ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 39,No. 2,Mar. 2024,pp. 424－432
DOI:10. 16337/j. 1004-9037. 2024. 02. 015

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于多任务学习的语音情感识别

李云峰[1]，闫祖龙[1]，高　天[2]，方　昕[2]，邹　亮[1]

(1. 中国矿业大学信息与控制工程学院,徐州 221116;2. 科大讯飞股份有限公司核心研发平台,合肥 230088)

摘　要：在近期的语音情感识别研究中,研究人员尝试利用深度学习模型从语音信号中识别情感。然而,传统基于单任务学习的模型对语音的声学情感信息关注度不足,导致情感识别的准确率较低。鉴于此,本文提出了一种基于多任务学习、端到端的语音情感识别网络,以挖掘语音中的声学情感,提升情感识别的准确率。为避免采用频域特征造成的信息损失,本文利用基于时域信号的Wav2vec2.0自监督网络作为模型的主干网络,提取语音的声学特征和语义特征,并利用注意力机制将两类特征进行融合作为自监督特征。为了充分利用语音中的声学情感信息,使用与情感有关的音素识别作为辅助任务,通过多任务学习挖掘自监督特征中的声学情感。在公开数据集IEMOCAP上的实验结果表明,本文提出的多任务学习模型实现了76.0%的加权准确率和76.9%的非加权准确率,相比传统单任务学习模型性能得到了明显提升。同时,消融实验验证了辅助任务和自监督网络微调策略的有效性。

关键词：深度学习;多任务学习;语音情感识别;自监督模型;微调策略

中图分类号：TP183　　**文献标志码**:A

Speech Emotion Recognition with Multi-task Learning

LI Yunfeng[1], YAN Zulong[1], GAO Tian[2], FANG Xin[2], ZOU Liang[1]

(1. School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China;2. Research & Development Group, iFLYTEK Co. Ltd., Hefei 230088, China)

Abstract: In recent speech emotion recognition, researchers attempt to identify emotion from speech signals using deep learning models. However, traditional single-task learning-based models do not pay enough attention to speech acoustic emotional information, resulting in low accuracy of emotion recognition. In view of this, this paper proposes a multi-task learning, end-to-end speech emotion recognition network to mine acoustic emotion in speech and improve the accuracy of emotion recognition. In order to avoid the loss of information caused by using frequency domain features, this paper adopts the Wav2vec2.0 as the backbone network of the model to extract the acoustic and semantic features of speech, and the attention mechanism is used to integrate the two kinds of features as self-supervised features. To make full use of the acoustic sentiment information in speech, using emotion-related phoneme recognition as an auxiliary task, a multi-task learning model is used to mine acoustic sentiment in self-supervised features. Experimental results on the public dataset IEMOCAP show that, the proposed multi-task learning model achieves a weighted accuracy rate of 76.0% and an unweighted accuracy rate of 76.9%, with significantly improved model performance compared to the traditional single-task learning model.

基金项目:科技创新2030——“新一代人工智能”重大项目(2020AAA0107300);徐州市基础研究计划项目(KC22020)。
收稿日期:2022-12-11;**修订日期**:2023-04-25

Meanwhile, ablation experiments verify the effectiveness of auxiliary task and self-supervised network fine-tuning strategy.

Key words: deep learning; multi-task learning; speech emotion recognition; self-supervised model; fine-tuning strategy

引　言

情感在日常生活交流中扮演重要的角色。语音情感可以赋予文字更多的含义,从而在对话中更有效的传达说话人表达的信息。对人类来说,可以在对话中轻松地识别一个人的情感,然而基于人工智能的语音情感自动识别,仍是一项艰巨的任务。语音情感识别研究对于智能化的人机交互具有重要意义。

传统方法中,将语音情感识别系统分为3个重要的模块,语音的特征提取、特征选择和情感分类。研究人员多利用语音的频谱特征[1]、韵律特征[2-4]和音质特征[5]等其他语音任务常用的特征作为系统的输入,这些手工选取的特征需要具有一定的语音信号处理的专业知识,然而语音信号处理在语音情感方面理论基础薄弱,设计适用于情感的特征难度高,这也是语音情感识别的难点之一。

近些年来,随着深度学习的发展,基于端到端的深度神经网络模型逐步替代了传统的特征提取工作[6]。端到端的神经网络通过可训练的卷积层等其他神经网络模型,提取可以表征语音的特征向量,一定程度上解决了传统方法中情感特征设计困难的问题。凭借有效的学习算法和强大的特征提取能力,深度神经网络已经成为了语音情感识别的主流。

在深度学习领域,处理图像的卷积神经网络和处理序列数据的循环神经网络成为深度学习模型的主要工具。文献[7]验证了一维卷积神经网络、二维卷积神经网络和长短期记忆网络的性能优于传统的语音情感分类器。文献[8]使用残差网络作为主干网络,通过增加模型的深度进一步提升了模型的语音情感特征抽取能力。近年来,利用注意力机制来分析文本中相对重要的信息,已经成为机器翻译、信息提取等任务重要的技术手段。目前,注意力机制在视觉和语音领域的研究也广受关注。文献[9]提出双向长短期记忆网络结合自注意力机制网络,利用语音帧的自相关性来处理信息缺失的问题,同时识别隐藏在句子中的情感信息。考虑到不同说话人的声纹差异和其独特的情感表达习惯,文献[10]提出了一种个体标准化网络(Individual standardization network)以缓解个体差异引起的个体间情感混乱问题。文献[11]提出了一种新型的协同多视角关系网络(Collective multi-view relation network)以利用多视角语音表征的内在特征进行语音情感识别。

最近,自监督模型Wav2vec 2.0在语音识别和说话人验证取得了巨大成功[12]。在语音情感识别方向,文献[13]研究发现,在语音情感识别中Wav2vec 2.0特征优于传统的语音频谱特征;文献[14]比较了不同时间跨度的特征,结果表明时域背景信息越多的自监督特征在情感识别上的性能越好;文献[15]证明Wav2vec 2.0不同层表征向量的线性组合优于最后单层的特征表示。

虽然端到端的深度学习模型在语音情感识别上取得了显著进展,但研究人员通常只使用语音频谱特征、自监督特征等语音识别中常用的特征。这些特征中包含着丰富的语义信息,却忽视了语音中其他有助于情感识别的信息,例如声学情感信息。语音中蕴含的情感信息主要源于说话人的发音和说话内容,而发音情感指同一句话可以用不同的发音表达不同的情感,然而语音的发音声学特征不易量化和提取。语音信号中除蕴含情感信息外还包含丰富的语义信息(主要由音素组成)。在语音识别研究中,往往先从语音中识别出音素信息,然后利用搜索算法和语言模型组合成对应的文本。对于语音情感识别任务,不同音素的声学变异性是干扰情感识别的重要原因。为了克服这种声学变异性,文

献[16]提出利用情感相关的音素识别来对说话人的情感进行推理,通过实验证明了音素对于不同的情感具备一定的区分性。文献[17-18]的研究证明了音素有助于语音的情感识别。因此,本文在情感识别中引入音素。为了在情感识别中利用到语音的音素信息,设计了一种多任务学习模型,利用与情感有关的音素识别作为语音情感识别的辅助任务,通过使用多任务学习学得两个任务的共享表征,使得其包含有助于情感识别的声学特征。此外,语音情感识别的数据集较小,经过时频域变换的频谱特征在特征提取的过程中会出现信息丢失问题,而时域信号保留更多的语音信息,有助于减轻模型过拟合风险。鉴于此,本文采用在大规模语音数据集下预训练的基于时域信号的Wav2vec 2.0模型[19]提取语音特征,并在多任务学习模型下进行微调。最后,在IEMOCAP数据集上验证了所提方法的可行性和有效性。本文的主要贡献如下:

(1)提出多任务学习模型,情感识别作为其主任务,利用与情感有关的音素识别作为辅助任务,提取语音中的声学情感特征。

(2)结合自监督模型微调策略,使得共享表征中包含更多有助于情感识别的声学情感信息。

1 模型结构

1.1 网络整体结构

网络整体结构如图1所示,模型主干网络为自监督模型Wav2vec 2.0,主要包含由7层卷积堆叠的特征编码模块和12层Transformer堆叠的上下文网络。主干网络为两个子任务语音情感识别和音素识别的特征提取共享模块,其中音素识别网络突出自监督特征中的声学情感信息,情感识别网络用于对情感进行分类。情感识别网络和音素识别网络一样,为由单层的全连接神经网络组成,用于组合无监督特征和对情感进行分类。

为了将无监督特征向量映射到情感特征空间,本文在时间轴上对自监督特征向量H_X进行平均池化。语音情感识别网络用交叉熵(Cross entropy, CE)损失函数计算损失来优化网络的参数。在训练阶段,多任务学习模型同时计算两个子任务的损失L_{ER}和L_{PR},模型参数共享的部分由两个损失函数共同优化。考虑到语音情感识别任务与音素识别任务的难易不同,导致两个任务收敛的速度不同,本文设置了一个权衡系数λ来平衡两个任务的训练,则多任务学习的损失函数为

$$L = L_{\mathrm{ER}} + \lambda L_{\mathrm{PR}} \tag{1}$$

1.2 自监督网络和特征融合

本文采用Wav2vec 2.0自监督学习模型,作为语音特征提取器和多任务学习模型共享参数的主干网络。在深度学习中,监督学习和无监督学习是其两种基本的学习范式。目前监督学习受限于有标签的数据量,在数据

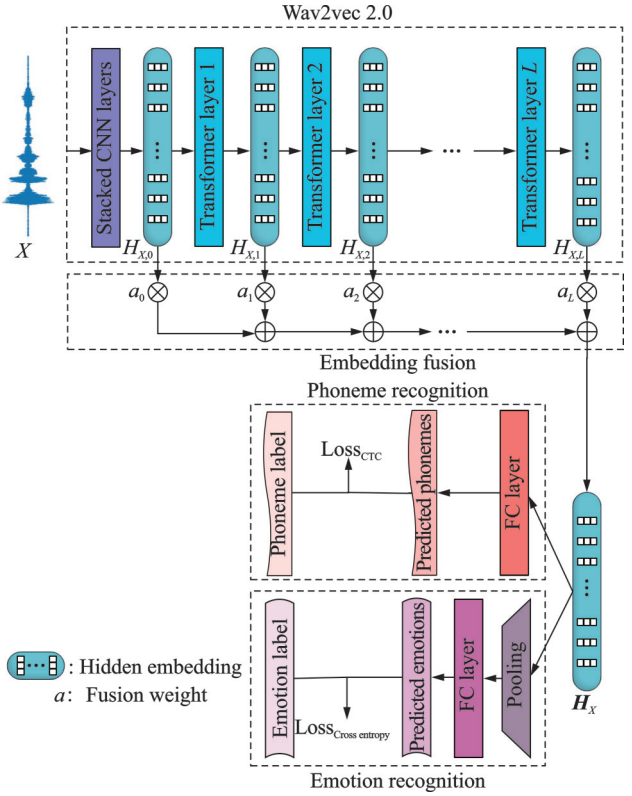

图1 多任务学习网络

Fig.1 Multi-task learning network

不足时模型的性能受到了极大的限制。无监督学习不需要任何人工标注的信息,通过挖掘数据本身的特征完成相关的任务。目前,自动编码器是无监督学习广泛应用的方式。然而,其在学习隐层表征向量上仅仅是将特征进行降维,特征在时序上不包含相互关系。基于无监督学习改进的自监督学习较好地解决了这个问题。自监督学习利用辅助任务生成伪标签作为训练的目标,运用伪标签和模型的输出来计算模型的损失,通过这种方式训练出的隐层表征向量可以很好地挖掘数据中的语义信息。

如图2所示,Wav2vec 2.0主要包含3个主要模块,特征编码网络、量化模块和上下文网络。在预训练阶段,语音经过特征编码网络生成多个时间步的低级特征向量,然后以一定的概率对每个时间步进行掩码操作。将掩码的低级特征向量输入到上下文网络中,输出包含上下文语义信息的高级特征向量。未掩码的低级特征向量经过量化模块得到伪标签作为上下文网络预测的目标,进而计算伪标签与上下文网络输出的预测向量间的相似度。通过计算相同时间步和不同临近时间步的低级特征向量与相同时间步高级特征向量的相似度,得到对比损失,再结合多样性损失保证量化模块中码本的码元出现概率均等。自监督网络利用这两个损失函数计算梯度,更新整个模型的参数。考虑到本文提出的模型在训练过程中属于监督学习,因此取消了原始网络的量化模块,低级特征向量在输入上下文网络之前并不进行掩码操作。使用预训练好的权重初始化网络参数,并利用多任务的损失函数结合反向传播算法对网络进行微调。

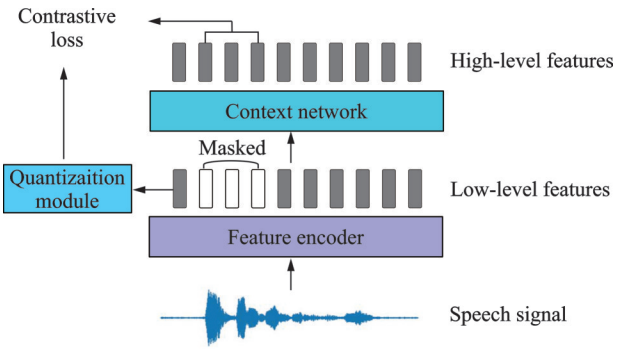

图2 Wav2vec 2.0自监督网络

Fig.2 Wav2vec 2.0 self-supervised network

在微调阶段,语音输入信号$\boldsymbol{X}\in\mathrm{R}^{1\times N}$($N$为语音采样点个数),经过特征编码网络得到低级特征向量$\boldsymbol{H}_{\mathrm{l}}=\left\{\boldsymbol{H}_{\mathrm{l}}^{1},\boldsymbol{H}_{\mathrm{l}}^{2},\cdots,\boldsymbol{H}_{\mathrm{l}}^{T}\right\}\in\mathrm{R}^{T\times 768}$($T$为时间帧个数,768为每个时间帧特征维度),$\boldsymbol{H}_{\mathrm{l}}$经过上下文网络可得到高级特征向量$\boldsymbol{H}_{\mathrm{h},i}=\left\{\boldsymbol{H}_{\mathrm{h},i}^{1},\boldsymbol{H}_{\mathrm{h},i}^{2},\cdots,\boldsymbol{H}_{\mathrm{h},i}^{T}\right\}\in\mathrm{R}^{T\times 768}$,$i\in[1,L]$($L$为上下文网络中Transformer层的个数,实验设置为12),模型前向传播计算过程如下

$$\boldsymbol{H}_{\mathrm{l}}=F_{\mathrm{FeatureEncoder}}(\boldsymbol{X}) \tag{2}$$

$$\boldsymbol{H}_{\mathrm{h},i}=F_{\mathrm{ContextNetwork}}(\boldsymbol{H}_{\mathrm{l}}) \tag{3}$$

$$\boldsymbol{H}_{X}=\sum_{i=1}^{L}\alpha_i\boldsymbol{H}_{\mathrm{h},i}+\alpha_0\boldsymbol{H}_{\mathrm{l}} \tag{4}$$

特征融合部分,本文将低级特征向量与高级特征向量通过注意力权重α_i融合得到语音信号X的自监督特征向量$\boldsymbol{H}_X\in\mathrm{R}^{T\times 768}$,其计算过程如式(4)所示。其中,低级特征向量主要包含声学特征信息,高级特征向量包含较多的语义信息,经过特征融合后的自监督特征向量能充分表征语音中的特征信息。为了充分利用这两种特征中的声学信息和语义信息,在实验中生成一组初始值为0的数,并通过Softmax函数获得初始的注意力权重。随着网络的训练,基于反向传播算法更新这些权重,从而使模型自适应地选择这两类特征的线性组合。

1.3 音素识别网络

音素是组成语音的最小识别单元。在语音识别研究中,先识别音素种类,然后再将音素通过字典组合成文本,因此音素识别属于语音识别的基础。语音中的情感信息主要由发音方式和发音内容决定,其中音素信息是发音的主要内容。为了提取语音中声学情感的特征,本文将情感和音素结合在一

起,将二者视作一个整体,通过神经网络获得语音的声学情感特征。

音素识别任务的标签来自于每条语音的文本标注,先将文本转换为发音的音标,其次将音标转换为音素,最后为了更好地描述语音中情感与音素的关系,将音素按照发音的方式分为6类,并且为每一类音素加上了情感标签。音素的划分方式如表1所示,其中音素情感标签后面的数字代表着情感的类别(0:Angry;1:Happy;2:Sad;3:Neutral),同一句话中的音素标签后缀为同一个数字。

表1 音素划分

Table 1 Phoneme division

音素类别	音素	音素情感标签
S(爆破音/塞擦音)	B P T D K…	S_0 S_1 S_2 S_3
F(摩擦音)	S Z SH ZH TH…	F_0 F_1 F_2 F_3
N(鼻音)	M N NG	N_0 N_1 N_2 N_3
L(流音/滑音)	L R Y W	L_0 L_1 L_2 L_3
V(元音)	AH0 AA0 AE0…	V_0 V_1 V_2 V_3…
VS(重音/次重音)	AA1 AH2 AE1 AE2…	Vs_0 Vs_1 Vs_2 …

音素识别网络如图3所示,其主要包含单层的全连接神经网络,模型参数采用随机初始化。自监督模型输出的特征$\boldsymbol{H}_X$只表征了语音的特征信息。因此,本文使用一层全连接神经网络实现音素的分类,再经过Softmax函数,输出得到预测音素的概率矩阵$\boldsymbol{P}_{\mathrm{M}}\in\mathbf{R}^{B\times T\times D}$($B$为批大小,$T$为时间帧个数,$D$为情感相关的音素个数),计算过程为

$$\boldsymbol{P}_{\mathrm{M}}=\mathrm{softmax}(F_{\mathrm{FC,PR}}(\boldsymbol{H}_X)) \tag{5}$$

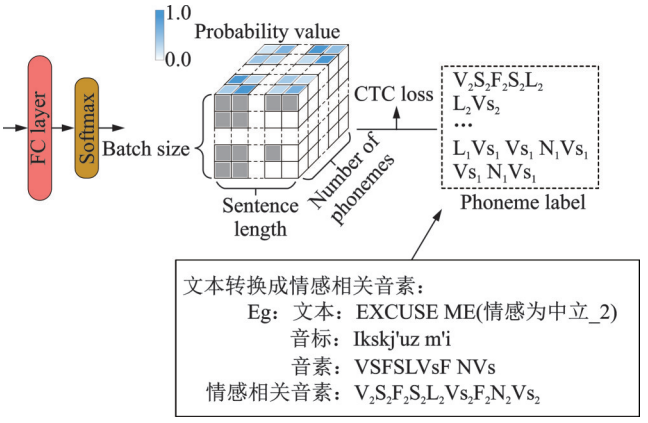

图3 音素识别网络

Fig.3 Phoneme recognition network

在微调过程中,本文通过连接时序分类(Connectionist temporal classification,CTC)损失函数不断优化音素识别网络,即

$$L_{\mathrm{PR}}=\mathrm{CTC}(\boldsymbol{P}_{\mathrm{M}},Y_{\mathrm{phonemes}}) \tag{6}$$

1.4 情感识别网络

考虑到自监督网络本身模型庞大,本文使用的语音情感数据集较小。为防止模型出现严重的过拟合,本文使用单层全连接神经网络对自监督特征进行分类。由于每条语音的时长不同,因此,自监督特征在输入全连接神经网络之前在时间轴上进行平均池化操作。情感识别网络如图4所示,其前向传播的计算过程为

$$\hat{y}=\mathrm{softmax}\Big(F_{\mathrm{FC,ER}}\big(F_{\mathrm{AveragePooling}}(\boldsymbol{H}_X)\big)\Big) \tag{7}$$

$$L_{\mathrm{ER}}=\mathrm{CE}(\hat{y},Y_{\mathrm{emotions}}) \tag{8}$$

式中:$F_{\mathrm{AveragePooling}}$和$F_{\mathrm{FC,ER}}$分别代表平均池化操作和全连接神经网络,$Y_{\mathrm{emotions}}$表示情感标签。

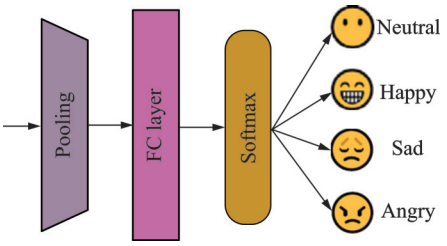

图4 情感识别网络

Fig.4 Emotion recognition network

2 实验与结果分析

2.1 数据集和实验设置

本文选用广泛用于情感识别的IEMOCAP(Interactive

emotional dyadic motion capture)[20]数据集。数据集记录了10个演员的5场会话,每场会话有一个男演员和一个女演员按照剧本和无剧本(即兴表演)进行会话表演,表演中包含中立、快乐和愤怒等9种情感,数据集时长约12 h。本文选用了数据量最多的4种情感:中立、快乐、悲伤和愤怒,与之前的研究保持一致,将激动情感也归纳到快乐当中。实验数据共有5 531条语音(其中1 103 Angry,1 636 Happy,1 084 Sad,1 708 Neutral),平均时长约为4.5 s。

使用留一法交叉验证对算法的性能进行评估,每次使用9个人的语音作为训练集,留下1个人的语音作为测试集,保证每次实验训练集和测试集没有说话人重叠。实验结果用加权准确率(Weighted accuracy,WA)和非加权准确率(Unweighted accuracy,UA)来衡量,计算公式分别为

$$\mathrm{WA}=\frac{1}{10}\left(\frac{N_{\mathrm{correct}}^{K}}{N_{\mathrm{total}}^{K}}\right) \tag{9}$$

$$\mathrm{UA}=\frac{1}{10}\left(\frac{1}{4}\sum_{i=1}^{4}\frac{N_{\mathrm{correct}}^{K(i)}}{N_{\mathrm{total}}^{K(i)}}\right) \tag{10}$$

式中:WA表示整个测试集的准确率;UA用来计算测试集中每个类别准确率的平均值。实验使用PyTorch作为深度学习训练框架,多任务学习模型主干网络Wav2vec 2.0初始化用Hugging Face①发布的预训练权重,主干网络后面子任务的网络权重随机初始化,训练整个模型迭代100轮,批(Batch)大小为16,使用Adam作为优化器,初始学习率设为5e−5,权重衰减设为1e−5。

2.2 实验结果对比

为了验证所提方法的有效性,本文将其与语音情感识别的最新研究进行了对比。这些研究与本文实验数据集的切分方式一致,本文简单总结了这些研究方法,实验结果对比如表2所示。通过比较可以发现,本文所提的多任务学习模型相较于前人基于单任务学习的深度模型在IEMOCAP数据集上取得了更高的性能,其中WA为76.0%(提升了约3.75%),UA为76.9%(提升了约4.3%)。所提出的单任务学

表2 IEMOCAP数据集语音情感识别实验结果

Table 2 Experimental results of speech emotion recognition in IEMOCAP dataset

文献	方法	WA/%	UA/%
Sajjad等[8]	用K均值聚类算法对语音帧进行聚类,从每类中选择一帧作为语音的特征帧拼接在一起,对此进行短时傅里叶变换计算得到频谱特征输入到ResNet101和BiLSTM网络中	72.25	—
Lu等[21]	采取端到端语音识别网络中的编码器提取包含声学和文本信息的特征向量,利用自注意力机制调整BiLSTM解码器网络输出的不定长的特征序列为固定长度的特征	71.7	72.6
Liu等[22]	多尺度的卷积神经网络学到语音频谱的局部特征,结合改进路由算法的CapsNet网络得到语音频谱的全局特征,将局部特征与全局特征结合在一起	70.34	70.78
Pappagari等[23]	将X-vector说话人预训练网络作为主干网络,微调语音情感识别	70.3	—
Liu等[24]	提出BiLSTM-GIN模型,用openSMILE工具提取语音特征,用BiLSTM网络编码特征,然后用GIN网络实现全局情感信息的整合	64.65	65.53
本文	单任务学习模型,Wav2vec 2.0作为主干网络,微调语音情感识别	72.4	73.9
本文	多任务学习模型,利用音素识别任务辅助主任务语音情感识别	76.0	76.9

①https://huggingface.co/.

习模型与之前的单任务研究工作相比取得了一个较好的结果(UA提升了1.3%)。

从表2可以看到,相对于文献[24]采用的语音统计值特征建模,本文使用的自监督特征在单任务学习模型上WA和UA均提升了约8%。相对于文献[8,22]使用的语音的频谱特征和倒谱特征,本文提出的多任务学习模型学到的自监督特征包含更多的情感信息。文献[23]使用了语音识别和说话人识别网络作为特征提取器,其特征本身包含了大量与情感无关的说话人信息,导致模型性能不佳。

2.3 消融实验

在本文提出的多任务学习网络中,音素识别作为辅助任务,用于提高主任务情感识别的性能。为了验证音素识别任务对于情感识别任务的重要性,本文通过调整多任务学习损失函数中λ的大小来控制音素识别任务的重要程度。λ的值为0到1之间,当λ等于零时,多任务学习模型为单任务学习模型,音素识别网络不参与训练。由于神经网络通过链式求导更新其参数,因此,可以将损失函数的权重看作是一个学习率调整因子,用于调整反向传播过程中梯度的大小,从而影响参数的更新速度和收敛性能。当λ的值越大,音素识别网络在反向传播时获得的梯度越大,收敛速度越快,其音素识别任务性能越好。从表3可以看到,随着λ的不断增大,语音情感识别的性能越好,说明与情感有关的音素信息有助于模型对语音中情感的识别。当λ等于1时,模型的性能变差。对于多任务学习来说,辅助任务获得的权重过大,会导致模型偏向于辅助任务,从而导致主任务性能变差。

表3 音素识别对于情感识别的重要性

Table 3 Importance of phoneme recognition for emotion recognition

λ	WA/%	UA/%
0	72.4	73.9
0.001	74.5	75.7
0.01	75.7	76.7
0.1	76.0	76.9
1	75.2	76.2

从图5和图6模型预测混淆矩阵可以看出本文提出的多任务学习模型有效地提高了“开心”和“中立”两种情感的预测,其中“开心”情感预测准确率提升了约8.5%,“中立”情感提升了约5%,剩余两种情感的准确率无明显变化。说明与情感有关的音素信息能提高模型对于语音中“开心”和“中立”两种情感的信息提取,证明了音素识别作为辅助任务对于主任务情感识别的有效性。

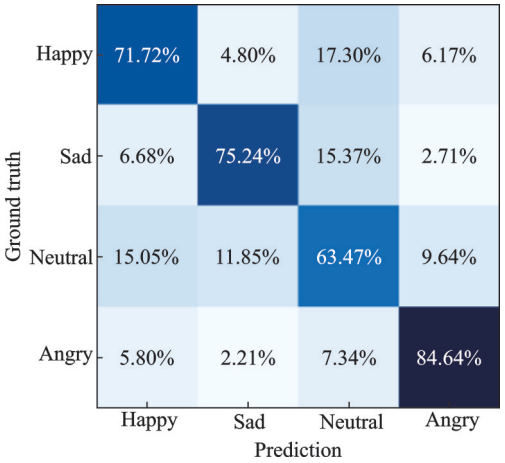

图5 $\lambda=0$时的单任务学习

Fig.5 Single task learning in the case of $\lambda=0$

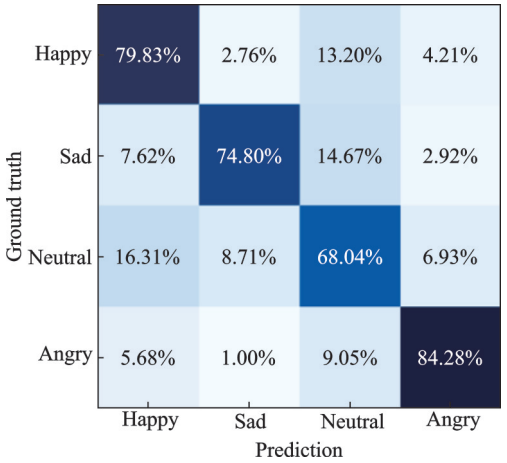

图6 $\lambda=0.1$时的多任务学习

Fig.6 Multi-task learning in the case of $\lambda=0.1$

3 结束语

针对语音情感识别中,语音的发音与语音情感高度相关,本文提出了一种多任务学习模型,与情感有关的音素识别作为辅助任务,让模型学到的自监督特征包含声学情感信息,使得模型在情感识别任务中可以利用语音中发音包含的情感。通过在IEMOCAP数据集上的训练和测试,相比较其他单任务语音情感识别方法,加权准确率和非加权准确率均有一定的提升。在后续研究中,语音中有很多与情感无关的信息,例如说话人信息,可以设计相关模型以剔除语音中的说话人相关信息,以提高模型的识别率。

参考文献:

[1] ZHOU P, LI X P, LI J, et al. Speech emotion recognition based on mixed MFCC[J]. Applied Mechanics and Materials, 2012, 249/250: 1252-1258.

[2] RAO K S, KOOLAGUDI S G, VEMPADA R R. Emotion recognition from speech using global and local prosodic features [J]. International Journal of Speech Technology, 2013, 16(2): 143-160.

[3] YAO Z, WANG Z, LIU W, et al. Speech emotion recognition using fusion of three multi-task learning-based classifiers: HSF-DNN, MS-CNN and LLD-RNN[J]. Speech Communication, 2020, 120: 11-19.

[4] ATMAJA B T, AKAGI M. The effect of silence feature in dimensional speech emotion recognition[EB/OL]. (2020-03-03). https://doi.org/10.21437/SpeechProsody.

[5] ANAGNOSTOPOULOS C N, ILIOU T, GIANNOUKOS I. Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011[J]. Artificial Intelligence Review, 2015, 43(2): 155-177.

[6] TZIRAKIS P, TRIGEORGIS G, NICOLAOU M A, et al. End-to-end multimodal emotion recognition using deep neural networks[J]. IEEE Journal of Selected Topics in Signal Processing, 2017, 11(8): 1301-1309.

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

[8] SAJJAD M, KWON S. Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM[J]. IEEE Access, 2020, 8: 79861-79875.

[9] LI D, LIU J, YANG Z, et al. Speech emotion recognition using recurrent neural networks with directional self-attention[J]. Expert Systems with Applications, 2021, 173: 114683.

[10] FAN W, XU X, CAI B, et al. ISNet: Individual standardization network for speech emotion recognition[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 30: 1803-1814.

[11] HOU M, ZHANG Z, CAO Q, et al. Multi-view speech emotion recognition via collective relation construction[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 30: 218-229.

[12] CHEN Z, CHEN S, WU Y, et al. Large-scale self-supervised speech representation learning for automatic speaker verification [C]//Proceedings of ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore: IEEE, 2022: 6147-6151.

[13] BOIGNE J, LIYANAGE B, ÖSTREM T. Recognizing more emotions with less data using self-supervised transfer learning [EB/OL]. (2020-11-11). https://doi.org/10.48550/arXiv.2011.05585.

[14] XIA Y, CHEN L W, RUDNICKY A, et al. Temporal context in speech emotion recognition[C]//Proceedings of Interspeech 2021: Conference of the International Speech Communication Association. Brno, The Czech Republic: [s.n.], 2021: 3370-3374.

[15] PEPINO L, RIERA P, FERRER L. Emotion recognition from speech using Wav2vec 2.0 embeddings[EB/OL].(2021-04-08). https://doi.org/10.48550/arXiv.2104.03502.

[16] YUAN J, CAI X, ZHENG R, et al. The role of phonetic units in speech emotion recognition[EB/OL].(2021-08-02). https://doi.org/10.48550/arXiv.2108.01132.

[17] SCHULLER B, VLASENKO B, ARSIC D, et al. Combining speech recognition and acoustic word emotion models for robust text-independent emotion recognition[C]//Proceedings of 2008 IEEE International Conference on Multimedia and Expo.

Hannover, Germany: IEEE, 2008: 1333-1336.

[18] DHAMYAL H, MEMON S A, RAJ B, et al. The phonetic bases of vocal expressed emotion: Natural versus acted[C]// Proceedings of INTERSPEECH 2020: Conference of the International Speech Communication Association. Shanghai, China: [s.n.], 2020: 3451-3455.

[19] CHEN L W, RUDNICKY A. Exploring Wav2vec 2.0 fine-tuning for improved speech emotion recognition[C]//Proceedings of 2023 IEEE International Conference on Acoustics, Speech and Signal Processing. [S.l.]: IEEE, 2023.

[20] 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.

[21] LU Z, CAO L, ZHANG Y, et al. FanSpeech sentiment analysis via pre-trained features from end-to-end ASR models[C]// Proceedings of ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE, 2020: 7149-7153.

[22] LIU J, LIU Z, WANG L, et al. Speech emotion recognition with local-global aware deep representation learning[C]// Proceedings of ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE, 2020: 7174-7178.

[23] PAPPAGARI R, WANG T, VILLALBA J, et al. X-Vectors meet emotions: A study on dependencies between emotion and speaker recognition[C]//Proceedings of ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE, 2020: 7169-7173.

[24] LIU J, WANG H. Graph isomorphism network for speech emotion recognition[C]//Proceedings of INTERSPEECH 2021: Conference of the International Speech Communication Association. Brno, the Czech Republic: [s.n.], 2021: 3405-3409.

作者简介:

李云峰(1998-),男,硕士,研究方向:深度学习、语音情感识别,E-mail: Liyunfeng7618@163.com。

闫祖龙(2000-),男,硕士,研究方向:机器学习、语音情感识别。

高天(1991-),男,博士,工程师,研究方向:说话人识别、语音信号处理。

方昕(1988-),男,博士,工程师,研究方向:语音识别、说话人识别。

邹亮(1987-),**通信作者**,男,博士,副教授,研究方向:深度学习、信号处理,E-mail: liangzou@cumt.edu.cn。

(编辑:王静)