基于改进 K 均值聚类的语音情感识别深度学习方法

李巧君[1]　郭　骥[2]

[1](河南工业职业技术学院电子信息工程学院　河南 南阳 473000)

[2](电子科技大学电子科学与工程学院　四川 成都 610054)

摘　要　针对当前语音情感识别(Speech Emotion Recognition,SER)方法中准确性低和时间复杂度高的问题,提出一种基于改进 K 均值聚类的语音情感识别深度学习方法。采用改进的 K-均值聚类算法从整个音频信号中选取反映情感特征的关键片段;使用短时傅里叶变换将所选序列转化为一个谱图;利用深度残差模型 ResNet 和深度双向长短时记忆 Bi-LSTM 网络从空间和时间上学习表征谱图中与情感相关的隐藏特征,基于 Softmax 分类器获得最终的情感分类。实验结果表明,所提方法比其他识别方法具有明显的优势,在改善情感识别率的同时,降低了模型的处理时间。

关键词　语音情感识别　深度双向长短时记忆　K-均值聚类　短时傅里叶变换

中图分类号　TP393　　**文献标志码**　A　　DOI:10.3969/j.issn.1000-386x.2024.09.032

DEEP LEARNING METHOD FOR SPEECH EMOTION RECOGNITION BASED ON IMPROVED K-MEAN CLUSTERING

Li Qiaojun[1]　Guo Guo[2]

[1](*School of Electronic Information Engineering*, *Henan Polytechnic Institute*, *Nanyang 473000*, *Henan*, *China*)

[2](*College of Electronic Science and Engineering*, *University of Electronic Science and Technology of China*, *Chengdu 610054*, *Sichuan*, *China*)

Abstract　Aimed at the problems of low accuracy and high time complexity in current speech emotion recognition (SRE) methods, a deep learning method for speech emotion recognition based on the improved k-mean clustering is proposed. The improved k-mean clustering algorithm was used to select the key segments which reflected the emotional features from the whole audio signal. The selected sequence was transformed into a spectrum by using short-time Fourier transform. The deep residual model ResNet and deep Bi-LSTM network were used to learn the hidden features related to emotion in the representation spectrum from space and time. The final sentiment classification was obtained based on Softmax classifier. Experimental results show that the proposed method has obvious advantages over other recognition methods, which improves the emotion recognition rate and reduces the processing time of the model.

Keywords　Speech emotion recognition　Deep Bi-LSTM　K-mean clustering　Short-time Fourier transform

0　引　言

语音信号是人类之间进行情感表达交流的主要手段之一,人们可以通过语音传达自己的情绪和心智状态等,聆听者也能根据自身对语言信号的理解识别来感知对方的情绪变化[1]。语音情感识别(SRE)就是利用计算机对人类语音情绪处理系统进行模拟,根据从语言信号中提取到的语言特征以及建立的映射关系来识别人类的基本情感状态[2]。SER 在信号处理、人工智能和模式识别等领域已成为一项重要而具有挑战性的任务,引起了人们极大关注[3-4]。

收稿日期:2021-04-07。河南省高等学校重点科研项目(19A520022);河南省高等职业学校青年骨干教师培养计划项目(教职成函〔2019〕326 号)。**李巧君**,教授,主研领域:人工智能,深度学习,数据挖掘,信息安全。**郭骥**,副研究员。

由于人类的情感可以以不同的语音形式进行表达,因此,语音情感识别是一个复杂性和多维性的问题。通常,SER 任务分为信号预处理、特征选择和分类 3 个主要步骤[5],其中语音特征提取是 SER 系统中的关键步骤,旨在提取表征人类情感表达的有效特征表示。研究人员在 SER 早期使用的大多数是低级的音频特征,如韵律特征(音调和强度)、语音质量特征(共振峰)和频谱特征(Mel 频率倒谱系数),虽然这类方法取得了一定的进展,但是由于语音信号是非平稳信号,但是也存在情感识别率低、泛化能力差的局限性[6]。随着深度学习的迅猛发展,SER 也迎来了新的突破方向。目前,大多数研究人员采用深度学习技术从大量的语音信号或频谱图中提取高级判别特征用于提高模型的识别率和泛化能力。Zhao 等[7]提出了一种基于多通道 CNN 特征融合的语音识别方法,分别采用 1 维和 2 维 CNN 两个通道来学习语言信号的深层特征,然后将学习到的 1D 和 2D 表征特征进行融合用于识别语音情感。Ma 等[8]提出了一种基于深度神经网络的可变长度语音的情感识别方法,将 CNN 和 RNN 有效地结合在一起,从频谱图中提取对情绪识别有用的综合的准语言信息,用于提高情感识别准确率。上述方法中仅从单一的时间域序列或者频率域序列中提取特征,忽略了频率特征或时间维度的影响,时频域间的潜在关联性被忽略了。因此,部分研究人员尝试将两者结合在一起,用于更好地学习表征与情感相关的特征。Chen 等[9]提出了一种基于注意力的三维 CNN-LSTM 网络,该网络采用 Mel 声谱图作为输入来学习 SER 的判别特征。Zhao 等[10]构建了两个 CNN-LSTM 网络,分别从语音信号和对数 Mel 声谱图中学习与局部和全局情感相关的特征。

目前,基于深度学习的语音情感识别方法提高了识别的准确性,但是随着网络权值的增加,计算量也随之增加。因此,针对传统 CNN-LSTM 网络结构中存在识别准确性低和计算代价高的问题,提出一种基于特征聚类和深度学习的 SER 方法,首先采用改进的 K-均值聚类算法从整个音频信号中选择反映情感特征的关键片段;其次利用短时傅里叶变换(Short-Time Fourier Transform,STFT)算法将所选序列转化为一个谱图,并将其传递到深度残差模型 ResNet 中;然后 ResNet 网络从谱图中提取出辨别特征和显著特征;最后采用深度双向长短时记忆(Bi-LSTM)学习时间信息用于识别情绪的最终状态。在所提方法中,为了降低整体模型的计算复杂度,仅仅选取了关键片段而不是整个信号长度用于情感识别,从而有效减少了处理时间。同时,为了保证模型具有精确的识别性能和能够轻松地识别时空信息,需要在输入前对 ResNet 特征进行规范化处理。

1　SER 模型设计

本文提出的基于特征聚类和深度学习的语音情感识别方法主要由 2 个模块组成,即基于特征聚类的语音谱图生成和基于深度学习的情感识别。图 1 给出了本文方法的具体流程。

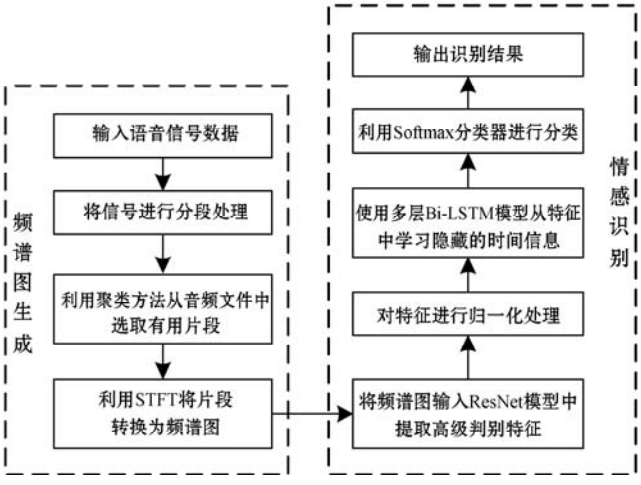


图 1　本文方法的基本流程

1.1　基于改进 K-均值聚类方法的谱图

由于频谱特征能够更好地描述情感细节的时频相关性,越来越多的研究人员将谱特征应用于语音情感识别。频谱特征的优点是将语音频谱建模为图像,利用图像特征描述符从频谱中提取情感信息。大多数研究学者是直接利用短时傅里叶变换将时域语音信号转换为时频视觉表示的谱图。但是,由于语音信号中存在大量冗余信息,计算量大,因此使用直接生成的谱图会影响模型的整体效率。

为了降低模型的计算复杂度,本文仅仅选取了信号的关键片段用于生成谱图,具体步骤如图 2 所示。第一步是将音频文件分成多个均匀的片段。在分段过程中,选择合适的时间来划分音频文件是一个具有挑战性的问题。本文通过对多帧持续时间的不同观察,优化选择 500 ms 窗口大小,将单个话语转换成多个片段,并分配一个单一标签给该话语的所有片段。第二步是基于采用 K-均值聚类算法对相似片段进行分组。K-均值聚类算法[11]是一种迭代的数据分割算法,在大数据分组中应用最为广泛。该方法随机选取 k 个观测值作为初始类簇中心点,然后计算每个采样点到簇中心点的距离,并将采样点指定给具有距离最近的簇。K-均值聚类算法一般采用欧几里得距离矩阵来度量样

本间的距离,在 SER 中,则表现为度量两帧之间的差值。

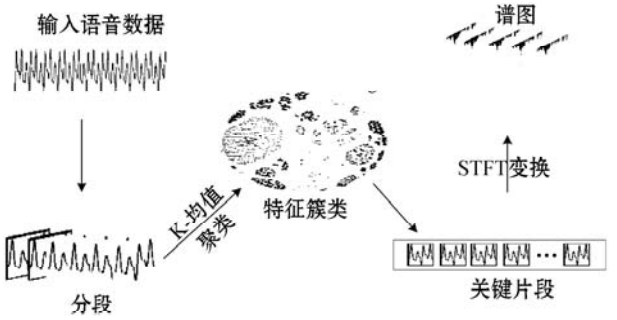


图 2 谱图生成过程

但是,为了优化算法性能,本文采用径向基函数 RBF 来计算两帧之间的差值。这是因为人脑的视觉感知部分是基于非线性处理系统进行语音处理和识别工作,而 RBF 恰恰是使用非线性方法来计算非线性片段之间的相似性,正适合寻找音频片段中的相似性度量。RBF 网络的基本思想是基于核函数将向量从线性不可分的低维度映射到线性可分的高维度的过程。RBF 网络的最终输出表示为:

$$y_j = \sum_{i=1}^{h} \omega_{ij} R(x_p - \mu_i) \tag{1}$$

式中:ω_{ij}为权重系数;$R(x_p - \mu_i)$为 RBF 网络的 1 维高斯激活函数。$R(x_p - \mu_i)$定义为:

$$R(x_p - \mu_i) = \exp\left(\frac{-\|x_p - \mu_i\|^2}{2\sigma^2}\right) \tag{2}$$

式中:μ_i为隐藏层第 i 个节点的高斯函数中心点;σ 为第 i 个节点的标准偏差。

在 RBF 中,使用 1 维高斯形状模型作为映射函数来确定音频片段之间的相似度。由于期望的标准偏差 σ 对距离的变化十分敏感,可作为判断片段相似性的依据:当语音信号的特定片段相关性大时,则 σ 在语音片段中较小;如果语音片段不相关,则 σ 的值很大。此外,传统的 K-均值算法使用随机初始化技术来选择 K 值,但在本文使用镜头边界检测方法(shot boundary detection)[12]动态地选择每个文件的 K 值来估计相似性。首先从 $K=1$ 开始,如果连续帧中的差值在阈值内,则估计为成对差异,当差值大于所选阈值,则将 K 值增加。在用 K-均值算法对所有的片段进行聚类后,从每个聚类中选出一个靠近聚类中心的片段作为关键片段。

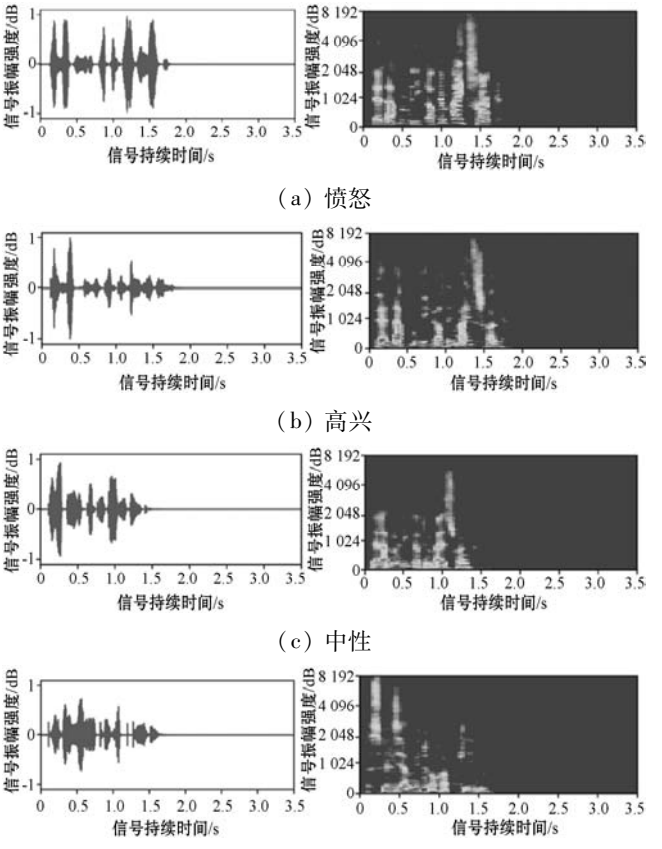
由于选取的关键片段是一维时域信号,很难分析频率变化规律,而基于傅里叶变换后频域信号又容易丢失时域信息。因此,本文采用基于 STFT 算法将选定的关键片段转化为谱图进行二维表示。STFT 选择汉明窗函数进行实现,帧长为 25 ms,速率为 10 ms,其定义如下:

$$h_i(f,m) = |STFT\{x_i\}(f,m)|^2 \tag{3}$$

式中:x_i 表示一个关键片段;f 代表频率;m 代表窗口位置。最后,通过式(4)将 f Hz 信号缩放到 m mel 标度波段来生成 mel 谱图。

$$m = 2\,595\log_{10}\left(1 + \frac{f}{700}\right) \tag{4}$$

图 3 给出了将声音信号转换为 mel 谱图的过程。



(a) 愤怒

(b) 高兴

(c) 中性

(d) 悲伤

图 3 4 种情绪的声音信号和 mel 谱图

1.2 基于深度学习的语音情感识别

当谱图生成之后,将其输入到深度学习模型中用于音频数据的特征提取和情感识别。在深度学习框架中,包含 ResNet 网络和多层 Bi-LSTM 网络两个部分。CNN 网络能够有效地提取输入数据中隐藏的特征信息,而 ResNet 网络作为典型的 CNN 架构之一,在继承了 CNN 优势的基础上,利用残差模块解决深度学习网络中网络退化问题,为了提高模型的性能,本文选择 ResNet101 模型[13]。LSTM 是分析时间序列数据中隐藏信息特征的最有效的网络,但是简单的 LSTM 网络无法正确识别训练出的海量、复杂序列的数据,因此采用一种多层深 Bi-LSTM 模型来学习和识别音频数据中的长时序列。深度双向长短时记忆 Bi-LSTM 是一种特殊的 RNN 网络,是由两个上下叠加在一起的反向 LSTM 组成。该模型的输出也基于这两个 LSTM 的状态共同决定:

$$\begin{cases} \overrightarrow{\boldsymbol{h}}_t = LSTM(x_t, \overrightarrow{\boldsymbol{h}}_{t-1}) \\ \overleftarrow{\boldsymbol{h}}_t = LSTM(x_t, \overleftarrow{\boldsymbol{h}}_{t-1}) \\ \boldsymbol{h}_t = \boldsymbol{\omega}_t \overrightarrow{\boldsymbol{h}}_t + \boldsymbol{\vartheta}_t \overleftarrow{\boldsymbol{h}}_t + \boldsymbol{b}_t \end{cases} \tag{5}$$

式中:$\boldsymbol{h}_t$ 表示经过深度双向长短时记忆模型(Bi-LSTM)处理后的结果;$\boldsymbol{\omega}_t$ 和 $\boldsymbol{\vartheta}_t$ 为权重系数;$\boldsymbol{b}_t$ 表示偏置项。与单个 LSTM 模型相比,Bi-LSTM 模型不仅利用时间序列中前后时刻数据之间的正向关联信息,还考虑了前后时刻的反向关联信息,因此在时间序列的分类问题中展现出了优越的性能。

在基于深度学习的情感识别中,首先需要将谱图序列输入到 ResNet101 模型进行训练,表征每个片段的隐藏特征,并且通过使用全连接层的转移学习策略来提取高级判别特征。ResNet-101 是 101 层残余网络,该网络的设计灵感来自 VGG-19 模型。通常,在 CNN 模型中,几个层相互连接,并被训练来执行各种任务,网络的最后部分负责学习多个层次的特征。ResNet 模型中卷积层的大小主要有 33 个滤波器。对于相同的输出特征图大小,层具有相同数量的滤波器,如果特征图大小减半,则滤波器的数量将增加一倍,以保持每个层的时间复杂度。ResNet 通过将步长为 2 的卷积层进行卷积来直接执行下采样,然后以一个全局平均池化层和一个 Softmax 激活的全连接层终止。实验证明,训练 ResNet 网络比训练简单的深度卷积神经网络更容易,并且该模型还解决了精度下降的问题。为了保证模型具有精确的识别性能和能够轻松地识别时空信息,采用全局平均值和标准差将提取到的特征进行归一化处理,如图 4 所示。

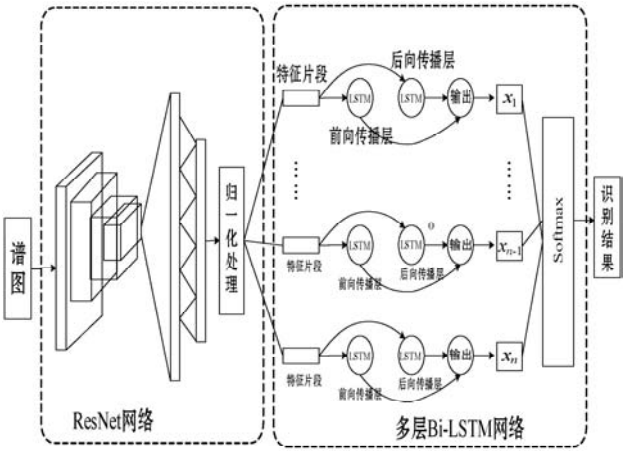


图 4　基于深度学习的情绪识别

然后,将处理后的特征输入多层 Bi-LSTM 模型中。在 Bi-LSTM 模型中,输入以正向传递和反向传递进行处理,允许网络在每个时间步合并未来和过去的信息,因此,经过训练后的输出充分结合了前向和后向传递隐藏的状态。最后将多层 Bi-LSTM 的输出状态反馈给 Softmax 分类器,以获得网络的最终分类结果。本文使用了 100 个 LSTM 单元,然后采用丢失概率为 0.5 速率的 Dropout 层和 64 个隐藏单元的密集层。最后一个密集层包含的神经元 n 等于相应数据集中的输出类数(IEMOCAP 数据集时 $n=4$,EmoDB 数据集时 $n=7$)。对网络进行端到端训练,使分类交叉熵损失最小,作为模型的目标函数。这样,模型可以从数据序列中学习提取的特征,并将学习到的内部特征映射到不同的情感类型。

2　实验与结果分析

为了验证本文方法的有效性,在 2 个公开的基准语音情感数据集上进行测试,并将测试结果与 3D-ACRNN[4]、ADRNN[5]、CNN-KELM[14]、PCRN[15] 和 Att-BLSTM-FCN[14] 等方法进行比较。尽管数据集数的情绪类别是不平衡的,但是在测试过程中的所有结果均为非加权平均召回率情况下得到的。所有实验在 MATLAB2019b 中实现,使用神经网络工具箱进行特征提取、模型训练和评估。实验环境为 Windows 10 系统下,i7-4790 @ 3.60 GHz 处理器和 32 GB RAM。

在训练阶段,我们对模型进行了参数调整,使其足够最优,并对不同批次和学习率进行了不同的实验,以选择最优解。选择 Adam 优化算法进行模型优化,并选择最佳偏差修正,以获得更好的效果。通过对比发现,本文采用将批量尺寸设置为 512,学习率为 0.001。模型从零开始训练,并进行微调以适应训练数据,

2.1　实验数据集

本文采用 IEMOCAP[15] 和 EMO-DB[16] 数据集评价所提模型的性能,数据集的详细内容为:

(1) IEMOCAP 是一个著名的数据集,由视听资料和两位专业演员之间的对话录音记录组成。数据集包含有脚本式和即兴式两种类型的对话,共计 12 小时的视听数据,包括音频、视频、面部动作、语音和文本转录。数据集有五个会话,每个会话以 16 kHz 的采样率记录 3 到 15 秒长的情感脚本,其中包含愤怒、悲伤、快乐、中立、惊讶、厌恶、沮丧、兴奋和恐惧等不同的情绪类别。本文选择以文学作品中最常用的愤怒、悲伤、快乐和中性四种情绪来评价模型。具体细节描述如表 1 所示。

表 1　两个数据集的具体描述

分类	IEMOCAP		EMO-DB	
	话语总量	占比/%	话语总量	占比/%
愤怒	1 103	19.94	127	23.74
悲伤	1 084	19.96	62	11.59
高兴	1 636	29.58	71	13.27

续表 1

分类	IEMOCAP		EMO-DB	
	话语总量	占比/%	话语总量	占比/%
中性	1 708	30.88	79	14.77
厌恶	—	—	46	8.60
恐惧	—	—	69	12.90
无聊	—	—	81	15.14

（2）EMO-DB 数据集收录了 50 位男性演员和 5 位女性演员的情感。每个演员阅读预选的句子，带有不同的情绪，如愤怒、恐惧、无聊、厌恶、高兴、中立和悲伤。在 Emo-DB 中，采样率为 16 kHz 的语音大约为 2 到 3 秒。具体细节描述如表 1 所示。

2.2 结果分析

首先给出本文模型在两个数据集上的混淆矩阵，用于分析所提模型的识别精度。表 2 给出了 IEMOCAP 数据集的混淆矩阵，该矩阵表示了原始情绪标签和预测情绪标签识别结果。可以看出，模型对愤怒和悲伤情绪有很好的识别效果，两种情绪的识别率分别为 92% 和 89%。快乐情绪相对于其他情绪的识别率是最低的，识别率为 64%。

表 2 IEMOCAP 数据集的混淆矩阵(%)

情绪	愤怒	高兴	中性	悲伤
愤怒	92.5	7.5	0.0	0.0
高兴	7.3	64.8	1.6	26.3
中性	0.0	5.2	79.1	15.7
悲伤	5.0	6.0	0.0	89.0

表 3 给出了 Emo-DB 数据集的混淆矩阵，所提模型在 EMO-DB 数据集上取得了较好的效果，除去快乐情绪外，其他情绪的识别率均在 90% 以上，其中对愤怒、恐惧和悲伤情绪的识别率较高。本文方法在 EMO-DB 数据集中被混淆在快乐和中性情绪之间，大多数快乐的情绪被认为是中性的。

表 3 EMO-DB 数据集的混淆矩阵(%)

情绪	愤怒	无聊	厌恶	恐惧	高兴	中性	悲伤
愤怒	96.2	0.0	0.0	1.9	1.9	0.0	0.0
无聊	0.0	91.4	0.0	0.0	0.0	4.1	4.5
厌恶	0.0	0.0	90.0	4.0	4.0	0.0	2.0
恐惧	0.0	4.0	0.0	96.0	0.0	0.0	0.0
高兴	7.2	0.0	0.0	0.0	75.4	17.4	0.0
中性	4.3	0.0	0.0	0.0	0.0	94.5	1.2
悲伤	0.0	0.0	0.0	0.0	0.0	3.1	96.9

其次，为了进一步验证本文模型的性能，将在 IEMOCAP 和 EMO-DB 数据集中的测试结果与其他识别方法进行对比。图 5 和图 6 分别给出了不同方法在两个数据集的分类精度和处理时间。可以清楚地看出，本文方法与其他传统的 3D-ACRNN、CNN-KELM 和 ADRNN 相比，该 SER 方法有效提高了分类精度和降低了处理时间，并在 IEMOCAP 和 EMO-DB 基准数据集上获得了更好的结果。这是因为本文方法通过基于 RBF 的 K 均值聚类技术从语音序列形成的每个簇中选取关键片段，并应用 STFT 将所有关键片段转换为二维谱图。然后利用 Resnet101 CNN 模型的 FC-1000 层，从光谱图中提取获得高层次的特征信息，并传递给 deep-BiLSTM 进行分类。此外，图 4 还对比了采用不同度量距离的 K-均值聚类算法的分类结果。可以清楚地看到，采用径向基函数 RBF 来计算两帧之间的差值的方式在识别率方面有较大的改善。

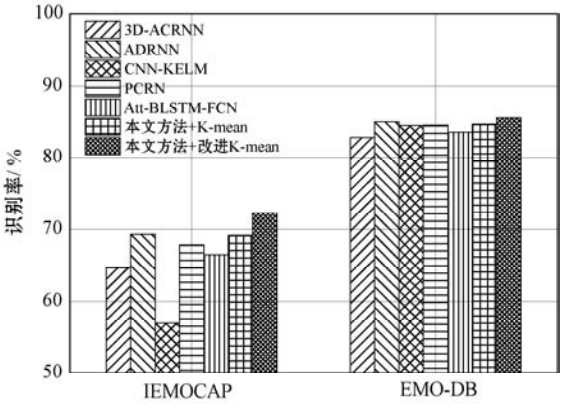


图 5 不同方法的识别率对比

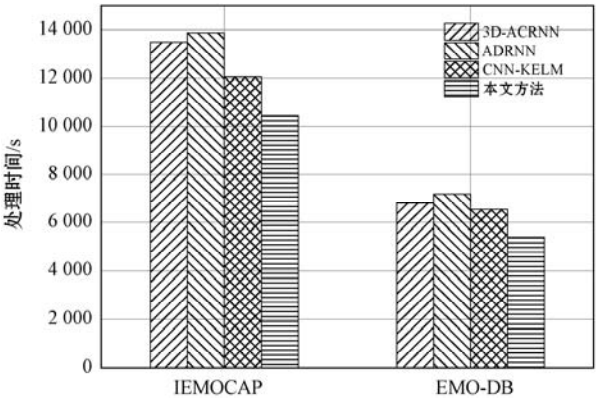


图 6 不同方法的处理时间对比

3 结 语

本文提出一种基于特征聚类和深度学习的 SER 方法，用于解决基于深度学习语音情感识别系统的计算复杂度高和精度不够的问题。所提方法主要分为 4 个步骤：首先采用改进的 K-均值聚类算法对分段的音

频信号进行动态聚类,从每个聚类中选出一个靠近聚类中心的序列作为关键片段;其次使用 STFT 变换将所选序列转化为一个二维谱图,输入到深度学习模型中;然后使用 ResNet 和 Bi-LSTM 网络的组合模型来学习谱图中的时空信息;最后利用 Softmax 分类器对输出特征进行分类。该方法通过特征聚类和深度学习提高情感识别的准确度,降低整体模型的处理时间。实验结果表明,相较于其他的识别方法,本文方法的性能最优。

参 考 文 献

[1] 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[J]. 计算机工程,2019,45(8):248 - 254.

[2] Zamil A, Hasan S, Baki S, et al. Emotion detection from speech signals using voting mechanism on classified frames [C]//2019 International Conference on Robotics, Electrical and Signal Processing Techniques,2019:281 - 285.

[3] Badshah A, Rahim N, Ullah N, et al. Deep features-based speech emotion recognition for smart affective services[J]. Multimedia Tools and Applications, 2019, 78(5): 5571 - 5589.

[4] Liu Z, Wu M, Cao W, et al. Speech emotion recognition based on feature selection and extreme learning machine decision tree[J]. Neurocomputing,2018,273:271 - 280.

[5] Hao M, Yan T, Fei Y, et al. Speech emotion recognition from 3D log-mel spectrograms with deep learning network [J]. IEEE Access,2019,7:125868 - 125881.

[6] 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.

[7] Zhao J, Mao X, Chen L, et al. Learning deep features to recognise speech emotion using merged deep CNN[J]. IET Signal Processing,2018,12(6):713 - 721.

[8] Ma X, Wu Z, Jia J, et al. Emotion recognition from variable-length speech segments using deep learning on spectrograms[C]//2018 Conference of the International Speech Communication Association,2018:3683 - 3687.

[9] Chen M, He X, Yang J, et al. 3-D convolutional recurrent neural networks with attention model for speech emotion recognition[J]. IEEE Signal Processing Letters,2018,25(10): 1440 - 1444.

[10] Zhao J, Mao X, Chen L, et al. Speech emotion recognition using deep 1D & 2D CNN LSTM networks[J]. Biomedical Signal Processing and Control,2019,47:312 - 323.

[11] Hajarolasvadi N, Demirel H. 3D CNN-based speech emotion recognition using K-means clustering and spectrograms[J]. Entropy,2019,21(5):479.

[12] Wu L, Zhang S, Jian M, et al. Two stage shot boundary detection via feature fusion and spatial-temporal convolutional neural networks[J]. IEEE Access,2019,7:77268 - 77276.

[13] Xu Z, Sun K, Mao J. Research on ResNet101 network chemical reagent label image classification based on transfer learning[C]//2020 IEEE 2nd International Conference on Civil Aviation Safety and Information Technology,2020:354 - 358.

[14] Guo L, Wang L, Dang J, et al. Exploration of complementary features for speech emotion recognition based on kernel extreme learning machine[J]. IEEE Access,2019,7:75798 - 75809.

[15] Jiang P, Fu H, Tao H, et al. Parallelized convolutional recurrent neural network with spectral features for speech emotion recognition[J]. IEEE Access,2019,7:90368 - 90377.

[16] Zhao Z, Bao Z, Zhao Y, et al. Exploring deep spectrum representations via attention-based recurrent and convolutional neural networks for speech emotion recognition[J]. IEEE Access,2019,7:97515 - 97525.

(上接第 223 页)

[13] Xu H. Hierarchical cost-sensitive techniques for class imbalance learning[C]//2021 4th International Conference on Artificial Intelligence and Big Data (ICAIBD),2021.

[14] Pasupa K, Vatathanavaro S, Tungjitnob S. Convolutional neural networks based focal loss for class imbalance problem: A case study of canine red blood cells morphology classification[J]. Journal of Ambient Intelligence and Humanized Computing,2023,14:15259 - 15275.

[15] Miao L, Liu M, Zhang D. Cost-sensitive feature selection with application in software defect prediction[C]//21st International Conference on Pattern Recognition (ICPR2012), 2012.

[16] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[EB]. arXiv:1706.03762,2017.

[17] Cho K, Merrienboer B V, Gulcehre C, et al. Learning phrase representations using RNN Encoder-Decoder for statistical machine translation[EB]. arXiv:1406.1078,2014.

[18] 林伟. 基于 BiGRU-CNN 的网络舆情情感识别模型[J]. 中国人民公安大学学报(自然科学版),2023,29(2):61 - 66.

[19] 梁越,刘晓峰,李权树,等. 面向司法文本的不均衡小样本数据分类方法[J]. 计算机应用,2022,42(S2):118 - 122.

[20] 宋明,刘彦隆. Bert 在微博短文本情感分类中的应用与优化[J]. 小型微型计算机系统,2021,42(4):714 - 718.

[21] 王雯慧,靳大尉. 基于改进 Focal Loss 和 EDA 技术的 UT 分类算法[J]. 计算机仿真,2023,40(4):346 - 349.