计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(11):31−48 [doi: 10.15888/j.cnki.csa.008811]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

跨库语音情感识别研究进展①

张石清[1,2], 刘瑞欣[1], 赵小明[2]

[1](浙江科技学院 理学院, 杭州 310023)
[2](台州学院 智能信息处理研究所, 台州 317000)
通信作者: 张石清, E-mail: tzczsq@163.com

摘　要: 语音情感识别在人机交互过程中发挥极为重要的作用, 近年来备受关注. 目前, 大多数的语音情感识别方法主要在单一情感数据库上进行训练和测试 . 然而, 在实际应用中训练集和测试集可能来自不同的情感数据库. 由于这种不同情感数据库的分布存在巨大差异性, 导致大多数的语音情感识别方法取得的跨库识别性能不尽人意. 为此, 近年来不少研究者开始聚焦跨库语音情感识别方法的研究. 本文系统性综述了近年来跨库语音情感识别方法的研究现状与进展, 尤其对新发展起来的深度学习技术在跨库语音情感识别中的应用进行了重点分析与归纳. 首先, 介绍了语音情感识别中常用的情感数据库, 然后结合深度学习技术, 从监督、无监督和半监督学习角度出发, 总结和比较了现有基于手工特征和深度特征的跨库语音情感识别方法的研究进展情况, 最后对当前跨库语音情感识别领域存在的挑战和机遇进行了讨论与展望.

关键词: 语音情感识别; 跨库; 深度学习; 手工特征; 深度特征; 语音情感

引用格式: 张石清,刘瑞欣,赵小明.跨库语音情感识别研究进展.计算机系统应用,2022,31(11):31–48. http://www.c-s-a.org.cn/1003-3254/8811.html

Research Advance of Cross-corpus Speech Emotion Recognition

ZHANG Shi-Qing[1,2], LIU Rui-Xin[1], ZHAO Xiao-Ming[2]

[1](School of Science, Zhejiang University of Science and Technology, Hangzhou 310023, China)
[2](Institute of Intelligent Information Processing, Taizhou University, Taizhou 317000, China)

Abstract: Speech emotion recognition (SER) plays an extremely important role in the process of human-computer interaction (HCI), which has attracted much attention in recent years. At present, most SER approaches are mainly trained and tested on a single emotion corpus. In practical applications, however, the training set and testing set may come from different emotion corpora. Due to the huge difference in the distribution of different emotion corpora, the cross-corpus recognition performance achieved by most SER methods is unsatisfactory. To address this issue, many researchers have started focusing on the studies of cross-corpus SER methods in recent years. This study systematically reviews the research status and progress of cross-corpus SER methods in recent years. In particular, the application of the newly developed deep learning techniques on cross-corpus SER tasks is analyzed and summarized. Firstly, the emotion corpora commonly used in SER are introduced. Then, on the basis of deep learning techniques, the research progress of existing cross-corpus SER methods based on hand-designed features and deep features is summarized and compared from the perspectives of supervised, unsupervised, and semi-supervised learning. Finally, the challenges and opportunities in the field of cross-corpus SER are discussed and predicted.

Key words: speech emotion recognition; cross-corpus; deep learning; hand-designed features; deep features; speech emotion

① 基金项目: 国家自然科学基金 (61976149); 浙江省自然科学基金 (LZ20F020002)
收稿时间: 2022-03-05; 修改时间: 2022-04-02; 采用时间: 2022-04-22; csa 在线出版时间: 2022-07-14

情感识别是心理学、生物学、计算机科学的一个重要研究方向. 由于情感识别在人机交互 (human-computer interaction, HCI) 中的关键作用, 近年来受到了工程研究领域的广泛关注[1–3].

语音作为最重要的交流方式之一, 一直是人机交互领域关注的焦点, 为了让计算机更好的理解人的情感, 对语音中的情感信息进行分析十分关键[4–8]. 语音情感识别是通过对语音信号进行特征提取来分析语音与情感之间的规律, 进而判断语音情感状态的过程[9].

语音情感识别的应用十分广泛. 例如, 对驾驶员驾驶车辆时的状态进行检测可以有效减少交通事故[10,11]; 在一些长时间任务中, 及时监控工作人员是否产生负面情绪, 可以预防发生意外[4]; 语音情感识别程序还可用于辅助残疾人讲话[12]. 除此之外, 语音情感识别在其他方面也有良好的应用, 如病人异常情绪检测[13]、抑郁症诊断[14]、自闭症谱系障碍 (autism spectrum disorder, ASD) 检测[15] 等.

语音情感识别的主要步骤包括语音情感特征提取和情感分类器设计. 近年来, 研究人员对语音情感特征提取和情感分类器设计方面的研究做出了巨大的努力[16].

语音情感特征主要有 3 类: 韵律特征、音质特征和谱特征[17,18]. 韵律特征包括基音频率[19]、短时能量等, 一般通过与韵母、语调相关的韵律来表达; 音质特征包括共振峰[20]、谐波噪声比 (harmonics-to-noise ratio, HNR) 等与发声声道的物理性质相关的声学特征; 谱特征包括梅尔频率倒谱系数 (Mel frequency cepstral coefficient, MFCC)[21] 等以及一些高级谱特征, 如类级 (class-level) 谱特征[22]、语谱图等. 其中, 共振峰和 MFCC 是语音情感识别中两种常用的特征, 共振峰是音质的决定因素, 可以反映声道的物理特征; MFCC 可以在很大程度上模拟人的听觉感知系统, 从而提高语音情感识别的性能[23].

语音情感识别的最终目的是对情感进行分类, 大量的机器学习方法被应用于语音情感识别领域, 如 K 近邻算法[24], BP 神经网络[25], 支持向量机 (support vector machines, SVM)[26]、隐马尔可夫模型 (hidden Markov model, HMM)[27] 等. 由于现有的机器学习方法大都可以用于语音情感识别, 因此本文对于语音情感分类器的设计不再详细介绍.

目前对语音情感识别的研究[5,28,29] 大多集中在单一语料库或语言的情况下进行, 即训练数据和测试数据具有相同的数据分布特点, 并获得了较好的性能. 但在实际生活中, 训练数据和测试数据往往来自不同的语料库, 这些不同的语料库之间的特征分布存在较大的差异, 从而使得现有的语音情感识别模型在跨语料库的情况下性能大幅下降. 针对此问题的研究出现了一个新的热点研究课题——跨库语音情感识别.

类似于普通的语音情感识别, 跨库语音情感识别一般也包括两个步骤: 跨库语音情感特征提取和跨库情感分类器的设计, 即对源数据库和目标数据库的语音样本提取域不变特征, 然后训练语音情感分类器.

传统的域不变特征提取方法侧重于手工特征的提取, 然后试图将源数据库和目标数据库映射到同一个特征空间, 并采用不同的方法最小化特征空间中两者的分布差异, 从而使在源数据库上训练得到的模型在目标数据库上有较好的效果. 但是, 手工提取的特征往往是低层次的, 与高层次的人类“情感”表达存在较大的“情感鸿沟”问题. 因此, 研究高层次的自动化特征提取方法用于跨库语音情感识别, 近年来备受关注.

近年来, 深度学习技术的不断发展为跨库语音情感识别带来了新的方向, 即利用具有多层结构的深度神经网络模型可以从大量的数据中自动学习出高层次的特征用于跨库语音情感识别. 因此, 很多学者开始将深度学习方法应用于跨库语音情感识别, 其中最具代表性的有卷积神经网络 (convolutional neural networks, CNNs)[30]、递归神经网络 (recurrent neural networks, RNNs)[31] 以及长短期记忆网络 (long short-term memory, LSTM)[32]、深度信念网络 (deep belief networks, DBNs)[9,33] 等.

考虑到现有跨库语音情感识别研究方面的综述较少, 为此本文结合新发展起来的深度学习技术, 从情感语音数据库、跨库语音情感特征提取等方面综述当前跨库语音情感识别关键技术和前沿进展, 并总结跨库语音情感识别任务面临的困难与挑战, 指出未来发展方向.

1 语音情感识别数据库

研究语音情感识别首先需要相应情感数据库的支撑. 目前情感语音数据库的数量较多, 附录中表 A1 给出了这些语料库的简要概述, 语料库的详细内容如下.

DES[34]: 该数据集包含 400 多个音频话语, 由 4 名专业演员 (2 名男性和 2 名女性) 模拟. 模拟的语音包括 5 种情绪状态: 愤怒、快乐、中性、悲伤和惊讶. 每个演员的录音由两个孤立的单词、9 句和两段流畅的语音材料组成. 整个音频话语持续时间约 30 min. 为了进行听力测试, 雇用了 20 名听众.

SUSAS[35]: 该数据集是一个包含 4 种压力和感受的压力数据集下的语音. 它包含了高度混乱的 35 个飞机通信词汇的集合. 研究人员邀请 32 名发言者 (13 名女性, 19 名男性) 发表 16 000 多篇演讲. 压力下的模拟语音由说话风格、单一跟踪任务、Lombard 效应域等 10 种压力类型组成. 音频采样率为 8 kHz.

EMO-DB[36]: 该数据集是由 10 名演员模拟 7 种情绪产生的 10 个德语语句, 7 种情绪分别是中性、愤怒、恐惧、喜悦、悲伤、厌恶和厌倦, 数据库共 536 个样本, 总时长约 0.38 小时. 录音在柏林技术大学技术声学系的消声室进行, 以 48 kHz 的采样频率进行记录, 然后下采样到 16 kHz. 每个采样用 16 位数字表示. 该数据库已经成为许多研究的基础.

eNTERFACE[37]: 该数据集由愤怒、厌恶、恐惧、快乐、悲伤、惊讶和中性 6 种情绪组成, 约 1 277 个样本, 总时长约 1 小时. 所有的实验都是用英语进行的. 录音共 42 名受试者, 每个受试者被要求连续听 6 个短篇故事, 每个故事都引发一种特殊的情绪, 他们对每种情况做出反应, 由两位专家判断反应是否表达了这种情绪. 收集过程使用数码摄像机和高质量麦克风记录, 音频采样率为 48 kHz, 每个采样用 16 位数字表示, 视频序列使用 720×576 Microsoft AVI 格式进行处理.

MASC[38]: 该数据集包含 68 名汉语使用者表达的中性、愤怒、兴奋、恐慌和悲伤 5 种情绪的 25 636 条录音, 其中每人说出 5 个短语, 10 个句子. 录音在一个安静无干扰的办公室里, 所有的数据都记录在奥林巴斯 DM-20 数字录音机上, 采样频率为 22 050 Hz. 随后, 录音文件通过 USB 传输到个人电脑上. 获得的录音被转换成单声道 Windows PCM 格式, 采样频率为 8 kHz, 分辨率为 16 位. 该数据库用于汉语情感表达的韵律和语言学研究以及为受情绪因素影响的说话人识别系统提供了训练集和测试数据集.

ABC[39]: 该数据集是一个面向公共交通特定应用的视听情感数据库. 为了引发一定的情感, 利用剧本使主体进入导演故事情节的语境. 选取的公共交通包含节假日航班和与错食、乱流、入睡、与邻居交谈等相关的回程航班. 邀请 8 名年龄在 25–48 岁之间性别均衡的参与者参加德语录音. 经过 3 位经验丰富的男性注释者预分割后, 共采集了 11.5 小时的视频, 共 431 个片段. 所有 431 个视频剪辑的平均持续时间为 8.4 s.

IEMOCAP[40]: 该数据集记录了 10 名演员的 5 种情绪, 分别为快乐、愤怒、悲伤、沮丧和中性, 共包含约 5 531 条语音样本, 样本均为英语, 时间约 12 小时. 标记数据使用带有 8 个摄像机的 VICON 运动捕捉系统和高质量麦克风. 受试者被要求在录制过程中就座, 尽可能自然地做手势, 同时避免用手遮住脸. 音频的采样率设置为 48 kHz, 每个采样用 16 位数字表示. 该数据库是现有数据库的宝贵补充, 用于多模态和表达性人类交流的研究和建模.

FAU Aibo[41]: 该数据集录制了 51 名儿童与机器人 Aibo 游戏过程中的自然语音, 总时长为 9.2 小时, 约 17 074 条德语语音样本. 语音通过无线耳麦传输, 并由 DAT-recorder 录制, 信号的采样率为 48 kHz, 而后压缩到 16 kHz, 每个采样用 16 位数字表示. 5 名语言学的学生按顺序听语音文件, 并相互独立地标注每个单词, 通过投票方式决定最终标注结果, 数据集共包括喜悦、恼怒、愤怒、中性等 11 个情感标签. 该数据库中的 18 216 个单词被选定为 INTERSPEECH 2009 年情感识别竞赛用数据库.

CASIA[42]: 该数据集由中国科学院自动化研究所开发, 共包含 9 600 个音频文件. 这个数据包含 6 个情绪状态: 快乐、悲伤、愤怒、惊讶、恐惧和中性. 4 个专业演员 (两名女性和两名男性) 模拟了这些情绪, 在 6 个不同的情绪类别中呈现了 400 个话语.

SIMIS[43]: 该数据集包括 35 个手术的现场记录, 其中 29 个记录被一个男性以 5 种情感状态进行了文本转录和注释, 其余的 6 个记录由一个女性在同一组情感中转录和注释. 记录在德国慕尼黑工业大学的临床研究中心进行, 使用无线耳机进行录音, 信号的采样率为 16 kHz, 每个采样用 16 位数字表示. 为了消除背景噪音, 对录音进行了自动分割和去噪, 共获得了 11 077 个语音, 这些语音被手动划分为愤怒、困惑、快乐、不耐烦和中性 5 种情绪之一. 迄今为止该数据库主要用于改进用于辅助手术机器人控制的自动语音识别.

IITKGP-SEHSC[44]: 该数据集由 Gyanavani FM radio station, Varanasi, India 的 10 名专业艺术家录制,

包含愤怒、厌恶、恐惧、高兴、中立、悲伤、讽刺和惊讶 8 种情绪. 为了记录情绪, 考虑了 15 个情感中性的印地语文本提示. 每个艺术家必须在一个环节中用 8 种基本情绪说出这 15 个句子, 共 15 段会话. 数据库中的话语总数为 12 000 个. 数据库的总持续时间约为 9 小时. 语音使用高级麦克风录制, 信号以 16 kHz 采样, 每个采样用 16 位数字表示. 该数据库在情感、说话者和文本方面具有广泛的特点.

CVE[45]: 该数据集由 4 个以普通话为母语的人制作了一组汉语伪句 (即与真实汉语相似的语义无意义的句子), 以表达愤怒、厌恶、恐惧、悲伤、快乐、惊喜和中立 7 种情绪, 之后被一组以母语为普通话的听众在一个 7 个选项强迫选择任务中识别出来. 录音使用数字记录仪和高质量麦克风记录, 并使用 Praat 语音分析软件为每一个话语编辑成单独的 wav 声音文件. 该数据库有助于关于普通话声音情绪的行为、神经心理学和神经影像学研究以及声音情绪交流的跨文化/跨语言研究.

SAVEE[46]: 该数据集是一个使用英国英语的多模态情感数据集. 它总共包含了 480 条语音以及 7 种不同的情感: 中性、快乐、悲伤、愤怒、惊讶、恐惧和厌恶. 这些话语由 4 个专业的男性演员产生. 为了保持情感表演的良好质量, 本数据集的所有录音均由 10 位不同的评价者在音频、视觉和视听条件下进行验证. 这些录音中的脚本选自常规 TIMIT 语料库[47].

CREMA-D[48]: 该数据集由在高兴、悲伤、愤怒、恐惧、厌恶和中性 6 种基本情绪状态下说出的 12 个句子中的面部和声音情绪表达组成. 来自不同种族背景的 91 名演员共生成了 7 442 个片段, 由多名评分员以 3 种方式进行评分: 音频、视频和视听. 录制时段通常持续大约 3 个小时, 录制过程在专业灯箱的声音衰减环境中进行, 信号以 48 kHz 采样. 在考虑情绪的视听感知问题时该数据集是一个极好的资源.

EMOVO[49]: 该数据集包含 6 名意大利演员表演 14 个情绪中性的短句模拟的厌恶、恐惧、愤怒、高兴、惊讶、悲伤和中性 7 种情绪状态, 共 588 条录音, 由两组不同的 24 个注释者进行注释. 录音是用 Ugo Bordoni 基金会实验室的专业设备录制的, 录用的采样频率为 48 kHz, 16 位立体声, 保存为 wav 格式. 整个数据库约为 1 小时. EMOVO 为科学界提供了第一个意大利语情感声音的语料库.

MSP-IMPROV [50]: 该数据集由 12 名英语专业的学生进行 6 次对话, 共约 8 438 条语音样本, 总时间超过 9 小时, 表达了高兴、悲伤、愤怒、中性 4 种情绪. 收集者为每个句子定义假设的场景, 两个演员在每个场景即兴发挥说出语境化的句子, 使用 crowdsourcing 的方式对情感内容进行标注. 录音在单壁音箱中收集, 音频由两个项圈麦克风记录, 信号以 48 kHz 和 32 位 PCM 采样, 视频以数码相机记录, 分辨率为 29.97 帧/s. 该语料库是情感识别研究的宝贵资源.

CIT[51]: 该数据集是一个开放的、可用的电子游戏语料库, 它包括基于既定戏剧表演技巧的二人即兴表演, 称为 active analysis, 以帮助激发自然的情感互动. 该数据库由 16 名演员 (8 名女性和 8 名男性) 组成, 50 次面对面互动约 3 min. 每个演员都从麦克风中录制. 每个会话的标记至少有 3 个 [1, 5] 之间的情感属性 (会话级) 的激活、支配和效价.

BAUM-1[52]: 该数据集是一个包含 8 种情绪 (喜悦、愤怒、悲伤、厌恶、恐惧、惊讶、厌烦和蔑视) 和 4 种精神状态 (不确定、思考、集中和烦恼) 的自发情绪数据集. 该数据集由来自 31 名土耳其被试者 (17 名女性, 14 名男性) 的 1 222 个听视觉样本组成. 整个样本的平均持续时间约为 3 s. 采用多数票表决的方式, 邀请 5 位注释者对每个样本进行标注. 录制的音频文件采样率为 48 kHz.

NTUA[53]: 该数据集是新近收集到的汉语情感语料库. 22 对演员参与面对面互动约 3 min, 表演专业导演预设的情感场景. 对于每一个场景, 他们都会自发地产生一种情绪, 即快乐、悲伤、神经、愤怒、惊讶和挫折. 每个会话包括每个演员从 lapel 麦克风中收集的录音. 在每个会话中, 42 名评分员根据 [1, 5] 之间的比例标记会话级激活和效价的情绪属性.

MSP-Podcast[54]: 该数据集收集了从网站下载的“创意共享空间”授权录音中多个发言者的自然发言, 音频分为 2.75–11 s 的发言轮流. 然后, 利用现有数据库训练的情感模型, 检索出包含目标情感内容的语音转折. 每个言语句子至少由 5 名评价者进行注释, 评价者对情感唤醒、效价和优势度进行评分, 分值在–1~1 之间.

EmoFilm[55]: 该数据集从 43 部电影中的 207 名说话者中提取的 1 115 个平均长度为 3.5 s 的英语、西班牙语和意大利语的情感话语, 包含了愤怒、悲伤、快

乐、恐惧和轻蔑 5 种情绪. 10 名意大利听众对整个数据库进行了评估, 然后删除所有评分协议低于 6 的剪辑. 随后以频率为 44.1 kHz 提取音频, 每个采样用 16 位数字表示.

MELD[56]: 该数据集包含了电视剧《老友记》中 1 433 段对话中的 13 707 个句子, 总时长约 12.1 小时, 包含音频、视觉和文本形式. 使用 3 个注释器对每个话语进行标注, 然后通过多数投票决定话语的最终标注. 音频文件格式化为 16 位 PCM WAV 文件以进行进一步处理. 数据集共包含愤怒、厌恶、悲伤、喜悦、中立、惊讶和恐惧 7 种情绪并且对每个话语都给出正面、负面和中性注释. 该数据集是可用于多模态情感对话系统, 并且是现有多模态会话数据集的两倍.

RAVDESS[57]: 该数据集是情感语音和歌曲的多模态语料. 该数据集是性别均衡的, 由 24 名专门演员组成, 他们以中性的北美发音产生语音和歌曲样本. 对于情感性言语, 它由平静、欢乐、悲伤、愤怒、恐惧、惊讶、厌恶构成. 对于情感性的歌曲, 它由平静、欢乐、悲伤、愤怒、恐惧、惊讶、厌恶和恐惧组成. 每个表情都是在情感强度的两个层次上产生的, 带有一个附加的中性表情. 最后收集的 7 356 份录音在情感有效性、强度和真实性方面分别被评为 10 次. 对于这些收视率, 雇佣了来自北美的 247 名未经培训的研究对象.

DEMoS[58]: 该数据集包含了 68 名母语为意大利语的受试者在愤怒、悲伤、快乐、恐惧、惊讶、厌恶, 以及次级情绪内疚 7 种情绪状态下产生的 9 697 个语音样本. 数据收集使用了两个工作站 (一个用于录音, 另一个用于诱导程序)、一个高心近距离话筒、耳机和一个专业声卡, 以 pcm 波单声道格式和 48 kHz 的采样率以及 16 位进行记录, 随后, 人工将情感语音分割为样本. 该数据集包含了对其他情感 (如内疚) 的考虑, 目前在大多数可用的情感语言语料库中未得到充分体现.

1.1 小结

依据情感描述不同, 现有情感数据集可分为离散型和维度型两种. 其中, 离散型情感数据集主要包含快乐、愤怒、悲伤、中性等离散情感标签. 维度型情感数据集大多是由注释者根据唤醒、效价和支配三维情感进行打分. 此外, 为更准确的描述情感, 少量的情感数据集, 如 FAU Aibo, 同时对离散型情感和维度型情感进行标注.

根据数据集的激发方式不同, 现有情感数据集又可分为表演型、诱发型以及自然型 3 种. 其中, 表演型数据集由专业演员根据给出台词说出符合规定情感的语音并进行采集, 如 DES、SAVEE. 诱发型数据集是利用剧本引导受试者进入情境并表达相应的情感, 如 ABC. 而自然型数据集多为自然环境下进行交流并收集的语音样本, 这种数据集因为环境的噪音而难以采集并注释, 是目前数据集制作的重点和难点.

由于人类表达情感的方式不仅是语音, 还有面部表情、文本和肢体动作等, 因此除了只包含语音的单模态数据集之外, 还有包含语音、文本或视频的多模态数据集, 如 IEMOCAP、MSP-Imporv 等. 此外, 部分数据集不仅记录了语音情感标签, 而且保留了说话人和性别相关信息. 这便于研究者利用这些信息进行与说话人独立、与性别独立的语音情感识别以及说话人识别、性别识别或将两者作为语音情感识别的辅助任务进行多任务学习.

2 跨库语音情感特征提取

对域不变语音情感特征的提取在跨库语音情感识别中极为重要, 如何提取有效的域不变特征会直接影响跨库语音情感识别的性能, 因此本节对近年来的手工域不变特征提取方法和面向深度学习方法的深度域不变特征提取方法分别进行详细分析与归纳. 根据使用的样本类别信息, 又可以分为监督、半监督、无监督 3 种类型. 监督学习是用标记的训练数据来训练网络, 使其可以根据输入得到相应的输出[59]. 与监督学习不同的是, 无监督学习是用完全无标记的训练数据训练网络[60]. 半监督学习介于两者之间, 其训练数据包含标记数据和未标记数据, 通常假设所有训练数据来自相同或相似分布[61,62]. 下面结合监督、半监督和无监督思想来阐述手工域不变特征提取方法和深度域不变特征提取方法的各自研究进展情况.

2.1 手工域不变特征提取方法

语音的手工特征指手工提取的低层描述 (low-level description, LLD) 特征, 主要有基因频率、共振峰、MFCC 等. 结合不同的 LLD 特征及其统计量当做域不变特征是跨库语音情感特征提取的重要方向. 目前典型的声学特征集包括 INTERSPEECH 2009 情感挑战[63]、INTERSPEECH 2010 副语言挑战[64] 和

INTERSPEECH 2013 计算副语言特征[65] 以及日内瓦最小声学参数集 (Geneva minimum acoustic parameter set, GeMAPS)[66].

2.1.1 面向监督的手工域不变特征提取方法

Kaya 等[67] 提出了融合线性说话人水平、非线性值水平和特征向量水平的归一化方法, 使用 z-归一化用作简单的跨语料库适应策略, 从而最小化不同语料库之间的差异, 同时利用线性核分类器来提高类分离性. 提取 INTERSPEECH 2013 特征集在 EMO-DB, DES, eNTERFACE 等 5 个数据集上使用极限学习机 (extreme learning machine, ELM) 进行情感识别.

Zhang 等[68] 提出了联合迁移子空间学习与回归 (joint transfer subspace learning and regression, JTSLR) 的迁移方法, 通过最大平均差异 (maximum mean discrepancy, MMD) 作为偏差度量测量语料库之间特征分布差异, 同时使用真伪标签构建基于监督图的差异度量, 可以在测量语料库之间特征分布差异的同时保持标签上的局部结构一致性. 提取 INTERSPEECH 2010 的 1 582 个声学特征, 在 EMO-DB, eNTERFAC 和 BAUM-1 数据库上用 SVM 对语音情感进行分类.

Zhang 等[69] 提出了联合分布自适应回归 (joint distribution adaptive regression, JDAR) 方法, 联合考虑源语料库和目标语料库之间的边际概率分布和条件概率分布来学习回归矩阵, 减轻他们之间的特征分布差异. 该方法提取了 IS09 和 IS10 特征集, 在 EMO-DB、eNTERFACE 和 CASIA 数据集上进行实验, 并与如 DaLSR、DoSL 等方法进行比较, 取得了更好的性能.

针对语料库之间的固有差异, Kaya 等[67] 使用了简单的归一化方法进行差异消除. 但简单的归一化方法只能在一定程度上缓解语料库差异, 因此 Zhang 等[68] 提出了用 MMD 作为偏差度量测量语料库之间的分布差异; Zhang 等[69] 联合考虑源语料库和目标语料库之间的边际概率分布和条件概率分布来减轻他们之间的特征分布差异. 以上监督学习是在已知目标域标签的基础上学习语料库的差异, 但现有的有标签数据集有限, 因此仅使用监督学习进行跨库语音情感识别是不够的.

2.1.2 面向无监督的手工域不变特征提取方法

Zong 等[70] 提出了基于域自适应最小二乘回归 (domain adaptive least squares regression, DaLSR) 模型的跨库语音情感识别方法. 先提取包括韵律特征和谱特征等 384 个特征, 然后从目标语料库中选择一组未标记样本, 与源语料库中标记样本共同训练最小二乘回归 (least squares regression, LSR) 模型. 同时, 在 LSR 中引入正则化约束来缓解两个语料库的分布差异. 使用 EMO-DB, eNTERFACE 等语料库上进行了实验, 最后通过线性 SVM 对语音情感进行分类.

Song 等[71] 提取包含韵律特征、音质特征和谱特征等 1 582 个声学特征, 然后利用非负矩阵分解 (non-negative matrix factorization, NMF) 方法获得源语料库和目标语料库的低维情感特征. 同时采用 MMD 进行相似度度量测量两个语料库之间的特征分布差异, 最后提出联合优化 NMF 和 MMD 的转移非负矩阵分解 (transfer non-negative matrix factorization, TNMF) 方法以最小化特征分布差异, 在 FAU Aibo, eNTERFACE 和 EMO-DB 三个数据集上进行实验, 获得了优于线性 SVM 的性能.

Mao 等[72] 提出了情感差异性和域不变特征学习 (emotion-discriminative and domain-invariant feature learning method, EDFLM) 方法. 他们提取 INTERSPEECH-2009 特征集作为输入, 通过情感辨别器将输入分为情感相关和情感无关, 随后将情感相关特征输入域鉴别器, 同时引入梯度反转层混淆域鉴别器, 从而得到领域不变特征. 最后在 FAU Aibo, ABC 和 EMO-DB 数据集上使用 SVM 对语音情感进行分类.

Liu 等[73] 提取了与文献 [70] 相同的特征, 结合域自适应子空间学习 (domain-adaptive subspace learning, DoSL) 方法来学习投影矩阵, 利用 MMD 准则测量平均投影源语音特征向量和平均投影目标语音特征向量之间的距离差, 随后使用文献 [71] 中的联合优化方法使源语音信号和目标语音信号在标签空间中的特征分布相似, 通过线性 SVM 对目标语音信号的情绪状态进行准确预测, 获得了比最新的跨库语音情感识别方法更有前景的结果.

Liu 等[74] 提取 INTERSPEECH 2009 特征集, 通过转移子空间学习方法学习投影矩阵, 将源和目标语音信号从原始特征空间转换到标签空间, 在此空间中利用 MMD 准测度量两个语料库之间差异并通过 TNMF 方法最小化语料库差异, 最后, 在标记的源语音信号上训练的分类器可以有效地预测未标记目标语音信号的情感状态.

针对无标签数据集, 最具挑战的方法是使用无监督学习方法在不访问目标域数据集标签的前提下提高跨库 SER 的性能. Zong 等[70] 在 LSR 中引入正则化约束来缓解两个语料库的分布差异. Song 等[71] 使用 MMD 进行语料库差异的测量. 而语料库之间存在差异的根本原因是两个语料库的特征分布不同, 因此学习不同语料库的特征分布并找到不同语料库相似的部分, 即域不变特征, 是非常重要的. 为此, Mao 等[72] 将情感相关特征输入域鉴别器, 从而得到领域不变特征. 进一步将源域和目标域数据集的语音样本投影到同一子空间中, 然后使用 MMD 可以更加准确有效的测量数据集间的差异. 因此, Liu 等[73] 结合 DoSL 方法来学习投影矩阵, Liu 等[74] 通过转移子空间学习方法学习投影矩阵, 并将语音信号转换至标签空间.

无监督学习在跨库语音情感识别任务上被证明是有效的, 但由于无监督学习不访问任何目标域数据集标签, 因此对跨库语音情感识别性能的提升有限. 若在模型训练时可以访问部分目标域标签, 这将使模型更容易对目标域进行情感分类.

2.1.3 面向半监督的手工域不变特征提取方法

金赟等[75] 采用半监督判别分析方法来减小不同语料库之间的差异. 先对每条语句提取 988 个特征, 包括韵律特征 (基频和基音频率包络)、谱特征 (MFCC) 等 26 个 LLDs 及其相应的一阶差分, 然后结合线性判别分析的思想, 使类间散度矩阵和类内散度矩阵的比值达到最大, 寻找有标签样本与目标语料库部分无标签样本之间的最优投影方向得到投影向量, 将训练样本向此投影方向投影得到情感分类面, 随后将测试样本向投影向量方向投影, 由训练样本得到的分类面同样适用于测试样本, 从而提高跨库识别率. 在 EMO-DB 和 eNTERFACE 数据库上进行实验, 最后使用 SVM 作为情感分类器实现跨库语音情感识别.

宋鹏等[76] 提出一种基于特征迁移学习的跨库语音情感识别方法. 利用 openSMILE 工具对每个语音样本提取出 INTERSPEECH 2010 竞赛中使用的特征集, 共 1 582 维特征, 引入映射函数将源语料库的有标签样本和目标语料库的无标签样本映射到低维空间, 然后用 MMD 判断低维空间中不同数据库情感特征之间的相似度并通过半定规划进行优化从而得到域不变特征. 为了更好地保证情感信息的类别区分度, 进一步引入半监督判别分析方法用于特征降维. 最后, 采用传统 SVM 方法作为情感分类器, 在 EMO-DB 和 eNTERFACE 数据库上进行了跨库语音情感识别实验.

Luo 等[77] 提出了半监督自适应正则化转移非负矩阵分解 (semi-supervised adaptive regularization transfer non-negative matrix factorization, SATNMF) 方法, 将训练数据的标签信息与 NMF 相结合, 寻找一个潜在的低秩维特征空间, 在这个特征空间中, 利用 MMD 测量两个语料库之间的差异以及两个语料库中每个类的差异, 同时最小化这两种差异使两个语料库的边际和条件分布相似. 他们提取了与文献 [67] 相同的特征, 在 CASIA, EMO-DB 和 eNTERFACE 等数据集上使用线性 SVM 对情感进行识别.

Luo 等[78] 提出了基于非负矩阵分解的迁移子空间学习 (nonnegative matrix factorization based transfer subspace learning, NMFTSL) 方法, 为源语料库和目标语料库找到一个共享特征子空间, 在子空间中, 使边际分布之间的距离以及条件分布之间的距离最小, 从而消除两个分布之间的差异. 该方法提取包含韵律特征、音质特征和谱特征等 1 582 个声学特征, 在 CASIA、SAVEE、IEMOCAP 等 6 个数据集上使用线性 SVM 对情感进行分类.

使用半监督的方法学习源域数据集和部分目标域数据集的公共信息是跨库语音情感识别的另一种方法. 在跨库的条件下, 影响语音情感识别率的因素不仅有域间差异, 还有源域和目标域的类区分度. 因此, 在类区分度上, 金赟等[75] 结合线性判别分析的思想来提升源域和目标域的类区分度. 而同时考虑类区分度和领域对齐是解决跨库语音情感识别问题的一个有效方法. 为此, 宋鹏等[76] 用 MMD 判断低维空间中不同数据库情感特征之间的相似度, 并引入半监督判别分析方法保证情感信息的类别区分度. 为了估计目标语料库的条件分布, Luo 等[78] 将目标标签的预测和特征表示的学习整合到一个联合学习模型中, 最后提出了一种区分损失来将标签引入共享子空间, 从而提高特征表示的区分能力.

2.1.4 小结

手工域不变语音情感特征提取方法的总结与比较见附录表 A2. 在结合手工特征的域不变语音情感特征提取方法中, 基础的方法是利用语料库归一化方法, 将不同的语料库按照同样的方法进行归一化从而降低两个语料库之间的差异. 这种方法虽然在一定程度上缓

解了语料库之间的差异, 但不同语料库之间的特征分布依然大不相同. 之后出现了不同的测量语料库数据分布差异的方法, 如 MMD 准则. 因此许多研究者使用不同的方法将源语料库和目标语料库的样本映射到同一特征空间并在这个特征空间中利用 MMD 准则测量两个数据库之间的特征分布差异, 最后最小化这个差异以得到域不变特征. 受 MMD 准则的启发, 一些方法联合考虑边际概率分布和条件分布, 使边际分布之间的距离以及条件分布之间的距离最小, 从而消除两个数据库特征分布之间的差异. 除此之外, 一些方法引入正则化约束来缓解两个语料库的分布差异或者利用梯度反转层形成对抗训练得到域不变特征.

手工语音情感特征多为简单的浅层特征, 只能包含语音信号的部分信息. 探索使用语音信号中更深层次的特征进行跨库语音情感识别是进一步提升模型对目标域情感识别率是一个重要研究方向.

2.2 深度域不变特征提取方法

由于手工提取特征耗资耗时且不能完整表示语音信号的特征, 因此研究者尝试将深度学习技术应用于跨库语音情感识别, 常用方法有 CNNs、RNNs、LSTM 以及 DBNs 等.

CNNs 首先由 LeCun 等[79] 在 1989 年首次提出, 用于手写邮编识别. CNNs 的一个重要特性是可以自动从输入数据中学习特征, 并具有一定的泛化能力, 因此关于解决此问题的应用均可以使用 CNNs. 一个基础的 CNNs 包括卷积层、激活层和池化层, 有些网络还包含全连接层. 在跨库语音情感识别中主要用于提取语音片段的局部特征.

为了更好地捕捉深层连接, 很多学者开始将 RNNs 应用于语音情感识别中. 一个简单的 RNNs 由输入层、隐藏层和输出层组成, 核心为一个全连接的循环单元, 采用递归连接来捕获数据的历史信息 , 使其可以在网络内部循环.

然而 RNNs 的一个重要缺陷是容易出现梯度消失和爆炸问题, 1997 年 Hochreiter 等[32] 为了缓解这个问题提出了 LSTM. LSTM 是 RNNs 的扩展, 在其内部加入了遗忘门, 输入门, 候选细胞单元和输出门, 通过门控状态来控制传输状态, 记住需要长时间记忆的, 忘记不重要的信息.

DBNs 是神经网络的一种, 由一系列叠加受限玻尔兹曼机 (restricted Boltzmann machine, RBM) 组成, RBM 由用于输入训练数据的显层和用作特征检测器的隐层构成. DBNs 中的相邻层可以看作是多个独立的 RBM 的堆积, 上一个 RBM 的隐层的输出即为下一个 RBM 的输入. DBNs 既可以作为自编码器, 也可以作为分类器.

2.2.1 面向监督的深度域不变特征提取方法

Zhang 等[80] 提出一种基于时频原子的听觉注意特征提取模型, 该模型模拟人类听觉系统, 首先计算语音声谱图, 之后从声谱图中提取多尺度情感对比特征, 然后对多尺度情感特征进行有效探测, 之后引入 Chirplet 时频原子, 通过形成的过完备原子库提高语谱图特征的信息量. 在 eNTERFACE、 EMO-DB 等数据库上的实验结果表明在跨库的情况下具有更好的鲁棒性.

Marczewski 等[81] 提出了一种由 2 个一维 CNN 层、1 个 LSTM 层和 2 个全连接层组成的深度学习网络体系结构进行语音情感识别, 其中 CNN 层获取不同抽象层次的空间特征, LSTM 层学习与情绪随时间演化相关的时间信息. 该网络从输入音频样本接收 54 000 维数据点, 之后联合利用 CNNs 提取领域共享特征和 LSTMs 识别具有领域特定特征的情绪. 在 6 个不同数据库上的实验表明, 它们可以学习可转移的特征, 使模型能够从多个源域适应.

Parry 等[82] 对 CNN、LSTM 和 CNN-LSTM 等深度学习模型的泛化能力进行了比较分析, 探索 3 种模型在跨库语音情感识别方面的性能. 其中 CNN 由一维卷积层和一个 max-pooling 层组成; LSTMs 为双层双向 LSTMs; CNN-LSTM 包含 3 个 CNN 和两个双层双向 LSTMs. 提取 40 个 Mel 滤波器组系数后, 他们在 IEMOCAP, EMOVO, EMO-DB 等 6 个数据集上进行了实验, 结果表明, CNN 和 CNN-LSTM 模型的性能非常接近, 但优于 LSTM.

Rehman 等[83] 提出了一种由 LSTM 和分支层组成的 RNN 网络, 该方法以 MFCCs 为输入, 分支层从语音信号的 MFCCs 中提取信息并对提取的表征进行分类, LSTM 处理语音的时序性和时间动态, 最后使用 Softmax 层对目标语料库的语音情感进行预测, 并且获得了 40%–45% 的 UAR.

Seo 等[84] 提出了融合视觉注意 CNN 和视觉词袋的跨语料库语音情感识别方法, 使用具有 2D CNN 的视觉注意卷积神经网络 (visual attention convolutional neural network, VACNN) 在一个大的语音数据集进行

预训练, 并对一个小的语音数据集进行微调, 以识别话语中的情绪, 并使用视觉词包 (bag of visual words, BOVW) 提取特征向量帮助 VACNN 学习 log-Mel 谱图中的全局和局部特征, 最后, 利用 Softmax 对小数据集进行情感分类. 实验结果表明, 与现有最先进的跨语料库 SER 方法相比, 分别提高了 7.73%、15.12% 和 2.34%.

Lee[85] 提出了一个基于三元网络的计算框架, 该网络包括具有共享参数的同一前馈网络的 3 个特征, 利用三元网络进行特征变换, 将归一化特征映射到一维欧几里得空间从而减少源和目标语料库之间的内在差异, 学习在多个语料库中不变的情感语言的更一般化的特征. 实验使用 1 582 个静态特性, 然后使用三重网络对特征进行特征变换, 并在 IEMOCAP、MSP-IMPROV 等数据集上进行实验, 最后使用作为分类器的前馈网络对情感进行分类.

庄志豪等[86] 提出一种基于深度自编码器子域自适应的跨库语音情感识别算法, 采用两个深度自编码器分别获取源域和目标域表征性强的低维情感特征, 然后, 利用基于局部最大均值差异 (local maximum mean discrepancy, LMMD) 的子域自适应模块, 实现源域和目标域在不同低维情感类别空间中的特征分布对齐. 在 eNTERFACE 和 EMO-DB 数据库上进行了跨库实验, 该方法的识别准确率得到了一定程度的提高.

深度学习在语音情感识别中已经取得良好的进展, 但将其应用到跨库语音情感识别的研究还不够深入. Zhang 等[80] 提出一种基于时频原子的听觉注意特征提取模型用于跨库语音情感识别, 证明了深度学习可以提高跨库语音情感识别的性能. 之后, 一些研究者开始对比各种深度模型在跨库语音情感识别任务中的表现. Marczewski 等[81] 提出了一种由 2 个一维 CNN 层、1 个 LSTM 层和 2 个全连接层组成的深度学习网络; Parry 等[82] 比较了 CNN、LSTM 和 CNN-LSTM 等深度学习模型在跨库语音情感识别方面的性能, 结果表明, CNN 和 CNN-LSTM 模型的性能优于 LSTM. Seo 等[84] 将视觉模型应用于跨库语音情感识别中. 这是一种新颖的方法, 也是未来跨库语音情感识别可以继续探索研究的方向.

与结合手工特征的方法类似, Lee[85] 利用三元网络进行特征变换得到语料库归一化特征. 归一化方法只能初步降低语料库差异, 但不能解决其根本原因. 而自编码器因其可以消除语音信号中多余的信息, 在跨库语音情感识别中受到广泛应用. 庄志豪等[86] 用两个深度自编码器分别获取源域和目标域的低维情感特征, 然后利用 LMMD 实现源域和目标域的特征分布对齐. 综上所述, 深度学习可以显著改善不同语料库条件下的语音情感识别率, 进一步设计其他更先进的深度模型以减小不同语料库特征分布差异需要更深层次的探索.

2.2.2 面向无监督的深度域不变特征提取方法

张昕然等[87] 利用深度学习领域的深度信念模型, 提出了基于深度信念网络的特征层融合方法, 将语音频谱图中隐含的情感信息作为图像特征, 与传统情感特征融合. 通过在 ABC 数据库和多个中文数据库上的实验验证, 特征融合后的新特征子集相比传统的语音情感特征, 其跨数据库识别结果获得了明显提升.

Deng 等[88] 提出了一种新的端到端域自适应方法, 称为 Universum 自动编码器 (Universum autoencoder, U-AE). 该方法旨在使无监督学习自编码器具有监督学习能力, 从而提高语音情感识别的性能. 该方法将未标记样本用作 SVM 目标函数的惩罚项, 并经基于边际的损失引入深度自动编码器中, 通过同时从标记和未标记数据中学习公共知识以减少训练数据和测试数据之间的不匹配. 实验使用 INTERSPEECH 2009 特征集作为输入, 结果表明, 该方法优于其他领域自适应方法, 如核均值匹配[89] 和共享隐藏层自编码[90].

Abdelwahab 等[91] 将领域对抗神经网络 (domain adversarial neural network, DANN) 用于跨语料库语音情感识别中, 该网络旨在学习未标记源数据和目标数据之间灵活且具有判别性的特征表示. 引入领域鉴别器区分源域和目标域, 同时使用梯度反转层混淆领域鉴别器, 经过对抗训练得到领域不变特征. 提取 INTERSPEECH 2013 特征集作为 DANN 的输入进行实验, 结果表明, 基于未标记训练数据的对抗训练相对于仅使用源数据的训练, 获得了约 27.3% 的性能提升.

Neumann 等[92] 在未标记数据上训练一个递归的序列到序列自编码器, 然后采用它对标记目标数据产生特征表示, 这些产生的特征表征随后在使用的注意型 CNN 的训练过程中被整合为额外的源信息用于情感识别, 从而提高了语音情感识别的性能.

Liu 等[93] 提出了一种新的深度域自适应卷积神经网络 (deep domain-adaptive convolutional neural network, DDACNN) 模型, 模型以源语料库的有标记谱图和目

标语料库的无标记谱图作为输入, 用深度卷积神经网络 (deep convolutional neural network, DCNN) 进行特征提取并利用 MMD 准则在 DCNN 中添加域自适应层以缩小源语料库与目标语料库特征分布的差异, 学习语料库不变特征, 最后由一个 Softmax 层执行最终的情感分类任务. 实验结果表明 DDACNN 通过添加一层域自适应在最初的全连接层能有效处理跨库语音情感识别问题.

Su 等[94] 提出了使用条件循环情感生成对抗网络 (conditional cycle emotion generative adversarial network, CCEmoGAN) 合成源域样本, 这些样本是目标域感知的. 该网络学习源语料库和目标语料库之间的双向映射函数, 此外利用一个情感条件向量约束生成对抗训练, 该方法使用目标和源双向数据增强作为一种策略, 以情绪一致的方式增加可变性, 以改善从源到目标的情绪转移性. 以 IEMOCAP 数据库为源语料库, 以 MSP-IMPROV 和 CIT 数据库为目标语料库, 提取 1 582 个维度的话语级功能特性进行实验, 结果表明, 增加源域的目标域感知可变性可以提高跨语料情感识别中的情感可分辨性 .

Chang 等[95] 提出了最大回归差异 (maximum regression difference, MRD) 网络, 使用编码器对样本进行编码, 用两个回归器对样本标签进行回归预测, 最大化两个预测回归器的分布差异, 最小化来自编码器的回归器的分布差异形成对抗训练以增强源和目标的语义一致性. 该方法提取 IS10 特征集作为 MRD 网络的输入, 在 IEMOCAP、MSP-Podcast、MSP-IMPROV 三个数据集上进行跨库实验, 并取得了比 DANN 网络更好的结果.

Ahn 等[96] 提出了一种基于少样本学习和无监督域自适应 (few-shot learning and unsupervised domain adaptation, FLUDA) 的跨语料库语音情感识别方法, 该方法通过训练从源域样本中学习适应于目标域的类相似度, 采用少样本学习进行跨语料库的语义搜索, 并且使用无监督域自适应提取独立于领域的情感特征. 实验使用 1 582 维声学特征集 IS10, 以 IEMOCAP 和 CREMA-D 数据库作为源语料库, 以 MSP-IMPROV、EMO-DB 数据库作为目标语料库进行实验, 最后由最后一层为 Softmax 激活函数的 3 个全连接层作为分类器对 4 种种音情感进行分类, 实验结果表明, 该算法能够有效地提高交叉语料搜索的性能.

语音频谱图不仅包含类似手工特征的全局信息, 并且具有语音信号的时间信息, 张昕然等[87] 将其与其他特征进行融合可以得到更适用于跨库语音情感识别的输入特征. 由于无监督方法不能访问目标域数据集标签, 这会造成模型学习到与情感无关的信息. 为了解决这个问题, Deng 等[88] 使用无监督自编码器提取与情感高度相关的特征, 并将未标记样本用作 SVM 目标函数的惩罚项. 这种基于分类器对齐的方法通过将目标域标签信息迁移到源域, 并共同训练分类器以减少对目标域情感分类时产生的错误识别率. 与此对应的, 将目标域特征分布的知识迁移到源域也是极为重要的.

Abdelwahab 等[91] 将 DANN 用于跨语料库语音情感识别中, 这种基于对抗学习的方法提取域不变特征是如今跨库语音情感识别的一个热点. 进一步, Liu 等[93] 将 MMD 与深度学习模型相结合以缓解域差异; Su 等[94] 使用 CCEmoGAN 合成源域样本来增强源语料库的可变性而非学习语料库之间的域不变特征. 在面对目标数据集标签稀少的问题, Ahn 等[96] 采用少样本学习进行跨语料库的语义搜索.

2.2.3 面向半监督的深度域不变特征提取方法

Chang 等[97] 提出了一种多任务深度卷积生成对抗网络 (deep convolutional generative adversarial networks, DCGAN), 用于从未标记数据上的计算谱图中学习强特征表示, 并以情绪效价为主要目标, 情绪激活为次要目标进行多任务学习.

Deng 等[98] 提出了一种半监督自编码器来提高跨库语音情感识别性能. 在监督学习中添加一个除情感类外额外的类, 当监督分类器从给定的标记数据中学习时将所有未标记的数据预测为这个额外的类. 使模型可以从标记数据和未标记数据的组合中获益. 该方法将无监督自编码器与深度前馈网络的监督学习目标连接, 构造联合优化目标函数, 确保在标记和未标记数据上最小化无监督目标的重建误差以及由监督目标测量的预测误差. 提取 INTERSPEECH 2009 特征集作为输入. 实验结果表明, 该方法以极少的标记数据获得了最先进的性能.

Latif 等[99] 提出了一种基于 DBNs 的迁移学习技术, 使用的 DBN 由 3 个 RBM 层组成, 其中前 2 个 RBM 包含 1 000 个隐藏神经元, 第 3 个 RBM 包含 2 000 个隐藏神经元. 采用包括韵律特征和谱特征等 88 个特征的 eGeMAPS 特征集作为 DBNs 的输入. 结果表明, 与

稀疏自编码器和 SVM 相比, DBNs 在跨库语音情感识别上表现出更好的性能.

Gideon 等[100] 提出了一个对抗鉴别域泛化 (adversarial discriminative domain generalization, ADDoG) 算法, 该算法在特征编码和情感分类的基础上增加了测量不同数据集之间距离的模块, 并迭代的将每个数据集学习到的表示移动的更近, 采用生成对抗网络 (generative adversarial networks, GANs)[101] 的思想, 该算法可以充分利用未标记的测试数据在不同数据集上获得域不变特征表示. 他们提取 40 维 Mel 滤波器组作为算法输入, 得到的实验结果表明, 该算法表现优于 CNNs.

Latif 等[102] 提出了一种多任务半监督对抗自编码 (adversarial autoencoding, AAE) 方法, 利用短时傅里叶变换 (short time Fourier transform, STFT) 得到的谱图作为 AAE 的输入, 在对抗性自编码器中生成潜在表示, 之后构建一个使用大量可用数据的多任务学习框架, 将情感、说话人和性别识别作为辅助任务, 利用一些有限的可用数据最终可以获得比 CNN、CNN+LSTM 和 DBN 更好的性能.

Parthasarathy 等[103] 提出了一种结合无监督辅助任务的阶梯网络半监督方法, 其中首要任务是预测维度情感属性, 辅助任务用去噪自编码器产生中间特征表示的重构并以半监督的方式对来自目标域的大量未标记数据进行训练. 以 INTERSPEECH 2013 特征集作为阶梯网络的输入, 研究表明, 与完全监督单任务学习 (single-task learning, STL) 和多任务学习基线相比, 所提方法取得了优越的性能.

在结合深度特征的域不变特征提取方法中, 也可以应用半监督学习来提升模型的泛化性. 在半监督学习中, 常用的是基于对抗学习或者基于自编码器的方法, 这两种方法各有其优缺点. 基于自编码器的方法可以得到与情感高度相关的特征, 但在特征分布对齐上效果不佳. 与之相对, 基于对抗学习的方法可以有效的对齐特征分布以减少域差异, 但训练所需的时间较多且模型复杂. 这是未来跨库语音情感识别要解决的问题.

2.2.4 小结

深度域不变特征提取方法的总结与比较见附录中的表 A3. 在跨库语音情感识别任务的初期, 许多研究者开始探索不同的深度学习模型在跨库语音情感识别任务上的性能, 如 CNN, LSTM 等. 除此之外, 一些研究者在针对单数据库语音情感识别时也会研究他们的模型在跨库语音情感识别方面的性能. 与结合手工特征的域不变语音情感特征提取方法相似, 之后出现了将提取到的深度特征进行类似于特征变换的归一化以缓解不同数据库之间的差异. 但这种方法同样没有考虑数据库之间的特征分布差异. 随后, 一些研究者将提取到的深度低维特征运用 MMD 准则测量不同数据库之间的特征分布差异并最小化这种差异. 在跨库语音情感识别任务中另一个常用的方法是使用自编码器对输入特征进行编码-解码, 并最小化重构损失以减少域间差异. 随着对抗学习的出现, 一些研究者开始将对抗训练运用于跨库语音情感识别域不变特征. 在此基础上有些方法将对抗训练与 MMD 准则相结合以减少数据库之间的特征分布差异. 除以上方法之外, 一些学者还探索了其他适用于跨库语音情感识别任务的方法, 比如将目标域未标记样本作为惩罚项或者运用少样本学习的思想进行跨库语音情感识别.

3 结论与展望

语音情感识别因为其简单、交互直观且在人机交互中的重要性使得其应用十分广泛, 现有的语音情感识别技术已经实现了较高的性能, 但在跨库语音情感识别方面还存在很多问题. 本文从语音情感数据库、特征提取等方面对近年来的跨库语音情感识别技术进行了详细的介绍与归纳.

在语音情感识别任务提出以来, 已经建立了大量的语音情感数据集, 但由于在自然环境下收集语音样本会受嘈杂的环境影响, 因此现有的自然型野外数据集稀少. 如何在自然环境中收集到干净的语音信号以及对嘈杂的语音信号进行数据增强是在数据集建立方面的一个研究方向. 此外因为对数据集进行注释难度较大, 如何使用现有的数据集对未知数据集进行情感注释是需要进一步探讨的问题.

在跨库语音情感识别任务中, 先前的工作是结合手工特征并利用归一化方法、MMD 准则等方法消除数据集间的差异, 对齐数据集间特征分布. 之后, 深度学习被证明可以有效提高模型对语音情感的识别精度而被广泛应用. 因此, 越来越多的利用深度学习模型进行跨库语音情感识别的方法被提出. 其中最主要的方法是利用自编码器、基于对抗学习等域自适应方法得到域不变特征来提高分类器对目标域数据集的识别精

度. 但深度学习主要提取更深层次的情感特征, 而浅层特征同样重要, 应全面考虑. 此外, 如对抗训练等深度模型参数量较大, 不易训练.

综上所述, 尽管基于深度学习的跨库语音情感识别近年来取得了很大的进展, 但仍存在一些问题与挑战: 1) 大部分数据集只由少量的受试者进行录音, 并且大多数为未经过训练的非专业演员进行模拟不同情感类型得到的, 但这种模拟情感的真实度不高. 此外, 进行人工标注的数据集, 参与者较少, 更多的参与者可能会提高数据集的可靠性[45]; 2) 目前的方法集中在识别情绪标签上的差异, 但识别除情绪标签外的其他差异也可能会提高跨库语音情感识别的精度[83]; 3) 针对变异性较大的源域, 应该使用更多的情感数据库, 建立一个庞大的情感库来评估域自适应的效率[95]; 4) 现有深度学习方法主要用于学习高层次的特征进行情感标签预测而忽略了与情绪相关的低级特征, 因此, 未来的工作可以考虑融合面向跨库的低层次特征和高级特征来改善语音情感识别的性能[84]; 5) 尽管如今的深度学习技术日渐成熟, 但仍存在一些问题, 如网络的参数较多, 计算量大, 并且需要大量的样本训练进行训练, 因此未来对深度网络进行压缩是一个重要的研究方向; 6) 对于现有数据集标签稀疏问题, 除少样本学习之外, 零样本学习等元学习策略同样可以应用于跨库语音情感识别中.

参考文献

1 Ramakrishnan S, El Emary IMM. Speech emotion recognition approaches in human computer interaction. Telecommunication Systems, 2013, 52(3): 1467–1478. [doi: 10.1007/s11235-011-9624-z]

2 张石清, 李乐民, 赵知劲. 基于一种改进的监督流形学习算法的语音情感识别. 电子与信息学报, 2010, 32(11): 2724–2729.

3 张石清, 李乐民, 赵知劲. 人机交互中的语音情感识别研究进展. 电路与系统学报, 2013, 18(2): 440–451, 434.

4 Zhang SQ, Zhang SL, Huang TJ, *et al*. Learning affective features with a hybrid deep model for audio-visual emotion recognition. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 28(10): 3030–3043. [doi: 10.1109/TCSVT.2017.2719043]

5 Zhang SQ, Zhang SL, Huang TJ, *et al*. Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. IEEE Transactions on Multimedia, 2018, 20(6): 1576–1590. [doi: 10.1109/TMM.2017.2766843]

6 陶华伟, 张昕然, 梁瑞宇, 等. 面向语音情感识别的改进可辨别完全局部二值模式. 声学学报, 2016, 41(6): 905–912.

7 刘振焘, 徐建平, 吴敏, 等. 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]

8 赵力, 王治平, 卢韦, 等. 全局和时序结构特征并用的语音信号情感特征识别方法. 自动化学报, 2004, 30(3): 423–429.

9 黄晨晨, 巩微, 伏文龙, 等. 基于深度信念网络的语音情感识别的研究. 计算机研究与发展, 2014, 51(S1): 75–80.

10 Pravena D, Govind D. Significance of incorporating excitation source parameters for improved emotion recognition from speech and electroglottographic signals. International Journal of Speech Technology, 2017, 20(4): 787–797. [doi: 10.1007/s10772-017-9445-x]

11 Schuller B, Rigoll G, Lang M. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. Proceedings of 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing. Montreal: IEEE, 2004. 577–580.

12 Bertenstam J, Granström B, Gustafson K, *et al*. The VAESS communicator: A portable communication aid with new voice types and emotions. Proceedings of the Swedish Phonetics Conference Fonetik'97. Umeå: PHONUM, 1997. 57–60.

13 Schelinski S, von Kriegstein K. The relation between vocal pitch and vocal emotion recognition abilities in people with autism spectrum disorder and typical development. Journal of Autism and Developmental Disorders, 2019, 49(1): 68–82. [doi: 10.1007/s10803-018-3681-z]

14 Harati S, Crowell A, Mayberg H, *et al*. Depression severity classification from speech emotion. Proceedings of the 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). Honolulu: IEEE, 2018. 5763–5766.

15 Lin YS, Gau SSF, Lee CC. A multimodal interlocutor-modulated attentional BLSTM for classifying autism subgroups during clinical interviews. IEEE Journal of Selected Topics in Signal Processing, 2020, 14(2): 299–311. [doi: 10.1109/JSTSP.2020.2970578]

16 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]
17 Özseven T. Investigation of the effect of spectrogram images and different texture analysis methods on speech emotion recognition. Applied Acoustics, 2018, 142: 70–77. [doi: 10.1016/j.apacoust.2018.08.003]
18 梁瑞宇, 赵力, 陶华伟, 等. 仿选择性注意机制的语音情感识别算法. 声学学报, 2016, 41(4): 537–544.
19 王治平, 赵力, 邹采荣. 基于基音参数规整及统计分布模型距离的语音情感识别. 声学学报, 2006, 31(1): 28–34. [doi: 10.3321/j.issn:0371-0025.2006.01.005]
20 金琴, 陈师哲, 李锡荣, 等. 基于声学特征的语言情感识别. 计算机科学, 2015, 42(9): 24–28. [doi: 10.11896/j.issn.1002-137X.2015.09.005]
21 王忠民, 刘戈, 宋辉. 基于多核学习特征融合的语音情感识别方法. 计算机工程, 2019, 45(8): 248–254.
22 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]
23 Ahmad KS, Thosar AS, Nirmal JH, *et al*. A unique approach in text independent speaker recognition using MFCC feature sets and probabilistic neural network. Proceedings of 2015 8th International Conference on Advances in Pattern Recognition (ICAPR). Kolkata: IEEE, 2015. 1–6.
24 Dellaert F, Polzin T, Waibel A. Recognizing emotion in speech. Proceeding of the 4th International Conference on Spoken Language Processing. Philadelphia: IEEE, 1996. 1970–1973.
25 Zhang GB, Song QH, Fei SM. Speech emotion recognition system based on BP neural network in Matlab environment. Proceedings of the 5th International Symposium on Neural Networks: Advances in Neural Networks, Part II. Beijing: Springer, 2008. 801–808.
26 Kwon OW, Chan K, Hao JC, *et al*. Emotion recognition by speech signals. Proceedings of the 8th European Conference on Speech Communication and Technology. Geneva: ISCA, 2003. 1–4.
27 Nwe TL, Foo SW, De Silva LC. Speech emotion recognition using hidden Markov models. Speech Communication, 2003, 41(4): 603–623. [doi: 10.1016/S0167-6393(03)00099-2]
28 高庆吉, 赵志华, 徐达, 等. 语音情感识别研究综述. 智能系统学报, 2020, 15(1): 1–13.
29 张会云, 黄鹤鸣, 李伟, 等. 语音情感识别研究综述. 计算机仿真, 2021, 38(8): 7–17. [doi: 10.3969/j.issn.1006-9348.2021.08.002]
30 Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe: Curran Associates Inc., 2012. 1097–1105.
31 Elman JL. Finding structure in time. Cognitive Science, 1990, 14(2): 179–211. [doi: 10.1207/s15516709cog1402_1]
32 Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735–1780. [doi: 10.1162/neco.1997.9.8.1735]
33 Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science, 2006, 313(5786): 504–507. [doi: 10.1126/science.1127647]
34 Engberg IS, Hansen AV, Andersen O, *et al*. Design, recording and verification of a Danish emotional speech database. Proceedings of the 5th European Conference on Speech Communication and Technology. Rhodes: ISCA, 1997. 1–4.
35 Swain M, Routray A, Kabisatpathy P. Databases, features and classifiers for speech emotion recognition: A review. International Journal of Speech Technology, 2018, 21(1): 93–120. [doi: 10.1007/s10772-018-9491-z]
36 Burkhardt F, Paeschke A, Rolfes M, *et al*. A database of German emotional speech. Proceedings of the 9th European Conference on Speech Communication and Technology. Lisbon: ISCA, 2005. 1517–1520.
37 Martin O, Kotsia I, Macq B, *et al*. The eNTERFACE'05 audio-visual emotion database. Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW'06). Atlanta: IEEE, 2006. 8.
38 Wu T, Yang YC, Wu ZH, *et al*. MASC: A speech corpus in Mandarin for emotion analysis and affective speaker recognition. Proceedings of 2006 IEEE Odyssey—The Speaker and Language Recognition Workshop. San Juan: IEEE, 2006. 1–5.
39 Schuller B, Arsic D, Rigoll G, *et al*. Audiovisual behavior modeling by combined feature spaces. Proceedings of 2007 IEEE International Conference on Acoustics, Speech and Signal Processing. Honolulu: IEEE, 2007. II-733–II-736.
40 Busso C, Bulut M, Lee CC, *et al*. IEMOCAP: Interactive emotional dyadic motion capture database. Language Resources and Evaluation, 2008, 42(4): 335–359. [doi: 10.1007/s10579-008-9076-6]
41 Batliner A, Steidl S, Nöth E. Releasing a thoroughly annotated and processed spontaneous emotional database: The FAU Aibo emotion corpus. Proceedings of the Satellite Workshop of LREC 2008 on Corpora for Research on Emotion and Affect Marrakesh. 2008. 1–4.

42 Tao JH, Liu FZ, Zhang M, *et al*. Design of speech corpus for mandarin text to speech. Proceedings of the Blizzard Challenge 2008 Workshop. Hyderabad, 2008. 1–4.

43 Schuller B, Eyben F, Can S, *et al*. Speech in minimal invasive surgery—Towards an affective language resource of real-life medical operations. Proceedings of LREC 2010, Workshop on EMOTION. Valletta, 2010. 5–9.

44 Koolagudi SG, Reddy R, Yadav J, *et al*. IITKGP-SEHSC: Hindi speech corpus for emotion analysis. Proceedings of 2011 International Conference on Devices and Communications (ICDeCom). Mesra: IEEE, 2011. 1–5.

45 Liu P, Pell MD. Recognizing vocal emotions in Mandarin Chinese: A validated database of Chinese vocal emotional stimuli. Behavior Research Methods, 2012, 44(4): 1042–1051. [doi: 10.3758/s13428-012-0203-3]

46 Sugan N, Srinivas NSS, Kar N, *et al*. Performance comparison of different cepstral features for speech emotion recognition. Proceedings of 2018 International CET Conference on Control, Communication, and Computing (IC4). Thiruvananthapuram: IEEE, 2018. 266–271.

47 Ali Z, Talha M. Innovative method for unsupervised voice activity detection and classification of audio segments. IEEE Access, 2018, 6: 15494–15504. [doi: 10.1109/ACCESS.2018.2805845]

48 Cao HW, Cooper DG, Keutmann MK, *et al*. CREMA-D: Crowd-sourced emotional multimodal actors dataset. IEEE Transactions on Affective Computing, 2014, 5(4): 377–390. [doi: 10.1109/TAFFC.2014.2336244]

49 Costantini G, Iaderola I, Paoloni A, *et al*. EMOVO corpus: An Italian emotional speech database. Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014). Reykjavik: European Language Resources Association, 2014. 3501–3504.

50 Busso C, Parthasarathy S, Burmania A, *et al*. MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception. IEEE Transactions on Affective Computing, 2017, 8(1): 67–80. [doi: 10.1109/TAFFC.2016.2515617]

51 Metallinou A, Yang ZJ, Lee CC, *et al*. The USC CreativeIT database of multimodal dyadic interactions: From speech and full body motion capture to continuous emotional annotations. Language Resources and Evaluation, 2016, 50(3): 497–521. [doi: 10.1007/s10579-015-9300-0]

52 Zhalehpour S, Onder O, Akhtar Z, *et al*. BAUM-1: A spontaneous audio-visual face database of affective and mental states. IEEE Transactions on Affective Computing, 2017, 8(3): 300–313. [doi: 10.1109/TAFFC.2016.2553038]

53 Chang CM, Lee CC. Fusion of multiple emotion perspectives: Improving affect recognition through integrating cross-lingual emotion information. Proceedings of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New Orleans: IEEE, 2017. 5820–5824.

54 Lotfian R, Busso C. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. IEEE Transactions on Affective Computing, 2019, 10(4): 471–483. [doi: 10.1109/TAFFC.2017.2736999]

55 Parada-Cabaleiro E, Costantini G, Batliner A, *et al*. Categorical vs. dimensional perception of Italian emotional speech. Proceedings of the 19th Annual Conference of the International Speech Communication Association. Hyderabad: ISCA, 2018. 3638–3642.

56 Poria S, Hazarika D, Majumder N, *et al*. MELD: A multimodal multi-party dataset for emotion recognition in conversations. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2018. 527–536.

57 Livingstone SR, Russo FA. The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS One, 2018, 13(5): e0196391. [doi: 10.1371/journal.pone.0196391]

58 Parada-Cabaleiro E, Costantini G, Batliner A, *et al*. DEMoS: An Italian emotional speech corpus. Language Resources and Evaluation, 2020, 54(2): 341–383. [doi: 10.1007/s10579-019-09450-y]

59 Caruana R, Niculescu-Mizil A. An empirical comparison of supervised learning algorithms. Proceedings of the 23rd International Conference on Machine Learning. New York: ACM, 2006. 161–168.

60 Barlow HB. Unsupervised learning. Neural Computation, 1989, 1(3): 295–311. [doi: 10.1162/neco.1989.1.3.295]

61 Zhu XJ, Goldberg AB. Introduction to Semi-supervised Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning. San Rafael: Morgan & Claypool Publishers. 1–84.

62 Zhu XJ. Semi-supervised learning literature survey. Technical Report, Madison: University of Wisconsin-Madison, 2005.

63 Schuller BW, Steidl S, Batliner A. The INTERSPEECH 2009 emotion challenge. Proceedings of the 10th Annual Conference of the International Speech Communication Association. Brighton: ISCA, 2009. 312–315.

64 Schuller BW, Steidl S, Batliner A, *et al*. The INTERSPEECH 2010 paralinguistic challenge. Proceedings of the 11th Annual Conference of the International Speech Communication Association. Makuhari: ISCA, 2010. 2794–2797.

65 Schuller BW, Steidl S, Batliner A, *et al*. The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. Proceedings of the 14th Annual Conference of the International Speech Communication Association. Lyon: ISCA, 2013. 148–152.

66 Eyben F, Scherer KR, Schuller BW, *et al*. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Transactions on Affective Computing, 2016, 7(2): 190–202. [doi: 10.1109/TAFFC.2015.2457417]

67 Kaya H, Karpov AA. Efficient and effective strategies for cross-corpus acoustic emotion recognition. Neurocomputing, 2018, 275: 1028–1034. [doi: 10.1016/j.neucom.2017.09.049]

68 Zhang WJ, Song P, Chen DL, *et al*. Cross-corpus speech emotion recognition based on joint transfer subspace learning and regression. IEEE Transactions on Cognitive and Developmental Systems, 2022, 14(2): 588–598. [doi: 10.1109/TCDS.2021.3055524]

69 Zhang JC, Jiang L, Zong Y, *et al*. Cross-corpus speech emotion recognition using joint distribution adaptive regression. Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto: IEEE, 2021. 3790–3794.

70 Zong Y, Zheng WM, Zhang T, *et al*. Cross-corpus speech emotion recognition based on domain-adaptive least-squares regression. IEEE Signal Processing Letters, 2016, 23(5): 585–589. [doi: 10.1109/LSP.2016.2537926]

71 Song P, Zheng WM, Ou SF, *et al*. Cross-corpus speech emotion recognition based on transfer non-negative matrix factorization. Speech Communication, 2016, 83: 34–41. [doi: 10.1016/j.specom.2016.07.010]

72 Mao QR, Xu GP, Xue WT, *et al*. Learning emotion-discriminative and domain-invariant features for domain adaptation in speech emotion recognition. Speech Communication, 2017, 93: 1–10. [doi: 10.1016/j.specom.2017.06.006]

73 Liu N, Zong Y, Zhang BF, *et al*. Unsupervised cross-corpus speech emotion recognition using domain-adaptive subspace learning. Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary: IEEE, 2018. 5144–5148.

74 Liu N, Zhang BF, Liu B, *et al*. Transfer subspace learning for unsupervised cross-corpus speech emotion recognition. IEEE Access, 2021, 9: 95925–95937. [doi: 10.1109/ACCESS.2021.3094355]

75 金赟, 宋鹏, 郑文明, 等. 半监督判别分析的跨库语音情感识别. 声学学报, 2015, 40(1): 20–27.

76 宋鹏, 郑文明, 赵力. 基于特征迁移学习方法的跨库语音情感识别. 清华大学学报 (自然科学版), 2016, 56(11): 1179–1183.

77 Luo H, Han JQ. Cross-corpus speech emotion recognition using semi-supervised transfer non-negative matrix factorization with adaptation regularization. Proceedings of the 20th Annual Conference of the International Speech Communication Association. Graz: ISCA, 2019. 3247–3251.

78 Luo H, Han JQ. Nonnegative matrix factorization based transfer subspace learning for cross-corpus speech emotion recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 2047–2060. [doi: 10.1109/TASLP.2020.3006331]

79 LeCun Y, Boser B, Denker JS, *et al*. Backpropagation applied to handwritten zip code recognition. Neural Computation, 1989, 1(4): 541–551. [doi: 10.1162/neco.1989.1.4.541]

80 Zhang XR, Song P, Zha C, *et al*. Auditory attention model based on Chirplet for cross-corpus speech emotion recognition. Journal of Southeast University, 2016, 32(4): 402–407.

81 Marczewski A, Veloso A, Ziviani N. Learning transferable features for speech emotion recognition. Proceedings of the on Thematic Workshops of ACM Multimedia 2017. Mountain: ACM, 2017. 529–536.

82 Parry J, Palaz D, Clarke G, *et al*. Analysis of deep learning architectures for cross-corpus speech emotion recognition. Proceedings of the 20th Annual Conference of the International Speech Communication Association. Graz: ISCA, 2019. 1656–1660.

83 Rehman A, Liu ZT, Li DY, *et al*. Cross-corpus speech emotion recognition based on hybrid neural networks. Proceedings of the 39th Chinese Control Conference (CCC). Shenyang: IEEE, 2020. 7464–7468.

84 Seo M, Kim M. Fusing visual attention CNN and bag of visual words for cross-corpus speech emotion recognition. Sensors, 2020, 20(19): 5559. [doi: 10.3390/s20195559]

85 Lee SW. Domain generalization with triplet network for cross-corpus speech emotion recognition. Proceedings of

2021 IEEE Spoken Language Technology Workshop (SLT). Shenzhen: IEEE, 2021. 389–396.

86 庄志豪, 傅洪亮, 陶华伟, 等. 基于深度自编码器子域自适应的跨库语音情感识别. 计算机应用研究, 2021, 38(11): 3279–3282, 3348.

87 张昕然, 巨晓正, 宋鹏, 等. 用于跨库语音情感识别的 DBN 特征融合方法. 信号处理, 2017, 33(5): 649–660.

88 Deng J, Xu XZ, Zhang ZX, *et al*. Universum autoencoder-based domain adaptation for speech emotion recognition. IEEE Signal Processing Letters, 2017, 24(4): 500–504. [doi: 10.1109/LSP.2017.2672753]

89 Gretton A, Smola A, Huang JY, *et al*. Covariate shift by kernel mean matching. Dataset Shift in Machine Learning. In: Quiñonero-Candela J, Sugiyama M, Schwaighofer A, *et al*., eds. Cambridge: MIT Press, 2009.

90 Deng J, Xia R, Zhang ZX, *et al*. Introducing shared-hidden-layer autoencoders for transfer learning and their application in acoustic emotion recognition. Proceedings of 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Florence: IEEE, 2014. 4818–4822.

91 Abdelwahab M, Busso C. Domain adversarial for acoustic emotion recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(12): 2423–2435. [doi: 10.1109/TASLP.2018.2867099]

92 Neumann M, Vu NT. Improving speech emotion recognition with unsupervised representation learning on unlabeled speech. Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton: IEEE, 2019. 7390–7394.

93 Liu JT, Zheng WM, Zong Y, *et al*. Cross-corpus speech emotion recognition based on deep domain-adaptive convolutional neural network. IEICE Transactions on Information and Systems, 2020, E103.D(2): 459–463. [doi: 10.1587/transinf.2019EDL8136]

94 Su BH, Lee CC. A conditional cycle emotion GAN for cross corpus speech emotion recognition. Proceedings of 2021 IEEE Spoken Language Technology Workshop (SLT). Shenzhen: IEEE, 2021. 351–357.

95 Chang CM, Chao GY, Lee CC. Enforcing semantic consistency for cross corpus emotion prediction using adversarial discrepancy learning. IEEE Transactions on Affective Computing, 2021: 1–12.

96 Ahn Y, Lee SJ, Shin JW. Cross-corpus speech emotion recognition based on few-shot learning and domain adaptation. IEEE Signal Processing Letters, 2021, 28: 1190–1194. [doi: 10.1109/LSP.2021.3086395]

97 Chang J, Scherer S. Learning representations of emotional speech with deep convolutional generative adversarial networks. Proceedings of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New Orleans: IEEE, 2017. 2746–2750.

98 Deng J, Xu XZ, Zhang ZX, *et al*. Semisupervised autoencoders for speech emotion recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(1): 31–43. [doi: 10.1109/TASLP.2017.2759338]

99 Latif S, Rana R, Younis S, *et al*. Transfer learning for improving speech emotion classification accuracy. Proceedings of the 19th Annual Conference of the International Speech Communication Association. Hyderabad: ISCA, 2018. 257–261.

100 Gideon J, McInnis MG, Provost EM. Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (ADDoG). IEEE Transactions on Affective Computing, 2021, 12(4): 1055–1068. [doi: 10.1109/TAFFC.2019.2916092]

101 Goodfellow IJ, Pouget-Abadie J, Mirza M, *et al*. Generative adversarial nets. Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal: MIT Press, 2014. 2672–2680.

102 Latif S, Rana R, Khalifa S, *et al*. Multi-task semi-supervised adversarial autoencoding for speech emotion recognition. IEEE Transactions on Affective Computing, 2022, 13(2): 992–1004. [doi: 10.1109/TAFFC.2020.2983669]

103 Parthasarathy S, Busso C. Semi-supervised speech emotion recognition with ladder networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 2697–2709. [doi: 10.1109/TASLP.2020.3023632]

附录

表 A1 代表性语音情感数据库的总结与比较

数据库名称	年份	语言	情感类别	样本总数	说话者总数	模态
DES	1997	丹麦语	愤怒、快乐、中性、悲伤、惊讶	419	4 (2f)	听觉
SUSAS	1997	英语	中立, 愤怒, 大声	16000	32 (13f)	听觉
EMO-DB	2005	德语	中性、愤怒、恐惧、喜悦、悲伤、厌恶、厌倦	536	10 (5f)	听觉
eNTERFACE	2006	英语	愤怒、厌恶、恐惧、快乐、悲伤、惊讶、中性	1277	42 (8f)	视听
MASC	2006	汉语	中性、愤怒、兴奋、恐慌、悲伤	25636	68 (23f)	听觉
ABC	2007	德语	攻击、愉快、陶醉、紧张、中性、疲倦	431	8 (4f)	视听
IEMOCAP	2008	英语	快乐、愤怒、悲伤、沮丧、中性	5531	10 (5f)	多模态
FAU Aibo	2008	德语	喜悦、恼怒、愤怒、中性等11个情感标签	17074	51 (30f)	听觉
CASIA	2008	普通话	快乐、悲伤、愤怒、惊讶、恐惧、中性	9600	4 (2f)	听觉
SIMIS	2010	德语、土耳其语	愤怒、困惑、快乐、不耐烦、中性	11077	10 (1f)	听觉
IITKGP-SEHSC	2011	印地语	愤怒、厌恶、恐惧、高兴、中立、悲伤、讽刺、惊讶	12000	10 (5f)	听觉
CVE	2012	汉语	愤怒、厌恶、恐惧、悲伤、快乐、惊喜、中立	874	4 (2f)	听觉
SAVEE	2014	英语	中性、快乐、悲伤、愤怒、惊讶、恐惧、厌恶	480	4 (–)	视听
CREMA-D	2014	英语、西班牙语等	高兴、悲伤、愤怒、恐惧、厌恶、中性	7442	91 (43f)	视听
EMOVO	2014	意大利语	厌恶、恐惧、愤怒、高兴、惊讶、悲伤、中性	588	6 (3f)	听觉
MSP-IMPROV	2016	英语	高兴、悲伤、愤怒、中性	8438	12 (6f)	多模态
CIT	2016	英语	激活、效价、支配	90	16 (8f)	多模态
BAUM-1	2016	土耳其语	喜悦、愤怒、悲伤、厌恶、恐惧、惊讶、厌烦、蔑视	1222	31 (17f)	视听
NTUA	2017	普通话	快乐、悲伤、神经、愤怒、惊讶、挫折	204	44 (22f)	多模态
MSP-Podcast	2017	英语	愤怒, 快乐, 悲伤	84125	—	听觉
EmoFilm	2018	英语、西班牙语、意大利语	愤怒、悲伤、快乐、恐惧、轻蔑	1115	207 (94f)	听觉
MELD	2018	英语	愤怒、厌恶、悲伤、喜悦、中立、惊讶、恐惧	13707	6+ (–)	多模态
RAVDESS	2018	英语	平静、欢乐、悲伤、愤怒、恐惧、惊讶、厌恶	7356	24 (12f)	多模态
DEMoS	2020	意大利语	愤怒、悲伤、快乐、恐惧、惊讶、厌恶、内疚	9697	68 (23f)	听觉

表 A2 手工域不变语音情感特征提取方法的总结与比较

时间	文献	类别	输入特征	跨库方法	实验数据集
2018	[67]	监督学习	INTERSPEECH 2013	融合归一化方法	EMO-DB, DES, eNTERFACE
2021	[68]	监督学习	INTERSPEECH 2010	联合迁移子空间学习与回归	EMO-DB, eNTERFACE, BAUM-1
2021	[69]	监督学习	INTERSPEECH 2009 INTERSPEECH 2010	联合分布自适应回归	EMO-DB, eNTERFACE, CASIA
2016	[70]	无监督学习	INTERSPEECH 2009	域自适应最小二乘回归	EMO-DB, eNTERFACE
2016	[71]	无监督学习	1582个声学特征	转移非负矩阵分解	FAU Aibo, eNTERFACE, EMO-DB
2017	[72]	无监督学习	INTERSPEECH 2009	情感差异性和域不变特征学习	ABC, EMO-DB, FAU Aibo
2018	[73]	无监督学习	INTERSPEECH 2009	域自适应子空间学习	EMO-DB, eNTERFACE
2021	[74]	无监督学习	INTERSPEECH 2009	转移子空间学习	IEMOCAP, EMO-DB, eNTERFACE
2015	[75]	半监督学习	26个LLDs及其一阶差分	半监督判别分析	EMO-DB, eNTERFACE
2016	[76]	半监督学习	INTERSPEECH 2010	特征迁移学习	EMO-DB, eNTERFACE
2019	[77]	半监督学习	INTERSPEECH 2013	半监督自适应正则化 转移非负矩阵分解	CASIA, EMO-DB, eNTERFACE
2020	[78]	半监督学习	INTERSPEECH 2010	非负矩阵分解的迁移子空间学习	CASIA, SAVEE, EMO-DB, IEMOCAP, eNTERFACE

表 A3 深度域不变特征提取方法的总结与比较

时间	文献	类别	输入特征	跨库方法	实验数据集
2016	[80]	监督学习	频谱图	基于时频原子的听觉注意特征提取模型	eNTERFACE, EMO-DB
2017	[81]	监督学习	输入音频样本的54 000维数据点	2个一维CNN层、1个LSTM层和2个全连接层组成的深度学习网络体系结构	EMO-DB, EMOVO, eNTERFACE, IEMOCAP
2019	[82]	监督学习	40个Mel滤波器组系数	CNN、LSTM和CNN-LSTM	IEMOCAP, EMOVO, EMO-DB, RAVDESS, SAVEE
2020	[83]	监督学习	MFCCs	LSTM和分支层组成的RNN网络	IEMOCAP, RAVDESS, EMO-DB
2020	[84]	监督学习	log-Mel谱图	融合视觉注意卷积神经网络和视觉词袋的网络	EMO-DB, RAVDESS, SAVEE
2021	[85]	监督学习	1 582个静态特性	三重网络	IEMOCAP, MSP-IMPROV
2021	[86]	监督学习	INTERSPEECH 2010	包括DAE和子域自适应的深度自编码器子域自适应模型	SAVEE, EMO-DB, eNTERFACE
2017	[87]	无监督学习	频谱图和LLDs	DBNs和注意力机制	ABC, CASIA
2017	[88]	无监督学习	INTERSPEECH 2009	无监督自编码器	ABC, EMO-DB, SUSAS
2018	[91]	无监督学习	INTERSPEECH 2013	领域对抗神经网络	IEMOCAP, MSP-IMPROV, MSP-Podcast
2019	[92]	无监督学习	128 Mel频段	无监督自动编码器与注意力卷积神经网络	IEMOCAP, MSP-IMPROV
2020	[93]	无监督学习	谱图	深度域自适应卷积神经网络	EMO-DB, eNTERFACE, CASIA
2021	[94]	无监督学习	1 582维特征	条件循环情感生成对抗网络	IEMOCAP, MSP-IMPROV, CIT
2021	[95]	无监督学习	INTERSPEECH 2010	最大回归差异网络	IEMOCAP, MSP-Podcast, MSP-IMPROV
2021	[96]	无监督学习	INTERSPEECH 2010	少样本学习与无监督域自适应	IEMOCAP, CREMA-D, MSP-IMPROV, EMO-DB
2017	[97]	半监督学习	谱图	多任务深度卷积生成对抗网络	IEMOCAP
2017	[98]	半监督学习	INTERSPEECH 2009	与深度前馈网络的监督学习连接的无监督自编码器	FAU Aibo, ABC, EMO-DB, SUSAS
2018	[99]	半监督学习	eGeMAPS特征集	3个RBM层组成的DBNs网络	FAU Aibo, IEMOCAP, EMO-DB, SAVEE, EMOVO
2019	[100]	半监督学习	40维Mel滤波器组	基于GANs的反对称区分域泛化算法	IEMOCAP, MSP-IMPROV
2020	[102]	半监督学习	谱图	多任务半监督对抗自编码	IEMOCAP, MSP-IMPROV
2020	[103]	半监督学习	INTERSPEECH 2013	结合无监督辅助任务的阶梯网络	MSP-Podcast, IEMOCAP, MSP-IMPROV

(校对责编: 孙君艳)