Evaluation. Reykjavik: ACL, 2014. 3501–3504.

[36] Li HF, Chen J, Ma L, Bo HJ, Xu C, Li HW. Dimensional speech emotion recognition review. Ruan Jian Xue Bao/Journal of Software, 2020, 31(8): 2465–2491 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6078.htm [doi: 10.13328/j.cnki.jos.006078]

[37] Lin JC, Wu CH, Wei WL. Error weighted semi-coupled hidden Markov model for audio-visual emotion recognition. IEEE Trans. on Multimedia, 2012, 14(1): 142–156. [doi: 10.1109/TMM.2011.2171334]

[38] Frick RW. Communicating emotion: The role of prosodic features. Psychological Bulletin, 1985, 97(3): 412–429. [doi: 10.1037/0033-2909.97.3.412]

[39] Busso C, Lee S, Narayanan S. Analysis of emotionally salient aspects of fundamental frequency for emotion detection. IEEE Trans. on Audio, Speech, and Language Processing, 2009, 17(4): 582–596. [doi: 10.1109/TASL.2008.2009578]

[40] Origlia A, Galata V, Ludusan B. Automatic classification of emotions via global and local prosodic features on a multilingual emotional database. In: Proc. of the 2010 Speech Prosody. Chicago: ISCA, 2010. 213.

[41] Seppänen T, Väyrynen E, Toivanen J. Prosody-based classification of emotions in spoken finnish. In: Proc. of the 8th European Conf. on Speech Communication and Technology. Geneval: ISCA, 2003. 717−720.

[42] Iliou T, Anagnostopoulos CN. Statistical evaluation of speech features for emotion recognition. In: Proc. of the 4th Int'l Conf. on Digital Telecommunications. Colmar: IEEE, 2009. 121−126. [doi: 10.1109/ICDT.2009.30]

[43] Luengo I, Navas E, Hernáez I, Sánchez J. Automatic emotion recognition using prosodic parameters. In: Proc. of the 9th European Conf. on Speech Communication and Technology. Lisbon: ISCA, 2005. 493−496.

[44] Nogueiras A, Moreno A, Bonafonte A, Mariño JB. Speech emotion recognition using hidden Markov models. In: Proc. of the 7th European Conf. on Speech Communication and Technology. Aalborg: ISCA, 2001. 2679−2682.

[45] Koolagudi SG, Rao KS. Emotion recognition from speech: A review. Int'l Journal of Speech Technology, 2012, 15(2): 99–117. [doi: 10.1007/s10772-011-9125-1]

[46] Kuchibhotla S, Vankayalapati HD, Vaddi RS, Anne KR. A comparative analysis of classifiers in emotion recognition through acoustic features. Int'l Journal of Speech Technology, 2014, 17(4): 401–408. [doi: 10.1007/s10772-014-9239-3]

[47] Kim EH, Hyun KH, Kim SH, Kwak YK. Improved emotion recognition with a novel speaker-independent feature. IEEE/ASME Trans. on Mechatronics, 2009, 14(3): 317–325. [doi: 10.1109/TMECH.2008.2008644]

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

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

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

[51] Sato N, Obuchi Y. Emotion recognition using Mel-frequency cepstral coefficients. Information and Media Technologies, 2007, 2(3): 835–848. [doi: 10.5715/jnlp.14.4_83]

[52] Wang KX, An N, Li BN, Zhang YY, Li L. Speech emotion recognition using fourier parameters. IEEE Trans. on Affective Computing, 2015, 6(1): 69–75. [doi: 10.1109/TAFFC.2015.2392101]

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

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

[55] Zhang SQ. Emotion recognition in Chinese natural speech by combining prosody and voice quality features. In: Proc. of the 5th Int'l Symp. on Neural Networks. Beijing: Springer, 2008. 457–464. [doi: 10.1007/978-3-540-87734-9_52]

[56] Kächele M, Zharkov D, Meudt S, Schwenker F. Prosodic, spectral and voice quality feature selection using a long-term stopping criterion for audio-based emotion recognition. In: Proc. of the 22nd Int'l Conf. on Pattern Recognition. Stockholm: IEEE, 2014. 803–808. [doi: 10.1109/ICPR.2014.148]

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

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

[59] Borchert M, Dusterhoft A. Emotions in speech-experiments with prosody and quality features in speech for use in categorical and dimensional emotion recognition environments. In: Proc. of the 2005 Int'l Conf. on Natural Language Processing and Knowledge Engineering. Wuhan: IEEE, 2005. 147–151. [doi: 10.1109/NLPKE.2005.1598724]

[60] Eyben F, Wöllme M, Schuller B. openSMILE: The munich versatile and fast open-source audio feature extractor. In: Proc. of the 18th ACM Int'l Conf. on Multimedia. Firenze: ACM, 2010. 1459–1462. [doi: 10.1145/1873951.1874246]

[61] Schuller BW, Steidl S, Batliner A. The INTERSPEECH 2009 emotion challenge. In: Proc. of the 10th Annual Conf. of the Int'l Speech Communication Association. Brighton: ISCA, 2009. 312–315.

[62] Schuller BW, Steidl S, Batliner A, Burkhardt F, Devillers L, Müller CA, Narayanan SS. The INTERSPEECH 2010 paralinguistic challenge. In: Proc. of the 11th Annual Conf. of the Int'l Speech Communication Association. Makuhari: ISCA, 2010. 2794–2797.

[63] Schuller BW, Steidl S, Batliner A, Vinciarelli A, Scherer KR, Ringeval F, Chetouani M, Weninger F, Eyben F, Marchi E, Mortillaro M, Salamin H, Polychroniou A, Valente F, Kim S. The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In: Proc. of the 14th Annual Conf. of the Int'l Speech Communication Association. Lyon: ISCA, 2013. 148–152.

[64] Mowlaee P, Saeidi R, Stylianou Y. Phase importance in speech processing applications. In: Proc. of the 15th Annual Conf. of the Int'l Speech Communication Association. Singapore: ISCA, 2014. 1623–1627.

[65] Yegnanarayana B, Sreekanth J, Rangarajan A. Waveform estimation using group delay processing. IEEE Trans. on Acoustics, Speech, and Signal Processing, 1985, 33(4): 832–836. [doi: 10.1109/TASSP.1985.1164651]

[66] Kua JMK, Epps J, Ambikairajah E, Choi EHC. LS regularization of group delay features for speaker recognition. In: Proc. of the 10th Annual Conf. of the Int'l Speech Communication Association. Brighton: ISCA, 2009. 2887–2890. [doi: 10.21437/Interspeech.2009-46]

[67] Hegde RM, Murthy HA, Rao GVR. Application of the modified group delay function to speaker identification and discrimination. In: Proc. of the 2004 IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing. Montreal: IEEE, 2004. 517–520. [doi: 10.1109/ICASSP.2004.1326036]

[68] Nakagawa S, Asakawa K, Wang LB. Speaker recognition by combining MFCC and phase information. In: Proc. of the 8th Annual Conf. of the 2007 Int'l Speech Communication Association. Antwerp: ISCA, 2007. 2005–2008.

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

[70] Wang LB, Nakagawa S, Zhang ZF, Yoshida Y, Kawakami Y. Spoofing speech detection using modified relative phase information. IEEE Journal of Selected Topics in Signal Processing, 2017, 11(4): 660–670. [doi: 10.1109/JSTSP.2017.2694139]

[71] Gerkmann T, Krawczyk-Becker M, Le Roux J. Phase processing for single-channel speech enhancement: History and recent advances. IEEE Signal Processing Magazine, 2015, 32(2): 55–66. [doi: 10.1109/MSP.2014.2369251]

[72] Guo LL, Wang LB, Dang JW, Zhang LJ, Guan HT, Li XG. Speech emotion recognition by combining amplitude and phase information using convolutional neural network. In: Proc. of the 19th Annual Conf. of the Int'l Speech Communication Association. Hyderabad: ISCA, 2018. 1611–1615. [doi: 10.21437/Interspeech.2018-2156]

[73] Guo LL, Wang LB, Dang JW, Chng ES, Nakagawa S. Learning affective representations based on magnitude and dynamic relative phase information for speech emotion recognition. Speech Communication, 2022, 136: 118–127. [doi: 10.1016/j.specom.2021.11.005]

[74] Ntalampiras S, Fakotakis N. Modeling the temporal evolution of acoustic parameters for speech emotion recognition. IEEE Trans. on Affective Computing, 2012, 3(1): 116–125. [doi: 10.1109/T-AFFC.2011.31]

[75] Neiberg D, Elenius K, Laskowski K. Emotion recognition in spontaneous speech using GMMs. In: Proc. of the 9th Int'l Conf. on Spoken Language Processing. Pittsburgh: ISCA, 2006. 809–812.

[76] Albornoz EM, Milone DH, Rufiner HL. Spoken emotion recognition using hierarchical classifiers. Computer Speech & Language, 2011, 25(3): 556–570. [doi: 10.1016/j.csl.2010.10.001]

[77] Seehapoch T, Wongthanavasu S. Speech emotion recognition using support vector machine. In: Proc. of the 5th Int'l Conf. on Knowledge and Smart Technology (KST). Chonburi: IEEE, 2013. 86–91. [doi: 10.1109/KST.2013.6512793]

[78] Shen PP, Zhou CJ, Chen X. Automatic speech emotion recognition using support vector machine. In: Proc. of the 2011 IEEE Int'l Conf. on Electronic & Mechanical Engineering and Information Technology. Harbin: IEEE, 2011. 621–625. [doi: 10.1109/EMEIT.2011.6023178]

[79] Hinton GE, Osindero S, Teh YW. A fast learning algorithm for deep belief nets. Neural Computation, 2006, 18(7): 1527–1554. [doi: 10.1162/neco.2006.18.7.1527]

[80] Dahl GE, Yu D, Deng L, Acero A. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans. on Audio, Speech, and Language Processing, 2012, 20(1): 30–42. [doi: 10.1109/TASL.2011.2134090]

[81] Han K, Yu D, Tashev I. Speech emotion recognition using deep neural network and extreme learning machine. In: Proc. of the 15th Annual Conf. of the Int'l Speech Communication Association. Singapore: ISCA, 2014. 223–227.

[82] Wang ZQ, Tashev I. Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks. In: Proc. of the 2017 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing. New Orleans: IEEE, 2017. 5150–5154. [doi: 10.1109/ICASSP.2017.7953138]

[83] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735–1780. [doi: 10.1162/neco.1997.9.8.1735]

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

[85] Lee J, Tashev I. High-level feature representation using recurrent neural network for speech emotion recognition. In: Proc. of the 16th Annual Conf. of the Int'l Speech Communication Association. Dresden: ISCA, 2015. 1537–1540.

[86] Thireou T, Reczko M. Bidirectional long short-term memory networks for predicting the subcellular localization of eukaryotic proteins. IEEE/ACM Trans. on Computational Biology and Bioinformatics, 2007, 4(3): 441–446. [doi: 10.1109/tcbb.2007.1015]

[87] Hsiao PW, Chen CP. Effective attention mechanism in dynamic models for speech emotion recognition. In: Proc. of the 2018 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing. Calgary: IEEE, 2018. 2526–2530. [doi: 10.1109/ICASSP.2018.8461431]

[88] LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc. of the IEEE, 1998, 86(11): 2278–2324. [doi: 10.1109/5.726791]

[89] Huang Z, Dong M, Mao Q, Zhan Y. Speech emotion recognition using CNN. In: Proc. of the 22nd ACM Int'l Conf. on Multimedia. Orlando: ACM, 2014. 801–804. [doi: 10.1145/2647868.2654984]

[90] Guo LL, Wang LB, Dang JW, Zhang LJ, Guan HT. A feature fusion method based on extreme learning machine for speech emotion recognition. In: Proc. of the 2018 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing. Calgary: IEEE, 2018. 2666–2670. [doi: 10.1109/ICASSP.2018.8462219]

[91] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: ACM, 2017. 6000–6010.

[92] Li PC, Song Y, McLoughlin I, Guo W, Dai LR. An attention pooling based representation learning method for speech emotion recognition. In: Proc. of the 19th Annual Conf. of the Int'l Speech Communication Association. Hyderabad: ISCA, 2018. 3087–3091. [doi: 10.21437/Interspeech.2018-1242]

[93] Liu JX, Liu ZL, Wang LB, Guo LL, Dang JW. Time-frequency deep representation learning for speech emotion recognition integrating self-attention. In: Proc. of the 26th Int'l Conf. on Neural Information Processing. Sydney: Springer, 2019. 681–689. [doi: 10.1007/978-3-030-36808-1_74]

[94] Zhao SC, Ma YS, Gu Y, Yang JF, Xing TF, Xu PF, Hu RB, Chai H, Keutzer K. An end-to-end visual-audio attention network for emotion recognition in user-generated videos. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. New York: AAAI, 2020.

303–311. [doi: 10.1609/aaai.v34i01.5364]

[95] Guo LL, Ding SF, Wang LB, Dang JW. DSTCNet: Deep spectro-temporal-channel attention network for speech emotion recognition. IEEE Trans. on Neural Networks and Learning Systems, 2023. [doi: 10.1109/TNNLS.2023.3304516]

[96] Deng J, Xu XZ, Zhang ZX, Frühholz S, Schuller B. Exploitation of phase-based features for whispered speech emotion recognition. IEEE Access, 2016, 4: 4299–4309. [doi: 10.1109/ACCESS.2016.2591442]

[97] Guo LL, Wang LB, Dang JW, Liu ZL, Guan HT. Speaker-aware speech emotion recognition by fusing amplitude and phase information. In: Proc. of the 26th Int'l Conf. on Multimedia Modeling, 2020. 14–25. [doi: 10.1007/978-3-030-37731-1_2]

[98] Prabhakar GA, Basel B, Dutta A, Rao CVR. Multichannel CNN-BLSTM architecture for speech emotion recognition system by fusion of magnitude and phase spectral features using DCCA for consumer applications. IEEE Trans. on Consumer Electronics, 2023, 69(2): 226–235. [doi: 10.1109/TCE.2023.3236972]

[99] Dai ZH, Yang ZL, Yang YM, Carbonell J, Le Q, Carbonell R. Transformer-XL: Attentive language models beyond a fixed-length context. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019. 2978–2988. [doi: 10.18653/v1/P19-1285]

[100] Arnab A, Dehghani M, Heigold G, Sun C, Lučić M, Schmid C. ViViT: A video vision Transformer. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Montreal: IEEE, 2021. 6816–6826. [doi: 10.1109/ICCV48922.2021.00676]

[101] Wang XF, Wang M, Qi WB, Su WQ, Wang XQ, Zhou H. A novel end-to-end speech emotion recognition network with stacked Transformer layers. In: Proc. of the 2021 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing. Toronto: IEEE, 2021. 6289–6293. [doi: 10.1109/ICASSP39728.2021.9414314]

[102] Andayani F, Theng LB, Tsun MT, Chua C. Hybrid LSTM-Transformer model for emotion recognition from speech audio files. IEEE Access, 2022, 10: 36018–36027. [doi: 10.1109/ACCESS.2022.3163856]

[103] Gumelar AB, Yuniarno EM, Adi DP, Setiawan R, Sugiarto I, Purnomo MH. Transformer-CNN automatic hyperparameter tuning for speech emotion recognition. In: Proc. of the 2022 IEEE Int'l Conf. on Imaging Systems and Techniques (IST). Kaohsiung: IEEE, 2022. 1–6. [doi: 10.1109/IST55454.2022.9827732]

[104] Huang J, Tao JH, Liu B, Lian Z, Niu MY. Multimodal Transformer fusion for continuous emotion recognition. In: Proc. of the 2020 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing. Barcelona: IEEE, 2020. 3507–3511. [doi: 10.1109/ICASSP40776.2020.9053762]

[105] Lian Z, Liu B, Tao JH. CTNet: Conversational Transformer network for emotion recognition. IEEE/ACM Trans. on Audio, Speech, and Language Processing, 2021, 29: 985–1000. [doi: 10.1109/TASLP.2021.3049898]

[106] Chen WD, Xing XF, Xu XM, Yang JC, Pang JX. Key-sparse Transformer for multimodal speech emotion recognition. In: Proc. of the 2022 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP). Singapore: IEEE, 2022. 6897–6901. [doi: 10.1109/ICASSP43922.2022.9746598]

[107] Tran M, Soleymani M. A pre-trained audio-visual Transformer for emotion recognition. In: Proc. of the 2022 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing. Singapore: IEEE, 2022. 4698–4702. [doi: 10.1109/ICASSP43922.2022.9747278]

[108] Wang YY, Gu Y, Yin YF, Han YP, Zhang H, Wang S, Li CY, Quan D. Multimodal Transformer augmented fusion for speech emotion recognition. Frontiers in Neurorobotics, 2023, 17: 1181598. [doi: 10.3389/fnbot.2023.1181598]

[109] Wagner J, Triantafyllopoulos A, Wierstorf H, Schmitt M, Burkhardt F, Eyben F, Schuller BW. Dawn of the Transformer era in speech emotion recognition: Closing the valence gap. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2023, 45(9): 10745–10759. [doi: 10.1109/TPAMI.2023.3263585]

[110] Liu Z, Kang X, Ren FJ. Dual-TBNet: Improving the robustness of speech features via dual-Transformer-BiLSTM for speech emotion recognition. IEEE/ACM Trans. on Audio, Speech, and Language Processing, 2023, 31: 2193–2203. [doi: 10.1109/TASLP.2023.3282092]

[111] Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. In: Proc. of the 5th Int'l Conf. on Learning Representations. Toulon: ICLR, 2017. 1–14.

[112] Yan SJ, Xiong YJ, Lin DH. Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence. New Orleans: AAAI, 2018. 7444–7452. [doi: 10.1609/aaai.v32i1.12328]

[113] Gao JY, Zhang TZ, Xu CS. Graph convolutional tracking. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019. 4644–4654. [doi: 10.1109/CVPR.2019.00478]

[114] Yao L, Mao CS, Luo Y. Graph convolutional networks for text classification. In: Proc. of the 33rd AAAI Conf. on Artificial Intelligence. Honolulu: AAAI, 2019. 7370–7377. [doi: 10.1609/aaai.v33i01.33017370]

[115] Shirian A, Guha T. Compact graph architecture for speech emotion recognition. In: Proc. of the 2021 IEEE Int'l Conf. on Acoustics,

Speech and Signal Processing (ICASSP). Toronto: IEEE, 2021. 6284–6288. [doi: 10.1109/ICASSP39728.2021.9413876]

[116] Liu JW, Wang HX, Sun MZ, Wei Y. Graph based emotion recognition with attention pooling for variable-length utterances. Neurocomputing, 2022, 496: 46–55. [doi: 10.1016/j.neucom.2022.05.007]

[117] Liu JX, Song YD, Wang LB, Dang JW, Yu RG. Time-frequency representation learning with graph convolutional network for dialogue-level speech emotion recognition. In: Proc. of the 22nd Annual Conf. of the Int'l Speech Communication Association. Brno: ISCA, 2021. 4523–4527. [doi: 10.21437/Interspeech.2021-2067]

[118] Ghosal D, Majumder N, Poria S, Chhaya N, Gelbukh A. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing (EMNLP-IJCNLP). Hong Kong: ACL, 2019. 154–164. [doi: 10.18653/v1/D19-1015]

[119] Fu YH, Okada S, Wang LB, Guo LL, Song YD, Liu JX, Dang JW. Context- and knowledge-aware graph convolutional network for multimodal emotion recognition. IEEE MultiMedia, 2022, 29(3): 91–100. [doi: 10.1109/MMUL.2022.3173430]

[120] Shen WZ, Wu SY, Yang YY, Quan XJ. Directed acyclic graph network for conversational emotion recognition. In: Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int'l Joint Conf. on Natural Language Processing. ACL, 2021. 1551–1560. [doi: 10.18653/v1/2021.acl-long.123]

[121] Chandola D, Altarawneh E, Jenkin M, PapagelisM. SERC-GCN: Speech emotion recognition in conversation using graph convolutional networks. In: Proc. of the 2024 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing. Seoul: IEEE, 2024. 76–80. [doi: 10.1109/ICASSP48485.2024.10446496]

[122] Hou MX, Zhang Z, Cao Q, Zhang D, Lu GM. Multi-view speech emotion recognition via collective relation construction. IEEE/ACM Trans. on Audio, Speech, and Language Processing, 2022, 30: 218–229. [doi: 10.1109/TASLP.2021.3133196]

[123] Liu LY, Liu WZ, Zhou J, Deng HY, Feng L. ATDA: Attentional temporal dynamic activation for speech emotion recognition. Knowledge-based Systems, 2022, 243: 108472. [doi: 10.1016/j.knosys.2022.108472]

[124] Li XK, Zhang ZF, Gan CQ, Xiang Y. Multi-label speech emotion recognition via inter-class difference loss under response residual network. IEEE Trans. on Multimedia, 2023, 25: 3230–3244. [doi: 10.1109/TMM.2022.3157485]

[125] Fan WQ, Xu XM, Cai BL, Xing XF. ISNet: Individual standardization network for speech emotion recognition. IEEE/ACM Trans. on Audio, Speech, and Language Processing, 2022, 30: 1803–1814. [doi: 10.1109/TASLP.2022.3171965]

[126] Lu C, Zong Y, Zheng WM, Li Y, Tang CG, Schuller BW. Domain invariant feature learning for speaker-independent speech emotion recognition. IEEE/ACM Trans. on Audio, Speech, and Language Processing, 2022, 30: 2217–2230. [doi: 10.1109/TASLP.2022.3178232]

[127] Morais E, Hoory R, Zhu WZ, Gat I, Damasceno M, Aronowitz H. Speech emotion recognition using self-supervised features. In: Proc. of the 2022 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP). Singapore: IEEE, 2022. 6922–6926. [doi: 10.1109/ICASSP43922.2022.9747870]

[128] Leem SG, Fulford D, Onnela JP, Gard D, Busso C. Adapting a self-supervised speech representation for noisy speech emotion recognition by using contrastive teacher-student learning. In: Proc. of the 2023 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island: IEEE, 2023. 1–5. [doi: 10.1109/ICASSP49357.2023.10097135]

[129] Li J, Zhang XY, Li FL, Duan SF, Huang LX. Acoustic-articulatory emotion recognition using multiple features and parameter-optimized cascaded deep learning network. Knowledge-based Systems, 2024, 284: 111276. [doi: 10.1016/j.knosys.2023.111276]

[130] Chen ZZ, Li JW, Liu H, Wang XY, Wang H, Zheng QY. Learning multi-scale features for speech emotion recognition with connection attention mechanism. Expert Systems with Applications, 2023, 214: 118943. [doi: 10.1016/j.eswa.2022.118943]

[131] De Lope J, Graña M. An ongoing review of speech emotion recognition. Neurocomputing, 2023, 528: 1–11. [doi: 10.1016/j.neucom.2023.01.002]

[132] Liu ZX, Xu JP, Wu M, Cao WH, Chen LF, Ding XW, Hao M, Xie Q. Review of emotional feature extraction and dimension reduction method for speech emotion recognition. Chinese Journal of Computers, 2018, 41(12): 2833–2851 (in Chinese with English abstract). [doi: 10.11897/SP.J.1016.2018.02833]

[133] Banse R, Scherer KR. Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology, 1996, 70(3): 614–636. [doi: 10.1037/0022-3514.70.3.614]

[134] Mauchand M, Vergis N, Pell MD. Irony, prosody, and social impressions of affective stance. Discourse Processes, 2020, 57(2): 141–157. [doi: 10.1080/0163853X.2019.1581588]

[135] Singh P, Sahidullah M, Saha G. Modulation spectral features for speech emotion recognition using deep neural networks. Speech Communication, 2023, 146: 53–69. [doi: 10.1016/j.specom.2022.11.005]

[136] Zhang Y, Tiňo P, Leonardis A, Tang K. A survey on neural network interpretability. IEEE Trans. on Emerging Topics in Computational

Intelligence, 2021, 5(5): 726–742. [doi: 10.1109/TETCI.2021.3100641]

附中文参考文献:

[1] 郭丽丽. 基于副语言信息和语言信息的情感识别研究 [博士学位论文]. 天津: 天津大学, 2021. [doi: 10.27356/d.cnki.gtjdu.2021.000107]

[11] 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 [doi: 10.13328/j.cnki.jos.004497]

[19] 宋鹏, 郑文明, 赵力. 基于子空间学习和特征选择融合的语音情感识别. 清华大学学报 (自然科学版), 2018, 58(4): 347–351. [doi: 10.16511/j.cnki.qhdxxb.2018.26.014]

[36] 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综述. 软件学报, 2020, 31(8): 2465–2491. http://www.jos.org.cn/1000-9825/6078.htm [doi: 10.13328/j.cnki.jos.006078]

[132] 刘振焘, 徐建平, 吴敏, 曹卫华, 陈略峰, 丁学文, 郝曼, 谢桥. 语音情感特征提取及其降维方法综述. 计算机学报, 2018, 41(12): 2833–2851. [doi: 10.11897/SP.J.1016.2018.02833]

郭丽丽(1990－), 女, 博士, 讲师, CCF 专业会员, 主要研究领域为语音情感识别, 情感计算, 深度学习.

党建武(1956－), 男, 博士, 教授, 博士生导师, 主要研究领域为语音生产, 语音合成, 语音识别, 口语理解.

王龙标(1979－), 男, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为语音信号处理, 自然语言处理, 人工智能.

丁世飞(1963－), 男, 博士, 教授, 博士生导师, CCF 杰出会员, 主要研究领域为智能信息处理, 人工智能与模式识别, 机器学习与数据挖掘, 粗糙集与软计算, 大数据分析与云计算.