引用格式：徐华南，周晓彦，姜万，等. 基于 3D 和 1D 多特征融合的语音情感识别算法[J]. 声学技术，2021，**40**(4): 496-502. [XU Huanan, ZHOU Xiaoyan, JIANG Wan, et al. Speech emotion recognition algorithm based on 3D and 1D multi-feature fusion[J]. Technical Acoustics, 2021, **40**(4): 496-502.] **DOI**: 10.16300/j.cnki.1000-3630.2021.04.009

基于 3D 和 1D 多特征融合的语音情感识别算法

徐华南，周晓彦，姜 万，李大鹏
(南京信息工程大学电子与信息工程学院，江苏 南京 210044)

摘要：针对语音情感识别任务中特征提取单一、分类准确率低等问题，提出一种 3D 和 1D 多特征融合的情感识别方法，对特征提取算法进行改进。在 3D 网络，综合考虑空间特征学习和时间依赖性构造，利用双线性卷积神经网络(Bilinear Convolutional Neural Network, BCNN)提取空间特征，长短期记忆网络(Short-Term Memory Network, LSTM)和注意力(attention)机制提取显著的时间依赖特征。为降低说话者差异的影响，计算语音的对数梅尔特征(Log-Mel)和一阶差分、二阶差分特征合成 3D Log-Mel 特征集。在 1D 网络，利用一维卷积和 LSTM 的框架。最后 3D 和 1D 多特征融合得到判别性强的情感特征，利用 softmax 函数进行情感分类。在 IEMOCAP 和 EMO-DB 数据库上实验，平均识别率分别为 61.22%和 85.69%，同时与提取单一特征的 3D 和 1D 算法相比，多特征融合算法具有更好的识别性能。

关键词：语音情感识别；双线性卷积网络；长短期记忆网络；注意力(attention)；多特征融合
中图分类号：TN912.34 **文献标志码**：A **文章编号**：1000-3630(2021)-04-0496-07

Speech emotion recognition algorithm based on 3D and 1D multi-feature fusion

XU Huanan, ZHOU Xiaoyan, JIANG Wan, LI Dapeng
(College of Electronic and Information Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, *Jiangsu, China)*

Abstract: In order to solve the problems of single feature extraction and low classification accuracy in speech emotion recognition task, a 3D and 1D multiple feature fusion method for emotion recognition is proposed in this paper to improve the feature extraction algorithm. In 3D network, the spatial feature learning and time-dependent construction are considered. The bilinear convolutional neural network (BCNN) is used to extract spatial features, the short-term memory network (LSTM) and the attention mechanism are used to extract significant time-dependent features. In order to reduce the influence of speaker differences, the Log-Mel features of speech signal and the first-order differential and the second- order differential features are computed to synthesize the 3D Log-Mel feature set. In 1D network, the 1D convolution and LSTM network are used. Finally, 3D and 1D features are fused to obtain discriminative emotional features, and the emotions are classified by using softmax functions. The average recognition rates are 61.22% and 85.69% respectively on IEMOCAP and EMO-DB databases, and the multi-feature fusion algorithm has better recognition performance than the 3D and 1D algorithm for single feature extraction.

Key words: speech emotion recognition; bilinear convolutional neural network (BCNN); long short-term memory (LSTM); attention mechanism; multi-feature fusion

0 引 言

情感识别是情感计算的重要组成部分，目的让计算机模拟与识别人类的情感感知和理解过程，而语音情感作为诸多情感中的最直接的部分，是实现自然人机交互中的重要前提[1-2]。近年来关于语音情感识别的研究，取得了一些令人瞩目的成绩，但由于情感表达的复杂性(比如说话者年龄、性别以及说话者所处的文化和环境背景)，语音情感识别仍然面临诸多挑战[3-4]。

随着深度神经网络的发展，大量文献表明深度神经网络比如卷积神经网络(Convolutional Neural Network, CNN)、长短期记忆网络(Long Short-Term Memory, LSTM)等在语音情感识别中能提取更有价值的信息[5-9]。Lim 等[5]主要用短时傅里叶变换将语音信号转换为二维信息，利用卷积神经网络和循环神经网络(Recurrent Neural Network, RNN)作为声学模型，在 EMO-DB 数据库上的准确率分别为

收稿日期：2020-03-31；**修回日期**：2020-07-02
基金项目：国家自然基金(61902064、81971282)；中央基本科研业务费(2242018K3DN01)。
作者简介：徐华南(1995－)，女，江苏南京人，硕士生，研究方向为语音情感识别。
通信作者：周晓彦，E-mail: 18326167806@163.com

86.06%和 78.31%，Basu 等[6]利用 13 阶梅尔频率倒谱系数(Mel Frequency Cepstrum Coefficient, MFCC)作为输入，依次输入到卷积神经网络和长短期记忆网络中，得到了将近 80%的识别率。文献[5-6]只考虑 MFCC、基频等个性化特征，忽略了非个性化特征的影响，并且人与人之间差异较大，携带了大量个人情感信息，不具有通性。Zhao 等[7]利用 1D 卷积神经网络和长短期记忆网络(CNN-LTSM)中和 2D CNN-LTSM 中进行语音情感识别，发现 2D CNN-LTSM 比 1D CNN-LTSM 的识别率高，Chen 等[8]提出基于 3D 卷积-循环神经网络 (Convolutional Recurrent Neural Network, CRNN)的语音情感识别的方法，并通过实验证明，3D 卷积-循环神经网络的识别率要比 2D 卷积-循环神经网络高。文献[7-8]可以看出 3D 网络能捕获更丰富的情感特征信息，但仅提取单一特征未能完全表达情感表征，Luo 等[9]提出 HSF-CRNN (High Level Statistics Functions-CRNN, HSF-CRNN)框架，将手工特征和深度学习特征级联输入到 softmax 后进行情感分类，识别率比单层卷积-循环神经网络和多层卷积-循环神经网络分别提高了 3.8 个百分点和 7.6 个百分点。但文献[7-9]都没有考虑到对关键的时空依赖关系进行建模。于是本文为消除个性化特征的影响和降低说话者年龄、性别以及说话者所处的文化和环境背景的影响，计算语音信号的对数梅尔特征(Log-Mel)和其一阶差分和二阶差分特征，合并成 3D Log-Mel 特征集，在文献[9]启发下提出利用双通道网络级联特征，其中一条通道采用 3D 卷积-循环神经网络网络，并提出对称型双线性卷积神经网络(Bilinear Convolutional Neural Network, BCNN)模型方案，以平移不变的方式，对局部特征交互进行建模，利用 BCNN 提取空间特征，LSTM-attention 提取判别性强的时间特征后融合得到时空特征，输入到原始的支持向量机(Support Vector Machines, SVM)分类器中分类，同时另外一条通道采用 1D 卷积-循环神经网络网络，最后将 3D 特征和 1D 特征融合到一起，增加每一个特征的信息量，提高分类精度。

1　基于多特征融合的语音情感识别

提出的语音情感识别系统总体框架如图 1 所示。在 3D 部分，首先对输入的语音进行分帧、加窗等预处理操作得到梅尔频率倒谱系数并计算进而其一阶差分、二阶差分，得到 3D Log-Mel 特征集，随后输入到双线性卷积神经网络中提取频域特征和短时域特征，由于网络共享机制，在 BCNN 的分流的输出再输入到长短期记忆网络中，提取长时域特征，随后输入到 attention 模块中得到显著的特征表示，最后连接 BCNN 的外积输出和 attention 输出得到 3D 特征。在 1D 部分，首先对输入的语音进行分帧、加窗等预处理操作后得到等长度的语音数据，随后输入到一维卷积神经网络中提取频域特征和短时域特征，再输入到长短期记忆网络中得

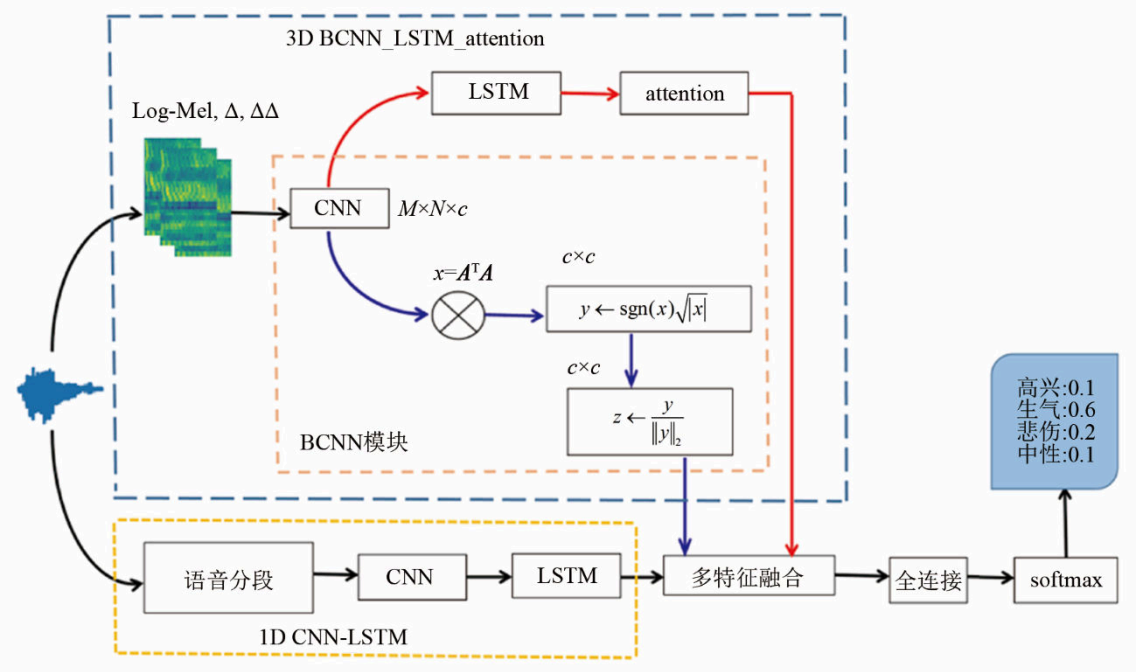


图 1　语音情感识别系统总体框架

Fig.1　General framework of speech emotion recognition system

到 1D 特征。将 3D 特征和 1D 特征进行融合后，选择出判别性强的情感特征输入到全连接层中，利用 softmax 进行语音情感分类。图 1 中 c 为 CNN 网络最后一层卷积层的通道数，M、N 为处理后的特征维度，$\boldsymbol{A}$ 表示经过 CNN 网络处理后的特征矩阵。

1.1 对数梅尔特征(Log-Mel)

为降低说话者年龄、性别以及说话者所处的文化和环境背景的影响，本文对给出的语音信号进行如下操作：

(1) 将语音信号通过高通滤波器进行预加重处理，高通滤波器表示为

$$H(z)=1-\mu z^{-1} \tag{1}$$

其中，μ 的取值范围为 0.9～1，通常取 0.97。

(2) 对预加重的信号进行零均值和单位方差处理；

(3) 对得到的语音信号进行分帧处理，汉明窗加窗，帧长为 25 ms，帧移为 10 ms；

(4) 对每一帧进行离散傅里叶(Discrete Fourier Transform, DFT)变换后得到各帧的频谱，并对频谱取模平方得到对应的功率谱，将时域信号转换为频域上的能量分布；

(5) 将功率谱输入到梅尔滤波器组中得到能量值，对于第 i 个滤波器 $(0<i\leqslant 40)$，能量为 p_i，对 p_i 进行对数变换后得到倒谱梅尔频率 $y_i=\lg(p_i)$；

(6) 为了更好地体现时域连续性，可在特征维度增加前后帧信息的维度，由 y_i 计算一阶差分 z_i^d 和二阶差分 z_i^{dd}：

$$z_i^{\mathrm{d}}=\frac{\sum_{n=1}^{v} n(y_{i+n}-y_{i-n})}{2\sum_{n=1}^{N} n^2} \tag{2}$$

$$z_i^{\mathrm{dd}}=\frac{\sum_{n=1}^{v} n(z_{i+n}^{\mathrm{d}}-z_{i-n}^{\mathrm{d}})}{2\sum_{n=1}^{N} n^2} \tag{3}$$

其中：$v=2$，经过式(2)～(3)得到 3D 的特征表示 $H\in\mathbf{R}^{t\times f\times k}$，$t$ 表示时间长度，f 表示梅尔滤波器的个数，k 表示特征的通道数，分别为静态特征、一阶差分和二阶差分特征，静态特征结合前两阶动态信息足够提高语音情感识别的性能。这里，t=300，f=40，k=3。

1.2 BCNN 模型

BCNN 模型是一个端对端的训练过程，具有优异的泛化能力，可以产生不同的无序的文字描述，如费希尔向量(Fisher vector)、局部特征聚合描述符(Vector of Locally Aggregated Descriptors, VLAD)和二阶池化(Second-order Pooling, O2P)等，不仅在细粒度图像分类上取得了优异效果，还被用于其他分类任务，如图像识别、语音情感识别等[10]。在语音情感识别领域中，对于前面提到的 3D Log-MEL 特征和一阶、二阶差分特征，通过双通道 CNN 网络后，会得到双路情感特征表示，传统上对于不同的情感特征的融合，常用的方法是进行串联、求和或者最大池化等一阶池化方法，假设利用平均二阶池化情感特征矩阵进行池化改进，对特征图 M_0 中相同位置的情感特征向量与自身的转置求外积从而直接得到两两特征维度之间的相关性：

$$G_{\mathrm{avg}}(M_0)=\frac{1}{M}\sum_{i=1}^{M}\boldsymbol{x}_i\cdot\boldsymbol{x}_i^{\mathrm{T}} \tag{4}$$

其中：$\boldsymbol{X}=[x_1\ x_2\ \cdots\ x_M]$ 是以空间位置 $\boldsymbol{F}=[f_1\ f_2\ \cdots\ f_M]$ 为中心点的局部特征向量；M 为局部特征的个数。而双线性汇合计算两路情感特征的外积，并对不同空间位置计算平均汇合得到双线性特征，外积捕获了特征通道之间成对的相关关系，提供了比线性模型更强的特征表示。为了简化计算，这里 BCNN 的前半部分使用相同的 CNN 模型，成为对称的 BCNN，因此只需要训练一个 CNN 模型即可。BCNN 模型的表达形式为

$$B=(F_A,F_A,P,Q) \tag{5}$$

其中：F_A 为双线性卷积神经网络两个分流的特征提取函数；P 是池化函数；Q 是分类函数。特征提取函数可看成一个函数映射：$f:\boldsymbol{L}\times\boldsymbol{I}\rightarrow\boldsymbol{R}^{c\times D}$，3D Log-MEL 特征数据经过卷积操作后，再进行最大池化操作，有效减少了网络的参数个数，并保存了有用的情感特征，因此将输入语音 $\boldsymbol{I}$ 与位置区域 $\boldsymbol{L}$ 映射为一个 $c\times D$ 维的情感特征向量。特征向量在 $\boldsymbol{L}$ 位置处使用矩阵外积进行特征组合，选择出判别性强的情感特征，即双线性(bilinear)特征：

$$F_b(\boldsymbol{L},\boldsymbol{I},F_A,F_A)=F_A(\boldsymbol{L},\boldsymbol{I})^T F_A(\boldsymbol{L},\boldsymbol{I}) \tag{6}$$

由式(4)、(6)可知，两者是等价的。因此二阶池化(对称型 BCNN)中的外积运算将特征图相同位置的输出拼接作为局部特征，然后对这个局部特征进行外积运算，将这些结果矩阵转化为特征图。这里 bilinear 特征为 $c\times c$ 的双线性特征，其中 c 为 CNN 模型的通道数，利用池化函数 P 将所有位置的 bilinear 特征进行累加汇聚成一个双线性特征 x_{bcnn}，函数表达式为

$$x_{\mathrm{bcnn}}=\sum_{l\in L}F_b(\boldsymbol{L},\boldsymbol{I},F_A,F_A) \tag{7}$$

最后将得到的双线性特征 x_{bcnn} 进行开平方 $y \leftarrow \text{sgn}(x_{\text{bcnn}})\sqrt{x_{\text{bcnn}}}$ 操作，$z \leftarrow y/\|y\|_2$ 归一化后输出。本文中，BCNN 模型中包括四层卷积层和两层池化层，卷积层中第一层卷积层有 128 个输出通道，其他卷积层的输出通道为 256，卷积核大小为 5×3，池化层大小为 1×2，经过多层卷积之后，双线性特征大小为 256×256。

1.3　LSTM 模型

为了解决梯度消失和梯度爆炸问题，LSTM 模型在循环神经网络 RNN 中对循环层进行改进，是一种特殊的时间递归神经网络[8]，适合处理和预测时间序列。LSTM 网络使用输入门、遗忘门和输出门来控制记忆过程。语音信号是时间序列信号，且序列之间的信息是相互关联的，本文选用 BLSTM 网络对每一个训练序列分别应用一个向前和向后的 LSTM 网络，使网络充分学习到序列的上下文信息。这里，BLSTM 层的单元大小设置为 128。

1.4　attention 机制

在语音情感识别中，不仅要关注具有情感信息的语音帧，也要考虑到每个情感语音帧的重要程度。因此对于 LSTM 输出的情感特征，本文并没有对其执行平均或最大池化等操作，而是利用 attention 机制去寻找显著的话语情感表征特征。在语音情感分类问题上，attention 机制已被大量使用在序列对序列的任务中[11]。每一步只关注特定的小区域，抽取区域表征信息，再整合到之前步骤所积累的信息中。attention 的任务是对于 LSTM 网络得到隐层输出序列 $\boldsymbol{o}_i$，在每个时间步，模型会根据上一时刻的隐层输出与情感编码序列进行逐一比较得到一个对齐权重，然后按照权重大小将编码序列中的每个编码向量加权求和得到最终的 attention 数值，即当前的情感语音向量。经过 T 次时间步后，模型会输出语音数据库中各类情感的判别概率。其中 attention 层的注意力权重为

$$\alpha_i = \frac{\exp[f(\boldsymbol{o}_i)]}{\sum_{i=1}^{T}\exp[f(\boldsymbol{o}_i)]} \tag{8}$$

其中，$f(\boldsymbol{o}_i)$ 是评价函数，表达式为 $f(\boldsymbol{o}_i)=\boldsymbol{W}^{\mathrm{T}}\cdot\boldsymbol{o}_i$，$\boldsymbol{W}$ 为训练参数。对 attention 层权重求和得到最后的特征向量为

$$\boldsymbol{u}=\sum_{i=1}^{T}\alpha_i\cdot\boldsymbol{o}_i \tag{9}$$

通过交替迭代训练，attention 机制更聚焦目标上细微的有区分性的部分，提取出判别性强的特征表征。

1.5　1D 特征表示

为增加特征向量的信息，对给定的语音进行 1D CNN 和 LSTM 操作[10]。首先将语音按照 300 帧划分成等长度的语音片段，对于不足 300 帧的语音用补 0 的方式填充，处理后的数据格式为片段个数与维度信息，将处理好的数据输入到 CNN 和 LSTM 网络中提取特征。本文选用四层卷积层、四层池化层和一层双向 LSTM 层，第一层卷积层有 64 个输出通道，第二层卷积层输出通道有 128 个，其他卷积层的输出通道为 256，卷积核的大小为 5，池化层大小为 2，步长为 2。LSTM 层的输出单元大小设为 128。

2　实验设置与分析

2.1　情感数据库

为了验证多特征融合模型的有效性，本文选用 IEMOCAP 和 EMO-DB 情感数据库进行实验。IEMOCAP 是由南加州大学 Sail 实验室录制的英语数据库，由 10 名专业演员(5 男 5 女)表演组成[12]。数据库包括 5 节(session)，分别为 session1、session2、session3、session4、session5，每一节分为即兴表演和剧本表演，由 1 男 1 女表演得到，即每一节包含两位说话者。本文选择即兴表演部分，采用中性、高兴、生气、悲伤四种情感，分别有 1 099，284，289 和 608 条语音。EMO-DB 数据库是由柏林工业大学录制的德语情感语音库，共 535 条语音[13]，由 10 名专业的演员(5 男 5 女，数据标注分别为 03、08、09、10、11、12、13、14、15、16)分别对 7 类情感(中性、高兴、生气、悲伤、厌恶、无聊和恐惧)表演得到，采样率为 16 kHz。

2.2　参数设置

给定的语音在 3D 模块中对 Log-MEL 特征按照 300 帧分段，不足 300 帧的通过补 0 填充；在 1D 模块中，将语音划分为等长度的片段。本文基于 Tensorflow 平台来实现 1D 和 3D 多特征融合网络，网络参数中，迭代次数(epoch)为 500，单次训练用的样本数(batch_size)为 40，学习率(learning_rate)为 10^{-4}，动量(momentum)为 0.99，权重衰减(decay_rate)为 0.99，丢弃率(dropout)为 0.1。

2.3　实验设置

本文采用非加权平均召回率(Unweighted Average Recall, UAR)作为评价指标，实验协议为“leave one subject out”[14-15]，将数据库中的 8 位说话者作为训练集，一位作为验证集，剩下一位为测

试集，为了使实验数据不具有偶然性，对每一人实验 5 次，求出均值与标准差，最后将 10 个人的数据求均值得到最后结果。为了验证本文提出的 1D 和 3D 融合算法的有效性，本文将本算法和其他方案进行了对比：

(1) CNN-LSTM[16]：对已经提取的特征进行归一化后，输入到局部卷积 CNN 层、全局卷积 CNN 层、LSTM 层，然后通过反馈层进行情感分类。

(2) DCNN_DTPM[17]：先提取三个通道的梅尔频谱图作为深度卷积神经网络(Deep Convolutional Neural Networks, DCNN)输入，然后使用预先训练的 DCNN 模型来学习每个片段的高级特征表示。利用判别时间金字塔匹配(Discriminant Temporal Pyramid Matching, DTPM)策略对学习到的分段级特征进行聚类。

(3) ML ELM_AE[18]：用 openEAR 工具提取特征[19]，通过训练 ML_ELM_AE 层(Multi-Layer Extreme Learning Machines based Auto-Encoder, ML ELM_AE)的级联来学习参数的多层神经网络，权衡系数为 100。

2.4 实验结果

本文对提出的 1D 和 3D 多特征融合算法进行验证，每人实验 5 次，求出均值与标准差，然后将 10 个人的数据求均值。表 1 为 IEMOCAP 数据库中每个人的识别率(以 session1,2,3,4 为例，将数据 session1，session2，session3，session4 作为训练集，session5 中的女性数据作为验证集，男性数据作为测试集时，平均识别率为 55.44%，若 session5 中的女性数据作为测试集时，平均识别率为 60.08%)。表 2 为 EMO-DB 数据库中每个人的识别率(以验证集为 08、测试集为 03 为例，将 08 说话者作为验证集，03 说话者为测试集，其他为训练集，平均识别率为 87.02%)，图 2 为 IEMOCAP 和 EMO-DB 数据库的混淆矩阵。

由实验结果可知，首先，在 IEMOCAP 数据库上，由于数据集不平衡，会导致识别率相差很大，在将 session1，session3，session4，session5 作为训练集，session2 中的男性作为测试集、女性为验证集时，识别率最高能有 69.63%，而最低的识别率在 session1，2，3，4 作训练集，session5 中男性作为测试集时为 55.14%，在 EMO-DB 数据库上，识别率最好的是 09 序号的说话者作为测试集时，为 94.64%，而最低的识别率在 10 序号的说话者作为测试集时，为 76.77%，将每个人作为测试集分别实验 5 次，求均值和标准差后再求均值，能消除数据集不平衡的影响，IEMOCAP 和 EMO-DB 数据库最后的平均识别率为 61.22%和 85.69%。

表 1 IEMOCAP 数据库中不同人的识别率

Table 1 The recognition rates of different speakers in IEMOCAP database

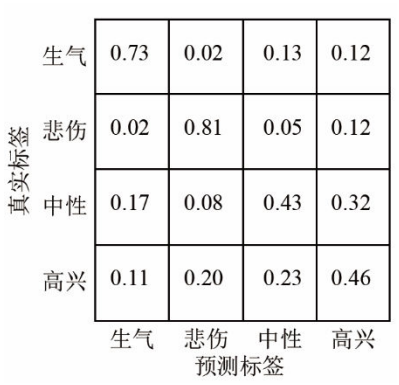
训练集	测试者(男/女)	准确率(UAR)/%					(均值±标准差)/%
		第 1 次	第 2 次	第 3 次	第 4 次	第 5 次	
session 1,2,3,4	男	55.78	55.14	55.88	55.24	55.17	55.44±0.320
	女	60.57	60.09	59.36	60.01	60.36	60.08±0.410
session 1,2,3,5	男	59.93	60.66	60.11	59.94	60.26	60.18±0.270
	女	57.86	58.77	56.30	59.16	58.61	58.14±1.012
session 1,2,4,5	男	59.84	58.97	59.31	60.13	59.03	59.45±0.451
	女	55.52	56.47	56.41	55.48	57.63	56.30±0.786
session 1,3,4,5	男	68.45	69.59	68.10	69.63	67.68	68.69±0.790
	女	66.37	66.95	66.17	68.10	66.43	66.80±0.697
session 2,3,4,5	男	64.58	64.18	64.56	64.17	64.41	64.38±0.177
	女	62.62	65.22	61.36	63.17	61.21	62.72±1.455
平均							61.22±0.637

表 2 EMO-DB 数据库中不同人的识别率

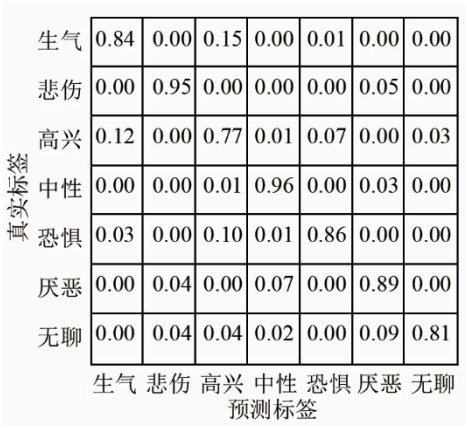
Table 2 The recognition rates of different speakers in EMO-DB database

验证集	测试集	准确率(UAR)/%					(均值±标准差)/%
		第 1 次	第 2 次	第 3 次	第 4 次	第 5 次	
08	03	85.16	88.78	86.73	86.46	87.96	87.02±1.252
03	08	85.33	84.34	79.18	84.47	83.50	83.36±2.171
10	09	91.07	94.64	91.07	92.86	91.07	92.14±1.429
09	10	77.46	76.77	78.83	79.04	80.62	78.54±1.339
12	11	79.77	80.71	84.42	85.59	82.14	82.53±2.192
11	12	84.05	86.90	84.05	86.43	84.76	85.24±1.203
14	13	85.63	84.29	89.05	85.44	86.90	86.26±1.621
13	14	89.29	88.99	88.51	87.86	89.07	88.74±0.510
16	15	88.98	91.05	89.76	88.95	90.66	89.88±0.856
15	16	81.73	83.77	83.77	82.00	84.51	83.16±1.092
平均							85.69±1.367

其次，通过混淆矩阵可以看出：(1) 情感标签为悲伤和生气在两个数据库上都能获得很高的识别率，在 IEMOCAP 在分别为 81%和 73%，在 EMO-DB 数据库上，分别为 95%和 84%。(2) 在 EMO-DB 数据库上，情感标签为中性、恐惧、厌恶、无聊识别率很高，分为达到了 96%，86%，89%和 81%。而在 IEMOCAP 数据库上中性情感的识别率很低，只有 43%，其中有 32%的情感被误判成高兴，17%的情感误判成生气。(3) 在两个数据库上，情感标签为高兴的识别率相比其他情感来说较低，分别为 46%和 77%。在 IEMOCAP 数据库上，有 20%的高兴情感被误判成悲伤，23%的高兴被误判成中性，在 EMO-DB 数据库上，有 12%的高兴情感被



(a) IEMOCAP 数据库



(b) EMO-DB 数据库

图 2　IEMOCAP 数据库和 EMO-DB 数据库的混淆矩阵

Fig.2　The confusion matrices of IEMOCAP database(left) and EMO-DB database(right)

误判成生气。

与其他方案比较，不同方案下的语音情感识别率如表 3 所示。

表 3　不同方案下的语音情感识别率

Table 3　Speech emotion recognition rate under different schemes

方法	识别率/%	
	IEMOCAP	EMO-DB
DCNN_LSTM[16]	-	80.6
DCNN_DTPM[17]	-	83.53
ML ELM_AE[18]	-	82.0
1D CNN	53.51	79.56
3D BCNN	58.63	82.93
1D+3D	61.22	85.69

通过表 3 可知，本文提出的算法与上述方案相比，准确率有了相应的提升。在 EMO-DB 语音库上，1D+3D 多特征融合网络与 DCNN_LSTM 网络相比，准确率提升了 5.09 个百分点；与 DCNN_DTPM 网络相比，准确率提升了 2.16 个百分点；与 ML ELM_AE 网络相比，准确率提升了 3.69 个百分点。本文分别对 1D CNN 网络和 3D BCNN 网络分别做了实验，由实验结果可知，1D CNN 网络在 IEMOCAP 和 EMO-DB 数据库的识别率分别为 53.51%、79.56%，3D BCNN 网络的识别率分别为 58.63%和 82.93%。两部分相比较可以发现，3D BCNN 网络的识别率比 1D CNN 网络的识别率分别提高 5.12 个百分点和 3.37 个百分点。这是因为在 3D 部分，本文对语音情感进行零均值和标准差处理并且提取对数梅尔特征减少了说话者之间的差异性。而将两种网络并联连接，在 IEMOCAP 数据库上识别率分别提高了 7.71 个百分点和 2.59 个百分点，在 EMO-DB 数据库上识别率分别提高了 6.13 个百分点和 2.76 个百分点，说明 1D 和 3D 特征融合网络提高了整个算法的性能。

3　结 论

目前多特征融合算法是解决语音情感识别问题的有效途径。本文提出的基于 3D 和 1D 多特征融合的语音情感识别算法，该算法通过将 3D 网络和 1D 网络输出特征融合，使得选择出判别性强的情感特征，相比仅提取单一特征的 3D 模型和 1D 模型，在语音情感识别中有更好的识别效果。同时也发现 BCNN 模型能提高语音情感识别的正确率。本文算法在训练样本的数量、网络模型的训练和 1D 模型的改进上还有待于进一步的理论和实验研究。

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 Wenjing, LI Haifeng, RUAN Huabin, et al. Review on speech emotion recognition[J]. Journal of Software, 2014, **25**(1): 37-50.

[2] SWAIN M, ROUTRAY A, KABISATPATHY P. Databases, features and classifiers for speech emotion recognition: a review[J]. International Journal of Speech Technology, 2018, **21**(1): 93-120.

[3] 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.

[4] DHALL A, GOECKE R, JOSHI J, et al. Emotion recognition in the wild challenge 2014: baseline, data and protocol[C]//Proceedings of the 16th International Conference on Multimodal Interaction. Istanbul Turkey. New York, NY, USA: ACM, 2014: 461-466.

[5] LIM W, JANG D, LEE T. Speech emotion recognition using convolutional and Recurrent Neural Networks[C]//2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA). Jeju, Korea (South). IEEE, 2016: 1-4.

[6] BASU S, CHAKRABORTY J, AFTABUDDIN M. Emotion recognition from speech using convolutional neural network with recurrent neural network architecture[C]//2017 2nd International Conference on Communication and Electronics Systems (ICCES). Coimbatore, India. IEEE, 2017: 333-336.

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

[8] CHEN M Y, HE X J, 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.

[9] LUO D , ZOU Y , HUANG D . Investigation on Joint Representation Learning for Robust Feature Extraction in Speech Emotion Recognition[C]//Interspeech 2018. Hyderabad: India, 2018: 152-156.

[10] LIN T Y, ROYCHOWDHURY A, MAJI S. Bilinear CNN models for fine-grained visual recognition[C]//2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile. IEEE, 2015: 1449-1457.

[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]//Interspeech 2018. Hyderabad, India, 2018: 272-276.

[12] 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.

[13] MILTON A, SHARMY ROY S, TAMIL SELVI S. SVM scheme for speech emotion recognition using MFCC feature[J]. International Journal of Computer Applications, 2013, **69**(9): 34-39.

[14] LIAO K, MOLLISON M V, CURRAN T, et al. Single-trial EEG predicts memory retrieval using leave-one-subject-out classification[C]//2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Madrid, Spain. IEEE, 2018: 2613-2620.

[15] ABDELWAHAB M, BUSSO C. Supervised domain adaptation for emotion recognition from speech[C]//2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). South Brisbane, QLD, Australia. IEEE, 2015: 5058-5062.

[16] KIM J, SAUROUS R A. Emotion recognition from human speech using temporal information and deep learning[C]//Interspeech 2018. ISCA: ISCA, 2018: 937-940.

[17] 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.

[18] GLÜGE S, BÖCK R, OTT T. Emotion recognition from speech using representation learning in extreme learning machines[C]// Proceedings of the 9th International Joint Conference on Computational Intelligence. Funchal, Madeira, Portugal. SCITEPRESS-Science and Technology Publications, 2017: 179- 185.

[19] EYBEN F, WÖLLMER M, SCHULLER B. OpenEAR—Introducing the Munich open-source emotion and affect recognition toolkit[C]//2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops. Amsterdam, Netherlands. IEEE, 2009: 1-6.