计算机科学
COMPUTER SCIENCE

基于语音节奏差异的情感识别方法

张家豪，章昭辉，严琦，王鹏伟

引用本文

张家豪，章昭辉，严琦，王鹏伟. 基于语音节奏差异的情感识别方法[J]. 计算机科学，2024，51(4)：262-269.

ZHANG Jiahao, ZHANG Zhaohui, YAN Qi, WANG Pengwei. Speech Emotion Recognition Based on Voice Rhythm Differences [J]. Computer Science, 2024, 51(4): 262-269.

相似文章推荐（请使用火狐或 IE 浏览器查看文章）

Similar articles recommended (Please use Firefox or IE to view the article)

改进MFCC和并行混合模型的语音情感识别

Speech Emotion Recognition Based on Improved MFCC and Parallel Hybrid Model

计算机科学, 2023, 50(6A): 220800211-7. https://doi.org/10.11896/jsjkx.220800211

一种基于离散变权网络的动态最短路径快速算法

Fast Algorithm of Dynamic Shortest Paths Based on Discrete Varying-weight Networks

计算机科学, 2010, 37(4): 238.

一种基于社区分类的社交网络用户推荐方法

Novel Method on Community-based User Recommendation on Social Network

计算机科学, 2016, 43(5): 198-203. https://doi.org/10.11896/j.issn.1002-137X.2016.05.036

基于情感计算的e—Learning系统建模

计算机科学, 2005, 32(8): 131-133.

基于声学特征的语言情感识别

Speech Emotion Recognition Based on Acoustic Features

计算机科学, 2015, 42(9): 24-28. https://doi.org/10.11896/j.issn.1002-137X.2015.09.005

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230200063

基于语音节奏差异的情感识别方法

张家豪　章昭辉　严　琦　王鹏伟
东华大学计算机科学与技术学院　上海 201620
(zhzhang@dhu.edu.cn)

摘　要　语音情感识别在金融反欺诈等领域有着重要的应用前景,但是语音情感识别的准确率提升变得越来越困难。现有基于语谱图的语音情感识别等方法难以捕捉节奏差异特征,从而影响识别效果。文中基于语音节奏特征的差异性,提出了能量帧时频融合的语音情感识别方法。其关键是,针对语音中高能量区域进行频谱筛选,以高能语音帧的分布和时频变化来体现个体的语音节奏差异。在此基础上建立基于卷积神经网络(CNN)和循环神经网络(RNN)的情感识别模型,实现对频谱的时域和频域变化特征的提取与融合。在公开数据集IEMOCAP上进行实验,结果表明,该基于语音节奏差异的语音情感识别与基于语谱图的方法相比,在加权准确率WA和非加权准确率UA指标上分别平均提升了1.05%和1.9%;同时也表明个体的语音节奏差异对提升语音情感识别效果具有重要作用。

关键词:语音情感识别;能量帧;频域谱线;时频融合;语音节奏差异
中图分类号　TP301

Speech Emotion Recognition Based on Voice Rhythm Differences

ZHANG Jiahao,ZHANG Zhaohui,YAN Qi and WANG Pengwei
School of Computer Science and Technology,Donghua University,Shanghai 201620,China

Abstract　Speech emotion recognition has an important application prospect in financial anti-fraud and other fields,but it is increasingly difficult to improve the accuracy of speech emotion recognition. The existing methods of speech emotion recognition based on spectrograms are difficult to capture the rhythm difference features,which affects the recognition effect. Based on the difference of speech rhythm features,this paper proposes a speech emotion recognition method based on energy frames and time-frequency fusion. The key is to screen high-energy regions of the spectrum in the speech,and reflect the individual voice rhythm differences with the distribution of high-energy speech frames and time-frequency changes. On this basis,an emotion recognition model based on convolutional neural network(CNN) and recurrent neural network(RNN) is established to realize the extraction and fusion of the time and frequency changes of the spectrum. On the open data set IEMOCAP,the experiment shows that compared with the method based on spectrogram,the weighted accuracy WA and the unweighted accuracy UA of the speech emotion recognition based on the difference of speech rhythm increases by 1.05% and 1.9% on average respectively. At the same time,it also shows that individual voice rhythm difference plays an important role in improving the effect of speech emotion recognition.

Keywords　Speech emotion recognition,Energy frames,Spectrum,Time-frequency fusion,Voice rhythm difference

1　引言

语音情感识别(Speech Emotion Recognition,SER)被广泛应用于人机交互、提升用户体验、欺诈检测等领域,因此如何提高语音情感识别的准确率得到了更加广泛的关注和研究。

目前语音情感识别主要分为基于统计的传统机器学习方法和端到端的神经网络方法,传统机器学习方法通常使用不同的特征集来做分类训练,包含了能量、音调、过零率、MFCC等统计特征[1]。2010年Zhang等采用改进SLLE算法,能够增强低维嵌入数据的判别力,对包含韵律和音质特征的48维语音情感特征数据进行了非线性降维,可以较好地改善语音情感识别结果[2]。2013年Busso等提出了迭代特征归一化(IFN)框架,减小了说话者之间的中性语音之间的声学差异,同时保留了表达性语音中的情绪间差异性,在IEMOCAP数据库取得了66.3%的准确率[3]。2015年Jin等从低层次的声学特征、倒谱声学特征、声学_词典特征和声学_高斯超向量特征这4个角度生成了更加全面的语音信号中的情感特征表示,在IEMOCAP数据集上取得了67.8%的准确率[4]。然而,尽管付出了大量的努力,但仍然没有找出最合适的特征集解决方案。

到稿日期:2023-02-09　返修日期:2023-04-26
基金项目:上海市科技创新行动技术高新技术领域项目(22511100700)
This work was supported by the Shanghai Science and Technology Innovation Action High-tech Field Project(22511100700).
通信作者:章昭辉(zhazhang@dhu.edu.cn)

近年来，自从神经网络 CNN，RNN，CapsNet 展现出惊人的能力，有研究者通过把语音信号生成语谱图的方式，将 SER 任务转到对图像特征的提取上。同时研究表明端到端的语音情感识别效果也不低于传统方式。2016－2017 年 Trigeorgis 等[5]和 Huang 等[6]利用 CNN 从二维语音光谱图、log-mel 光谱图，甚至原始语音信号中学习。Satt 等[7]提出了以光谱图为输入的 CNN-BiLSTM 结构，并研究了光谱图分辨率和模型识别准确率之间的关系。2018 年 Tzirakis 等[8]提出了一种从语音中进行连续情感识别的新模型。他们的模型经过端到端训练，由卷积神经网络（CNN）组成，它从原始信号中提取特征，并在其上堆叠了 2 层长短期记忆（LSTM），以便考虑数据中的上下文信息。2019 年 Wu 等[9]通过胶囊网络来解决 CNN 无法捕获光谱图中的空间信息的问题，同时可以捕获全局水平的话语水平的特征。由于在对比实验中，为了更好地比较方法模型，统一了语音的预处理，仅复现了其论文中的网络模型，在 IEMOCAP 数据集上加权准确率达到了70.1％。Zhao 等[10]分别比较了以原始语音和 log-mel 谱图为输入的一维和二维 CNN_LSTM 架构的性能。2D-CNN_LST 的性能优于 DBN 和 CNN 等方法。2020 年 Mustaqeem 等[11]提出了被称为深步幅的 CNN 架构，它使用步幅对输入特征图进行降采样，而不是池化层。通过新的自适应阈值进行预处理，简化了 CNN 结构，降低了计算复杂度。在 IEMOCAP 数据集上提高了 7.85％的精度，显著优于最先进的系统。Liu 等[12]采用胶囊神经网络来弥补 CNN 从频谱图中捕获浅层全局特征的不足，分别设计了用于学习语谱图局部时频特征的 TFCNN 与学习浅层和深层全局信息的 CapsNet。本文在做对比实验时，统一了语音的处理流程，复现了其网络模型，在 IEMOCAP 数据集上分类准确率分别达到 68.6％（WA）和 68.9％（UA）。

2021 年 Hu 等[13]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。以 BLSTM-Attention 网络为主网络，关注语音信号中的情感信息；并以 CNN-GAP 网络为辅助网络，使用主辅网络深度特征融合的 WA 和 UA 分别提升了 1.24％和 1.15％。2022 年 Wu 等[14]将探索神经架构搜索（Neural Architecture Search，NAS）与语音情感识别相结合，针对语音数据量少的特点，着重提升 DARTS 算法的收敛速度，提升识别准确率。在原有的 CNN_RNN_att 基线上，WA 提升了 0.58％，UA 提升了 1.39％。可见现在单模态的语音情感识别在提升准确率方面越来越困难。

现有这些基于语谱图和统计特征的方法不能较好地体现出对语音不同节奏的情感区分能力。由于语音存在较强的主观性和个体差异，每个人对情绪的表达语速、停顿和程度均不相同，导致了语谱图上的能量分布也存在差异，那么在相同的量纲下用卷积核来提取特征会有不准确的情况，导致误分类。并且语音具有稀疏性的特点，并非所有时刻都含有情感特征，语音情感特征的定位也具有个体差异性。

本文的具体贡献如下：

1）提出了一种蕴含个体差异的 K 能量帧频谱确定方法，能够提取语音中的强节拍类型的高能语音帧。

2）语音中的情绪内容随时间变化，因此利用有效的时间建模的技术是正确的，采用 CNN＋RNN 的模型结构可以将频域特征和时域变化特征进行有效融合，提高对不同语音情感的区分度。

本文第 2 章介绍了相关工作；第 3 章是问题的详细说明；第 4 章介绍了 K 能量帧的频谱确定方法；第 5 章介绍了时频融合网络模型；第 6 章为实验及分析；最后总结全文。

2 相关工作

用于 SER 任务具有代表性深度学习的技术有卷积神经网络（Convolutional Neural Network，CNN）、循环神经网络（Recurrent Neural Network，RNN）、胶囊网络（Capsual Network，CapsNet）等。SER 最近的研究更多地在于对各深度学习模型进行改进和整合。

2.1 CNN 网络建模

CNN 是专门用于处理图像数据这种具有类似网格拓扑结构数据的神经网络[15]，被广泛用作 SER 的基本框架。深度卷积神经网络（Deep Convolutional Neural Network，DCNN）被认为是传统卷积神经网络的扩展。Zhang 等[16]受到了 DCNN 在计算机视觉领域较好表现的启发，提出了判别时间金字塔匹配算法，用于汇集深度特征，实验结果表明了该模型与算法结合的有效性且在小型语音情感数据集上有着一定优势。2019 年 Heracleous[17]提出将 DCNN 与 i-vector 相结合的情感识别方法，实验结果显示了该方法的有效性。

2.2 时间序列建模

语音中的情绪内容随时间变化，因此利用有效的时间建模的方向是无误的，RNN 与 LSTM 是专门用于处理序列数据的神经网络。其中 Wang 等[18]提出了双序列 LSTM 模型，用来同时处理两个 Mel 谱图，在 IEMOCAP 上的准确率相比目前最优的单模型提高了 6％。Hsu[19]利用 SVM 检测语音和无语义的发声，使用韵律短语提取器将两种类型的声音进行分离，然后使用深度残差网络提取各自的特征进行决策级融合。之后输入基于注意力机制的 LSTM 的序列到序列模型进行分类，结果准确率优于基于特征级和模型级的融合方法。

RNN 或 LSTM 通常与 CNN 结合用于 SER 任务，两者的各种组合成为了 SER 领域的一种流行趋势，其组合结构通常优于单独的模型。Zhao 等[10]构建 1 维和 2 维的 CNN_LSTM 学习局部特征和长期上下文关系，2D_CNN_LSTM 在 EMO-DB，IEMOCAP 语料库上与说话人相关和无关的实验中均取得了较好的识别率，优于深度信念网络和 CNN 等传统方法。Atila 等[20]提出了基于注意力的 3 维 CNN LSTM，将语谱图、MFCC 图、耳蜗图和分形图拼接成 4 维作为该模型的输入，在 SAVEE，RAVDESS 和 RML 数据集的实验准确率相比以往文献在这 3 个数据集上的实验结果分别提高了 2.71％，8.75％和 7.81％。

2.3 胶囊网络建模

Sabour 提出 CapsNet[21]，其每一个胶囊都由许多神经元组成，输入和输出都是向量，而非 CNN 的标量，它具有平移同变性，因此其可以克服 CNN 捕捉空间信息能力不足的局限性。有研究人员将其用于提取语谱图空间信息，Wu 等[9]

通过胶囊网络来弥补 CNN 无法捕获光谱图中的空间信息的不足，同时可以捕获基音和共振峰频率等低层特征的位置和关系信息。Liu 等[12]分别设计了用于学习语谱图局部时频特征的 TFCNN 和学习浅层和深层全局信息的 CapsNet，在 IEMOCAP 数据集上达到了不错的效果。

3 问题阐述

本文通过复现了基于语谱图的胶囊网络模型（见文献[9]）的实验，发现了语音个体差异的问题。同一类情感对个体而言在语速、停顿、发音轻重等方面都是不一样的，从频域角度即能量在时域上的分布也不一样。从图 1 中可以看到，两个人的语谱图均为 sad 标签，左图是成功被模型分类正确的，右图是被误分为 nature 类别。不难发现左图与右图整体上的差别不大，如果将左图中红色方框的位置删除，重新拼接并适当压缩后基本与右图一致。

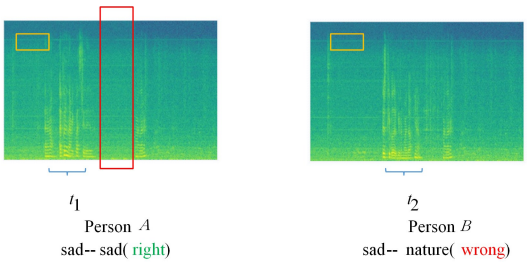

图 1 个体差异影响示意图（电子版为彩图）

Fig. 1 Schematic diagram of the impact of individual differences

分析了 A 能被准确分类，而 B 却被误判了的原因。由图 1 所示，在时序上面，A 的能量峰出在 t_1 时刻，而 B 的能量峰出现在 t_2 时刻。A 的语速稍平缓，那么其能量峰的宽度相比 B 来说会更宽，且 A 中的能量峰中存在间隔。基于上述情况，在同一量纲下，如图 1 所示，例如黄色方框均代表采用 $n*m$ 的卷积核来提取特征，会使特征提取不准确，从而造成误判的情况。因此不可忽视语音节奏差异特征，那么如何刻画和提取个体的语音节奏特征成为了关键问题。

在现有的诸多网络模型中，例如 CNN 具有很强的图像特征提取能力，RNN 或者 LSTM 可从时间层面建模，胶囊网络可以捕获更深层的特征信息，但是它们依旧难以处理语音节奏不同所带来的时频特征分布差异影响。在复现实验中，通过提取 CNN 卷积池化中间层的输出，如图 2 所示，可以看出高能区域更受关注。结合这一特点，本文提出了通过能量帧的方式来选取频域谱线，用高能语音帧的分布和时频特征来体现个体节奏差异，从而提升准确率。

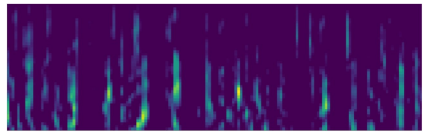

图 2 卷积网络对语谱图的中间层输出（电子版为彩图）

Fig. 2 Middle layer output of convolutional network to spectrogram

4 K 能量帧的频谱确定方法

4.1 整体流程

语谱图的特点就是二维图展示了三维信息，横坐标为时间，纵坐标为频率，颜色的深浅代表了能量的强弱。现有深度学习的模型就是从语谱图上提取时域和频域上的能量变化特征。因此，可以推断出能量变换的不同就是表达不同情感的所在，情感就是通过能量传递出来的。同时现有基于语谱图的模型并没有考虑到个体差异带来的影响。

因此，本文设计了 K 能量帧的选取，通过 K 能量帧的选取来蕴含个体语音节奏差异。K 值是语音帧中含有个体节奏差异特征的一种程度表示，本文所提出的基于语单节奏差异的情感识别方法的准确率会随着 K 值变化。整体流程如图 3 所示。

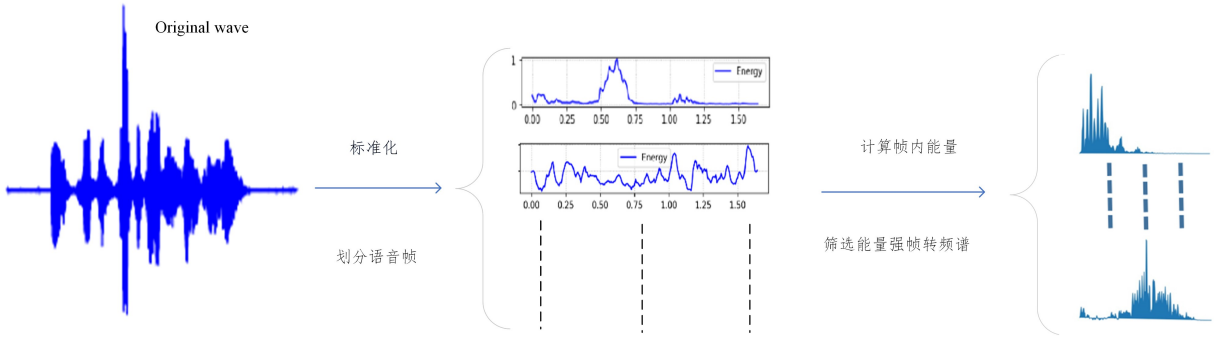

图 3 K 能量帧的频谱确定方法流程

Fig. 3 Flow of spectrum determination method for K energy frames

在读取音频文件后，需要先进行归一化操作，减小奇异样本数据导致的不良影响。然后根据帧长和帧移重合来划分语音信号，其中 nf 为总帧数，即该原始语音信号被划分成短信号的帧数。每个采样点都有对应的能量，我们通过累加的形式对每一帧信号求能量和 $E_i(i=1,2,3,\cdots,nf)$，nf 和 E_i 的计算式将在 3.2 节中介绍。由于能量大的帧包含了更多的情感信息，是整段语音情感的精髓所在。从 nf 帧最终选取能量最大的前 K 帧以及它们相邻的两帧，由于能量峰具有一定的宽度，把 K 帧前后两帧也找出来的目的是防止 K 帧首尾的信息丢失。最后对这个 $3*K$ 帧信号进行加窗，随后进行快速傅里叶变换得到这 $3*K$ 帧的频谱。

4.2 帧的划分

由于每条语音的长度不一，代表了每个音频的采样点数不一样，提前设置好帧长和帧移参数，通过式(1)可以计算出语音信号被分成的短信号帧数，即 nf 的值。其中，$signal\_length$ 为该音频的总采样点数，$wlen$ 为每一小帧的采样点

数，inc 为重叠帧移采样点数，$ceiling$ 为上取整函数。

$$nf=ceiling((signal\_length-wlen+inc)/inc) \tag{1}$$

其中有一点需要注意的是，由于很多情况下音频总采样点数是无法整除帧长和帧移之和的，最后一帧若不足 $wlen$ 个采样点则需要补零。

4.3 帧内短时能量求和

E_n 表示在信号的第 n 个点开始加窗函数时的短时能量，可通过式(2)计算得到，窗函数可选矩形窗和汉明窗等，在本文中选取了汉明窗如式(3)所示，短时能量可以看作语音信号的平方经过一个线性滤波器的输出，该线性滤波器的单位冲激响应为 $h(n)$[22]，其中 $h(n)=w(n)^2$。

$$\begin{aligned}E_n&=\sum_{m=-\infty}^{\infty}[x(m)w(n-m)]^2\\&=\sum_{m=-\infty}^{\infty}x^2(m)h(n-m)\\&=x^2(n)*h(n)\end{aligned} \tag{2}$$

$$w(n)=\begin{cases}0.54-0.46\cos\left(\dfrac{2\pi n}{N-1}\right), & 0\leqslant n\leqslant N-1\\0, & others\end{cases} \tag{3}$$

设第 i 帧语音信号的短时能量用 E_i 表示，则通过式(4)计算 E_i 值。

$$E_i=\sum_{m=0}^{M-1}x_n^2(m) \tag{4}$$

其中，M 为帧长即 $wlen$，$x_n(m)$ 为该帧中的样本点。

然后从 nf 帧中选取能量从高到低排序前 k 的帧，并把它们的前后相邻两帧也提取出来，一个音频文件共计 $3*K$ 帧，按时间顺序排列好。

4.4 频域谱线转化

获取到 $3*K$ 帧原始语音信号后，由于语音的本质就是各种波的组成，因此我们肯定需要从频域上着手，可以通过式(5)将其中一帧的信号 y 转变为频率谱线(横坐标为频率，纵坐标为幅值)。利用快速傅里叶变化(FFT)进行从时域到频域的转换，这里 FFT 的计算式不再展开介绍。经过变换后得到的是 N 个复数，每个复数值包含着一个特定频率的信息[23]，我们根据这 N 个复数(其中 N 为采样点个数)，取复数的绝对值(abs)即复数的模。从原始信号中获得各个频率信号和他们的幅度值，并进行归一化处理。

$$normalization\_half\_y=\frac{abs[fft(y)]}{N}\left[range\left(\frac{N}{2}\right)\right] \tag{5}$$

同时，由于对称性，仅需要取一般区间(单边频率即可)，因此式(5)中的后半个中括号并非是乘积的含义，而表示区间取半。如图 4(左)所示的示例，其中横坐标为频率，纵坐标为幅值。每一帧都对应一张频谱图，而且由于帧序之间存在时间关系，需要根据时间关系从前往后把图片排序好，以便后续模型学习。

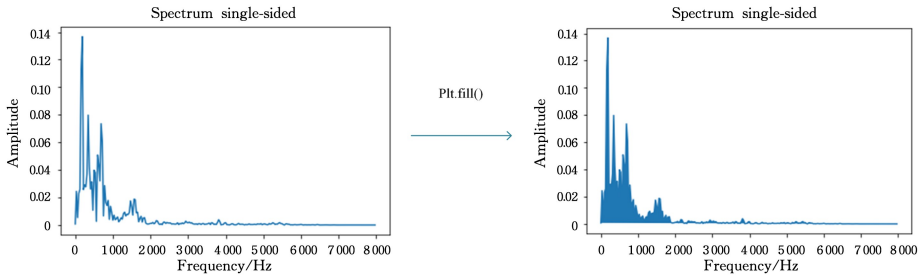

图 4 频谱图数据加强示例

Fig. 4 Example of spectrum diagram

由于生成的频域谱线在显示时仅有曲线，存在大面积的空白区域，通过 $plt.fill()$ 函数把 X 轴与曲线围成的区间填充上色，如图 4(右)所示，这样不仅减少了空白区域，同时对 CNN 的特征提取起到了增强作用。

4.5 K 值的选择

K 值的选择跟建模的数据有关，一般来说不能偏大也不能偏小，是个经验值。如果 K 偏小，很有可能导致选择的帧包含的特征信息量不够，导致分类不够准确；K 值若偏大，不仅在模型训练上会增加负担，也有可能导致信息的干扰和混乱，造成准确率下降。

5 时频融合网络模型

每一个人对情绪的表达，无论在语速或者停顿节奏上都是不一致的。语音情感特征会受到语速、停顿节奏等影响产生分布差异，从而导致模型特征提取不准确或者误判的情况。为了能够对 $3*K$ 语音频谱进行有效的特征提取，设计了如图 5 所示的 CNN-RNN 时频融合网络模型。语音是时序数据，那么 $3*K$ 语音帧可以看作时序序列。本文利用 RNN 来提取时间维度的变化特征，用 CNN 来提取频域维度特征。

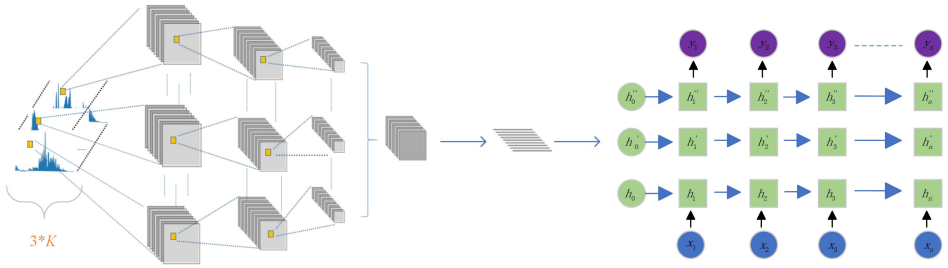

图 5 时频融合网络模型示意图

Fig. 5 Time frequency fusion network model

如图 5 所示，该模型主要由两个网络组成，前半部分的 CNN 主要用来提取该帧内不同频率上的能量特征，后半部分为多层的 RNN 模型，主要用来学习帧与帧之间的关联。因为当训练数据较多时，多层 RNN 效果会比单层 RNN 效果好，同时每一帧可被视为一个时间节点，以时序帧输入来获取时间维度上的语言情感特征。

5.1 CNN 网络设计

CNN 主要用来提取每一帧频谱在频域上的能量分布特征。本文采用了较大的感受野，因为随着 CNN 卷积的深入，为了压缩计算量，越到后面的层数往往 feature size 越小，但参与“决策”往往是最后几层网络。例如第一次卷积核可选取 8 * 8，后两次卷积核可为 5 * 5。输出层有足够大的理论感受野，有能力表达出不同尺度的目标[24]。与此同时，为了减少模型运算参数且能保留更多的纹理信息，采用最大池化层通过消除非极大值，降低了上层的计算复杂度。

CNN 的网络节点采用了整流线性单元(Rectified Linear Unit，ReLU)函数作为激活函数，ReLU 函数的定义如下：

$$h(y^{(i)})=\max(y^{(i)},0)=\begin{cases}0, & y^{(i)}<0\\ y^{(i)}, & y^{(i)}\geqslant 0\end{cases} \tag{6}$$

其中，$y^{(i)}$ 是通道的输出。ReLU 激活函数相比传统的激活函数可以更好地抑制梯度消失，加快模型的收敛[25]。在 CNN 提取每帧的能量分布特征的过程结束后，需要把这些特征图按照时序在通道层面上进行拼接。拼接完成后，为了统一 RNN 的输入尺寸，经过 reshape，将所有通道上的特征图按行序排列成一维特征向量，例如原通道上的特征图的行有 i 维度，列有 j 维，则经过 reshape 后就变为 $i*j$ 维的一维向量。

5.2 RNN 网络设计

为了学习到频谱能量在时域上的变化特征以及时间序列的趋势变化，本文采用了最经典的 RNN 模型，其结构如图 6 所示。为了建模序列问题，RNN 引入了隐状态 h(hidden state)的概念，h 可以对序列形的数据提取特征，接着将其转换为输出。因此 RNN 最大的优势是带有动态的记忆性，RNN 将每一个时间点的输出作为下一个时间点的输入，故在输入最后一个时间点的 X_n 时，RNN 可能记忆着前面 n 个输入的信息，但是这种记忆性和普通神经网络直接将 n 个 X 作为输入的不同之处在于，它不是将所有输入一视同仁，具有一定的动态性。

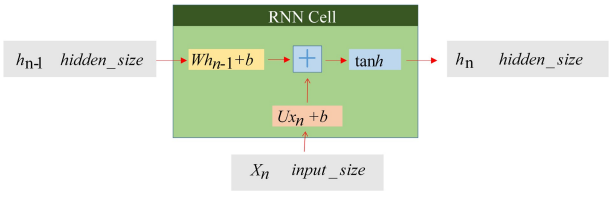

图 6 RNN 网络模型

Fig. 6 RNN network model

第 i 隐含层的隐含状态 h_i 的计算式如下：

$$h_i=\tanh(Ux_i+Wh_{i-1}+b)(1\leqslant i\leqslant n) \tag{7}$$

其中，x_i 表示第 i 隐含层的输入层的输入。要注意的是，在计算时，每一步使用的参数 U，W，b 都是一样的，也就是说每个步骤的参数都是共享的，能加快模型的训练[26]。第 i 隐含层的输出计算方法如式(8)所示，直接通过 h_i 进行计算，一个箭头就表示对对应的向量做一次类似于 $f(Wx+b)$ 的变换，这里的这个箭头就表示对 h_i 进行一次变换，得到输出 y_i。

$$y_i=\text{Softmax}(Vh_i+c)(1\leqslant i\leqslant n) \tag{8}$$

6 实验及分析

6.1 数据集介绍

本实验选取了 IEMOCAP 语音情感数据集，在这个数据集中，10 个演员被记录在二元会话中(5 个会话，每个会话有 2 个对象)。该数据集大小为 16.4 GB，包含大约 12 h 的视听数据语料库，总共包含 10039 轮会话，平均持续时间为 4.5 s，每轮单词的平均值为 11.4 s[27]。声道数为 1，采样频率为 16000 Hz，数据集一共包含 10 个情感(中立状态、高兴、悲伤、愤怒、惊讶、恐惧、厌恶、挫败感、兴奋、其他)，标注方式为人工标注。

由于目前语音情感识别的分类主要以 4 类为主(中立、悲伤、愤怒、高兴)，因此本文基于这 4 类音频文件筛选了共计 4490 条音频。对原始语音分别进行了归一化，并进行了语音帧的划分、计算与选取，最后将选出的语音帧转为频域谱线和数据增强，具体内容可见本文第 3 章，这里不做重复叙述。

6.2 评价指标与实验设置

1)评价指标

目前评价语音情感识别模型最广泛使用的两个评价指标分别为加权准确率(Weighted Accuracy，WA)和非加权准确率(Unweighted Accuracy，UA)。它们的计算式如式(9)、式(10)所示：

$$WA=\frac{\sum_{1}^{k}n_i}{\sum_{1}^{k}N_i},1\leqslant i\leqslant k \tag{9}$$

$$UA=\frac{\sum_{1}^{k}\frac{n_i}{N_i}}{k},1\leqslant i\leqslant k \tag{10}$$

其中，k 表示语音情感类别数，n_i 表示第 i 类情感中正确识别的个数，N_i 表示第 i 类情感的总样本数。

2)参数设置

CNN 的首个卷积层采用 8 * 8 大小的滤波器，同时为了防止丢失边缘信息，设置零填充为有效。其他的卷积层均采用 5 * 5 尺寸大小的滤波器。网络的较高层均采用了更多的卷积核，因为更多的卷积核能提取到更多抽象的特征，虽然提高了高层网络的计算复杂度，但模型的特征描述能力也得到了提高。

对 RNN 选择交叉熵作为损失函数，将 Adadelta 作为优化器，初始学习率为 0.001，最后一层以 softmax 作为输出层函数，采用交叉熵作为损失函数。数据集被随机分成训练集(80%)和测试集(20%)，4 个情感类别在训练/测试集中的比例仍然与整个语料库中的比例相同。RNN 的 time_step 为 9，因为一个语音用 9 帧来表示，batch_size 为 3592，因为训练集中有 3592 个音频，RNN 的层数为 3，堆叠了 3 层 simple_rnn。

6.3 实验结果分析

首先本文在 IEMOCAP 数据集上，通过对参数 K 选取 1，2，3，4 不同的数值进行对比实验，由表 1 可以发现，K 值从 1 到 3 不断增大，WA 和 UA 指标也是不断提升的，但是当 K 值为 4 时，该模型的 WA 和 UA 指标不升反降，而且模型的训练时间也随着 K 值的增大而迅速增加。

表 1 不同 K 值的模型效果

Table 1 Model effects with different K values

(%)

指标	$K=1$	$K=2$	$K=3$	$K=4$
WA	66.2	68.1	70.4	69.3
UA	62.3	65.5	69.1	67.1

那么可以看出，K 值是需要根据实验的数据和模型来确定的，一般来说不能太大也不能太小。

在确定 K 值后，将基于能量帧的时频融合模型与现有基于语谱图的模型进行对比，结果如表 2 所列。为方便叙述，将本文模型称为 En_frame_CNN_RNN。由于模型 1(CNN_GRU-SeqCap)[9] 与模型 2(TFCNN_DenseCap_ELM)[12] 的数据处理方式不一样，为了便于对比，统一了语音的数据处理方式，因此复现结果与原参考文献会有所不同。

表 2 能量帧时频融合的语音情感识别方法的实验结果

Table 2 Experimental results of speech emotion recognition based on energy frame time-frequency fusion

(%)

	IEMOCAP	
	WA	UA
CNN_GRU-SeqCap[9]	70.1	65.5
TFCNN_DenseCap_ELM[12]	68.6	68.9
En_frame_CNN_RNN	70.4	69.1

从表 2 中可以看出，本文模型相比模型 1 在 WA 和 UA 上分别有 0.3%和 3.6%的提升，对比模型 2 在 WA 和 UA 上分别有 1.8%和 0.2%的提升。因此，本文模型在 WA 和 UA 指标上平均分别提升了 1.05%和 1.9%。虽然胶囊网络能够提取更深层次的特征信息，例如位置关系，能够进一步提升模型的准确率，但由于存在个体语音节奏的差异，语速节奏变化快慢不一，对语音情感特征在时域和频域上的分布都会造成偏差，从而导致语音特征提取不准确。虽然个体存在差异，但是每类情感的变化是有规律的，它的能量分布以及变化特征总体上看是有一致性的。因此本文方法考虑到了个体语音节奏差异的影响，对语音中高能量区域进行频谱筛选，以高能语音帧的分布和时频变化来体现个体的语音节奏差异可以进一步提高语音情感识别的效果。

为了进一步验证 K 能量帧的频谱筛选方法是否起到了蕴含个体语音节奏特征的作用，本文在相同的网络模型上对比了等间隔采样的频谱选取方法。由于每段语音的长度都不一致，那么所划分成的语音帧数也不同，因此这里的等间隔确切来说是等比例间隔，并非是传统意义上的等间隔。

从整段语音的初始帧到结束帧，本文分别等比例地选取了 5 帧和 10 帧来作为对比实验。

从图 7 中可以看出，K 能量帧的频谱选取方式相比 5 帧等比例间隔的频谱选取方式，在 WA 和 UA 指标上分别高出了 10.8%和 11.8%；相比 10 帧等比例间隔的频谱选取方式，在 WU 和 UA 的指标上分别高出了 5.3%和 4.7%。实验结果表明，K 能量帧的频谱选取方法是蕴含个体语音节奏的关键基础，同时个体语音节奏差异特性对提升语音情感识别有着重要意义。

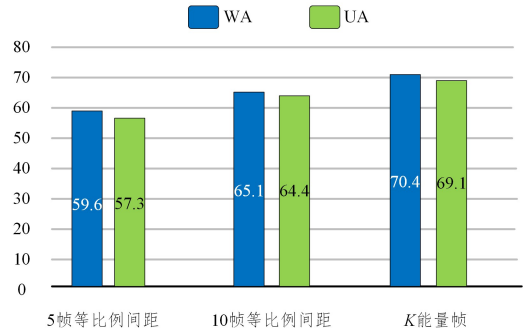

图 7 消融实验结果对比图

Fig. 7 Comparison of ablation experiment results

本文方法在测试集上的表现如图 8 所示，4 种情绪(生气、高兴、悲伤、中立)的准确率分别为 78.18%，55.5%，74.1%，68.91%。从图中可以看到 happy 的准确率是偏低的，没有达到 60%以上。本文初步分析其中的原因是 happy 的语料在整个语料库的占比偏低，如图 9 所示，给出了 4 类情感的占比，4 种情绪的总语音文件有 4490 个，而 happy 类的占比仅有 13.25%。很有可能由于样本不均匀，在训练 happy 情感时造成了过拟合的情况，即在训练集上表现好，但在测试集上表现不好。

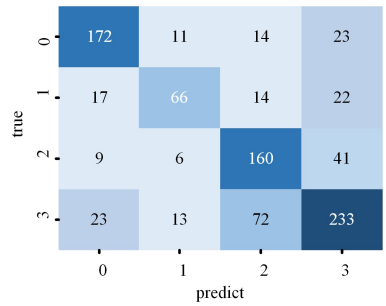

图 8 混淆矩阵

Fig. 8 Confusion matrix

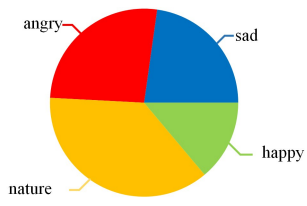

图 9 4 类情感占比的示例

Fig. 9 Proportion of four types of emotions

为了进一步推测 happy 情感类识别不高的原因，本文做了数据平衡实验，通过减少其他情感类别的语音数量，以数量最少的 happy 情感为基准，将 4 类情感的语音数均控制在 595 条，从而达到 4 个类别的语料均衡并进行模型训练，实验结果如图 10 所示，生气、高兴、悲伤、中立这 4 类情感的准确率分别为 61.3%，57.1%，59.6%，52.1%。

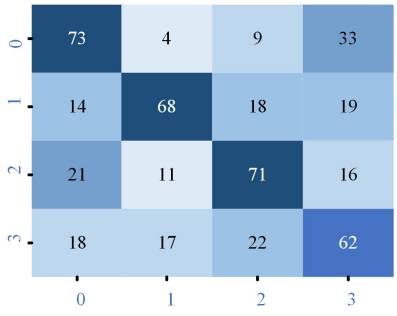

图 10 数据平衡后的混淆矩阵

Fig. 10 Confusion matrix after data balancing

因此减少其他情感类别的语音数量，在对其他情感的准确率造成严重影响的同时，happy 类别的准确率也没有较大的提升。那么本文进一步推测 happy 准确率不高的原因是 happy 语料的覆盖面不全，同时可能存在情绪程度的差异，例如小开心和很开心的程度是不同的，而小开心可能会被误判为 nature 等情绪类别。

实验结果表明，本文提出的通过能量帧的方式来选取频域谱线，在提取语音节奏特征和提高语音情感识别准确率方面是有效的。因此语音节奏的差异性成为了不可忽视的因素，后续的工作可继续围绕语音情感的个体差异进行研究。考虑在语谱图中能够融入语音的节奏特性来提高识别准确率，同时也可以考虑用胶囊网络来代替本文的网络。对 happy 语料的生成学习能否提高识别准确率也是十分值得探究的方向。

结束语 本文提出的基于 K 能量帧的时频融合网络方法是解决语音节奏特征难提取的一种有效手段。该方法表明个体的语音节奏差异对提升语音情感识别率具有重要的影响，其核心是针对语音中的高能量区域进行频谱筛选，以高能语音帧的分布和时频变化来体现个体的语音节奏差异。通过选取 $3*K$ 能量帧来蕴含语音节奏差异性，并设计了 CNN-RNN 网络模型将频域特征和时域变化特征有效融合，提高了对不同语音节奏的区分能力。

下一步工作是在语谱图中融入个体的语音节奏差异，以进一步提升识别效果，用胶囊网络来代替本文的网络，并尝试采用对抗生成语料的方式来提高识别准确率。

参 考 文 献

[1] SONG Y K, XIE J. Lightweight speech emotion recognition model based on multitask learning [J/OL]. Computer Engineering: 1-8. [2023-03-06]. https://doi. org/10. 19678/j. issn. 1000-3428. 0064430.

[2] ZHANG S Q, LI L M, ZHAO Z J. Speech emotion recognition based on an improved supervised manifold learning algorithm [J]. Journal of Electronics and Information, 2010, 32(11): 2724-2729.

[3] BUSSO C, MARIOORYAD S, METALLINOU A, et al. Iterative Feature Normalization Scheme for Automatic Emotion Detection from Speech[J]. IEEE Transactions on Affective Computing, 2013, 4(4): 386-397.

[4] JIN Q, CHEN S Z, LI X R, et al. Speech emotion recognition based on acoustic features [J]. Computer Science, 2015, 42(9): 24-28.

[5] TRIGEORGIS G, RINGEVAL F, BRUECKNER R, et al. Adieu Features? End-To-End Speech Emotion Recognition Using A Deep Convolutional Recurrent Network[C]// International Conference on Acoustics, Speech, and Signal Processing. 2016: 5200-5204.

[6] HUANG C W, NARAYANAN S S. Deep Convolutional Recurrent Neural Network With Attention Mechanism For Robust Speech Emotion Recognition[C]// International Conference on Multimedia Computing and Systems. 2017: 583-588.

[7] SATT A, ROZENBERG S, HOORY R. Efficient Emotion Recognition From Speech Using Deep Learning On Spectrograms [C]// Conference of the International Speech Communication Association. 2017: 1089-1093.

[8] TZIRAKIS P, ZHANG J H, SCHULLER B. End-To-End Speech Emotion Recognition Using Deep Neural Networks [C]// International Conference on Acoustics, Speech, and Signal Processing. 2018: 5089-5093.

[9] WU X X, LIU S X, CAO Y W, et al. Speech Emotion Recognition Using Capsule Networks[C]// IEEE ICASSP 2019. IEEE, 2019.

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

[11] MUSTAQEEM, KWON S. A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition[J]. Sensors, 2020, 20(1.0): 183.

[12] LIU J, LIU Z, WANG L, et al. Speech Emotion Recognition with Local-Global Aware Deep Representation Learning [C]// 2020 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP 2020). IEEE, 2020.

[13] HU D S, ZHANG X Y, ZHANG J, et al. Speech emotion recognition based on feature fusion of primary and secondary networks [J]. Journal of Taiyuan University of Technology, 2021, 52(5): 769-774.

[14] WU X X, HU S K, WU Z Y, et al. Neural Architecture Search for Speech Emotion Recognition[C]// International Conference on Acoustics, Speech, and Signal Processing. 2022: 6902-6906.

[15] LU G M, YUAN L, YANG W J, et al. Speech emotion recognition based on short-term memory and convolutional neural network[J]. Journal of Nanjing University of Posts and Telecommunications: Natural Science Edition, 2018, 38(5): 63-69.

[16] ZHANG S, ZHANG S, HUANG T, 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] HERACLEOUS P, MOHAMMAD Y, YONEVAMA A. Deep Convolutional Neural Networks for Feature Extraction in Speech Emotion Recognition[C]// International Conference on

Human-Computer Interaction(HCII). 2019：117-132.

[18] WANG J，XUE M，CULHANE R，et al. Speech emotion recognition with dual-sequence LSTM architecture[C]// 2020 IEEE International Conference on Acoustics，Speech and Signal Processing(ICASSP 2020). IEEE，2020：6474-6478.

[19] HSU J，SU M，WU C，et al. Speech Emotion Recognition Considering Nonverbal Vocalization in Affective Conversations[J]. IEEE/ACM Transactions on Audio，Speech，and Language Processing，2021，29：1675-1686.

[20] ATILA O，ŞENGÜR A. Attention guided 3D CNN-LSTM model for accurate speech based emotion recognition[J]. Applied Acoustics，2021，182：108260.

[21] SABOUR S，FROSST N，HINTON G E. Dynamic routing between capsules[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems(NIPS'17). 2017：3859-3869.

[22] LI W H. Research on speech emotion recognition based on spectrum sensing feature[D]. Nanchang：Donghua University of Technology，2018.

[23] YANG X J，WANG H Y，CHEN J H，et al Application of Fast Fourier Transform Algorithm in Audio Power Amplifier[J]. Electronic Technology，2015，44(7)：33-35.

[24] CHEN J. Speech emotion recognition based on convolutional neural network[C]// 2021 International Conference on Networking，Communications and Information Technology. 2021：106-109.

[25] KAVITHA S，SANJANA N，YOGAJEEVA K，et al. Speech Emotion Recognition Using Different Activation Function[C]// 2021 International Confe-rence on Advancements in Electrical，Electronics，Communication，Computing and Automation(ICAECA). 2021：1-5.

[26] LIESKOVSKA E，JAKUBEC M，JARINA R. RNN with Improved Temporal Modeling for Speech Emotion Recognition [C]// 2022 32nd International Conference RADIOELEKTRONIKA. 2022：1-5.

[27] 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.

ZHANG Jiahao，born in 1998，postgraduate. His main research interest includes speech emotion recognition.

ZHANG Zhaohui，professor，Ph. D supervisor，is a member of CCF（No. 15065M）. His main research interests include risk control technology of Internet financial transaction，big data and artificial intelligence.

（责任编辑：喻藜）