文章编号: 1671-6833(2021)06-0028-06

基于时空特征的语音情感识别模型 TSTNet

薛均晓[1,2], 黄世博[1], 王亚博[1], 张朝阳[3], 石 磊[1,2]

(1.郑州大学 软件学院,河南 郑州 450002; 2.郑州大学 网络空间安全学院,河南 郑州 450002; 3.郑州大学 信息工程学院,河南 郑州 450001)

摘 要: 针对社交语音由于语气、音调、语速等差异以及填充信息丢失或冗余等问题,提出一种基于时空特征的语音情感识别方法。该方法利用卷积神经网络(CNN)和双向循环神经网络(BiGRU)技术,包含空间特征提取、时间特征提取和特征融合 3 个模块。考虑到音频数据内容长短不一,首先对音频数据进行预处理,应用 3 种补零填充方法,得到不同尺度的语谱图。设计了空间特征提取方法捕获音频的局部特征,并利用时间特征提取方法获取音频数据的时间特征和前后语义关系,从而得到 3 个时空特征向量。此外,融合了时空特征向量并通过全连接层进行语音情感分类。利用科大讯飞语音情感数据集进行了数值实验,实验结果与传统语音情感识别模型的实验结果相比,在准确率、精确率、召回率和 *F*1 值等 4 项指标上均取得了较好结果。

关键词: 语音情感识别; 语谱图; 时空特征

中图分类号: TP39 **文献标志码**: A **doi**: 10.13705/j.issn.1671-6833.2021.06.008

0 引言

语音情感识别是人机交互领域的重要技术,在安全驾驶、采集病人情绪状态、结合情感辅助发言等方面都有广泛的应用。现实生活中,由于语音多样性、环境多样性,以及说话者的说话习惯、性别、语气、音调、语速等问题,导致语音的情感识别成为一项具有挑战性的工作。

近年来,随着深度学习的迅速发展,研究人员在语音情感识别领域运用深度学习技术,取得了很好的成果[1-4],但仍存在一些需要改进的地方:①对于语音的分析中并没有全部关注到语音的空间特征、时序特征以及前后语义关系;②对于语音样本长度参差不齐的问题,填充长度过长会导致每个样本中增添很多冗余信息,过短则会导致数据丢失。

针对上述问题,本文提出一种基于时空特征的语音情感识别方法。该方法由空间特征提取模块、时间特征提取模块以及特征融合模块组成。空间特征提取模块关注语音的空间特征,时间特征提取模块关注语音的时间特征和语音信号中前后语义关系。为了解决语音长度不一导致填充时信息丢失或冗余问题,模型采用 3 种补零填充长度得到 3 个不同尺度的语谱图,分别提取它们的空间特征、时间特征以及前后语义关系,在特征融合模块中将提取得到的 3 个特征向量融合到一起。

1 相关工作

1.1 情感描述方式

目前主要有 2 种描述情感的方法: 基于离散的方法和基于维度的方法。

情感的离散描述方法是将情感离散化,并进一步类别化。陈炜亮等[5]提出一种新的情感识别模型 MFCCG-PCA,实现生气、高兴、害怕、悲伤、惊讶和中性 6 种情感的分类。离散的描述方式简单并且应用广泛,但是情感描述单一。

情感的维度描述方法是将情感状态描述为一种笛卡尔空间,空间的每个维度对应 1 种情感属性。Schlosberg[6]提出倒圆锥三维情感空间,从 3 个维度对情感进行描述,将情感描述成 1 个倒立圆锥形的空间模型。基于维度的情感描述方法利用多维的数值来表示情感,能够描述情感的微妙变化。

收稿日期: 2021-03-22; **修订日期**: 2021-06-18

基金项目: 河南省高等学校青年骨干教师培养计划(22020GGJS014)

通信作者: 石磊(1967—),男,河南郑州人,郑州大学教授,博士,博士生导师,主要从事人工智能、网络空间安全方面的研究,E-mail: shilei@zzu.edu.cn。

1.2　语音情感识别分类器

早期的语音情感识别模型主要有隐马尔可夫模型、支持向量机等传统的模型。Lin 等[7]利用隐马尔可夫模型和支持向量机识别 5 种情绪。Pan 等[8]探究线性预测频谱编码(LPCC)、梅尔频谱系数(MFCC)等特征，并在相关数据集上训练支持向量机。

近年来，基于深度学习的方法成为语音情感识别的研究热点。Mao 等[9]提出使用卷积神经网络学习情感显著性特征；Trigeorgis 等[10]结合卷积神经网络和长短期记忆网络，提出解决“情境感知”情感相关特征的方法；Badshah 等[11]提出 3 个卷积神经网络结合 3 个全连接层的模型从语谱图中提取特征，并预测 7 种情感；Tzirakis 等[12]利用卷积神经网络和长短期记忆网络，提出一种端到端的连续语音情感识别方法；Zhang 等[13]利用预训练的 AlexNet 模型以及支持向量机预测话语级情绪。

2　方法

本文提出了一种基于卷积神经网络 CNN 和双向循环神经网络(BiGRU)的语音情感识别模型 TSTNet，模型结构如图 1 所示。在数据预处理部分，首先对一个语音信号样本进行傅里叶变换，针对 3 种补零填充长度得到 3 个不同尺度的语谱图，然后将其依次输入空间特征提取模块和时间特征提取模块中，得到 3 个特征向量，最后将这 3 个特征向量进行特征融合和情感分类。

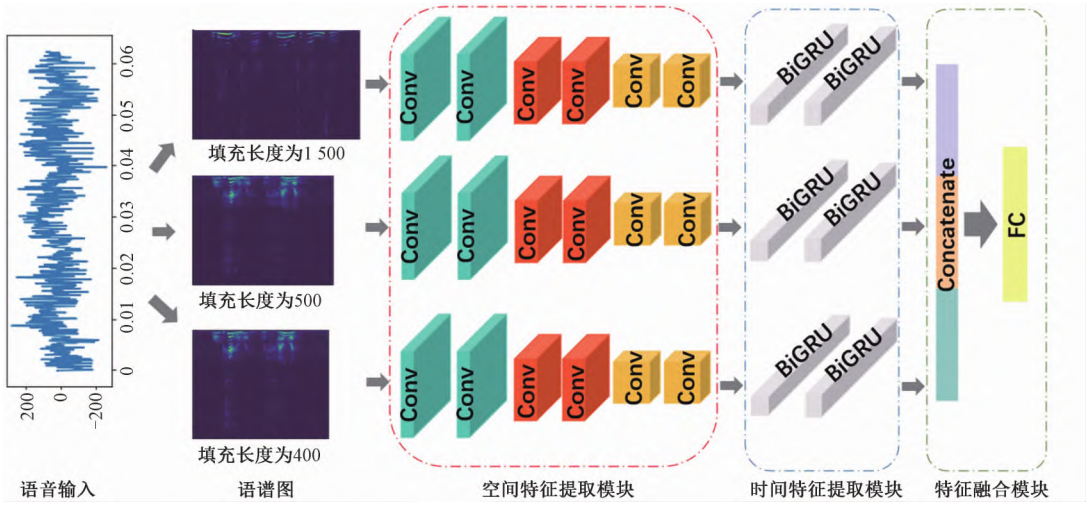

图 1　TSTNet 模型结构

Figure 1　TSTNet model structure

2.1　语谱图

在预处理部分考虑到语音长度相差很大的问题，首先将普通的 WAV 语音信号采用 3 种补零填充长度进行填充，并转换为语谱图。基于对数据信号长度分布情况的分析，选择的 3 种填充长度分别为 400、800、1 500。语谱图的转换过程如图 2 所示。

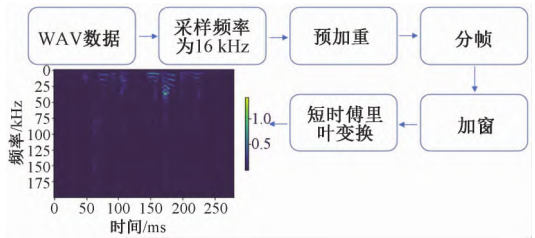

图 2　语谱图转换过程

Figure 2　Spectrogram conversion process

首先对语音信号进行采样、量化、编码处理，使之转变成数字信号。通过下采样，将语音信号的采样率由 44.1 kHz 转化为 16 kHz。为避免在傅里叶变换操作期间出现数值问题，对模数转换后的数据帧预加重，并进行分帧、加窗以及短时傅里叶变换，得到需要的语谱图(spectrogram)。

2.2　空间特征提取模块

卷积神经网络(convolutional neural networks，CNN)对于图像和语音的特征提取有出色的表现。将 2.1 节中的 3 个不同尺度的语谱图其中之一(维度为 $[L, 200, 1]$，$L \in (400, 800, 1\,500)$)送入 CNN 中，利用 CNN 去捕获音频的局部特征，其他 2 个语谱图处理过程与此相同。卷积层的计算式为

$$\boldsymbol{Y}_i = f(\boldsymbol{W}_i \otimes \boldsymbol{X} + b_i)。\tag{1}$$

式中：$\boldsymbol{X} \in \mathbf{R}^{L\times 200\times 1}$ 为语谱图矩阵；$\boldsymbol{W}_i$ 为卷积核的权重值；$\otimes$ 为卷积操作；b_i 为卷积核的偏置值，i

为卷积核数;$f(\cdot)$ 表示 ReLU 函数,其定义为

$$\boldsymbol{Y}_i = \max(0, \boldsymbol{Z}_i)。 \tag{2}$$

式中: $\boldsymbol{Z}_i = \boldsymbol{W}_i \otimes \boldsymbol{X} + b_i$。将得到的特征 $\boldsymbol{Y}_i$ 输入平均池化层,一个池化区的计算式为

$$S_j = \frac{1}{|R_j| \sum_{i \in R_j} \boldsymbol{P}_i}。 \tag{3}$$

式中: R_j 为池化区的像素点数; j 为区域数; $\boldsymbol{P}_i$ 为 $\boldsymbol{Y}_i$ 一个通道中的池化区; i 为池化区第 i 个像素点。

在空间特征提取模块,模型使用 6 层卷积神经网络,卷积核通道分别为 32、32、64、64、128、128,卷积核大小均为 3×3。3 个语谱图经过空间特征提取模块得到 3 个特征向量,送入时间特征提取模块中。

2.3 时间特征提取模块

GRU[14](gate recurrent unit) 是循环神经网络(recurrent neural networks, RNN) 的变体。将空间特征提取模块中提取的 3 个特征向量展开,分别输入 GRU,一个 GRU 单元的计算式为

$$z_t = \sigma(\boldsymbol{W}_t \cdot [x_t, h_{t-1}]); \tag{4}$$

$$r_t = \sigma(\boldsymbol{W}_t \cdot [x_t, h_{t-1}]); \tag{5}$$

$$\widetilde{h}_t = \tanh(\boldsymbol{W} \cdot [x_t, r_t \cdot h_{t-1}]); \tag{6}$$

$$h_t = z_t \cdot \widetilde{h}_t + (1 - z_t) \cdot h_{t-1}。 \tag{7}$$

式中: z_t、r_t 分别为更新门和重置门; $\boldsymbol{W}_t$ 为 t 时刻的权重值; x_t 为 t 时刻的输入; h_{t-1} 为 $(t-1)$ 时刻的状态输入; $\sigma(\cdot)$ 为 Sigmoid 函数; $\widetilde{h}_t$ 为计算当前节点状态; $\tanh(\cdot)$ 为激活函数; h_t 为计算当前节点输出。GRU 单元如图 3 所示。

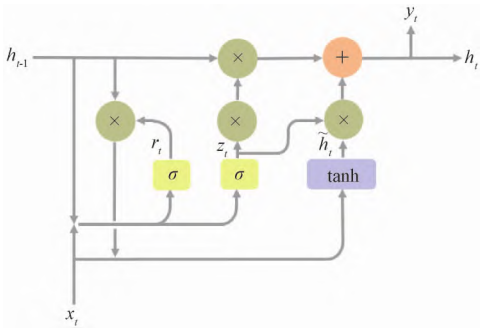

图 3 GRU 单元

Figure 3 GRU unit

文中使用双向循环神经网络(BiGRU) 提取音频的时间特征以及前后语义关系。BiGRU 模型结构如图 4 所示,圆圈表示 GRU 单元,BiGRU 模型是 2 个单向 GRU 的结合。在前向传播的信息流中,输入数据 x_t 和 $(t-1)$ 时刻的状态输入 h_{t-1} 计算出当前时刻 t 的输出 A_t,如式(8) 所示。在后向传播的信息流中,输入数据 x_t 和时刻 $(t+1)$ 的状态输入 h_{t+1} 计算出当前时刻 t 的输出 A'_t,如式(9) 所示,最终的输出 y_t 取决于 A_t 和 A'_t。

$$A_t = f(h_{t-1}, x_t); \tag{8}$$

$$A'_t = f'(h_{t+1}, x_t)。 \tag{9}$$

式中: f、f'分别表示 GRU 单元前向、后向传播; h_{t-1}、h_{t+1} 分别表示前向传播中时刻 $(t-1)$ 状态输入、后向传播中时刻 $(t+1)$ 状态输入。

时间特征提取模块中 BiGRU 层数为 2 层,中间设置一层 Dropout 层,BiGRU 序列长度设置为 128。

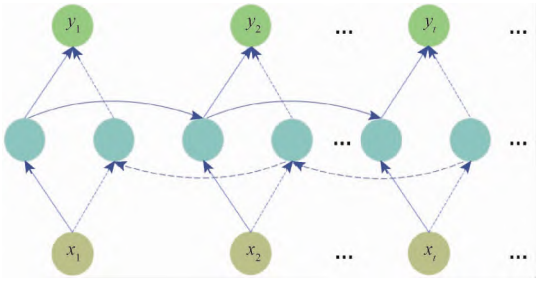

图 4 BiGRU 模型结构

Figure 4 BiGRU model structure

2.4 特征融合模块

TSTNet 模型利用 CNN 处理由语音生成的语谱图,提取出语音中的局部区域特征; BiGRU 关注语音的时间特征以及前后语义关系,故将 CNN 与 BiGRU 结合搭建 TSTNet 模型。

3 种尺度的语谱图经过空间特征提取模块、时间特征提取模块之后,得到 3 个特征向量。在特征融合模块中,将这 3 个特征向量拼接在一起,得到 1 个新的特征向量。将该特征向量输入 1 个 FC 层和 1 个 Softmax 函数,得到最终的语音情感识别结果。如图 1 中的特征融合模块所示。

3 实验分析

3.1 数据集

实验数据集来自科大讯飞,数据集总共有 7 004个音频样本,详细描述如表 1 所示。样本标签分布如图 5 所示,可知标签数量分布均匀。实验按照 8∶2的比例随机划分数据集,80%的样本作为训练集,20%的样本作为测试集。

3.2 实验环境

实验使用 Keras 框架搭建 TSTNet 模型,所用到的硬件设备为 NVIDIA RTX2080Ti。模型参数配置的详细情况如表 2 所示。

表 1　科大讯飞数据集

Table 1　HKUST IFLYTEK data set

数据类型	情绪类型	采样环境	数据格式	采样率/kHz
语音数据	开心、伤心、愤怒、平和	自然环境、低噪声环境	WAV 格式	44.1

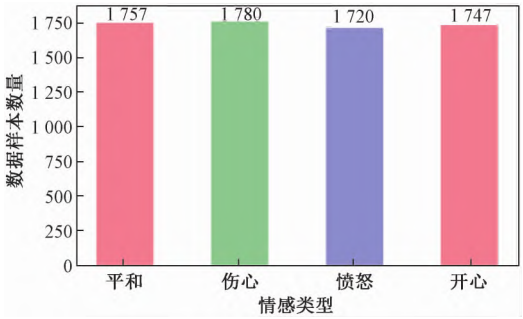

图 5　数据集标签分布情况

Figure 5　Data set label distribution

表 2　模型参数配置

Table 2　Model parameter configuration

学习率/10^{-4}	批处理	最大训练轮数	损失函数	优化器
1	8	40	交叉熵损失函数	Adam

3.3　实验结果分析

将 TSTNet 模型的实验结果和以下 5 个已有的情感识别模型的实验结果进行对比，实验指标为准确率、精确率、召回率和 *F*1 值。

(1) MFCC+随机森林。提取语音数据中的 MFCC 特征，将提取的 MFCC 归一化并求最大值得到语音特征向量，用随机森林去拟合提取的特征向量。

(2) 语谱图+CNN。通过傅里叶变换将语音转化为语谱图，用 CNN 网络提取语谱图特征。

(3) MFCC+CNN。基于 MFCC+CNN 的方法已经被应用于多种领域中，比如在语音识别[15]领域，此方法获得了很好的效果。在语音情感识别任务中，提取语音中的 MFCC 特征，然后输入 CNN 网络中对情感进行识别。

(4) 语谱图+CNN+RNN。在语谱图和 CNN 的基础上加上 RNN，去捕捉语音的时序特征。

(5) 语谱图+CNN+LSTM。LSTM 广泛应用于语音识别[16]、文本情感分析[17]中。这里将 CNN 和 LSTM 结合应用于情感识别中。

TSTNet 模型与后 4 种模型对比，得到实验的训练准确率曲线和损失值曲线，分别如图 6、7 所示。从图 6、7 中可知，相比于其他方法的模型，TSTNet 模型在准确率和损失值上都表现良好，得到了较好的准确率；TSTNet 模型训练的波动幅度相对平稳，对数据拟合程度较好。

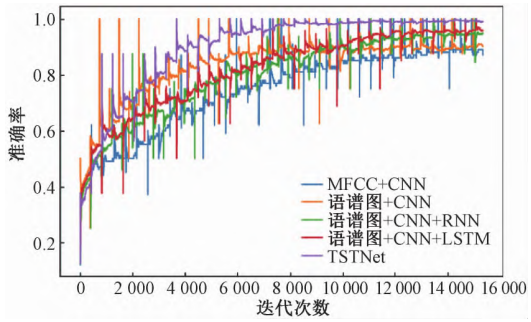

图 6　准确率曲线

Figure 6　Accuracy curve

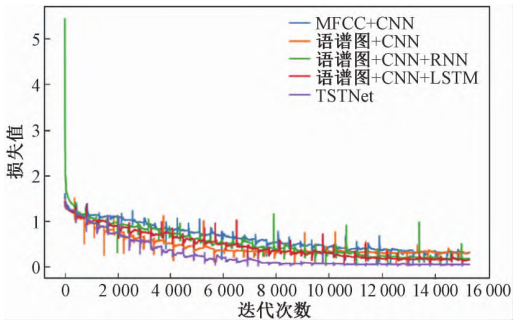

图 7　损失值曲线

Figure 7　Loss value curve

TSTNet 模型和以上模型在准确率、精确率、召回率和 *F*1 上的测试集表现情况如表 3 所示。从表 3 中可以看出，基于深度学习的方法比传统方法效果好，并且 TSTNet 模型在准确率、精确率、召回率、*F*1 值上都得到了较好的结果。

表 3　不同模型在准确率、精确率、召回率、*F*1 值上的表现

Table 3　Performance of different models on accuracy, precision, recall, and *F*1 values　%

方法	准确率	精确率	召回率	*F*1 值
MFCC+随机森林[18]	53.24	51.39	46.23	46.20
语谱图+CNN[19]	90.17	88.55	88.41	88.48
MFCC+CNN[20]	90.37	88.77	88.44	88.56
语谱图+CNN+RNN[21]	91.12	89.59	89.13	90.43
语谱图+CNN+LSTM[9]	91.72	90.79	90.90	91.65
TSTNet	**94.69**	**93.89**	**94.34**	**94.08**

本文方法采用不同的语音填充长度，分别为 400、800、1 500，最后在特征融合模块将它们集成到一起。为了验证模型中集成方法的有效性以及 BiGRU 特征提取的有效性，训练了 4 个实验模型，填充长度分别为 400、800、1 500 以及填充长度为 800 但没有使用 BiGRU，对比结果如表 4 所示，训练过程准确率曲线如图 8 所示。

表4 TSTNet模型消融实验

Table 4 TSTNet model ablation experiment

实验模型	准确率/%
填充长度为400	92.02
填充长度为800	92.71
填充长度为1 500	92.12
没有BiGRU(填充长度为800)	88.87
TSTNet	**94.69**

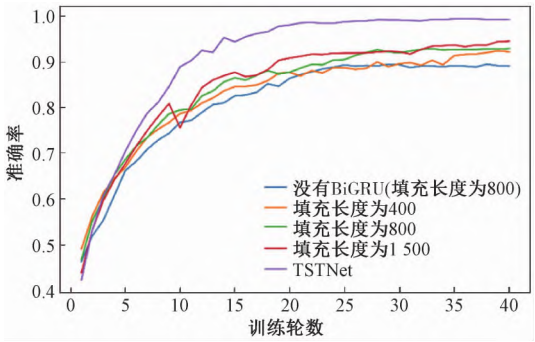

图8 模型训练准确率曲线

Figure 8 Model training accuracy curve

由表4和图8可知,填充长度为800的模型比没有使用BiGRU(填充长度为800)的模型的准确率高,集成3种填充长度的TSTNet比3个单一填充的实验效果明显。由此可验证TSTNet模型中的BiGRU可以关注到语音的前后语义关系。前后语义关系以及不同填充长度的集成方法对于语音情感识别准确率的提高有重要的意义。

4 结论

本文提出了一种语音情感识别模型TSTNet,该模型结合CNN和BiGRU,能够关注语音信号中的前后双向语义关系(two-way semantic relationship)以及时空特征(spatial-temporal features)。采用3种不同的填充长度进行特征融合,能较好缓解语音长度相差大导致填充时信息丢失或冗余的问题。本文方法在实验数据集上能够得到94.69%的识别准确率,相对于基于MFCC和随机森林等语音情感识别方法,本文方法在多项实验指标上效果显著。

参考文献:

[1] QAYYUM A B A,AREFEEN A,SHAHNAZ C.Convolutional neural network (CNN) based speech-emotion recognition[C]//2019 IEEE International Conference on Signal Processing, Information, Communication & Systems (SPICSCON). Piscataway: IEEE, 2019: 122-125.

[2] DAVIS S,MERMELSTEIN P.Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences[J].IEEE transactions on acoustics, speech, and signal processing, 1980, 28(4): 357-366.

[3] HUANG Z W,DONG M,MAO Q R,et al.Speech emotion recognition using CNN[C]// Proceedings of the 22nd ACM International Conference on Multimedia. New York: ACM,2014: 801-804.

[4] 王蔚,胡婷婷,冯亚琴.基于深度学习的自然与表演语音情感识别[J].南京大学学报(自然科学版),2019,55(4):660-666.

[5] 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.

[6] SCHLOSBERG H. Three dimensions of emotion[J]. Psychological review,1954,61(2):81-88.

[7] LIN Y L,WEI G.Speech emotion recognition based on HMM and SVM[C]//2005 International Conference on Machine Learning and Cybernetics. Piscataway: IEEE,2005: 4898-4901.

[8] PAN Y, SHEN P, SHEN L. Speech emotion recognition using support vector machine[J]. International journal of smart home, 2012, 6(2): 101-108.

[9] MAO Q R, DONG M, HUANG Z W, et al. Learning salient features for speech emotion recognition using convolutional neural networks[J].IEEE transactions on multimedia,2014,16(8):2203-2213.

[10] TRIGEORGIS G, RINGEVAL F, BRUECKNER R, et al. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network[C]//2016 IEEE International Conference on Acoustics,Speech and Signal Processing (ICASSP). Piscataway: IEEE,2016: 5200-5204.

[11] BADSHAH A M, AHMAD J, RAHIM N, et al. Speech emotion recognition from spectrograms with deep convolutional neural network[C]//2017 International Conference on Platform Technology and Service (PlatCon). Piscataway: IEEE,2017: 1-5.

[12] TZIRAKIS P,ZHANG J H,SCHULLER B W.End-to-end speech emotion recognition using deep neural networks[C]//2018 IEEE International Conference on Acoustics,Speech and Signal Processing (ICASSP). Piscataway: IEEE,2018: 5089-5093.

[13] ZHANG S Q,ZHANG S L,HUANG T J,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.

[14] CHUNG J, GULCEHRE C, CHO K, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling [EB/OL]. (2014-12-11) [2021-03-10]. https://arxiv.org/abs/1412.3555.

[15] CHOWDHURY A, ROSS A. Fusing MFCC and LPC features using 1D triplet CNN for speaker recognition in severely degraded audio signals [J]. IEEE transactions on information forensics and security, 2020, 15: 1616-1629.

[16] 赵淑芳,董小雨.基于改进的 LSTM 深度神经网络语音识别研究[J].郑州大学学报(工学版),2018,39(5):63-67.

[17] 李勇,金庆雨,张青川.融合位置注意力机制和改进 BLSTM 的食品评论情感分析[J].郑州大学学报(工学版),2020,41(1):58-62.

[18] BREIMAN L. Random forests [J]. Machine learning, 2001, 45: 5-36.

[19] 张雄,刘蓉,刘明.基于卷积特征提取与融合的语音情感识别研究[J].电子测量技术,2018,41(16):138-142.

[20] ZISAD S N, HOSSAIN M S, ANDERSSON K. Speech emotion recognition in neurological disorders using convolutional neural network [C]// International Conference on Brain Informatics. Cham: Springer, 2020: 287-296.

[21] 王金华,应娜,朱辰都,等.基于语谱图提取深度空间注意特征的语音情感识别算法[J].电信科学,2019,35(7):100-108.

Speech Emotion Recognition TSTNet Based on Spatial-temporal Features

XUE Junxiao[1,2], HUANG Shibo[1], WANG Yabo[1], ZHANG Chaoyang[3], SHI Lei[1,2]

(1.School of Software, Zhengzhou University, Zhengzhou 450002, China; 2.School of Cyberspace Security, Zhengzhou University, Zhengzhou 450002, China; 3.School of Information Engineering, Zhengzhou University, Zhengzhou 450002, China)

Abstract: For differences in tone, pitch, speaking speed, etc. of social speech and information loss or redundancy during filling, a speech emotional recognition method was proposed based on spatial-temporal features. The method applied convolutional neural network (CNN) and bilateral recurrent neural network (BiGRU), including spatial feature extraction module, temporal feature extraction module and feature fusion module. Considering the different lengths of audio data content, the audio data was preprocessed first, and three zero-padded padding lengths were applied to obtain spectrograms of different scales. Then the spatial feature extraction module was designed to capture the local feature of the audio, and used the temporal feature extraction module to obtain the temporal feature and the semantic relationship of the audio data, thus obtained three spatial-temporal feature vectors. In addition, these temporal feature vectors were fused and input full connection layer for classification of speech emotion. With the numerical experiment using IFLYTEK speech emotion data sets, the experiment achieved better results in the accuracy, precision, recall, and $F1$ value than those of the experiment of traditional speech emotion recognition model.

Keywords: speech emotion recognition; spectrogram; spatial-temporal features