DOI: 10.11992/tis.201904065

语音情感识别研究综述

高庆吉，赵志华，徐达，邢志伟
（中国民航大学 电子信息与自动化学院，天津 300300）

摘　要：针对语音情感识别研究体系进行综述。这一体系包括情感描述模型、情感语音数据库、特征提取与降维、情感分类与回归算法 4 个方面的内容。本文总结离散情感模型、维度情感模型和两模型间单向映射的情感描述方法；归纳出情感语音数据库选择的依据；细化了语音情感特征分类并列出了常用特征提取工具；最后对特征提取和情感分类与回归的常用算法特点进行凝练并总结深度学习研究进展，并提出情感语音识别领域需要解决的新问题、预测了发展趋势。

关键词：深度学习；情感语音数据库；情感描述模型；语音情感特征；特征提取；特征降维；情感分类；情感回归
中图分类号：TP391　　**文献标志码：**A　　**文章编号：**1673−4785(2020)01−0001−13

中文引用格式：高庆吉，赵志华，徐达，等. 语音情感识别研究综述 [J]. 智能系统学报，2020，15(1)：1–13.
英文引用格式：GAO Qingji，ZHAO Zhihua，XU Da，et al. Review on speech emotion recognition research[J]. CAAI transactions on intelligent systems，2020，15(1)：1–13.

Review on speech emotion recognition research

GAO Qingji, ZHAO Zhihua, XU Da, XING Zhiwei
(College of Electronic Information and Automation, Civil Aviation University of China, Tianjin 300300, China)

Abstract: In this paper, the research system of speech emotion recognition is summarized. The system includes four aspects: emotion description models, emotion speech database, feature extraction and dimensionality reduction, sentiment classification and regression algorithms. Firstly, we sum up the emotional description method of discrete emotion model, dimensional emotion model and one-way mapping between two models, then conclude the basis of emotional speech database selection, and then refine the classification of speech emotion features and list common tools for extracting the characteristics, and finally, extract the features of common algorithms, such as feature extraction, emotion classification and regression, and make a conclusion of the progress made in deep-learning research. In addition, we also propose some problems that need to be solved in this field and predict development trend.

Keywords: deep learning; sentiment speech databases; sentiment description models; acoustic sentiment features; feature extraction; feature reduction; sentiment classification; sentiment regression

语音情感计算包括语音情感识别、表达和合成等内容，近年受到广泛关注。其中，语音情感识别应用广泛，具有不可替代的作用。如结合驾驶员的语音[1]、表情[2-3]和行为[4]信息检测其精神状态，提醒驾驶员控制情绪、安全驾驶；依据可穿戴设备采集病人的语音信号实时检测其异常情感状态[5-6]，提高治疗效率；结合语音情感信息和自动翻译结果来帮助各方发言者顺畅交流[7]等。

近年来，研究者们就语音情感识别做了大量研究。韩文静等[8]从情感描述模型、情感语音数据库、特征提取和识别算法 4 个角度总结了 2014 年为止的语音情感识别的研究进展，并重点分析 SVM、GMM 等传统机器学习算法对离散情感的分类效果。随着深度学习技术逐步完善，在海量复杂数据建模上有很大优势，多用于解决数据分类。同时，部分研究者将其应用于语音特征的提取，取得了一定的成果。2018 年，刘振焘等[9]介

收稿日期：2019−04−27.
基金项目：国家自然科学基金委员会−中国民航局民航联合研究基金项目 (U1533203).
通信作者：赵志华. E-mail: 657902648@qq.com.

绍了语音情感特征提取和降维的方法，其中，重点描述了基于 BN-DBN、CNN 等深度学习方法的语音特征提取相关研究。

随着研究者深入探索，语音情感识别在以下几方面进展突出：维度情感和离散情感到维度情感的映射使情感描述更精确；情感语音数据库联合使用；采用深度学习方法进行特征提取和情感分类与回归；情感识别算法向更深层网络、多方法融合角度演变。

本文将从情感描述模型、情感语音数据库、特征提取与降维、情感分类与回归算法四个环节综述当前主流技术和前沿进展，然后总结深度学习研究的难点，指出未来的研究趋势。其中，着重分析深度学习算法在特征提取、情感分类与回归算法方面的研究进展。

1　情感描述模型

完整的语音情感识别包括采集语音片段、预处理、语音特征提取与降维、情感分类与回归等流程,如图 1 所示。

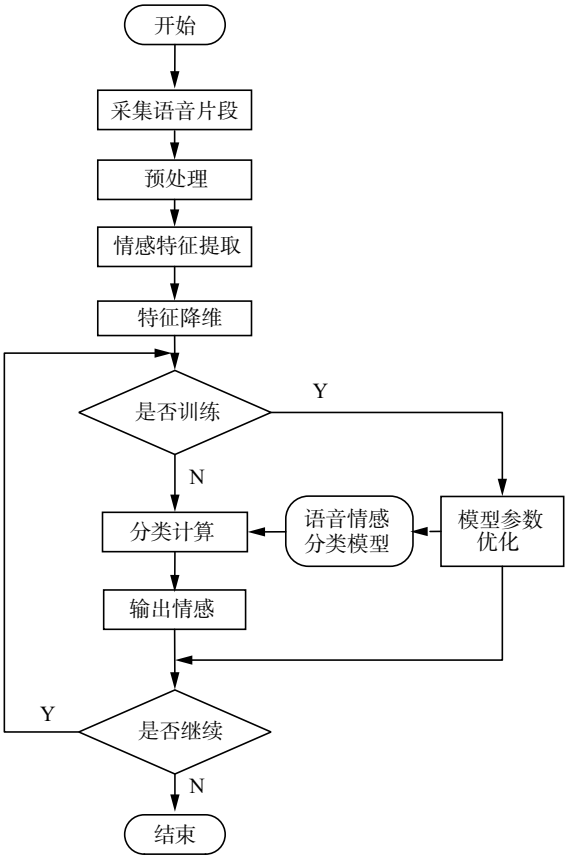

图 1　语音情感识别流程图

Fig. 1　Speech emotion recognition flow chart

实现语音情感识别，首先需要定义情感。情感描述模型将情感表征为一组互斥的离散情感类别或数字维度组合[10](空间坐标值)。根据表征方式不同，分为离散情感模型和维度情感模型。

离散情感模型使用形容词标签表示情感。Mover 简单地将离散情感分为痛苦、快乐两类基本情感[11]；进一步，Ekman[12] 将离散情感分成愤怒、厌恶、恐惧、快乐、悲伤和惊喜 6 种基本情感。目前常用 10 余种情感描述模型中，使用最广泛的是 Ekman 提出的 6 种基本情感和在此基础上加入中性的 7 种基本情感[13]。离散情感模型简单直观，运用广泛，但描述精确度不高、连续性不好，表征情感能力有限。为克服以上不足，学者们建立了维度情感模型。

维度情感模型可以在二维或多维空间中构造，用以描述连续情感。可利用效价−唤醒二维模型 (valence-arousal, VA) 描述情感的极性和度量情感程度，能够表示大部分情感；愉悦−唤醒−支配三维模型 (pleasure-arousal-dominance, PAD) 在 VA 模型上添加支配维，用以描述周围环境对自身的影响，如高支配度是一种主宰感，低支配度是一种软弱感[14]，理论上可以表示无穷多种情感，但难以表述惊讶[15]。Fontaine 等[16] 研究表明，在 PAD 模型基础上添加期望维，度量个体对情感出现的准备性，可以描述惊讶。维度情感模型表征情感能力强 (情感类别多、精确性高)，可连续表征情感变化，但维度情感理解困难且操作复杂，目前研究者较少。

因此，为合理利用现有基于离散情感模型建立的数据库，研究者们开始关注离散情感和维度情感间的映射。Russell 提出环形模型[17]，将 VA 模型空间中的某些区域解释为 28 种离散情感。而后，Yik 等[18] 将维度情感聚类为 8 种基本离散情感，即情绪轮。再后，又将该模型聚类情感扩展到 12 种。Plutchik[19] 引入更加复杂的混合情感模型，即情绪“沙漏”。目前，环形模型和情绪“沙漏”的关联性是大家关注的焦点，实用性有待研究[10]。

离散情感模型简单直接易搭建，但描述精度不足；维度情感模型描述精度高，情感表征能力强，但操作复杂，文献不多；离散情感和维度情感间的映射有研究价值，但实用性有待研究。故如何更好地将离散情感与维度情感结合来提升情感状态表征精度，受到广泛关注。

2　情感语音数据库与数据标注

2.1　情感语音数据库

研究情感分类与回归算法，需要数据库的支撑。目前情感语音数据库数量较多，但没有统一

的划分标准。为方便理解，学者们从情感描述角度，将语音数据库划分为离散情感数据库和维度情感数据库两类。表 1 和表 2 从参与人数、语言、标注人数、离散情感种类数、样本量及激发方式 6 个方面分别分析常用离散和维度情感语音数据库。

表 1　常用离散情感语音数据库统计表

Table 1　Statistical table of the frequently used discrete emotion speech databases

数据库	参与人数	语言	标注者数	离散情感	样本量	激发方式
BAUM-1s[20]	17男、14女	土耳其语	5	8	1 222个	自然
SAVEE[21]	4男	英语	10	7	480个	表演
RML[22]	8	7种	2	6	720个	表演
EMO-DB[23]	5男、5女	德语	20	7	535个	表演
eNTERFACE'05[24]	34男、8女	英语	2	6	1 166个	表演
RAVDESS[25]	12男、12女	英语	247	8	1 440个	引导
FAU AIBO[26]	21男、30女	德语	5	9	9.2 h	自然
CASIA汉语数据库[8]	2男、2女	汉语	—	5	9 600个	表演
ACCorpus _ SR[8]	25男、25女	汉语	—	5	—	表演

表 2　常用维度情感语音数据库统计表

Table 2　Statistical table of the frequently used dimensional emotion speech databases

数据库	参与人数	语言	标注者数	维度情感	样本量	激发方式
VAM[27]	47	德语	6~34	V、A、D	1 018个	自然
FAU AIBO[26]	30男、21女	德语	5	P、A、D	9.2 h	自然
IEMOCAP[28]	5男、5女	英语	≥2	V、A、D	1 150个	引导
RECOLA[29]	46	法语	6	V、A	9.5 h	引导
CreativeIT[30]	7男、9女	英语	3~4	V、A、D	8 h	引导
SEMAINE[31]	150	德语	2~8	V、A、D、E、I	80 h	引导

注：P为愉悦维；V为效价维；A为唤醒维；D为支配维；E为期望维；I为强度维. 愉悦(Pleasure)维和效价(Valence)维同义.

表 1 和表 2 根据激发方式将数据库分为自然型、表演型和引导型数据库。自然型数据库采集的语音样本最接近自然交流，但是其制作难度高，目前数量较少，常用 FAU AIBO 数据库[26] 和 VAM 数据库[27]；表演型数据库要求专业演员在安静环境中根据指定语料进行表演并采集语音样本；引导型数据库数量最多，常通过视频或对话诱导安静环境中的参与者表达相应情感以获取样本。进一步可看出，常见离散情感语音数据库多属于表演型，常见维度情感语音数据库多属于引导型。

此外，由表 1 和表 2 可知，多数数据库区分性别和语言种类。研究者可利用此信息度量不同性别、跨文化等语境特征对情感识别的影响并建立智能推理模型，为实现个性化人机交互提供可能[32]。

进一步，离散情感与维度情感联合建立数据库，便于研究者理解和使用，如 FAU AIBO；结合面部表情、语音和姿态等[33] 建立多模态数据库，拓展识别算法的信息维度，如 SAVEE、IEMOCAP。而并且嘈杂环境最接近自然环境，在其中采集样本，数据库建立的难度较大，也对识别算法鲁棒性提出严峻挑战。

2.2　数据标注

离散情感数据库主要采用标注者投票判别情感种类和准确性[20,24]，使用专业工具辅助判别较少。维度情感数据库借助 SAM 系统[34] 或 MAAT 工具[35] 量化 PAD 模型维度取值；FEELTRACE[30] 量化 VA 模型维度取值；ANNEMO[36] 一次仅标记一个维度，结果更精确[15]。

同时，情感标注也要求标注者有一定的经验，同时标注过程中精神高度集中。多数数据库采用对多标注者标注的数据进行插值、标准化等处理，以降低标注者自身因素对标注结果的干扰。

语音情感数据库不断丰富，情感描述能力不断提升的同时，对数据标注的新需求也不断扩充，如何通过模块设计等方法集成各优秀的数据标注工具的性能集成，是一个研究方向。此外，研究者开始探索弱标注，即采用半监督的方法提

取无标注和有标注样本的公共信息，学习无标注的样本，充分利用数据库。

3　特征提取与降维

如何提取语音中的丰富情感信息并凝练，直接影响情感分类和回归算法的运算效率和准确性。因此，提取有代表性的语音情感特征并进行降维，便显得十分必要。

3.1　预处理

为消除人体语音器官和声音采集设备的差异、混叠、高次谐波失真等影响，在特征提取前需进行预处理。预处理包括：提取语音信号的起始点和终止点的端点检测、将语音信号转化为短时平稳分析帧的加窗分帧、对高频部分进行加重，增强分辨率的预加重等[37-38]。

3.2　特征提取

深度特征是深度学习提取的高级特征，在语音情感识别应用中表现突出[39]，故将特征在原来 4 类[40] 基础上拓展为声学、语言、语境、深度和混合共 5 类特征。

1) 声学特征提取

声学特征分为 3 类：韵律学特征、基于谱的相关特征和声音质量特征，描述语音的音调、幅度、音色等信息。如共振峰、梅尔频率倒谱系数[41-45] (Mel frequency cepstrum coefficient, MFCC) 和抖动和谐波噪声比 (harmonic to noise ratio, HNR)。常规提取方法，如自相关函数法和小波法，可参考文献 [9]。

此外，为减少手工提取的复杂性和盲目性，学者们采用深度学习提取声学特征[46]。如分层稀疏编码[47] 以无人监督的方式，自动挖掘情绪语音数据的非线性特征，语音情感区分特性更强；基于双层神经网络的域自适应方法[48] 可共享源域和目标域中相关类的公共先验知识，实现源域和目标域的共享特征表示，可有效传递知识和提高分类性能。

2) 语言、语境和混合特征提取

语言特征通过对语义信息进行分析提取获得。研究者提出对语义按固定长度分段，比对码本将其转化为特征向量，如 BoW(Bag of Words)[49]，BoNG(Bag of N-grams)[50]、BoCNG(Bag of Character N-grams)[50] 和小波特征[51] 等。

语境特征主要描述不同说话者的性别和文化背景的差异[40]。区分性别能改善分类效果[52]。通过对比分析文化内、多元文化和跨文化情况下情感识别效果，其中，文化内、多元文化背景下情感识别效果最佳[53]。

混合特征融合两种及两种以上的特征。金琴等[49] 为每个情感类构建一个情感词典，其中包含特定情绪的词汇和分配的权重，用以表明这种情感的倾向。然后，使用此情感词典为每个话语生成矢量特征表示即情感向量词汇特征。最后，融合声学特征、情感向量和 BoW，提升情感识别准确率。Ashish 等[52] 以语境特征和声学特征为切入点，系统分析了区分和忽略性别信息对情感识别率的影响。实验结果表明，区分性别的情感识别更准确。

3) 深度情感特征提取

因低级特征数量有限、提取耗资且不能完整描述语音信号，所以研究者尝试从低级特征中进一步提取高级特征或直接批量处理原始音频，自动提取高级特征，例如深度特征[39]。深度学习可从每层网络和网络层次结构中提取复杂特征—深度特征，常用方法有卷积神经网络 (convolutional neural network, CNN)、深度信念网络 (deep belief network, DBN)、深度神经网络 (deep neural network, DNN) 等。

王忠民等[43] 采用 CNN 从语谱图中提取图像特征，改善 MFCC 丢失信息识别准确率不高的问题。但 CNN 无法准确捕捉语谱图中特征的空间信息，为此 Wu 等[54] 采用两个循环连接的胶囊网络提取特征，增强时空敏感度。此外，Zhang 等[40] 以类似于 RGB 图像表示的 3 个对数梅尔光谱图作为 DCNN 的输入，然后通过 ImageNet[55] 预训练的 AlexNet DCNN 模型学习光谱图通道中的高级特征表示，最后将学习的特征由时间金字塔匹配 (DTPM) 策略聚合得到全局深度特征，进一步提升对有限样本特征提取的有效性。为有效描述情感连续性变化，Zhao 等[56] 采用局部特征学习块从 log-mel 谱图提取的局部特征，重构为时序形式后输入至长短期记忆网络 (long and short term memory network, LSTM)，以进一步提取全局上下文特征。

张丽等[57] 采用贪婪算法进行无监督学习，通过 BP 神经网络反向微调，找到全局最优点，再将 DBN 算法的输出参数作为深度特征，并在此过程中，采用随机隐退思想防止过拟合。

进一步，为解决基于多样本库的源域和目标域中数据分布差异，Abdelwahab 等[58] 采用域对抗神经网络创建源域 (USC-IEMOCAP 和 MSP-IMPROV 数据库) 和目标域 (MSP-Podcast 数据库) 的共同特征表示——深度特征，然后通过梯度反转

层将域分类器生成的梯度在传播回共享层时乘以负值，使训练集和测试集的特征收敛，提升泛化能力。同时使用 t-SNE 数据可视化技术[59]，通过创建不同层的特征分布 2D 投影，直观检查模型学习特征表示的全过程。

此外，说话者无关训练 (speaker-invariant training, SIT，模型的学习结果与说话者自身无关，即要求模型有较强的泛化能力)[60] 通过对抗性学习减少声学建模过程中说话者差异的影响，再联合 DNN，来提取与说话者无关且辨别力强的深度特征。

4) 常用特征提取工具

目前 Praat[61] 和 OpenSMILE[62] 两种工具使用最广泛。Praat 是一款语音学专业软件，其 GUI 界面简洁且指导手册持续更新，便于学习。可对语音文件进行特征提取、标注等工作，结果可导出。OpenSMILE 使用命令行和 GUI 结合的方式进行使用。常用配置文件 config/IS09/10/11/12/13 paraling. conf，分别提取 384、1 582、4 368、6 125 和 6 373 维特征。此外，在 Tensorflow 框架中，可以调用 Librosa 工具包提取频谱图、MFCC 等特征，便于后续识别。表 3 整理了更多的提取工具可供学习。

表 3　常用语音特征提取工具统计表[63]

Table 3　Statistical table of common speech feature extraction tools

工具箱	平台	提取特征
Praat	C++	信号能量、FFT频谱、倒频谱、语音质量、LPC、共振峰等
OpenSMILE	C++	波形、信号能量、FFT光谱、语音质量、Mel/Bark光谱、共振峰等
HTK	C	信号能量、Mel/Bark光谱、LPC、波形等
Voicebox	MATLAB	信号能量、F0、LPC、倒谱、Mel/Bark光谱等
COLEA	MATLAB	F0、共振峰、频谱、信号能量等
SPEFT	MATLAB	波形、信号能量、语音质量、共振峰、倒谱、Mel/Bark频谱等
SPAC	MATLAB	F0、共振峰、语音质量、LPCC、MFCC、信号能量、语速、小波等

3.3　特征降维

上述特征提取方法得到的语音情感特征一般维数较高，直接处理易导致维度灾难。为保障识别准确率和效率，采用主成分分析 (principle component analysis, PCA)[64]、Fisher 准则[38]、线性判别分析 (linear discriminate analysis, LDA)[65] 和 FCBF(fast correlation-based filter solution) [66] 等方法进行特征降维。如 BP 神经网络[67] 可进行特征选择，检测冗余的同时，通过节点信号变化的敏感度挑选对网络贡献度大的特征得到组合特征。

声学特征因提取算法和提取工具丰富，使用广泛；深度学习框架环境日益发展，被更多研究者用于提取情感特征。此外，声学和语义是语音信号的两个主要部分。随着文本情感研究深入，从语义中提取的语言特征将会成为混合特征中的重要组成部分。故如何有效利用句子含义与转折词，精简语言特征并提升特征的有效性，将成为研究热点。

4　情感分类与回归

根据情感表征方式不同，将目前主流识别算法分为情感分类算法和情感回归算法两类。

4.1　情感分类算法

情感分类算法将测试集样本归类为不同离散情感类别，常使用支持向量机 (support vector machines, SVM)、隐马尔可夫模型 (hidden Markov model, HMM) 和 DCNN。

SVM[68-71] 在求解非线性、小样本和高维模式识别等问题具有优越性，且泛化能力强，在情感分类中广泛使用[38]。半定规划多核 SVM[72] 来提高分类算法的鲁棒性。

Zheng 等[73] 采用 DCNN 对通过 PCA 白化处理的光谱图学习处理并进行情感分类，结果表明该方法优于 SVM。进一步，Shahin 等[74] 级联高斯混合模型和深度神经网络 (gaussian mixture model-deep neural network, GMM-DNN) 构建混合分类器，其分类性能优于 SVM、MLP(multi-layer perception)、GMM 和 DNN，并且在嘈杂谈话背景下，情感分类效果良好。

Sagha 等[75] 以 OpenSMILE 提取 384 个特征为基于核典型相关分析的域自适应方法的输入，在 EMODB、SAVEE、EMOVO 和 Polish 等 4 个不同语言的语音数据库上实现跨语料库迁移学习，学习速度快且有效克服过拟合，明显降低陷入局部最小值的风险。

以上算法均针对语音信号来提升情感分类准确性。此外，融合其他模态的特征，如面部表情[2]、姿态和生理信号[1]，可提升情感分类的鲁棒

性和可信度。陈师哲等[76]结合面部表情和语音模态，采用 SVM 和随机森林来减弱文化差异对情感分类的影响。刘颖等[77]结合面部表情和语音双模态，采用 GMM，有效提升分类的准确性。

4.2　情感回归算法

情感回归算法将测试样本的连续维度情感值映射到二维或多维坐标空间。情感回归传统方法为 SVR(support vector regression)[41,47]。

近年来，研究者将深度学习技术引入情感回归中，取得良好效果，如 LSTM 和 DANN。Zhao 等[56]采用二维 CNN LSTM 网络学习局部特征学习块和 LSTM 提取的局部特征和全局特征并在全连接层实现 VA 模型情感预测，改变 DBN、CNN 等算法模型只能学习一种深度特征的现状。在 IEMOCAP 数据库中，本方法在说话者相关和说话者无关中的识别正确率分别为 89.16% 和 52.14%，远高于分别采用 DBN 和 CNN 获得的 73.78% 和 40.02% 的正确率。Abdelwahab 等[58]在合并情感语音数据库中采用域对抗神经网络 (domain adversarial neural network, DANN) 根据提取的源域和目标域的共同特征表示，实现 PAD 模型下的情感回归，提升鲁棒性。

一般情感回归主要解决维度情感问题，为有效利用仅标注离散情感的语音数据库，Ma 等[78]采用 SVM 分类，通过离散情感和维度情感间的单向映射，实现 PAD 模型下的情感回归；Eyben 等[70]将离散情感映射至 VA 模型中对应维度情感坐标，拓展情感描述能力。基于此，表 4 描述了数据库特定离散情感类别至 VA 模型中维度情感间的映射。

表 4　数据库离散情感到 VA 模型中的维度情感的映射统计表[70]

Table 4　Mapping statistical table of dataset specific emotion categories to dimensional emotions in VA models

数据库	激活维		效价维	
	低	高	消极	积极
FAU AIBO	—	—	消极	积极
TUM AVIC	无聊	中立、快乐	无聊	中立、快乐
EMO-DB	无聊、厌恶、中性、悲伤	生气、害怕、开心	生气、悲伤	开心、中性、吃惊
GEMAP	快乐、宽慰、兴趣、烦躁、焦虑、悲伤	喜悦、娱乐、骄傲、生气、恐惧、绝望	生气、恐惧、绝望、烦躁、焦虑、悲伤	喜悦、娱乐、骄傲、快乐、宽慰、兴趣
VAM	q2、q3	q1、q4	q3、q4	q1、q2

注：将VAM数据库离散情感映射到代表VA模型中的4个象限(q1、q2、q3和q4, 对应于高−积极、低−积极、低−消极、高−消极), 以便于评估.

离散情感简单易理解，可用于训练的数据库较多，受众广，其中，基于声学特征的情感分类研究成果丰硕而情感回归实现维度情感识别有待加强。

5　对比与分析

本部分首先总结当前研究者们在语音情感识别的研究过程，然后对比分析各特征提取和情感分类与回归算法的优缺点，最后重点评析深度学习的研究近况。

近年来，研究者们在语音情感特征、特征提取方法、降维或融合方法、数据库、情感类别和情感分类与回归算法等影响情感识别的方面做了不懈努力，取得了较好成就。故从这 6 个方面总结近年语音情感识别相关研究，如表 5 所示。由于不同文献所使用的数据库、特征和使用的情感类别等方面都不尽相同，暂不比较各分类与回归算法准确率。

此外，表 6 概要分析了典型特征提取和情感分类与回归算法的特点。DNN、CNN 和 DBN 等深度学习算法因表征能力好、学习能力强，能有效提取情感特征，但参数调整直接影响提取效果。情感分类与回归算法中，SVM 等传统算法简单并能有效解决小样本、高维、非线性等问题，但对缺失数据问题敏感且无法处理大样本；各类深度学习算法优点较多，如：KNN 处理大样本能力强；RNN、LSTM 适合处理样本序列；ELM 计算复杂但泛化能力强，对训练数据依赖性大，处理大样本耗时。鉴于算法特点的差异，采用前需全面分析。

总体而言，深度学习因其有着比传统机器学习方法更优越的性能，近期在特征提取和情感分类与回归中受到广泛关注。但是在不断研究和处理中，研究者也发现深度学习主要存在以下 3 个方面的不足：1) 作为典型的“黑箱”算法，不易描述网络具体实现且解释能力弱；2) 以输入为导向的神经网络算法仅根据现有样本进行学习，但自身无法验证输入样本的代表性和正确性；3) 隐藏层中节点个数设置缺乏客观性，主要凭借经验设置。

表 5　语音情感识别过程总结统计表

Table 5　Summary statistical table of speech emotion recognition's whole process

文献(时间)	语音情感特征	特征提取	降维和融合	数据库	情感类别		识别算法
					离散情感	维度情感	
[37] (2015)	声学特征	常规	MFCCG-PCA	C1	6	—	SVM
[38] (2019)	声学特征	—	Fisher准则	C1、E1	6/7	—	决策树SVM
[40] (2018)	深度特征	DCNN	DTPM	E1、R、e、B	6/7	—	SVM
[41] (2018)	声学特征	PCNN、常规	—	C1、E1、S1	—	P、A、D	SVR
[42] (2019)	声学特征	VMD+IMF、常规	—	E1、R2	5	—	ELM
[43] (2019)	声学特征	CNN、常规	多核学习方法	E1、C1	6/7	—	SVM
[44] (2018)	深度特征	LSTM+CNN	—	e、R1、A1	6	—	CNN
[45] (2018)	声学特征	Praat、GF、HOG、GLCM和WD	—	E1、e、S1	6/7	—	SVM
[47] (2018)	声学特征	Praat、Pysound	—	V、A2	—	V、A	HSC+SVR
[49] (2015)	声学特征、语言特征	OpenSMILE、BoW	—	I	4	—	SVM
[56] (2019)	深度特征	CNN+LSTM	—	E1、I	6/7	—	CNN+LSTM
[57] (2019)	深度特征	DBN	—	C1	6	—	ELM
[58] (2018)	声学特征	OpenSMILE	DANN	I、M1、M2	—	V、A、D	DANN
[64] (2015)	声学特征	OpenSMILE	PCA	C1、S1	6	—	IC-D、IC-S
[66] (2016)	声学特征	Praat	PCA、FCBF	S1	7	—	FAMNN
[68] (2015)	声学特征、非线性特征	C-C、G-P、重标极差法等	—	E1	4	—	SVM
[69] (2018)	声学特征	OpenSMILE	子空间学习、特征选择、MMD	E1、E2	5	—	SVM
[71] (2019)	声学特征	openSMILE	thSD、thMN、thMED、thCV	E1、e、E4、S1	6/7	—	SVM、MLP、KNN
[72] (2015)	声学特征	Voicebox	双输入对称相关算法	E1	5	—	半定规划多核SVM
[73] (2015)	深度特征	DCNN	PCA	I	5	—	DCNN
[74] (2019)	声学特征	常规	—	S2、E3	6	—	GMM-DNN
[79] (2019)	声学特征	WP滤波器组	DBN+BP神经网络	E1	6	—	SVM
[80] (2017)	声学特征	音调功率比、频谱通量、常规和音调色度提取	串联	E1、T	5/7	—	AFDBN
[81] (2019)	声学特征	openSMILE	LDE、GbFA、LPDA、FDA、LDP	G、A3、V、e	4/6/12	—	ELM+子空间学习+KNN
[82] (2019)	声学特征	常规	—	E1、S1	7	—	FAM-FIS
[83] (2019)	声学特征	openSMILE	—	U、R2、E1、e、V、A4	—	V、A	GMTL、MIT-KDG等
[84] (2019)	声学特征	openSMILE	—	C3、M3	6	—	CNN
[85] (2019)	声学特征	常规	—	自己建立	2	—	SVM

注：常规: MFCC特征提取的FFT+梅尔滤波器组+对数变换+DCT。数据库名称按出现顺序缩写、E1:EMO-DB、E2:Enteraface、E3:ESD、E4:EMOVO、R1:RML、R2:RAVDESS、e:eNTERAFACE'05、B:BAUM-1、C1:CASIA、C2:CHiME-3、C3: CHI-MEI、S1:SAVEE、S2:SUSAS、A1:AFEW -6.0、A2:AVEC2012、A3:ABC、A4: AVEC(2011)、V:VAM、U: UMSSED、I:IEMOCAP、M1:MSP-IMPROV、M2: MSP-Podcast、M3: MHMC、T: 泰卢固语数据库、G:GEMEP；"+"表明方法结合使用；表格中识别方法仅列举研究者主要使用方法。

表 6　特征提取和情感分类与回归算法特点统计表

Table 6　Statistical table of the feature extraction and sentiment classification and regression algorithms

算法名称	优点	缺点
SVM	解决小样本、高维、非线性等问题, 泛化能力强, 避免网络结构选择。	对缺失数据敏感, 对非线性问题没有通用解决方案, 训练时间较长。
KNN	简单有效, 重新训练的代价较低, 适用于样本容量较大的类域的自动分类。	样本容量较小的类域易误分, 输出解释性不强, 计算量较大。
DNN	可模拟任意函数, 情感表征能力强。	处理大量数据时, 耗时较严重。
CNN	提取抽象特征能力强, 具有类内收敛、类间发散的特点。	一定程度上解决梯度消失和缓解过拟合, 对空间信息不敏感。
RNN	引入记忆单元, 可用于语音序列建模。	长时间训练易出现梯度爆炸。
LSTM	具有记忆特性, 能有效学习帧间相关性。	仅使用历史信息。
GMM	对语音情感数据的拟合性能高。	需存储各维度各高斯分量的参数, 对训练数据的依据性强, 计算较复杂。
HMM	适用于分析短时平稳的语音信号。	处理海量数据能力有限。
GMM-HMM	可模拟任意函数, 情感表征能力强。	性能取决于混合高斯函数的个数, 具有一定的局限性。
ELM	泛化能力强, 学习速度快, 预测准确, 减轻计算负担。	输出层决策值的计算完全取决于标签。
GAN	对原始数据底层概率分布的感知能力强。	训练不稳定且调参难度大
DBN	情感表征能力强, 无监督特征学习能力强。	网络超参选择直接影响检测准确率。

为此, 研究者们主要做了如下努力: 1) 采用数据可视化技术, 如 t-SNE[59], 从各个层次理解数据分布, 有助于模型敏感性分析, 提升模型可解释性; 2) 采用弱标注、无监督[86] 或半监督[87] 方法, 在现有数据库基础上, 尽可能利用无标签数据, 提取域共同特征表示, 拓展模型泛化能力[79]。3) 在隐藏层数、节点和超参数调整时根据算法性能反馈设置实时优化设置[89], 如通过 LSTM 等类深度学习算法获取时序信息[90] 和通过采用多个模型检测隐藏变量。

鉴于上述分析, 概要深度学习在语音情感识别领域中的发展趋势:

1) 当前机器缺乏理论推理是因为没有常识, 故张钹院士提出将感知和认知投射到语义向量空间 (特征向量空间和符号向量投射到一个空间, 该空间称之为语义向量空间), 从而为感知和认知建立统一的理论框架, 统一处理, 解决理解问题。

2) 鉴于当前深度学习存在需要大量的数据集、知识无法积累等问题, 可采用迁移学习或基于度量、模型和优化的元学习方法, 实现小样本学习, 在一定程度上实现知识积累。

3) 使用更高算力的设备。如清华大学团队在《自然》上介绍的天机芯片, 灵活使用类脑计算和深度学习算法[91], 使语音情感识别运算结果尽可能达到模型理论推导的预期结果。

6　应用与发展趋势

语音情感识别源于简单、交互直观, 应用广泛。商务谈判中, 提取语音和表情进行情感计算并利用云计算远程存储, 实时获取情感信息, 辅助用户决策[91]; 远程教育系统对学员的语音、姿态和表情进行分析, 及时反馈学习状态, 调整教学计划, 实现个性化培养; 可穿戴设备监测用户情感状态、通过语音交互, 辅助医生进行精神分裂症患者[92]、情绪障碍患者[84]、自闭症儿童的状态监测和治疗[93]。公共服务中, 呼叫中心根据语音情感值筛选紧急电话[85]。机器人通过语音情感识别, 提升了人机交互[94] 的针对性和准确性, 为实现机器情感仿生奠定了基础。

语音情感识别显示出广阔的应用前景。研究者对语音情感识别深入探索, 推进了其理论研究和实际应用。目前已基本实现安静环境下的语音情感识别。而嘈杂环境下的语音情感识别尚有待深入; 此外, 现有情感语音数据库总体语料不足, 特别是自然型数据库。同时标注离散情感和维度情感的语音数据库数量较少, 如何对数据库同时标注, 尚未形成广泛认可的体系; 标注方法较少、典型特征近期没有得到重大突破、语音情感识别理论需进一步完善。综上所述, 语音情感识别尚未达到成熟阶段, 需进行语料库的丰富、理论的加强和方法的创新。尚待解决的问题和未来发展

趋势包括：

1) 数据库不足且缺少广泛认可的数据库。目前数据库多在实验室环境中采集，而情感在现实世界中比实验室环境中表达方式更复杂，自然型数据库可有效解决这一难题，但其语料目前较少，需进一步丰富。可考虑采用跨语音库、合并语音库等方法扩充语料或基于当前数据自动生成样本填充数据库，如 GAN。此外，建立广泛认可的数据库，如图像处理领域中的 ImageNet，进一步集结科研力量，推动深入研究。

2) 标注多样化。同时对数据库标注维度情感和离散情感，二者互为补充，相互验证。促进从离散情感研究转向更精确的维度情感研究，为人机高级交互奠定基础。此外，目前广泛使用的标注方法和专业辅助工具较少，需进一步丰富。

3) 特征挖掘是可提升的方向。语音情感识别领域中，情感特征丰富，但是现有典型特征较少，多为声学特征，且近期没有提出类似或更优的特征，故特征挖掘需要进一步加强。

4) 特征提取方法需改良。语音情感数据样本有限，而语音识别语料数量庞大，如何采用无监督或半监督方式，自动学习无标签语料提取有效情感特征，是一个难题。此外，基于深度学习的特征提取依赖所搭建网络的具体结构，对比点云领域的 PointNet 和目标检测领域的 YOLO，目前情感识别领域缺乏对语音情感特征敏感且被广泛认可的网络结构，如何搭建一个擅于提取情感特征的专用网络结构已成为新的研究热点。

5) 情感识别算法需深入研究。如何使用跨领域算法微调，获取更优的初始化参数，提升识别收敛性与准确性，将成为新的研究热点；同时，区分性别、文化差异等来优化识别效果，是研究的一个方向。目前实时连续语音识别很少成功，可通过模拟复杂背景环境、提升算法的鲁棒性、提高模型运算效率和改善资源分配等方法解决。此外，高级语音情感识别应能模拟人对语音信息的处理，实现机器情感仿生，这些功能目前暂未实现。该课题需神经学、脑科学等多学科支持，值得深入研究。

7　结束语

本文从情感描述模型、情感语音数据库、特征提取与降维、情感分类与回归算法 4 个环节对语音情感识别进行综述，并着重分析深度学习算法在特征提取、情感分类与回归算法方面的研究进展。总体而言，语音情感识别研究体系较为完整，深度学习因其优越的性能也在特征提取和情感分类与回归中受到了广泛关注、显示出广阔的应用前景。最后总结了语音情感识别领域中语音情感数据库、标注方法和专业辅助工具、语音情感特征挖掘、语音情感特征提取、情感识别算法 5 个方面的尚待解决的问题，预测了未来的发展趋势。

参考文献：

[1] PRAVENA D, GOVIND D. Significance of incorporating excitation source parameters for improved emotion recognition from speech and electroglottographic signals[J]. International journal of speech technology, 2017, 20(4): 787–797.

[2] MIXDORFF H, HÖNEMANN A, RILLIARD A, et al. Audio-visual expressions of attitude: how many different attitudes can perceivers decode?[J]. Speech communication, 2017, 95: 114–126.

[3] BUITELAAR P, WOOD I D, NEGI S, et al. MixedEmotions: an open-source toolbox for multimodal emotion analysis[J]. IEEE transactions on multimedia, 2018, 20(9): 2454–2465.

[4] SAPIŃSKI T, KAMIŃSKA D, PELIKANT A, et al. Emotion recognition from skeletal movements[J]. Entropy, 2019, 21(7): 646.

[5] PARIS M, MAHAJAN Y, KIM J, et al. Emotional speech processing deficits in bipolar disorder: the role of mismatch negativity and P3a[J]. Journal of affective disorders, 2018, 234: 261–269.

[6] SCHELINSKI S, VON KRIEGSTEIN K. The relation between vocal pitch and vocal emotion recognition abilities in people with autism spectrum disorder and typical development[J]. Journal of autism and developmental disorders, 2019, 49(1): 68–82.

[7] SWAIN M, ROUTRAY A, KABISATPATHY P. Databases, features and classifiers for speech emotion recognition: a review[J]. International journal of speech technology, 2018, 21(1): 93–120.

[8] 韩文静, 李海峰, 阮华斌, 等. 语音情感识别研究进展综述 [J]. 软件学报, 2014, 25(1): 37–50.
HAN Wenjing, LI Haifeng, RUAN Huabin, et al. Review on speech emotion recognition[J]. Journal of software, 2014, 25(1): 37–50.

[9] 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 [J]. 计算机学报, 2018, 41(12): 2833–2851.
LIU Zhentao, XU Jianping, WU Min, et al. Review of emotional feature extraction and dimension reduction method for speech emotion recognition[J]. Chinese journal of computers, 2018, 41(12): 2833–2851.

[10] KRATZWALD B, ILIĆ S, KRAUS M, et al. Deep learning for affective computing: text-based emotion recogni-

tion in decision support[J]. Decision support systems, 2018, 115: 24–35.

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

[12] EKMAN P, FRIESEN W V, O'SULLIVAN M, et al. Universals and cultural differences in the judgments of facial expressions of emotion[J]. Journal of personality and social psychology, 1987, 53(4): 712–717.

[13] SCHULLER B W. Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends[J]. Communications of the ACM, 2018, 61(5): 90–99.

[14] 乐国安, 董颖红. 情绪的基本结构: 争论、应用及其前瞻 [J]. 南开学报(哲学社会科学版), 2013(1): 140–150.
YUE Guoan, DONG Yinghong. On the categorical and dimensional approaches of the theories of the basic structure of emotions[J]. Nankai journal (philosophy, literature and social science edition), 2013(1): 140–150.

[15] 李霞, 卢官明, 闫静杰, 等. 多模态维度情感预测综述 [J]. 自动化学报, 2018, 44(12): 2142–2159.
LI Xia, LU Guanming, YAN Jingjie, et al. A survey of dimensional emotion prediction by multimodal cues[J]. Acta automatica sinica, 2018, 44(12): 2142–2159.

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

[17] RUSSELL J A. A circumplex model of affect[J]. Journal of personality and social psychology, 1980, 39(6): 1161–1178.

[18] YIK M S M, RUSSELL J A, BARRETT L F. Structure of self-reported current affect: integration and beyond[J]. Journal of personality and social psychology, 1999, 77(3): 600–619.

[19] PLUTCHIK R. The nature of emotions: human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice[J]. American scientist, 2001, 89(4): 344–350.

[20] ZHALEHPOUR S, ONDER O, AKHTAR Z, et al. BAUM-1: a spontaneous audio-visual face database of affective and mental states[J]. IEEE transactions on affective computing, 2017, 8(3): 300–313.

[21] WANG Wenwu. Machine audition: principles, algorithms and systems[M]. New York: Information Science Reference, 2010, 398–423.

[22] WANG Yongjin, GUAN Ling. Recognizing human emotional state from audiovisual signals[J]. IEEE transactions on multimedia, 2008, 10(4): 659–668.

[23] BURKHARDT F, PAESCHKE A, ROLFES M, et al. A database of German emotional speech[C]//INTERSPEECH 2005. Lisbon, Portugal, 2005: 1517–1520.

[24] MARTIN O, KOTSIA I, MACQ B, et al. The eNTERFACE' 05 audio-visual emotion database[C]//Proceedings of the 22nd International Conference on Data Engineering Workshops. Atlanta, USA, 2006: 1–8.

[25] LIVINGSTONE S R, RUSSO F A. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): a dynamic, multimodal set of facial and vocal expressions in north American English[J]. PLoS one, 2018, 13(5): e0196391.

[26] STEIDL S. Automatic classification of emotion-related user states in spontaneous children's speech[M]. Erlangen, Germany: University of Erlangen-Nuremberg, 2009: 1–250

[27] GRIMM M, KROSCHEL K, NARAYANAN S. The Vera am Mittag German audio-visual emotional speech database[C]//Proceedings of 2008 IEEE International Conference on Multimedia and Expo. Hannover, Germany, 2008: 865–868.

[28] BUSSO C, BULUT M, LEE C C, et al. IEMOCAP: interactive emotional dyadic motion capture database[J]. Language resources and evaluation, 2008, 42(4): 335–359.

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

[30] METALLINOU A, YANG Zhaojun, LEE C, et al. The USC CreativeIT database of multimodal dyadic interactions: from speech and full body motion capture to continuous emotional annotations[J]. Language resources and evaluation, 2016, 50(3): 497–521.

[31] MCKEOWN G, VALSTAR M, COWIE R, et al. The SEMAINE database: annotated multimodal records of emotionally colored conversations between a person and a limited agent[J]. IEEE transactions on affective computing, 2012, 3(1): 5–17.

[32] 饶元, 吴连伟, 王一鸣, 等. 基于语义分析的情感计算技术研究进展 [J]. 软件学报, 2018, 29(8): 2397–2426.
RAO Yuan, WU Lianwei, WANG Yiming, et al. Research progress on emotional computation technology based on semantic analysis[J]. Journal of software, 2018, 29(8): 2397–2426.

[33] WANG Yiming, RAO Yuan, WU Lianwei. A review of sentiment semantic analysis technology and progress[C]//Proceedings of 2017 13th International Conference on Computational Intelligence and Security. Hong Kong, China, 2017: 452–455.

[34] MORRIS J D. Observations: SAM: the self-assessment manikin—an efficient cross-cultural measurement of emotional response[J]. Journal of advertising research, 1995, 35(6): 63–68.

[35] 夏凡, 王宏. 多模态情感数据标注方法与实现 [C]//第一届建立和谐人机环境联合学术会议 (HHME2005) 论文集. 北京, 2005: 1481–1487.
XIA Fan, WANG Hong. Multi-modal affective annotation method and implementation[C]//The 1th Jiont Conference on Harmonious Human Machine Environment (HHME2005). Beijing, 2005: 1481–1487.

[36] COWIE R, DOUGLAS-COWIE E, SAVVIDOU S, et al. FEELTRACE: an instrument for recording perceived emotion in real time[C]//Proceedings of the 2000 ISCA Tutorial and Research Workshop on Speech and Emotion. Newcastle, United Kingdom, 2000: 19–24.

[37] 陈炜亮, 孙晓. 基于 MFCCG-PCA 的语音情感识别 [J]. 北京大学学报(自然科学版), 2015, 51(2): 269–274.
CHEN Weiliang, SUN Xiao. Mandarin speech emotion recognition based on MFCCG-PCA[J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2015, 51(2): 269–274.

[38] SUN Linhui, FU Sheng, WANG Fu. Decision tree SVM model with Fisher feature selection for speech emotion recognition[J]. EURASIP journal on audio, speech, and music processing, 2019, 2019: 2.

[39] NASSIF A B, SHAHIN I, ATTILI I, et al. Speech recognition using deep neural networks: a systematic review[J]. IEEE access, 2019, 7: 19143–19165.

[40] ZHANG Shiqing, ZHANG Shiliang, HUANG Tiejun, et al. Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching[J]. IEEE transactions on multimedia, 2018, 20(6): 1576–1590.

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

[42] 王玮蔚, 张秀再. 基于变分模态分解的语音情感识别方法 [J]. 应用声学, 2019, 38(2): 237–244.
WANG Weiwei, ZHANG Xiuzai. Speech emotion recognition based on variational mode decomposition[J]. Journal of applied acoustics, 2019, 38(2): 237–244.

[43] 王忠民, 刘戈, 宋辉. 基于多核学习特征融合的语音情感识别 [J]. 计算机工程, 2019, 45(08): 248–254.
WANG Zhongmin, LIU Ge, SONG Hui. Feature fusion based on multiple kernal learning for speech emotion recognition[J]. Computer engineering, 2019, 45(08): 248–254.

[44] 卢官明, 袁亮, 杨文娟, 等. 基于长短期记忆和卷积神经网络的语音情感识别 [J]. 南京邮电大学学报(自然科学版), 2018, 38(5): 63–69.
LU Guanming, YUAN Liang, YANG Wenjuan, et al. Speech emotion recognition based on long short-term memory and convolutional neural networks[J]. Journal of Nanjing University of Posts and Telecommunications (Natural Science Edition), 2018, 38(5): 63–69.

[45] ÖZSEVEN T. Investigation of the effect of spectrogram images and different texture analysis methods on speech emotion recognition[J]. Applied acoustics, 2018, 142: 70–77.

[46] JIANG Wei, WANG Zheng, JIN J S, et al. Speech emotion recognition with heterogeneous feature unification of deep neural network[J]. Sensors, 2019, 19(12): 2730.

[47] TORRES-BOZA D, OVENEKE M C, WANG Fengna, et al. Hierarchical sparse coding framework for speech emotion recognition[J]. Speech communication, 2018, 99: 80–89.

[48] MAO Qirong, XUE Wentao, RAO Qiru, et al. Domain adaptation for speech emotion recognition by sharing priors between related source and target classes[C]//Proceedings of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing. Shanghai, China, 2016: 2608–2612.

[49] JIN Qin, LI Chengxin, CHEN Shizhe, et al. Speech emotion recognition with acoustic and lexical features[C]//Proceedings of 2015 IEEE International Conference on Acoustics, Speech and Signal Processing. Brisbane, QLD, Australia, 2015: 4749–4753.

[50] SCHULLER B. Recognizing affect from linguistic information in 3D continuous space[J]. IEEE transactions on affective computing, 2011, 2(4): 192–205.

[51] DIMOULAS C A, KALLIRIS G M. Investigation of wavelet approaches for joint temporal, spectral and cepstral features in audio semantics[C]//Audio Engineering Society Convention. New York, USA, 2013.

[52] TAWARI A, TRIVEDI M M. Speech emotion analysis: exploring the role of context[J]. IEEE transactions on multimedia, 2010, 12(6): 502–509.

[53] QUIROS-RAMIREZ M A, ONISAWA T. Considering cross-cultural context in the automatic recognition of emotions[J]. International journal of machine learning and cybernetics, 2015, 6(1): 119–127.

[54] WU Xixin, LIU Songxiang, CAO Yuewen, et al. Speech emotion recognition using capsule networks[C]//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, United Kingdom, 2019: 6695–6699.

[55] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks [C]//Proceedings of the 25th International Conference on Neural Information Processing Systems. Red Hook, United States, 2012: 1097–1105.

[56] ZHAO Jianfeng, MAO Xia, CHEN Lijiang. Speech emotion recognition using deep 1D & 2D CNN LSTM networks[J]. Biomedical signal processing and control, 2019, 47: 312–323.

[57] 张丽, 吕军, 强彦, 等. 基于深度信念网络的语音情感识别 [J]. 太原理工大学学报, 2019, 50(1): 101–107.
ZHANG Li, LV Jun, QIANG Yan, et al. Emotion recognition based on deep belief network[J]. Journal of Taiyuan University of Technology, 2019, 50(1): 101–107.

[58] ABDELWAHAB M, BUSSO C. Domain adversarial for acoustic emotion recognition[J]. IEEE/ACM transactions on audio, speech, and language processing, 2018, 26(12): 2423–2435.

[59] MENG Zhong, LI Jinyu, CHEN Zhuo, et al. Speaker-invariant training via adversarial learning[C]//Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. Calgary, AB, Canada, 2018: 5969–5973.

[60] VAN DER MAATEN L, HINTON G. Visualizing data using t-SNE[J]. Journal of machine learning research, 2008, 9: 2579–2605.

[61] BOERSMA P, WEENINK D. Praat, a system for doing phonetics by computer[J]. Glot international, 2002, 5(9/10): 341–345.

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

[63] ÖZSEVEN T, DÜĞENCI M. SPeech ACoustic (SPAC): a novel tool for speech feature extraction and classification[J]. Applied acoustics, 2018, 136: 1–8.

[64] 孙凌云, 何博伟, 刘征, 等. 基于语义细胞的语音情感识别 [J]. 浙江大学学报(工学版), 2015, 49(6): 1001–1008.
SUN Lingyun, HE Bowei, LIU Zheng, et al. Speech emotion recognition based on information cell[J]. Journal of Zhejiang University (Engineering Science), 2015, 49(6): 1001–1008.

[65] SCHULLER B, BATLINER A, STEIDL S, et al. Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge[J]. Speech communication, 2011, 53(9/10): 1062–1087.

[66] GHARAVIAN D, BEJANI M, SHEIKHAN M. Audio-visual emotion recognition using FCBF feature selection method and particle swarm optimization for fuzzy ARTMAP neural networks[J]. Multimedia tools and applications, 2016, 76(2): 2331–2352.

[67] 王艳, 胡维平. 基于 BP 特征选择的语音情感识别 [J]. 微电子学与计算机, 2019, 36(5): 14–18.
WANG Yan, HU Weiping. Speech emotion recognition based on BP feature selection[J]. Microelectronics & computer, 2019, 36(5): 14–18.

[68] 孙颖, 姚慧, 张雪英, 等. 基于混沌特性的情感语音特征提取 [J]. 天津大学学报(自然科学与工程技术版), 2015, 48(8): 681–685.
SUN Ying, YAO Hui, ZHANG Xueying, et al. Feature extraction of emotional speech based on chaotic characteristics[J]. Journal of Tianjin University (Science and Technology), 2015, 48(8): 681–685.

[69] 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 [J]. 清华大学学报(自然科学版), 2018, 58(4): 347–351.
SONG Peng, ZHENG Wenming, ZHAO Li. Joint subspace learning and feature selection method for speech emotion recognition[J]. Journal of Tsinghua University (Science and Technology), 2018, 58(4): 347–351.

[70] EYBEN F, SCHERER K R, SCHULLER B W, et al. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing[J]. IEEE transactions on affective computing, 2016, 7(2): 190–202.

[71] ÖZSEVEN T. A novel feature selection method for speech emotion recognition[J]. Applied acoustics, 2019, 146: 320–326.

[72] 姜晓庆, 夏克文, 夏莘媛, 等. 采用半定规划多核 SVM 的语音情感识别 [J]. 北京邮电大学学报, 2015, 38(S1): 67–71.
JIANG Xiaoqing, XIA Kewen, XIA Xinyuan, et al. Speech emotion recognition using semi-definite programming multiple-kernel SVM[J]. Journal of Beijing University of Posts and Telecommunications, 2015, 38(S1): 67–71.

[73] ZHENG Weiqiao, YU Jiasheng, ZOU Yuexian. An experimental study of speech emotion recognition based on deep convolutional neural networks[C]//Proceedings of 2015 International Conference on Affective Computing and Intelligent Interaction. Xi'an, China, 2015: 827–831.

[74] SHAHIN I, NASSIF A B, HAMSA S. Emotion recognition using hybrid Gaussian mixture model and deep neural network[J]. IEEE access, 2019, 7: 26777–26787.

[75] SAGHA H, JUN Deng, GAVRYUKOVA M, et al. Cross lingual speech emotion recognition using canonical correlation analysis on principal component subspace[C]//Proceedings of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing. Shanghai, China, 2016: 5800–5804.

[76] 陈师哲, 王帅, 金琴. 多文化场景下的多模态情感识别 [J]. 软件学报, 2018, 29(4): 1060–1070.
CHEN Shizhe, WANG Shuai, JIN Qin. Multimodal emotion recognition in multi-cultural conditions[J]. Journal of software, 2018, 29(4): 1060–1070.

[77] 刘颖, 贺聪, 张清芳. 基于核相关分析算法的情感识别模型 [J]. 吉林大学学报(理学版), 2017, 55(6):

1539–1544.
LIU Ying, HE Cong, ZHANG Qingfang. Emotion recognition model based on kernel correlation analysis algorithm[J]. Journal of Jilin University (Science Edition), 2017, 55(6): 1539–1544.
[78] MA Yaxiong, HAO Yixue, CHEN Min, et al. Audio-Visual Emotion Fusion (AVEF): a deep efficient weighted approach[J]. Information fusion, 2019, 46: 184–192.
[79] HUANG Yongming, TIAN Kexin, WU Ao, et al. Feature fusion methods research based on deep belief networks for speech emotion recognition under noise condition[J]. Journal of ambient intelligence and humanized computing, 2019, 10(5): 1787–1798.
[80] MANNEPALLI K, SASTRY P N, SUMAN M. A novel adaptive fractional deep belief networks for speaker emotion recognition[J]. Alexandria engineering journal, 2017, 56(4): 485–497.
[81] XU Xinzhou, DENG Jun, COUTINHO E, et al. Connecting subspace learning and extreme learning machine in speech emotion recognition[J]. IEEE transactions on multimedia, 2019, 21(3): 795–808.
[82] TON-THAT A H, CAO N T. Speech emotion recognition using a fuzzy approach[J]. Journal of intelligent & fuzzy systems, 2019, 36(2): 1587–1597.
[83] ZHANG Biqiao, PROVOST E M, ESSL G. Cross-corpus acoustic emotion recognition with multi-task learning: seeking common ground while preserving differences[J]. IEEE transactions on affective computing, 2019, 10(1): 85–99.
[84] HUANG Kunyi, WU C H, SU M H. Attention-based convolutional neural network and long short-term memory for short-term detection of mood disorders based on elicited speech responses[J]. Pattern recognition, 2019, 88: 668–678.
[85] YOON S A, SON G, KWON S. Fear emotion classification in speech by acoustic and behavioral cues[J]. Multimedia tools and applications, 2019, 78(2): 2345–2366.
[86] 袁非牛, 章琳, 史劲亭, 等. 自编码神经网络理论及应用综述 [J]. 计算机学报, 2019, 42(1): 203–230.
YUAN Feiniu, ZHANG Lin, SHI Jinting, et al. Theories and applications of auto-encoder neural networks: a literature survey[J]. Chinese journal of computers, 2019, 42(1): 203–230.
[87] 林懿伦, 戴星原, 李力, 等. 人工智能研究的新前线: 生成式对抗网络 [J]. 自动化学报, 2018, 44(5): 775–792.
LIN Yilun, DAI Xingyuan, LI Li, et al. The new frontier of AI research: generative adversarial networks[J]. Acta automatica sinica, 2018, 44(5): 775–792.
[88] ZHOU Jie, HUANG J X, CHEN Qin, et al. Deep learning for aspect-level sentiment classification: survey, vision, and challenges[J]. IEEE access, 2019, 7: 78454–78483.
[89] O'SHAUGHNESSY D. Recognition and processing of speech signals using neural networks[J]. Circuits, systems, and signal processing, 2019, 38(8): 3454–3481.
[90] XIE Yue, LIANG Ruiyu, LIANG Zhenlin, et al. Attention-based dense LSTM for speech emotion recognition[J]. IEICE transactions on information and systems, 2019, E102.D(7): 1426–1429.
[91] PEI Jing, DENG Lei, SONG Sen, et al. Towards artificial general intelligence with hybrid Tianjic chip architecture. Nature, 2019, 572: 106−111
[92] QIAN Yongfeng, LU Jiayi, MIAO Yiming, et al. AIEM: AI-enabled affective experience management[J]. Future generation computer systems, 2018, 89: 438–445.
[93] LADO-CODESIDO M, PÉREZ C M, MATEOS R, et al. Improving emotion recognition in schizophrenia with “VOICES”: an on-line prosodic self-training[J]. PLoS one, 2019, 14(1): e0210816.
[94] CUMMINS N, BAIRD N, SCHULLER B W. Speech analysis for health: current state-of-the-art and the increasing impact of deep learning[J]. Methods, 2018, 151: 41–54.
[95] LIU Zhentao, XIE Qiao, WU Min, et al. Speech emotion recognition based on an improved brain emotion learning model[J]. Neurocomputing, 2018, 309: 145–156.

作者简介：

高庆吉，教授，博士，中国航空学会青年工作委员会副主任委员，制导导航与控制委员会委员，主要研究方向为人工智能、智能机器人，主持并完成国家自然科学基金项目、“863”计划项目及省部级科研项目 10 余项。发表学术论文百余篇。

赵志华，硕士研究生，主要研究方向为多模态情感计算、情感识别。

徐达，硕士研究生，主要研究方向为机器学习、情感识别。