DOI:10.16644/j.cnki.cn33-1094/tp.2025.03.001

# 基于深度学习的情感识别研究综述*

张艳倪，汤　敏，王坤赤

（南通大学信息科学技术学院，江苏 南通 226019）

**摘　要：**深度学习是基于深层神经网络的机器学习，在情感识别技术领域已取得了显著的进步。本文综述深度学习在情感识别上的应用，分析其技术框架与研究现状；通过研究常用数据库特点、语音特征提取和情感分类技术的应用，进一步讨论基于深度学习的情感识别模型的评价指标体系，并对相关文献进行分类综述。研究表明，基于深度学习的情感识别准确率已达80%以上，应用于现实生活可期，本文对未来研究方向进行了展望。

**关键词：**情感识别；语音特征提取；深度学习；神经网络

**中图分类号：**TP391.4　　**文献标识码：**A　　**文章编号：**1006-8228(2025)03-01-05

## Review of Emotion Recognition Based on Deep Learning

Zhang Yanni, Tang Min, Wang Kunchi

(*School of Information Science and Technology, Nantong University, Nantong, Jiangsu 226019, China*)

**Abstract:** Deep learning is a type of machine learning based on deep neural networks, and it has made remarkable progress in the field of emotion recognition technology. The application of deep learning in emotion recognition is reviewed in this paper, the framework of deep learning is introduced, the current situation of emotion recognition research is analyzed, including the characteristics of the database, speech feature extraction and emotion classification. Then, the relevant literature is reviewed and the evaluation indexes of emotion recognition are introduced. The accuracy rate of emotion recognition based on deep learning has exceeded 80%, and it is possible to be applied in real life after further improvement and optimization in the future, the future development direction of emotion recognition based on deep learning is looked forward to.

**Keywords:** Emotion recognition; Speech feature extraction; Deep learning; Neural network

## 0 引言

语音作为人类最常用的交流方式，蕴含丰富的情感信息。机器从语音自动识别人类情感及相关状态的过程，即语音情感识别(Speech Emotion Recognition, SER)，是实现自然人机交互的关键技术。该技术在心理健康评估、驾驶员情绪监测、医疗诊断等领域具有重要应用价值。

深度学习特指基于深层神经网络模型的机器学习。相较于传统人工神经网络，深度学习强调结构深度，通常有五层、六层，甚至十多层隐藏节点。深度学习具有从数据中自动提取特征的能力，已广泛应用于计算机视觉、自然语言处理、多模态数据分析、图像处理及语音识别等领域。

近年来，情感计算和语音情感识别技术引起学界高度关注。本文综述深度学习在情感识别领域的研究现状，包括：①介绍卷积神经网络(Convolutional Neural Network, CNN)[1]、深度残差收缩网络(Deep Residual Shrinkage Network, DRSN)[2]、循环网络(Recurrent Neural Network, RNN)[3]、长短时记忆网络(Long-Short-Term Memory, LSTM)[4]。②从传统声学特征与深度学习方法两方面，介绍语音特征提取、常用情感识别数据库及特点、语音情感分类。③情感识别模型应用和评价指标。④总结目前存在问题，展望未来发展趋势。

**收稿日期：**2024-12-24

***基金项目：**国家自然科学基金项目“大深度饱和潜水智能氦语音通信系统研究”(No.62371261)

**作者简介：**张艳倪(2005-)，女，江苏徐州人，本科在读，主要研究方向：信号处理。

**通讯作者：**王坤赤(1971-)，男，吉林四平人，博士研究生，副教授，主要研究方向：信号处理。

## 1 主流的深度学习框架

### 1.1 CNN

卷积神经网络是由卷积计算和深度网络组成的前馈神经网络，基本架构包括输入层、卷积层、激活层、池化层、全连接层和输出层。输入层将输入数据转换为网络可以处理的形式。卷积层是网络核心，通过局部映射将输入数据转换为特征图并进行特征提取。激活层是将特征提取结果通过非线性映射到高维非线性区间，增强网络表达能力。池化层通过下采样操作降低维度，从而减少计算复杂度并保留关键特征。全连接层将提取到的特征进行融合，由分类器输出结果。输出层则根据特征提取和转换，输出最终结果[5]。

### 1.2 DRSN

残差网络(Residual Network, ResNet)是基于残差结构的卷积神经网络，缓解深度卷积神经网络训练过程中出现梯度消失和梯度爆炸的问题，提升网络性能。而深度残差收缩网络(DRSN)是残差网络(ResNet)的改进，在数据包含噪声的情况下，通过结合深度残差网络、软阈值化、注意力机制，提高深度学习模型在处理含噪数据时的性能[2]。

### 1.3 RNN

循环网络(RNN)具有处理序列数据中时间依赖关系的能力，RNN的输出结果不但与当前输入信息以及网络权重有关，还与之前的输入信息相关。当权重初始化不当、使用饱和的激活函数、网络层数过多时，都会造成梯度消失和梯度爆炸，导致网络无法继续训练[3]。

### 1.4 LSTM

长短时记忆网络(LSTM)是循环神经网络的进阶版，通过引入单元状态和门控结构，解决RNN的长期依赖问题(梯度消失是其表现之一)。其中，遗忘门用来筛选上一刻的信息传递给当前时刻需要的信息；输入门决定当前时刻的信息有多少需要进行传递；输出门则决定单元状态在当前可以输出的信息。LSTM网络通过借助门控结构实现对信息的存储、更新和保护，门控结构通常由Sigmoid和点乘运算组成[4]。

## 2 情感识别研究现状

语音情感识别本质上是一项分类任务，训练计算机识别语音中的情感状态，进而获得一个具有区分情感能力的模型。语音情感识别流程图参见图1，主要包括三个模块：选取并建立合适的语音情感数据库；对语音信号进行预处理和特征提取；送入分类器进行训练，得到分类模型，实现情感分类[6-7]。

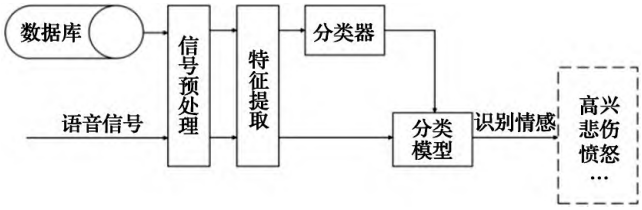


图1 语音情感识别流程图

### 2.1 情感识别常用数据库

常用的语音情感识别数据库及其特点参见表1。

表1 语音情感识别常用数据库

| 数据库名称 | 网址 | 数据库特点 |
|---|---|---|
| IEMOCAP | https://sail.usc.edu/iemocap/release_form.php | 10名专业演员在有台词或即兴对话场景下，引导出情感数据。每一段对话被切分为单句，每一句话至少由3个标注员进行离散情感和3个维度的标注。 |
| CASIA | https://gitcode.com/Open-source-toolkit/bc5e6 | 4个专业发音人，包括6种离散情感，共9600句不同发音。 |
| EMODB | http://emodb.bilderbar.info/docu/#emodb | 5男5女表演者将10句德文通过不同的情感表达出来。 |
| RAVDESS | https://zenodo.org/record/1188976 | 24名专业者用中性北美口音，展现平静、快乐、悲伤、愤怒、恐惧、惊讶和厌恶的情绪。 |
| eNTERFACE'05 | http://www.enterface.net/results | 42位不同国家的参与者(男性81%、女性19%)，听连续短篇故事，引发特定情感。 |
| MELD | http://web.eecs.umich.edu/mihalcea/downloads/MELD.Raw.tar.gz | 超过1400个对话和13,000个句子，均来自《老友记》。每句话都被标注为生气、厌恶、悲伤、快乐、中性、惊喜、恐惧等情感标签。 |
| SAVEE | http://kahlan.eps.surrey.ac.uk/savee/ | 4位男演员用7种不同情绪录音组成，共480条英式英语。 |
| TESS | https://tspace.library.utoronto.ca/handle/1807/24487 | 两位女演员在主词"Say the word"中说出一组200个目标词，并录制描绘愤怒、厌恶、恐惧、快乐、惊喜、悲伤和中性的录音，共2800个数据点。 |
| AFEW6.0 | https://cs.anu.edu.au/few/emotiw.html | 1750个短视频(训练集774个，验证集383个，测试集593个)，包含头部姿势、运动、光照、多主体和遮挡等条件，接近真实环境。 |
| RML | https://cs.anu.edu.au/few/emotiw.html | 具有多调制类型和多信噪比大样本量，可在各种条件下训练和测试模型，提高模型泛化能力和鲁棒性。 |

### 2.2 基于情感识别的特征提取

#### 2.2.1 基于传统声学的特征提取

传统声学特征主要包括韵律特征[8]、谱特征[9]以及音质特征[10]。其中，韵律特征主要表现在语音抑扬顿挫，包括能量、基频、时长等相关特征。谱特征指发声运动中声道形状的变化引起的能量变化，如线性预测系数(Linear Predictive Coding, LPC)[11]、线性预测倒谱系数(Linear Prediction Cepstral Coefficients, LPCC)[12]、Mel频率倒谱系数(Mel-frequency Cepstral Coefficients, MFCC)[13]等。音质特征表现语音是否清晰、是否易于识别，主要包括共振峰频率(format frequency)、声门参数(glottal parameter)、频率微扰(frequency disturbances)等。其中，梅尔频率倒谱系数较为常用，MFCC参数提取流程见图2。

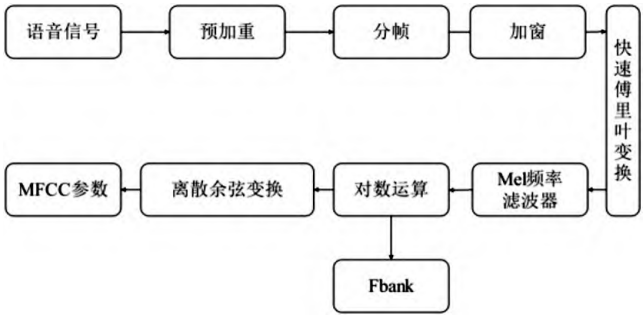


图2 MFCC参数提取流程图

#### 2.2.2 基于深度学习的特征提取

近年来，研究者尝试从浅层特征中进一步提取深层特征，或直接批量处理原始语音数据，自动提取深层特征。利用主流的深度学习框架，不仅可以从原始语音信号中提取更多、更丰富的语音情感特征，还可以自动学习特征表示，降低特征工程的难度。从语谱图(Spectrogram)中训练提取情感特征的过程参见图3。首先对语音信号进行预处理，包括预加重、分帧、加窗。然后经过傅里叶变换，生成语谱图。语谱图是语音信号在时间和频率两个维度上的表示形式，横轴表示时间，纵轴表示频率，信号强度通过颜色深浅或亮度表示。

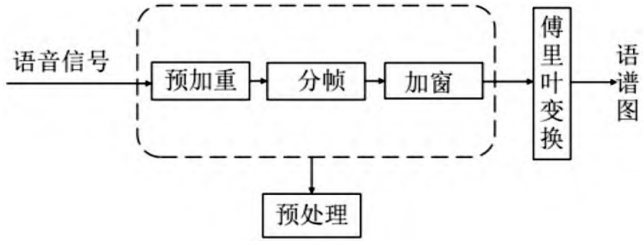


图3 语谱图的提取过程

### 2.3 基于情感识别的情感分类

语音情感识别的最终目的是分辨语音信号中的情感状态，在提取特征后，利用分类器对语音进行情感分类。传统的分类器主要包括决策树(Decision Tree)、极限学习机(Extreme Learning Machine, ELM)、支持向量机(Support Vector Machine, SVM)、高斯混合模型(Gaussian Mixture Model, GMM)、随机森林(Random Forest, RF)、K近邻(K-Nearest Neighbor, KNN)、隐马尔科夫模型(Hidden Markov Model, HMM)等[8,10,14,15]。

对语音情感分类，使用广泛的是离散情感模型[10]和连续情感模型[16]。离散型是用离散情感来描述高兴、悲伤、生气、恐惧、中性等情感，模型通俗易懂。连续型则是用连续的维度空间来描述情感，将情感定义在不同维度空间中的一点。

## 3 情感识别的文献综述和评价指标

### 3.1 基于CNN及其改进版的情感识别

文献[17]采用MFCC从音频数据中提取语音特征，结合Transformer、CNN、LSTM建立语音情感识别系统，在RAVDESS、EMODB、SAVEE和IEMOCAP数据库中准确率可达96.3%、99.86%、96.5%和85.3%。文献[18]通过CNN模型与卷积注意力模块相结合，采用MFCC进行特征提取，在RAVDESS、TESS、CREMA-D和IEMOCAP四个数据库中分别取得83%、100%、68%和63%的平均准确率。文献[19]针对现有方法对高层语音特征提取丢失大量原始信息、识别准确率不高的问题，采用CNN堆叠一个二层的LSTM，利用MFCC进行特征提取，在EMODB数据库中识别准确率达到91.74%。文献[20]采用语谱图和LLDs特征作为输入，引入自注意力机制，构建一种基于自注意力机制的双通道卷积门控循环网络模型，通过使用CCC-Loss和交叉熵损失函数共同训练模型，将两个网络的分类结果决策融合，在EMODB、RAVDESS、CASIA数据库上取得92.90%、88.54%、90.58%的识别结果。文献[21]采用随机森林进行特征选择，结合一维卷积(CNN)以及门控循环单元(GRU)构建CGRU模型，在EMODB、SAVEE、RAVDESS三个数据库上分别取得79%、69%、75%的识别准确率。

### 3.2 基于DRSN及其改进版的情感识别

文献[22]利用Mel谱图提取语音特征，采用双向门控递归单元的深度残差收缩网络(DRSN-BiGRU)和

注意力机制，自动忽略噪声信息，在CASIA、IEMOCAP和MELD数据库上准确率达到86.03%、86.07%和70.57%。文献[23]使用双向长短期记忆网络(Bi-directional Long Short-Term Memory, BiLSTM)提取语音和文本特征，采用改进后两层Transformer，将增强后的特征进行3D加权融合，在IEMOCAP和MELD数据库上取得75.9%和71.8%加权精确度。文献[24]结合CNN、LSTM、注意力机制建立基线模型，引入DRSN分配二维网络中的通道权重，提高语音情感识别的精度，采取特征层融合和决策层融合机制，在CASIA和EMODB数据库上分别取得84.93%和86.83%的平均召回率。

### 3.3 基于RNN及其改进版的情感识别

文献[25]利用RNN网络在小语音区间上计算一系列声学特征，并使用CTC损失函数考虑包含情绪化和中性成分的长话语，在IEMOCAP数据库上总体精确度和平均类别准确度均达到53%。文献[26]采用传统声学特征，在RNN中高效利用分段注意力机制，在CASIA数据库上的识别平均准确率达到84%以上。文献[27]通过音频信号提取底层特征，用一维CNN抽象提取深层特征，送入RNN捕捉时间维度上的语调变化，在CAVSR1.0数据库上达到41.51%的平均精确度。

### 3.4 基于LSTM及其改进版的情感识别

文献[13]提取信号梅尔频谱序列作为LSTM网络输入，利用LSTM网络提取语音信号时域特征，加入CNN网络提取更高层次情感特征，在eNTRAFACE′05、RML和AFEW6.0数据库上获得平均识别率分别为49.15%、5.38%和37.90%。文献[28]对含有冗余信息的手工声学特征进行选择和优化，采用CNN和LSTM对语谱图进行语音情感特征提取，在EMODB和IEMOCAP数据库中识别准确率分别达到91.24%和71.88%。文献[29]采用语音信号的梅尔频率倒谱系数MFCC作为LSTM的输入，将单元状态作为输入数据加入门限层中，将LSTM得到的情感特征输入注意力层，计算每一帧语音信号权重，在EMODB、CASIA和RAVDESS数据库上的准确率分别达到86.52%、87.98%和86.36%。文献[30]提取情感韵律、MFCC、非线性属性以及非线性几何特征，构建DMB-LSTM网络，利用深度受限玻尔兹曼机的特征重构，将不同情感特征进行融合，在EMODB数据库的平均识别率达到90%以上。

### 3.5 情感识别评价指标

在分类问题中，通常使用真正例(True Positive, TP)、假正例(False Positive, FP)、真负例(True Negative, TN)和假负例(False Negative, FN)来衡量模型分类性能，判断优劣。混淆矩阵参见表2。

表2 混淆矩阵

| 真实情况 | 预测情况 | |
|---|---|---|
| | 正例 | 负例 |
| 正例 | TP(True Position) | FN(False Negative) |
| 负例 | FP(False Position) | TN(True Negative) |

基于混淆矩阵，采用精确率(Precision)、准确率(Accuracy)、召回率(Recall)、F1-Score值、加权精确率(Weighted Accuracy, WA)、非加权精确率(Unweighted Accuracy, UA)等客观评价指标，计算公式分别如下[31]：

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \times 100\% \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{F1} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \tag{4}$$

$$\text{WA} = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N} \left(TP_i + FP_i\right)} \tag{5}$$

$$\text{UA} = \frac{\sum_{i=1}^{N} \text{Acc}_i}{N} \tag{6}$$

其中，$\text{Acc}_i = \frac{TP_i}{TP_i + FP_i}$。

## 4 结论

随着计算机技术和人工智能技术不断进步，语音情感识别技术从传统的声学特征到现在的深度神经网络，从离散情感识别到维度情感识别等多方面取得一定进展。但是，人的情感表达受到语言习惯、语言环境等多种因素影响，具有复杂性和多变性。现实环境中的噪声干扰和环境变化，也会影响情感识别的准确性。针对数据集的可靠性、情感特征提取不充分、提取无关特征过多等问题也亟待解决。

情感识别的未来发展方向包括：①语音情感标注难度大、成本高，有效利用海量的无标签语音数据，挖掘其中蕴含的情感模式，将是一种行之有效的解决方案，提高语音情感识别系统的性能。②当前语音情感识别，都是假定训练集和测试集的情感类别一致。但

在实际应用中，训练集和测试集可能出现相互并不包含的情感类别，这可能会成为语音情感识别的研究热点。

**参考文献(References):**


[1] ALI H, TRAN S N, BENETOS E, et al. Speaker recognition with hybrid features from a deep belief network[J]. Neural Computing and Applications, 2018, 29(6):13–19.

[2] 左全胜.基于深度残差收缩网络的自然情境下的多模态情感识别研究[D].哈尔滨:哈尔滨理工大学,2023.

[3] ELMAN J L. Finding Structure in Time[J]. Cognitive Science,1990,14(2):179–211.

[4] MA Yukun, PENG Haiyun, CAMBRIA E. Targeted aspect–based sentiment analysis via embedding commonsense knowledge into an attentive LSTM[A]. Proceedings of the AAAI conference on artificial intelligence[C]. New Orleans, LA, USA: Association for the Advancement of Artificial Intelligence,2018,32(1):5876–5883.

[5] WANG Hongxia, ZHOU Jiaqi, GU Chenghao, et al. Design of activation function in CNN for image classification[J]. Journal of Zhejiang University (Engineering Science),2019,53(7):1363–1373.

[6] AKCAY M B, OGUZ K. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers[J]. Speech Communication,2020,116:56–76.

[7] THIRIPURASUNDARI D, BHANGALE K, AASHRITHA V, et al. Speech emotion recognition for human–computer interaction[J]. International Journal of Speech Technology,2024,27(3):817–830.

[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.

[9] 肖宇锌.基于深度学习的语音情感分类器研究[D].长沙:湖南大学,2020.

[10] 郑婉璐.基于领域对抗训练的跨数据库语音情感识别[D].南京:东南大学,2022.

[11] KADIRI S, GANGAMOHAN P, GANGASHETTY S, et al. Excitation features of speech for emotion recognition using neutral speech as reference[J]. Circuits Systems and Signal Processing,2020,39(9):4459–4481.

[12] MAKOWSKI R, HOSSA R. Voice activity detection with quasi–quadrature filters and GMM decomposition for speech and noise[J]. Applied Acoustics, 2020, 166(10): 107344.

[13] 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,2018,38(5):63–69.

[14] 屠彦辉.复杂场景下基于深度学习的鲁棒性语音识别的研究[D].合肥:中国科学技术大学,2019.

[15] 徐新洲.基于情感特征信息增强的语音情感识别研究[D].南京:东南大学,2017.

[16] 徐华南.基于深度学习的语音情感识别研究[D].南京:南京信息工程大学,2021.

[17] ALKHAMALI E A, ALLINJAWI A, ASHARI R B. Combining transformer, convolutional neural network, and long short–term memory architectures: a novel ensemble learning technique that leverages multi–acoustic features for speech emotion recognition in distance education classrooms applied sciences[J]. Applied Sciences,2024,14(12):2076–3417.

[18] FRANCESCO A D R, FABIO C C, NICOLA C. Speech emotion recognition and deep learning: an extensive validation using convolutional neural networks[J]. IEEE Access,2023,11:116638–116649.

[19] 杨明极,张家彬.基于深度神经网络的语音情感识别方法[J].科学技术与工程,2019,19(8):127–131.

[20] 孙韩玉,黄丽霞,张雪英,等.基于双通道卷积门控循环网络的语音情感识别[J].计算机工程与应用,2023,59(2):170–177.

[21] 郑艳,陈家楠,吴凡,等.基于CGRU模型的语音情感识别研究与实现[J].东北大学学报(自然科学版), 2020,41(12):1680–1685.

[22] TIAN Han, ZHANG Zhu, REN Mingyuan, et al. Speech emotion recognition based on deep residual shrinkage network[J]. Electronics,2023,12(11):2079–9292.

[23] 张俊丰.基于多模态特征融合的语音情感识别研究[D].保定:河北大学,2022.

[24] 李瑞航,吴红兰,孙有朝,等.基于深度残差收缩网络多特征融合语音情感识别[J].数据采集与处理,2022,37(3):542–554.

[25] 余华,颜丙聪.基于CTC–RNN的语音情感识别方法[J].电子器件,2020,43(4):934–937.

[26] 蒯红权,吴建华,吴亮.基于注意力机制的深度循环神经网络的语音情感识别[J].电子器件,2022,45(1):139–142.

[27] YE Jiayin, ZHENG Wenming, LI Yang, 等. Multimodal emotion recognition based on deep neural network[J]. Journal of Southeast University (English Edition),2017, 33(4):444–447.

[28] 丁楠.基于特征学习的语音情感识别研究[D].南京:南京邮电大学,2023.

[29] 陈巧红,于泽源,孙麒,等.基于注意力机制与LSTM的语音情绪识别[J].浙江理工大学学报(自然科学版),2020,43(6):815–822.

[30] 高帆,张雪英,黄丽霞,等.基于DBM–LSTM的多特征语音情感识别[J].计算机工程与设计,2020,41(2):465–470.

[31] 章逸凡.基于语音和文本双模态融合的儿童情感识别研究[D].哈尔滨:哈尔滨理工大学,2023.