太原理工大学学报
Journal of Taiyuan University of Technology
ISSN 1007-9432,CN 14-1220/N

《太原理工大学学报》网络首发论文

题目：　基于主辅网络特征融合的语音情感识别
作者：　胡德生，张雪英，张静，李宝芸
网络首发日期：　2021-04-01
引用格式：　胡德生，张雪英，张静，李宝芸．基于主辅网络特征融合的语音情感识别．太原理工大学学报.
https://kns.cnki.net/kcms/detail/14.1220.N.20210401.1511.004.html

太原理工大学学报
Journal of Taiyuan University of Technology

基于主辅网络特征融合的语音情感识别

胡德生，张雪英，张静，李宝芸

(太原理工大学，信息与计算机学院，山西 太原 030024)

摘　要：语音情感识别是人机交互的重要研究方向，有效的特征提取与融合是提高语音情感识别率的关键因素之一。本文提出了一种使用主辅网络进行深度特征融合的语音情感识别算法。首先将段特征输入 BLSTM-Attention 网络作为主网络，注意力机制能够关注语音信号中的情感信息；然后，把 Mel 语谱图输入 CNN-GAP 网络作为辅助网络，GAP 可以减轻全连接层带来的过拟合；最后，将两个网络提取的深度特征以主辅网络方式进行特征融合，解决不同类型特征直接融合带来的识别结果不理想的问题。在 IEMOCAP 数据集上对比四种模型的实验结果表明：使用主辅网络深度特征融合的 WA 和 UA 均有不同程度上的提高。

关键词：语音情感识别；主辅网络；长短时记忆单元；卷积神经网络

中图分类号：TP181; TP399　**文献标识号：**A

Feature Fusion Based on Main-Auxiliary Network for Speech Emotion Recognition

Hu Desheng, Zhang Xueying, Zhang Jing, Li Baoyun

(College of Information and Computer, Taiyuan University of Technology, Taiyuan 030024, China)

Abstract: Speech emotion recognition is an important research direction of human-computer interaction. Effective feature extraction and fusion is one of the key factors to improve the rate of speech emotion recognition. In this paper, a speech emotion recognition algorithm using Main-auxiliary networks for deep feature fusion is proposed. First, segment features are input into BLSTM-attention network as the main network. The Attention mechanism can pay attention to the emotion information in speech signals. Then, the Mel spectrum features are input into Convolutional Neural Networks - Global Average Pooling (GAP) as auxiliary network. GAP can reduce the overfitting brought by the fully connected layer. Finally, the two are combined in the form of Main-auxiliary networks to solve the problem of unsatisfactory recognition results caused by direct fusion of different types of features. The experimental results of comparing four models on IEMOCAP dataset show that WA and UA using the depth feature fusion of the Main-Auxiliary network are improved to different degrees.

Key words: speech emotion recognition; main-auxiliary network; long-short term memory; convolutional neural network

---

基金项目：国家自然科学基金项目(61371193)；山西省回国留学人员科研资助项目(HGKY2019025)；山西省研究生教育创新计划项目(No.2020BY130)。

作者简介：胡德生（1996-），男，安徽亳州人，硕士研究生，主要研究方向为深度学习、语音情感识别等；通信作者：张雪英（1964-），女，河北石家庄人，教授，博导，博士，主要研究方向为语音信号处理、大数据分析及应用等（tyzhangxy@163.com）；张静（1993-），男，山西吕梁人，博士研究生，主要研究方向为信号处理，脑认知与情感识别等；李宝芸（1996-），女，山西吕梁人，硕士研究生，主要研究方向为深度学习、语音情感识别等。

语言是人类交流最方便、最快捷的方式，语言中包含的情感信息在交流时发挥着重要作用。让机器像人一样具备说话、思维和情感能力，是人工智能领域一直追求的目标。语音情感识别的研究，将推动这一目标的逐步实现。

典型的语音情感识别模型由语音情感数据库、特征提取和识别三部分组成[1]，提取有效情感特征是语音情感识别的关键。传统的语音情感识别首先分帧提取 Mel 频率倒谱系数(Mel Frequency Cepstral Coefficients, MFCC)等声学特征，然后提取所有帧的最大值、最小值、均值、方差等统计特征作为语音信号的全局特征[2-4]。由于全局特征是在句子级别上提取统计特征，所以其只能粗略反应语音情感随时间变化的特性。针对这个问题，近年来段特征的概念被提出，首先将语音信号分段，每段包含若干帧语音，对这若干帧语音各自提取声学特征后，再计算这段语音的多个统计特征作为段特征。文献[2]将段特征直接输入基于注意力机制的长短时记忆单元(Long Short-Term Memory, LSTM)网络提取深度特征并分类，与全局特征相比取得了较好的效果。

语谱图是一维语音信号在二维时频域的展开，能够充分反映语音信号在时频域大部分信息，卷积神经网络(Convolutional Neural Networks, CNN)由于其自动学习特征的能力和适用于二维图像数据的特点，目前被广泛用在语谱图中提取特征，进一步提高语音情感识别性能[6-8]。如，文献[8] 先将语谱图输入全卷积网络(Fully Convolutional Networks, FCN)，并在最后一层卷积层使用注意力机制，最后进行情感识别，在 IEMOCAP 数据集上其 WA(Weighted Accuracy, WA)和 UA(Unweighted Accuracy, UA)分别达到 70.4%，63.9%。

近年来，国内外学者提出多种混合网络模型用于将不同类型的特征进行特征融合，提升了语音情感识别的性能[13,15,17]。文献[13]提出 HSF-CRNN 模型，采用 CRNN 网络对语谱图提取深度特征，将全局特征输入全连接层，最后将两者拼接进行情感识别；文献[15]提出 Attention-BLSTM-FCN 模型，在 Mel 语谱图上分别应用 Attention-BLSTM 网络和 Attention-FCN 网络，然后将两个网络提取的深度情感特征以直接拼接的方式进行特征融合，最后输入全连接层进行分类识别。虽然这些方法取得了一定的效果，但将不同类型的特征简单拼接起来作为识别网络的输入，没有考虑不同特征的量纲和维度的差异，以及各类型特征实际物理意义的不同，会对识别结果带来不利影响。

针对上述问题，本文提出了通过主辅网络方式将不同类别特征进行融合的方法。首先将段特征输入 BLSTM-Attention 网络作为主网络，提取深度段特征；然后，把 Mel 语谱图输入 CNN-GAP 网络作为辅助网络，提取深度 Mel 语谱图特征；最后，用深度 Mel 语谱图特征辅助深度段特征，将两者以主辅网络方式进行特征融合。在 IEMOCAP 数据集上的实验结果表明：使用主辅网络深度特征融合的 WA 和 UA 分别达到 74.45%、72.50%，比特征直接拼接的 WA 和 UA 分别提高了 1.24%、1.15%。

1 不同类别的特征提取

本节首先描述段特征提取和 Mel 语谱图生成，然后将段特征输入 BLSTM-Attention 网络，得到深度段特征；将 Mel 语谱图输入 CNN-GAP 网络，得到深度 Mel 语谱图特征。本节工作将为后继主辅网络特征融合打下基础。

1.1 段特征提取

段特征的提取步骤：

第一步：语音信号采样率为 16kHz，分帧处理时取窗长 256，窗移 128；

第二步：使用截断或补零的方式使所有语音长度为 1000 帧；

第三步：计算每一帧信号的平均过零率、能量、基音频率、共振峰、MFCC，共 18 维声

学特征[18]；

第四步：把 5 帧组成一段[19]，共 200 段。计算一段内声学特征的最大值、最小值、平均值、中值和方差等统计特征，得到一段信号的 18×5=90 维特征；

第五步：标准化处理，得到每一句情感语音信号的 90×200 的段特征，段内的声学特征如表 1 所示。

表 1 段内的声学特征及其统计特征

Tab.1 Acoustic feature and statistical feature within a segment

声学特征	统计特征
平均过零率(1 维)	最大值，最小值，平均值，中值，方差，共 5 维
能量(1 维)	最大值，最小值，平均值，中值，方差，共 5 维
基音频率(1 维)	最大值，最小值，平均值，中值，方差，共 5 维
共振峰(3 维)	前三个共振峰的最大值，最小值，平均值，中值，方差，共 15 维
MFCC(12 维)	前 12 阶最大值，最小值，平均值，中值，方差，共 60 维

1.2 Mel 语谱图生成

大量实验表明人耳听到的声音高低和实际频率(Hz)不呈线性关系，Mel 频率更加符合人耳的听觉特性，Mel 频率 f_{Mel} 和 Hz 频率 f 的关系如公式(1)所示。

$$f_{Mel}=1125(1+\frac{f}{700Hz}) \tag{1}$$

$$H_m(k)=\begin{cases}\dfrac{k-f(m-1)}{f(m)-f(m-1)} & f(m-1)\le k\le f(m)\\ \dfrac{f(m+1)-k}{f(m+1)-f(m)} & f(m)\le k\le f(m+1)\\ 0 & \text{其他}\end{cases} \tag{2}$$

Mel 滤波器组的输出计算如公式(2)所示，每个滤波器具有三角滤波特性，其中心频率为 $f(m)$，m 表示 Mel 滤波器的阶数，k 表示 FFT 中点的编号。

Mel 语谱图计算如公式(3)所示，$|X(k)|^2$ 表示频谱能量。

$$MelSpec(m)=\sum_{k=f(m-1)}^{f(m+1)}\log(H_m(k)|X(k)|^2) \tag{3}$$

Mel 语谱图的生成：对语音信号进行 STFT 变换，使用汉明窗，窗长 256，窗移 128；通过 40 阶 Mel 滤波器组得到 $H_m(k)$，如公式(2)所示，再将 $H_m(k)$ 乘以 $|X(k)|^2$ 求和得到 Mel 语谱图，如公式(3)所示。最后使用截断补零的方式使所有 Mel 语谱图大小对齐，得到大小为 40×432 的 Mel 语谱图。

1.3 基于 BLSTM-Attention 的深度段特征提取

LSTM 适合对序列问题进行建模，因此被广泛应用在语音识别和语音情感识别中。BLSTM 由正向 LSTM 和反向 LSTM 组成。BLSTM 不仅可以考虑输入数据以前的信息，还可以考虑输入数据未来的信息，可以更好的对序列问题进行建模。采用 BLSTM 对段特征进行建模可以提取考虑上下文情感信息的深度情感特征。本文使用两层 BLSTM，隐藏神经元个数为 300，为了减轻过拟合使用 dropout，弃权率为 0.5。

通过对 BLSTM 输出应用注意力机制可以关注输入的情感语音信号中更显著的情感片段，增强 BLSTM 网络提取显著深度段特征的能力。更具体的说，以 LSTM 为核心的识别器在时间上展开的长度是 T，则 LSTM 在每一时刻都对应一个输出。相比于平均池化输出和最后时刻输出，注意力机制可以兼顾 LSTM 层每一时刻的输出，其对 LSTM 网络每一时刻的输出分配不同的权重来考虑上下文情感信息的深层特征。注意力机制的具体计算如公式(4)所示：

$$\boldsymbol{s}^{l}=\sum_{1}^{\mathrm{T}}\alpha_t\boldsymbol{s}_t \tag{4}$$

$\boldsymbol{s}_t\in\boldsymbol{R}^{D}$ 是输入序列的某一元素，T 表示 LSTM 的某一时刻 α_t 为加权系数，可以通过公式(5)、(6)式算出，并通过网络训练进行更新：

$$\boldsymbol{a}_t=\boldsymbol{s}\mathrm{o}\,\mathrm{f}\,\mathrm{t}\,\mathrm{m}\,\boldsymbol{\beta}_t \tag{5}$$

$$\boldsymbol{\beta}_t=\gamma_{\beta}^{T}\sigma(\boldsymbol{W}_{\beta}\,\boldsymbol{s} \tag{6}$$

σ为非线性映射函数，如 Sigmoid 函数。$W_b \in R^{D*D}$、$\gamma_\beta^T \in R^D$为系数矩阵与系数向量，它们都是网络学习参数。

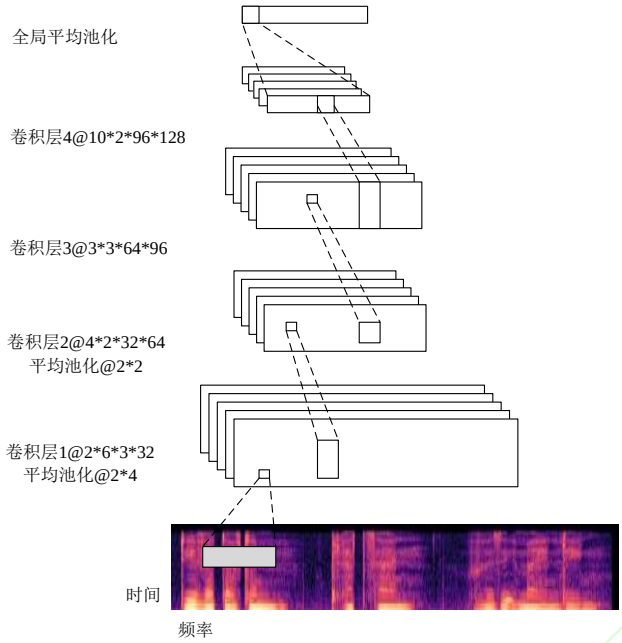

图 1 设计的卷积神经网络结构，卷积核大小表示为长×宽×输入通道×输出通道，池化层大小表示为长×宽

Fig.1 The design of the convolutional neural network structure, the convolutional kernel size is expressed as length × width × input channel × output channel, and the pooling layer size is expressed as length × width

1.4 基于 CNN-GAP 的深度 Mel 语谱图特征提取

由于 CNN 适合于二维图像数据，而 Mel 语谱图是一维语音信号在二维时频域的展开，因此 CNN 可以用来在 Mel 语谱图上提取深度特征。CNN 由卷积层和池化层组成，卷积层用来提取特征，池化层用来降低网络规模和过拟合，通常采用最大值池化或均值池化。

Mel 语谱图是一维语音信号在二维时频域的展开，能够充分反映语音信号在时频域大部分信息。针对 Mel 语谱图的这一特点，可以分别在时间轴和频率轴设计较大的卷积核，提取 Mel 语谱图的频率和时间特性，进而提取显著的情感特征。设计的卷积神经网络结构如图 1 所示。第一层卷积层在时间轴上设计较大的卷积核，提取 Mel 语谱图的时间特性；第二层卷积层在频率轴上设计较大的卷积核，提取 Mel 语谱图的频率特性；第三层卷积层使用 3×3 卷积核；第四层卷积层在频率轴上使用全卷积，最后在使用全局平均池化(Global Average Pooling, GAP)。用 GAP 代替全连接层可以减轻过拟合，使网络易于训练。每一层卷积层都使用了批归一化(Batch Normalization，BN)以及 Relu 激活函数。具体网络参数经过调参得到。

2 主辅网络特征融合模型

本节提出基于主辅网络特征融合的语音情感识别方法，详细叙述主辅网络的网络结构及主辅网络特征融合参数传递及更新过程。

2.1 主辅网络特征融合的网络结构

本文将段特征输入 BLSTM-Attention 网络提取了深度段特征，将 Mel 语谱图输入 CNN-GAP 网络提取深度 Mel 语谱图特征，通常将两者以直接拼接的方式进行特征融合，但是没有考虑不同特征的量纲和维度的差异，会对识别结果带来不利影响。因此，本文提出基于主辅网络特征融合的语音情感识别.

如图 2 所示，为主辅网络特征融合的网络结构。传统声学特征以时域特征为主，具有明确的物理意义，因此将其作为主网络输入特征。主网络是基于 BLSTM-Attention 的深度段特征提取模块，辅助网络是基于 CNN-GAP 的深度 Mel 语谱图特征提取模块。两者以主辅网络方式组成特征融合网络。主网络分为上、下两部分，M_U 代表上半部分，由全连接层构成；M_D 代表下半部分，由 BLSTM-Attention 网络构成。e_0 表示语音段特征，经标准化后作为主网络的输入，h^l 代表主网络 M_D 部分的输出，是深度段特征，维度是 600。辅助网络由 CNN-GAP 网络构成，v_0 表示 Mel 语谱图，作为辅助网络输入，v^M 表示辅助网络 GPA 的输

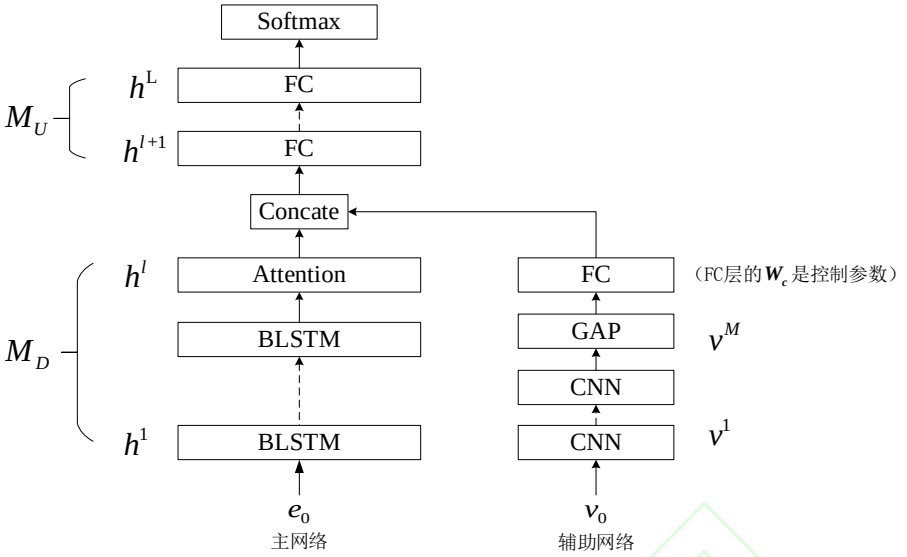

图 2 主辅网络特征融合的结构模型

Fig.2 A structural model of feature fusion of main-auxiliary networks

出，辅助网络 FC 层的W_c是控制参数(为了简化描述，省略了辅助网络 FC 层偏置项)，也是权重，维度是 128×200，一方面在主网络参数更新时可以控制辅助网络参数不更新，另一方面是对v^M进行特征变换。Concate 表示h^l与$W_c v^M$直接拼接，并输入主网络的M_U上半部分，其中h^l表示 BLSTM-Attention 网络提取的深度段特征，v^M表示 CNN-GAP 网络提取的深度 Mel 语谱图特征，W_c表示控制主辅网络训练的参数。然后将 Concate 拼接结果通过 FC 层做进一步特征融合，最后使用 Softmax 进行分类。

2.2 主辅网络特征融合参数传递及更新

本文提出的基于主辅网络特征融合的语音情感识别模型，最重要的是网络训练过程，也就是误差反传参数更新的过程，主辅网络特征融合参数传递示意图如图 3 所示。由于辅助网络的加入，网络的训练被分为三步：

第一步：参数初始化。将语音段特征e_0输入主网络，将 Mel 语谱图v_0输入辅助网络；然后将控制参数W_c初始化为0，主网络和辅助网络的权重和偏置通过截断正太分布随机初始化。

第二步：主网络训练。首先通过控制W_c和网络设置使辅助网络不起作用，然后使用梯度下降算法和反向传播算法训练主网络使主网络M_D和M_U参数更新。

第三步：辅助网络训练。首先将主网络M_D和M_U的权重和偏置设置为不更新，将辅助网络的权重、偏置和W_c设置为更新；使用梯度下降算法和反向传播算法训练辅助网络使辅助网络参数和W_c更新。

下面介绍辅助网络的一些参数更新。主辅网络最后一层输出拼接向量为$\hat{h}^l$，具体的拼接公式如(7)所示：

$$\hat{h}^l = h^l + W_c v^M \tag{7}$$

L为代价函数，根据标准的反向传播算法求得$\frac{\partial L}{\partial \hat{h}^l}$后，$\frac{\partial L}{\partial W_c}$、$\frac{\partial L}{\partial v^M}$可根据链式法则由(8)、(9)式求得：

$$\frac{\partial L}{\partial W_c} = \frac{\partial L}{\partial \hat{h}^l} v^M \tag{8}$$

$$\frac{\partial L}{\partial v^M} = \frac{\partial L}{\partial \hat{h}^l} W_c \tag{9}$$

根据标准的反向传播算法，辅助网络各层的参

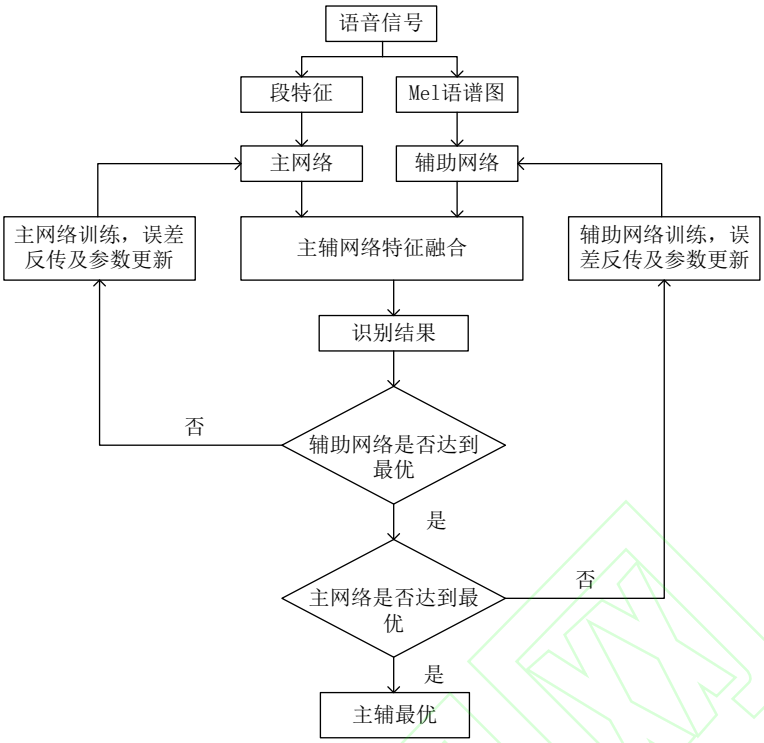

图 3 主辅网络特征融合参数传递及更新

Fig.3 Transfer and update of main-auxiliary network feature fusion parameters

数关于代价函数 L 的偏导数可以依据 $\frac{\partial L}{\partial v^{M}}$ 逐层推导得到。

3 实验及结果

3.1 实验数据库及网络参数设置

本文使用美国南加州大学发布的英文情感数据集 IEMOCAP(The Interactive Emotional Dyadic Motion Capture database)。该数据集由 5 个会话组成，每个会话由一对说话者(女性和男性)在预先设定的场景和即兴场景中对话。本文使用即兴场景对话中的语句，选取四种情感，分别是高兴、悲伤、愤怒、中性，共 2046 条语句。

本文使用 tensorflow 深度学习框架，以本文提出的网络模型为例，通过多次试验确定网络参数：主网络学习率为 0.0005，辅助网络学习率为 0.001，主网络 minibatch 大小为 96，辅助网络 minibatch 大小为 48，主辅网络均使用 Adam 优化器。

3.2 实验结果及分析

本文采用情感识别领域常见的两种评价指标：加权准确率 WA 和非加权准确率 UA。WA 衡量了语音情感识别系统的总体性能，其计算方式为正确分类的样本数量除以样本总数；UA 衡量所有类别的识别性能，其计算方式为各类的分类准确率再除以类别数。本文采用分层五折交叉方式验证模型预测效果，使用样本的 80%进行训练，20%进行测试，最后对五次预测结果取平均。为了评价本文所提算法的有效性，对五种识别模型在 IEMOCAP 数据集上进行了对比实验。

下面对表 2 中使用的模型及输入特征进行说明：

BLSTM：将段特征输入 BLSTM 网络进行特征提取，然后输入 Softmax 分类器进行语音情感识别。

BLSTM-Attention ： 将 段 特 征 输 入 BLSTM-Attention 网络进行特征提取，然后输入 Softmax 分类器进行语音情感识别。

CNN-GAP：将 Mel 语谱图输入 CNN-GAP 网络进行特征提取，然后输入 Softmax 分类器进行语音情感识别。

Concate Network：将段特征输入基于 BLSTM-Attention 网络，将 Mel 语谱图输入 CNN-GAP 网络，再将两者以直接拼接的方式进行特征融合，最后输入 Softmax 分类器进行语音情感识别。

Our Methods：本文提出的主辅网络特征融合识别模型，先将段特征输入基于 BLSTM-Attention 网络，将 Mel 语谱图输入 CNN-GAP 网络，再将两者以主辅网络特征融合的方式进行特征融合，最后输入 Softmax 分类器进行语音情感识别。

表 2 五种识别模型在 IEMOCAP 数据集上的识别结果

Tab.2 The recognition results of the four recognition models on IEMOCAP dataset

识别模型	WA(%)	UA(%)
BLSTM	69.44	67.38
BLSTM-Attention	71.25	69.42
CNN-GAP	70.77	68.48
Concate Network	73.21	71.35
Our Methods	**74.45**	**72.50**

识别结果如表 2 所示，Our Methods 的 WA 和 UA 比 BLSTM-Attention 分别提高 3.20%、3.08%，比 CNN-GAP 分别提高 3.68%、4.02%，比 Concate Network 分别提高 1.24%、1.15%，表明使用两种特征融合比单独使用一种特征更有效，且主辅网络特征融合方式的识别结果比直接拼接方式特征融合的识别结果更有效。

表 3 Our Methods 的混淆矩阵

Tab.3 The confusion matrix of Our Methods

准确性(%)	中性	高兴	悲伤	生气
中性	**60.02**	16.54	18.83	4.60
高兴	12.61	**78.15**	2.52	6.72
悲伤	5.43	5.43	**88.37**	0.78
生气	12.16	16.22	8.11	**63.51**

Our Methods 的混淆矩阵如表 3 所示，中性、高兴、悲伤、生气情感的准确率分别为 60.02%、78.15%、88.37%、63.51%，四种情感准确率均高于 60%，进一步证明了 Our Methods 的有效性。

表 4 中列出了 Our methods 和其他模型在 IEMOCAP 数据集上研究的识别结果，识别模型均采用即兴场景对话中的语句，选取了四种情感。从表 4 可以看出，Our methods 和其他模型相比取得了不错的效果，比 Attention-BLSTM-FCN 模型的 WA 和 UA 分别提高了 6.35%和 5.5%。

表 4 Our Methods 和其他模型在 IEMOCAP 数据集上的识别结果

Tab.4 The recognition results of Our Methods and other models on IEMOCAP datasets

识别模型	WA(%)	UA(%)
FCN-Attention[8]	70.4	63.9
Attention-BLSTM-FCN[15]	68.1	67.0
CNN_GRU_SeqCap[17]	72.73	59.71
Our Methods	**74.45**	**72.50**

4 结束语

情感的表达本身是一个很复杂的过程，涉及到心理以及生理方面的诸多因素，因此从语音信号中识别出情感信息是一个挑战性的课题。本文将段特征输入 BLSTM-Attention 网络作为主网络，把 Mel 语谱图输入 CNN-GAP 网络作为辅助网络，然后，将两个网络提取的深度特征以主辅网络方式进行特征融合，解决不同类型特征直接融合带来的识别结果不理想的问题，并用实验验证了所提出模型的有效性。在今后的研究过程中，拟改进 CNN-GAP 网络的最后一层的池化方式并将脑电信号作为辅助信号引入主辅网络结构进行语音情感识别。

参考文献

[1]张雪英, 孙颖, 张卫, 等. 语音情感识别的关键技术[J]. 太原理工大学学报, 2015, 46(6): 630-636.
ZHANG X Y, SUN Y, ZHANG W, et al. Key technologies in speech emotion recognition[J]. *Journal of Taiyuan University of Technology*, 2015, 46(6): 630-636.

[2]MIRSAMADI S, BARSOUM E, ZHANG C. Automatic speech emotion recognition using recurrent neural networks with Local attention[C] // *IEEE International Conference on Acoustics Speech and Signal Processing*. New Orleans, 2017: 2227-2231.

[3]KIM J W, SAUROUS R A. Emotion recognition from human speech using temporal information and deep learning[C] // *19th Annual Conference of the International Speech Communication Association*. Hyderabad, 2018: 937-940.

[4]HUANG C W, NARAYANAN S S. Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition[C] // *IEEE International Conference on Multimedia and Expo*. Hong Kong, 2017: 583-588.

[5]陈晓敏. 基于时序深度学习模型的语音情感识别方法研究[D]. 哈尔滨:哈尔滨工业大学, 2018.
CHEN XIAOMIN. Research on speech emotion recognition method based on time series deep learning model[D]. Harbin: Harbin Institute of Technology, 2018.

[6]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.

[7]李鹏程. 基于深度学习的语音情感识别研究[D]. 合肥: 中国科学技术大学, 2019.
LI PENGCHENG. Deep learning based on speech emotion recognition research[D]. Hefei: University of Science and Technology of China, 2019.

[8]ZHANG Y Y, DU J, WANG Z R, et al. Attention based fully convolutional network for speech emotion recognition[C] // *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*. Honolulu, 2018.

[9]GUO L L, WANG L B, DANG J W, et al. A feature fusion method based on extreme machine for speech emotion recognition[C] // *IEEE International Conference on Acoustics, Speech and Signal Processing*. Calgary, 2018: 2666-2670.

[10]LEE J, TASHEV I. High-level feature representation using recurrent neural network for speech emotion recognition[C] // *16th Annual Conference of the International Speech Communication Association*. Dresden, 2015: 1537-1540.

[11]ZHAO Z P, ZHENG Y, ZHANG Z X, et al. Exploring spatio-temporal representations by integrating attention-based Bidirectional-LSTM-RNNs and FCNs for speech emotion recognition[C] // *19th Annual Conference of the International Speech Communication Association*. Hyderabad, 2018: 272-276.

[12]ALDENEH Z, PROVOST E M. Using regional saliency for speech emotion recognition[C] // *IEEE International Conference on Acoustics, Speech and Signal Processing*. New Orleans, 2018: 2741-2745.

[13]LUO D Q, ZOU Y X, HUANG D Y. Investigation on joint representation learning for robust feature extraction in speech emotion recognition[C] // *19th Annual Conference of the International Speech Communication Association*. Hyderabad, 2018: 152-156.

[14]ZHANG C J, WANG C L, JIA N. An ensemble model for multi-level speech emotion recognition[J]. *Applied Sciences*, 2020, 10(1): 205.

[15]ZHAO Z P, BAO Z T, ZHAO Y Q, 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.
[16] 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.
[17] WU X X, LIU S X, CAO Y W, et al. Speech emotion recognition using capsule networks[C] // *IEEE International Conference on Acoustics, Speech and Signal Processing*. Brighton, 2019: 56-76.
[18] 张卫.基于模糊认知图的语音情感识别关键问题研究[D].太原: 太原理工大学，2017. ZHANG WEI. Research on the key problems of speech emotion recognition based on fuzzy cognitive maps[D]. Taiyuan: Taiyuan University of Technology, 2017.
[19] Eyben F, Scherer K R, Schuller B W, et al. The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing[J]. IEEE Transactions on Affective Computing, 2017, 7(2): 190-202.

本文创新点：提出了一种使用主辅网络进行深度特征融合的语音情感识别算法