



研究与开发

基于语谱图提取深度空间注意特征的 语音情感识别算法

王金华, 应娜, 朱辰都, 刘兆森, 蔡哲栋
(杭州电子科技大学, 浙江 杭州 310018)

摘要: 从语音情感特征的提取和分类建模出发, 以混合卷积神经网络模型为基础, 改进特征提取中的 Itti 模型, 包括: 增加通过局部二值模式提取的纹理特征; 结合听觉敏感度权重提取情感强相关特征。然后提出通过特征约束条件提取标定权重特征的约束挤压和激励网络结构; 最后形成以 VGGnet 和长短时记忆网络混合网络为基础的微调模型, 进一步提升了情感表征能力。通过在自然情感数据库和柏林德语数据库上进行验证, 该模型在情感识别率上有明显的上升, 相较于基准模型提升了 8.43%, 同时对比了本模型在自然数据库 (FAU-AEC) 和柏林数据库 (EMO-DB) 上的识别效果, 实验结果证明模型具有良好的泛化性。

关键词: 情感识别; 深度混合神经网络模型; 视觉注意机制

中图分类号: TP18

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2019052

Speech emotion recognition algorithm based on spectrogram feature extraction of deep space attention feature

WANG Jinhua, YING Na, ZHU Chendu, LIU Zhaosen, CAI Zhedong
Hangzhou Dianzi University, Hangzhou 310018, China

Abstract: Starts from the extraction and classification modeling of speech emotion features, based on the hybrid convolutional neural network model, the Itti model in feature extraction was improved, including increasing the extraction by local binary mode. The strong correlation features were extracted combining with the sensitivity of the auditory sensitivity. Then, the constrained extrusion and excitation network structure of the calibration weights were extracted by feature constraints. Finally, a fine-tuning model based on VGGnet and long-short-time memory network hybrid network was formed, further enhancing the ability to express emotions. By validating on the natural sentiment database and the German-German database, the model had a significant increase in the rate of sentiment recognition, which is 8.43% higher than the benchmark model. At the same time, the recognition effect of the model on the natural database (FAU-AEC) and the Berlin database (EMO-DB) were compared. The experimental results show that the model has a good generalization.

Key words: emotion recognition, deep hybrid neural network model, visual attention mechanism

收稿日期: 2018-09-03; 修回日期: 2019-03-06

基金项目: 国家自然科学基金资助项目 (No.61705055); 浙江省自然科学基金资助项目 (No.LY16F010013)

Foundation Items: The National Natural Science Foundation of China (No.61705055), The Natural Science Foundation of Zhejiang Province of China (No.LY16F010013)

1 引言

人们在沟通中能够听懂对方语音并感知对方情绪变化,是因为人脑具备理解语音信号和感知情绪状态的系统。语音情感识别(speech emotion recognition, SER)就是计算机对人脑系统处理语音情绪过程的模拟,主要包括从语音信号中提取语音特征,将语音特征通过一定映射关系建立情感间关联,这对人机交互有重要意义^[1-2]。一直以来,通过语音情感识别算法提取情感相关特征难度较大,因此寻找合适的情感识别模型是研究者关注的焦点。

语音情感识别率的提高需要克服识别情感信号所面临的各种困难,主要包括多样性(说话人的多样性、环境的多样性等)和深层次抽象特征(情感)等。深度学习的迅速发展给 SER 带来了更大的突破,利用合适的模型和算法训练出的深度模型可以从海量多源数据库中提取更有价值的信息。Yamada 等^[3]研究了一种传递信息的情感模型,利用 BP 神经网络作为声学模型,识别的基本情感包括悲伤、兴奋、欢乐和愤怒,取得了 70% 的识别率。Teng 等^[4]探讨了传统声学特征和新 i-vector 特征在情感识别中的表现,同时证明递归神经网络(recurrent neural network, RNN)可以将情绪识别率提升 5% 左右。Basu 等^[5]主要采用梅尔(Mel)频率倒谱系数作为输入特征,利用卷积神经网络(convolutional neural network, CNN)和 LSTM 的分类方法实现了接近 80% 的识别率。Shi 等^[6]提出用端对端混合神经网络模型来解决图片序列识别的问题,在场景文本识别中表现显著,同时在音乐配乐识别的应用中证实了很好的泛化性。综上所述可以看出,重构混合神经网络的方式可以对网络性能以及泛化能力产生实质性的改善。

目前对基于神经网络的语音情绪识别系统的研究不仅集中在模型的选取和改进上,还在识别系统中探索新的特征选取,Zazo 等人^[7]使用 Mel

频率倒谱系数和增量 Delta 系数(MFCC-SDC)特征作为单向 LSTM 的输入,然后直接连接到 Softmax 分类器,得到最后语言分类结果。Gelly 等人^[8]使用双向递归神经网络(bidirectional recurrent neural network, Bi-RNN)从输入中捕获语言信息(音频转换为感知线性预测(linear predictive coding, PLP)系数及其一阶和二阶导数),所得到的序列特征被融合在一起作为分类器输入样本的语言。两种方法都只考虑序列特征作为其网络的输入,但由于选取频域特征会忽略时间维度的影响,选取时域特征会忽略频率维度的影响,同时使得时频域间隐藏的潜在性质(关联性等)被忽略。因此有学者尝试将语谱图和图像处理的网络相结合,如 Lozano-Diez 等人^[9]将输入数据转换为包含 MFCC-SDC 特征的图像,利用 CNN 进行语言识别,获得了更好的性能;Zhang 等^[10]提出了基于 Chirplet 模型的特征提取方法,该模型以语谱图作为输入,模型输出的注意特征图作为新型特征直接送入 SVM 分类器,结果表明,该特征识别效果优于传统特征,且具有更好的泛化性和顽健性。综上所述可以看出,输入特征的选取对模型最终效果有很重要的影响,而语谱图作为特征输入相比传统特征具有更好的泛化性和顽健性,所以探索合适的新型特征意义重大。

根据上述研究现状,本文提出一种改进的深度空间注意模型,算法主要引入了特征增强策略:采用改进的 Itti 模型来提取 SEF 特征,增加纹理特征,依据听觉敏感度确定特征图优先级,选取适当权重,通过加权机制获取最终显著图;考虑到语音的时序性和空间性,算法以 SEnet(squeeze-and-excitation network)为基础,提出改进的 SEnet 模型(C_SEnet),该模型通过阈值机制增强甄选出 SEF 特征,赋予这类特征高权重,抑制其他弱相关特征;最后将得到的特征输入混合神经网络进行训练和分类。实验结果显示:本文提出的算法在增加较小模型复杂度的前提下,



识别率提升了 8.43%。

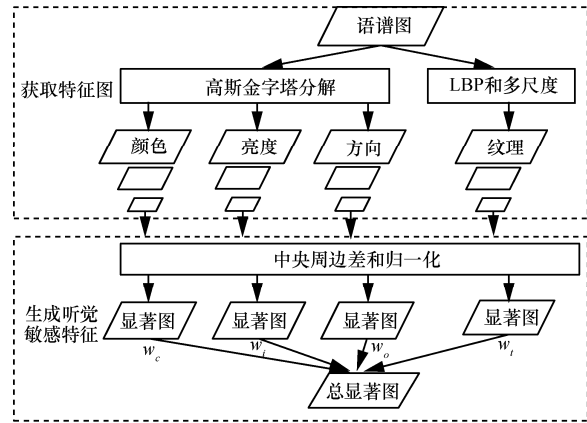
2 基于深度学习的空间注意模型

2.1 改进的 Itti 模型

Itti 模型是一种自底向上的基于人眼视觉机制的注意模型，能够关注主要信息，抑制干扰信息。对于一幅输入的图像，该模型首先提取 3 种视觉特征，分别是颜色、亮度和方向；然后在多种尺度算法下使用中央周边差（center-surround）和非线性归一化操作产生对应特征的显著图；最后将这些特征图进行线性合并，得到最终的显著图（saliency map）。研究表明，人眼对不同显著性特征的敏感度不同，最敏感的是颜色，其次是亮度，最后是方向^[11]，而颜色对应着语音中的能量信息，高能量和共振峰特征恰恰是区分不同音素的关键特征，亮度和方向对应着频率和相邻频率间的相关梯度特征，结合视觉理论研究和人耳听觉敏感度，本文在获取语谱图的 Itti 模型特征的基础上，结合 LBP 算子增加纹理特征，并根据听觉敏感特征设置权重，最终得到 SEF 特征。

针对语音情感特征的实验中发现原始的 Itti 模型的缺点^[12]：由于金字塔采样过程不停地使用邻近插值，造成边缘信息丢失，导致显著图的分辨率降低。模型中最终显著图是将几种特征最终显著图等量平均加权。这里通过增加纹理特征来弥补这一缺陷，纹理特征的提取通过局部二值模式，LBP 算子相比于其他算法具有旋转不变性、特征维度低和计算速度快等优点；通过情感敏感度权重方法来弥补第二个缺陷，对于情感信息的处理过程，通常先利用包含情感信息的局部帧，先获取主要音素位置；再获得相关帧附近的频率、频率间的相关梯度等特征。再根据本文提出情感敏感度权重方法求显著图。整体示意图如图 1 所示。

设训练语音 $x_{\text{train}}^i (i=1,2,3,\dots,N)$ ，将语音转化成语谱图 $s_{\text{train}}^i (i=1,2,3,\dots,N)$ 。再经过高斯滤波后



输出图为 $p_{\text{train}}^i (i=1,2,3,\dots,N)$ 。

算法具体实现步骤如下。

(1) 获取特征图

为了提取多尺度信息，需要多尺度多通道滤波，即高斯金字塔分解^[13]，对于输入的语谱图 s_{train}^i ，第 j 层图像计算式为：

$$p_{\text{train}}^j = \sum_{m=-4}^{m=4} \sum_{n=-4}^{n=4} w(m,n) p_{\text{train}}^{j-1} \quad (1)$$

其中， $w(m,n)$ 为下采样高斯核函数，实质上是一个二维可分离的 5×5 的窗口，这里模型采用的是 $\frac{1}{16}[1,4,6,4,1]$ 的平滑滤波器，经过 8 次高斯滤波下采样，加上原图共形成 9 层高斯金字塔结构。该滤波过程从颜色、亮度、纹理和方向 4 种角度提取出不同尺度的语音情感信息，从生理学角度看，这些信息反映了语音情感信息变化的敏感度。第 j 层中 4 种特征提取的主要步骤如下。

设 r 、 g 、 b 分别表示图像的红色、绿色和蓝色通道，则第 j 层颜色特征图包括红绿和黄蓝两组通道，表达式如式 (2)：

$$\begin{aligned} R-G: RG(j) &= \frac{r-g}{\max(r,g,b)} \\ B-Y: BY(j) &= \frac{b-\min(r,g)}{\max(r,g,b)} \end{aligned} \quad (2)$$

亮度特征图为归一化后的 r 、 g 、 b 的均值，表达式如式 (3)：

$$I: I(j) = \frac{(r+g+b)}{3} \quad (3)$$

纹理特征图处理过程是, 首先将语谱图 s_{train}^i 灰度化, 然后经过局部二值算法 (LBP) 得到 L_{train}^i , 最后经过梯度采样后获得纹理通道特征, 表达式如式 (4):

$$T: T(j) = L_{3 \times 3}(L_{\text{train}}^i) \quad (4)$$

其中, $L_{3 \times 3} = L(x_c, y_c) = \sum_{p=0}^{p=1} 2^p s(i_p - i_c)$, (x_c, y_c) 代表 3×3 邻域的中心元素, 它的像素为 i_c , 邻域其他像素值为 i_p , $s(x)$ 是符号函数, 定义如式 (5):

$$s(x) = \begin{cases} 0, x < 0 \\ 1, x \geq 0 \end{cases} \quad (5)$$

方向特征图为图像经过二维 Gabor 方向滤波器提取所得, 表达式如式 (6):

$$O: O(j) = |I(j) \times G_o(\theta)| + |I(j) \times G_{\frac{\pi}{2}}(\theta)| \theta = \{0^\circ, 45^\circ, 90^\circ, 135^\circ\} \quad (6)$$

从而获得 4 个方向上的特征图。

(2) 生成听觉敏感权重特征

经过中央周边差^[14]和归一化^[15]运算得到的颜色、亮度、方向和纹理的显著图分别为 $C_{\text{train}}^i, I_{\text{train}}^i, O_{\text{train}}^i, T_{\text{train}}^i, (i=1, 2, 3, \dots, N)$, S_{train}^i 为融合后的显著图, 传统的 Itti 视觉模型中采用均值求和方法获得最终显著图, 表达式如下:

$$S_{\text{train}}^i = \frac{1}{4}(C_{\text{train}}^i + I_{\text{train}}^i + O_{\text{train}}^i + T_{\text{train}}^i) \quad (7)$$

而研究表明, 通过上述方法提取的显著区域会存在较多噪声^[11], 影响图像的清晰度和完整度, 对最终情感识别有一定影响。根据听觉敏感度和情感信息处理过程, 本文提出听觉敏感权重调整方法来改进信息融合的方法, 结合视觉理论研究^[10]和人耳听觉敏感度, 以获得信息相对完整的显著图:

$$S_{\text{train}}^i = w_c \times C_{\text{train}}^i + w_i \times I_{\text{train}}^i + w_o \times O_{\text{train}}^i + w_t \times T_{\text{train}}^i \quad (8)$$

经过多次试验, 这里的 w_c 、 w_i 、 w_o 和 w_t 的比例为 3:2:1:4。

2.2 改进的空间权重结构

SEnet 是由 Hu 等人^[16]提出的压缩提取模块, 选取全局池化和全连接层相结合作为主要结构, 通过模型执行动态通道特征重新校准来提高网络的表示能力。该单元从空间维度入手, 将重要通道的特征强化, 非重要通道的特征弱化, 将降低容量的单元置于网络中, 很好地降低了网络复杂度和训练时间。考虑到语音空间关联性对情绪的影响及模型复杂度设计, 选择 SEnet 来生成标定权重特征。

针对语音情感识别的实际情况, 本文提出基于 SEnet 的改进结构 (C_SEnet), 目的是从全局特征中甄选出与 SEF 特征相近的特征, 解决这个特征约束问题可以通过阈值公式来实现。C_SEnet 在保留全局特征又不破坏特征空间关联性的前提下, 能够提高有用特征比例, 从而提高情感特征的利用率, 达到更好的分类效果。

C_SEnet 算法框架示意图如图 2 所示。

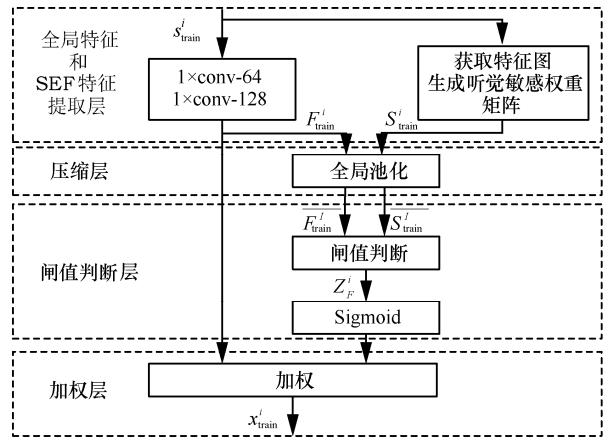


图 2 C_SEnet 算法框架示意图

C_SEnet 算法主要处理步骤如下。

(1) 全局特征和 SEF 特征提取层

语谱图 $s_{\text{train}}^i (i=1, 2, 3, \dots, N)$ 通过右边的 Itti 模型提取出 SEF 特征 $S_{\text{train}}^i (i=1, 2, 3, \dots, N)$, 其中第 i 个语音数据的 SEF 特征图为 S_{train}^i ; 通过左边



CRNN 的前两层结构提取出初层特征, 作为待标定权重全局特征 F_{train}^i , 其中第 i 个语音数据的待标定权重特征图为 F_{train}^i , 以下均以第 i 个数据做说明。

(2) 压缩和提取层^[15]

全局池化 (global average pooling) 作为压缩操作, 将输入的二维特征通道 (SEF 特征图 S_{train}^i 和全局特征 F_{train}^i) 变成实数, 记为 $\overline{S_{\text{train}}^i}$ 和 $\overline{F_{\text{train}}^i}$ 。

$$\begin{aligned}\overline{S_{\text{train}}^i} &= F_{\text{sq}}(S_{\text{train}}^i) = \frac{1}{W \times H} \sum_{i=1}^W \sum_{j=1}^H S_{\text{train}}^i \\ \overline{F_{\text{train}}^i} &= F_{\text{sq}}(F_{\text{train}}^i) = \frac{1}{W \times H} \sum_{i=1}^W \sum_{j=1}^H F_{\text{train}}^i\end{aligned}\quad (9)$$

其中, sq 为 squeeze 的简写。

通过全连接 (fully connected, FC) 层建模通道间的相关性, 首先将特征维度降低到输入的 1/16, 然后经过 ReLU 函数激活, 最后通过一个 FC 层升回原来的维度。

$$\begin{aligned}Z_{\text{S}}^i &= F_{\text{ex}}(\overline{S_{\text{train}}^i}, W) = \sigma(g(\overline{S_{\text{train}}^i}, W)) = \\ &\sigma(W_2 g(\overline{S_{\text{train}}^i} \times W_1)) \\ Z_{\text{F}}^i &= F_{\text{ex}}(\overline{F_{\text{train}}^i}, W) = \sigma(g(\overline{F_{\text{train}}^i}, W)) = \\ &\sigma(W_2 g(\overline{F_{\text{train}}^i} \times W_1))\end{aligned}\quad (10)$$

其中, ex 为 excitation 的简写, δ 指 ReLU 函数, $W_1 \in \mathbb{R}^{r \times c_{\text{xc}}}$ 和 $W_2 \in \mathbb{R}^{r \times c_{\text{xc}}}$, 这里降维的比例系数为 γ 。

(3) 阈值约束层

Z_{S}^i 即 SEF 特征, 将其与全局特征 Z_{F}^i 通过阈值公式进行比较, 得到最终待标定权重矩阵 Z_{F}^i 。

$$Z_{\text{F}}^i = \begin{cases} \max(Z_{\text{F}}^i), Z_{\text{S}}^i - \alpha < Z_{\text{F}}^i < Z_{\text{S}}^i + \alpha \\ Z, \text{其他} \end{cases}\quad (11)$$

如果 Z_{F}^i 在 Z_{S}^i 以为 α 邻域的半径内, 取 Z_{F}^i 的最大值, 即提高与 SEF 特征具有强相关性特征的

权重; 如果 Z_{F}^i 在 Z_{S}^i 以 α 为邻域的半径外, Z_{F}^i 值不变, 即相对抑制与 SEF 特征相关性小的特征权重。

(4) 加权层

将待标定权重矩阵 Z_{F}^i 与全局特征加权 (scale) 来获取标定权重特征 $\widetilde{x_{\text{train}}^i}$ ($i=1, 2, 3, \dots, N$)。

2.3 微调的 CRNN 模型

用于图像处理的 CNN, 能够捕获具有全局感受域的分层模式作为强大的图像描述, 同时由于其局部连接和权重共享的特点, 使得它具有很好的平移不变性。将卷积神经网络的思想应用到语音情感识别的声学建模中, 可以利用卷积的不变性来克服语音信号本身的多样性; 同时时序性对语音信号的情感识别具有重要意义, 卷积网络则更关注空间窗口, 考虑到 RNN 可以捕获数据的时序性, 以使数据保留声音和情绪的上下文环境, 所以这里采用 CNN 和 RNN 结合的混合神经网络 (convolutional recurrent neural network, CRNN)^[17] 作为主声学模型。

微调的 CRNN 网络模型如图 3 所示。经过预处理 (分帧、加窗和傅里叶变换) 生成语谱图, 语谱图通过浅层网络提取初级特征, 通过增强模型提取出标定权重特征, 将标定权重特征送入微调的深层混合网络 (fine-tuned convolutional neural network, F-CRNN) 提取深层特征, 最后送入 Softmax 分类器进行情绪分类。

在 CRNN 的顶部, 考虑到语音的序列中, 两个方向的上下文是相互有用且互补的。因此, 选择将一个深度双向循环神经网络 (bidirectional LSTM, BiLSTM) 置于网络顶层, 结构如图 4 所示。选择该结构的优点是: 首先, 具有更强的捕获序列内上下文信息的能力。以语音识别为例, 连续语音信号连续性可能需要一些连续的帧来完全描述整段涵盖

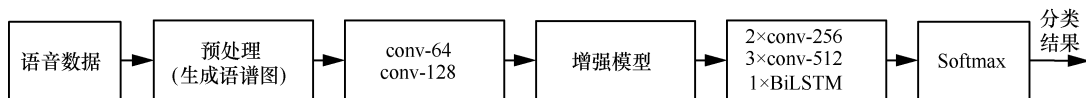


图 3 微调的 CRNN 网络模型

的信息。其次,可以将误差差值反向传播到卷积层,可以在同一个网络中共同训练循环层和卷积层;最后,堆叠多个双向 LSTM。深层网络结构允许比浅层网络结构提取更高层次的抽象特征(比如:情感特征、语义特征等),并且研究表明,已在语音识别任务中显著地改进了性能^[4]。

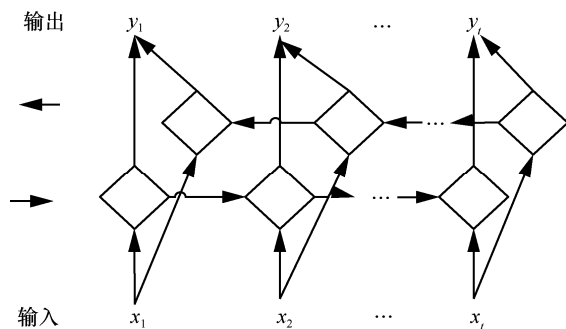


图4 BiLSTM 结构

如图4所示,输入数据(input)为卷积层处理后的数据,输出数据(output)送入 Softmax 进行分类,输出最终分类结果。

3 实验结果与分析

实验旨在通过语谱图和深度空间注意模型结合的算法提取出更具表征力的深度情感特征,以实现语音情绪的精确识别,为此分别设计3组实验来证明算法的有效性:通过对比基线模型和深度注意模型(I, II)来证明本模型的有效性,记为实验1;在基线模型背景下,通过测试不同的输入特征(传统特征、语谱图特征)证明语谱图时频域关联性等潜在特征更适合情绪的识别,再与本模型进行对比,证明本模型在少量增加模型复杂度的前提下,能够提取出更具有区分力的特征,从而提高了整体识别率,记为实验2;通过更换测试数据库,检测模型的泛化效果,记为实验3。

3.1 实验数据库

为了验证本文算法的有效性,本文采用普遍认同的情感数据库:自然情感数据库(FAU-AEC)和柏林德语数据库(EMO-DB)进行仿真。实验1

和实验2中选择的数据库是FAU-AEC,情感类别参考了国际语音情感挑战赛 Inter speech 2009 EC 对FAU-AEC的分类标准,挑选的5类情感分别是:愤怒(敏锐、谴责)、强调、中性、高兴(母语、快乐)和其他。在实验3的泛化效果评估中,数据集为EMO-DB,根据EMO-DB情感分类标准共选出7类,分别是:中性、恐惧、厌恶、高兴、烦躁、难过和愤怒。将80%的数据作为训练集;将其余数据(20%)作为测试集。算法均采用有监督训练,数据的类别标签只在训练时使用,实验结果以识别率的形式呈现。

采用 Opensmile 工具箱^[16]来提取传统特征,实验采用 Inter speech 2010 语音情感识别竞赛使用的特征集,共1582维特征,包含38个底层描述及其一阶差分,这里采用差值算法将其扩展成固定尺寸输入网络中。

3.2 实验设置

(1) 实验1

模型学习率 L 和输入批次 B 的设置要适当,过大过小都会加大重构误差,经测试,本实验设置的学习率为0.001,输入批次为400个 epochs(轮次)。根据参考文献[18]在最优性能下确定网络的层数,卷积部分以 VGGnet 为基础经过多次试验形成具体层设置,层参数(卷积层数、卷积层内核大小和滤波器数)为(1, 3×3, 64)、(1, 3×3, 128)、(2, 3×3, 256)、(3, 3×3, 512)。每个卷积层后加归一化层,卷积层后连接一个 BiLSTM。以下实验均以语谱图作为模型输入。3个算法模型依次是基线模型 CRNN 模型、基于改进 SEnet 结构的混合模型(CRNN_CSEnet)和基于改进 Itti 算法和改进 SEnet 结构的混合模型(CRNN_ISEnet),识别率见表1。

通过比较3个算法的识别率,从实验结果可以看出以下内容。

(1) 识别率 CRNN_ISEnet > CRNN_CSEnet > CRNN, 可以看出 CRNN_ISEnet 算法相对于基本



算法平均识别率上升 8.43%，说明 SEnet 本身的空间关联性对情绪的区分有一定帮助，通过目标情感特征来甄选强相关特征的方法更进一步增加了情感间区分性；CRNN_ISEnet 算法相比于改进算法 CRNN_CSEnet 平均识别率提升 2.43%，表明增加的纹理特征和基于声学敏感度的改进对识别效果有较小提升，该结果表明，在今后的研究中对于目标特征的选取，也可以尝试对其他特征进行探索。

表 1 CRNN、CRNN_CSEnet 和 CRNN_ISEnet 算法下的情绪识别率

情绪类别	CRNN	CRNN_CSEnet	CRNN_ISEnet
愤怒	67.89%	68%	64.5%
强调	47.56%	70.54%	80.17%
中性	89.16%	84.88%	90.94%
高兴	74.69%	75.79%	74.64%
其他	35.84%	58.06%	58.92%
平均识别率	63.02%	71.45%	73.83%

(2) CRNN_ISEnet 相比 CRNN，部分情感识别率有所上升，比如上升幅度较大的强调和其他这两种情感，识别率分别上升 22.98%和 22.22%，但愤怒和其他情感的识别率较低，分别为 68%和 58.06%，主要原因可能是两种情感的情感效价维和激活维比较接近。

表 1 中 CRNN_ISEnet 算法混淆矩阵如图 5 所示。这里定义：愤怒 (A)、强调 (E)、中性 (N)、高兴 (P) 和其他 (R)，对于混淆度较高的情绪，其中愤怒和强调、其他和强调混淆度较高，因为本数据库采用的是自然情感数据库 (FAU-AEC)，该数据库中强调发音特性与愤怒和其他较相似，导致模型误判率升高，可以看出总体上本模型对于情绪特征区分性较明显。

(2) 实验 2

以下实验的 3 个算法模型依次是基线模型 CRNN I 模型，输入为传统特征；基线模型 CRNN II，输入为语谱图；CRNN_ISEnet 模型，输入为语谱图。识别率见表 2。

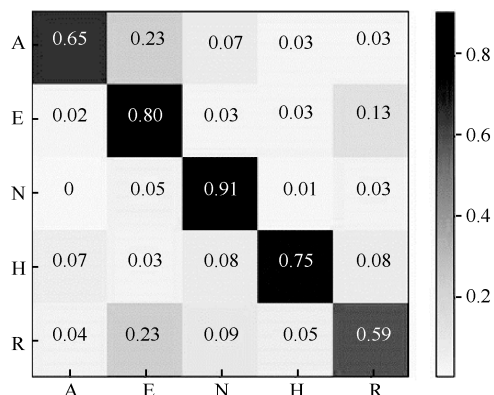


图 5 实验 1 中 CRNN_ISEnet 算法的混淆矩阵

表 2 CRNN1、CRNN2 和 CRNN_ISEnet 算法下的情绪识别率

情绪分类	CRNN I	CRNN II	CRNN_ASEnet
愤怒	53.63%	67.89%	64.5%
强调	40.56%	47.56%	80.17%
中性	76.71%	89.16%	90.94%
高兴	60.08%	74.69%	74.64%
其他	34.84%	35.84%	58.92%
平均识别率	53.16%	63.02%	73.83%

从 CRNN I 和 CRNN II 算法可以看出，语谱图特征相较于传统特征具有较强的情绪表达能力，这是因为传统特征只包含全局特征，而语谱图从时频域角度出发，包含相邻帧间相关信息，可以弥补传统特征单一域视角的不足；同时表明本模型在少量增加模型复杂度的前提下，优于传统特征加基准模型和语谱图加基线模型。

(3) 实验 3

以下实验数据集为 EMO-DB，实验模型为 CRNN_ISEnet 算法，这里选取输出的混淆矩阵进行分析。识别率见表 3。

从实验结果可以看出，在 EMO-DB 数据集下平均识别率为 79.97%，相较于 FAU-AEC 数据集提升了 6.14%。这一方面得益于本模型良好的泛化效果；另一方面得益于 EMO-DB 中包含的语音由专业演员所录制，相较于自然情绪数据库可能存在情绪放大效果，所以对识别效果具有良性影响。

表3 CRNN_ISEnet 算法下的情绪识别率

情绪分类	中性	恐惧	厌恶	高兴	烦躁	难过	愤怒
中性	75.27%	16.13%	1.08%	1.79%	2.15%	3.23%	0.36%
恐惧	0.75%	80.05%	1.13%	1.12%	5.64%	1.88%	9.40%
厌恶	0.62%	7.41%	82.10%	1.23%	5.56%	2.47%	0.62%
高兴	3.39%	1.69%	3.87%	85.71%	0.48%	0	4.84%
烦躁	7.40%	9.63%	0	2.22%	78.51%	3.70%	5.18%
难过	4.88%	2.93%	4.88%	3.41%	1.46%	80.49%	1.95%
愤怒	0.93%	1.40%	6.98%	6.97%	3.72%	2.32%	77.67%

4 结束语

本文主要研究了基于深度空间注意模型的语音情感识别算法,提出了CRNN_ISEnet算法模型。通过增加约束条件获取标定权重的SEF特征,标定权重的SEF特征送入微调的深层混合网络(F-CRNN)提取深层特征,最后进行情绪分类。通过在两个数据库上实验,仿真结果证明,在较少增加算法复杂度的前提下,相较于传统特征加基准模型算法和语谱图加基准模型算法,本算法识别结果更具有情感区分度。同时在EMO-DB上识别率效果也同样明显,证明了算法良好的泛化效果。本算法验证了C_SEnet结构能够甄选出更具有识别性的情感,同时有益于增强非同类情感间区分度;本文对Itti算法的改进对识别效果有一定提升,可以看出继续探索与情感具有强相关性特征的重要性。

参考文献:

- [1] 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 W J, LI H F, RUAN H B, et al. A review of research progress in speech emotion recognition[J]. Journal of Software, 2014, 25(1): 37-50.
- [2] 王海坤, 潘嘉, 刘聪. 语音识别技术的研究进展与展望[J]. 电信科学, 2018, 34(2): 1-11.
WANG H K, PAN J, LIU C. Research progress and prospect of speech recognition technology[J]. Telecommunications Science, 2018, 34(2): 1-11.
- [3] YAMADA T, HASHIMOTO H, TOSA N. Pattern recognition of emotion with neural network[C]//The 1995 IEEE IECON 21st International Conference on Industrial Electronics, Control, and Instrumentation, Nov 6-10, 1995, Orlando, FL, USA. Piscataway: IEEE Press, 1995: 183-187.
- [4] TENG Z, JI W. Speech emotion recognition with i-vector feature and rnn model[C]//2015 IEEE China Summit and International Conference on Signal and Information Processing (China SIP), July 12-15, 2015, Chengdu, China. Piscataway: IEEE Press, 2015: 524-528.
- [5] BASU A, CHAKRABORTY J, AFTABUDDIN M. Emotion recognition from speech using convolutional neural network with recurrent neural network architecture[C]//2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), Dec 13-16, 2016, Jeju, South Korea. Piscataway: IEEE Press, 2017: 333-336.
- [6] SHI B, BAI X, YAO C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(11).
- [7] ZAZO R, LOZANO-DIEZ A, GONZALEZ D J, et al. Language identification in short utterances using long short-term memory (LSTM)[J]. Recurrent Neural Networks, 2016(1).
- [8] GELLY G, GAUVAIN J L, LE V, et al. A divide-and-conquer approach for language identification based on recurrent neural networks[Z]. 2016.
- [9] LOZANO-DIEZ A, ZAZO C R, GONZALEZ D J, et al. An end-to-end approach to language identification in short utterances using convolutional neural networks[Z]. 2015.
- [10] ZHANG X R, SONG P, ZHA C, et al. Auditory attention model based on Chirplet for cross-corpus speech emotion recognition[J]. Journal of Southeast University, 2016, 32(4): 402-407.
- [11] 纪滨, 杨盼盼, 申元霞. 基于改进 ITTI 模型及粒子群优化算法的白细胞区域提取[J]. 安徽工业大学学报, 2016, 33(3): 284-288.

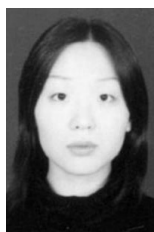


- JI B, YANG P P, SHEN Y X. Leukocyte region extraction based on improved ITTI model and particle swarm optimization algorithm[J]. Journal of Anhui University of Technology, 2016, 33(3): 284-288.
- [12] 刘兵, 霍键亮. 基于灰度概率统计的视觉注意改进算法[J]. 电子设计工程, 2013, 21(5): 54-57.
- LIU B, HUO J L. Improved visual attention algorithm based on gray probability statistics[J]. Electronic Design Engineering, 2013, 21(5): 54-57.
- [13] KALINLI O, CHEN R. Speech syllable/vowel /phone boundary detection using auditory attention cues: US20120253812[P]. 2016-02-02.
- [14] STEVENS C, HARN B, CHARD D J, et al. Examining the role of attention and instruction in at-risk kind ergarteners electrophysiological measures of selective auditory attention before and after an early literacy intervention[J]. Journal of Learning Disabilities, 2013, 46(1): 73-86.
- [15] 张欣然, 巨晓正, 宋鹏, 等. 用于垮库语音情感识别的 DBN 特征融合方法[J]. 信号处理, 2017, 33(5): 649-650.
- ZHANG X R, JU X Z, SONG P, et al. DBN feature fusion method for voice emotion recognition in library[J]. Signal Processing, 2017, 33(5): 649-650.
- [16] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[J]. arXiv:1709.01507, 2017.
- [17] EYBEN F, WOLLMER M, SCHULLER B. openSMILE—the Munich versatile and fast open-source audio feature extractor[C]//The 18th ACM International Conference on Multimedia, October 25-29, 2010, Firenze, Italy. New York: ACM Press, 2010: 1459-1462.
- [18] BARTZ C, HEROLD T, HAOJIN Y, et al. Language identification using deep convolutional recurrent neural networks[J]. arXiv: 1708.04811v1, 2017.

[作者简介]



王金华 (1992-), 女, 杭州电子科技大学硕士生, 主要研究方向为深度学习与语音处理。



应娜 (1978-), 女, 博士, 杭州电子科技大学副教授、硕士生导师, 主要研究方向为信号处理与人工智能。



朱辰都 (1995-), 男, 杭州电子科技大学硕士生, 主要研究方向为语音处理。

刘兆森 (1995-), 男, 杭州电子科技大学硕士生, 主要研究方向为深度学习与图像处理。

蔡哲栋 (1994-), 男, 杭州电子科技大学硕士生, 主要研究方向为深度学习与图像处理。