ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 36,No. 5,Sep. 2021,pp. 1062－1072
DOI:10. 16337/j. 1004-9037. 2021. 05. 021

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

融合图像显著性的声波动方程情感识别模型

贾　宁，郑纯军

（大连东软信息学院软件学院，大连 116023）

摘　要：语音情感识别（Speech emotion recognition，SER）是计算机理解人类情感的关键之处，也是人机交互的重要组成部分。当情感语音信号在不同的介质传播时，使用深度学习模型获得的识别精度不高，识别模型的迁移能力不强。为此，设计了一种融合图像显著性和门控循环的声波动方程情感识别（Image saliency gated recurrent acoustic wave equation emotion recognition，ISGR-AWEER）模型，该模型由图像显著性提取和基于门控循环的声波动模型构成。前者模拟注意力机制，用于提取语音中情感表达的有效区域，后者设计了一个声波动情感识别模型，该模型模拟循环神经网络的流程，可以有效提升跨介质下语音情感识别的精度，同时可快速地实现跨介质下的模型迁移。通过实验，在交互情感二元动作捕捉（Interactive emotional dyadic motion capture，IEMOCAP）情感语料库和自建多介质情感语音语料库上验证了当前模型的有效性，与传统的循环神经网络相比，情感识别精度获得了25％的改善，并且具有较强的跨媒介迁移能力。

关键词：语音情感识别；图像显著性和门控循环的声波动方程情感识别；图像显著性；声波动方程；门控循环；多介质情感语音语料库

中图分类号：TP183　　**文献标志码**：A

An Acoustic Wave Equation Emotion Recognition Model Based on Image Saliency

JIA Ning，ZHENG Chunjun

（School of Software，Dalian Neusoft University of Information，Dalian 116023，China）

Abstract：Speech emotion recognition（SER）is the key point for computer to understand human emotion，and it is also important in human-computer interaction. When the emotional speech signal transforms in the different media，the recognition accuracy of traditional deep learning model is not high enough，and the migration ability is not strong. Here，an acoustic wave equation emotion recognition model，i.e.，image saliency gated recurrent acoustic wave equation emotion recognition（ISGR-AWEER）model is designed. The model is composed of image saliency extraction and gated recurrent model. The first part simulates the attention mechanism，which is used to extract the salient regions in speech. An acoustic wave equation emotion recognition model is designed. The model simulates the recurrent neural network，which can effectively improve the accuracy of SER in cross-media，and can quickly realize the model migration in cross-media. The effectiveness of the current model is verified by the experiments on the interactive emotional dynamic motion capture emotional corpus and the self-built multi-media emotional speech corpus. Compared with recurrent neural network，the accuracy of emotion recognition is improved by

收稿日期：2021-05-24；**修订日期**：2021-09-11

25%, and it has a strong ability of cross-media migration.

Key words: speech emotion recognition (SER); image saliency gated recurrent acoustic wave equation emotion recognition (ISGR-AWEER); image saliency; acoustic wave equation; gated recurrent; multi-media emotional speech corpus

引　言

随着语音学领域及相关技术的日趋成熟,人们逐渐意识到语音中传达的信息远超出了文本中拟表达的内容[1]。作为区分智慧和智能的基本特征,情感是语音交互不可或缺的一部分,正确的检测语音中的情感表达具有深远的应用价值和实际意义。作为语音学的新兴研究方向之一,语音情感识别(Speech emotion recognition, SER)旨在实现语音信号至说话者情感表达的映射关系,帮助机器模拟情感理解、情感监测和反馈等过程,从而带来人机交互方式的变革。

现有的情感识别研究可分为单模态和多模态[2]等方向。前者专注于原始音频信号,后者需融合音频信号、词汇信息和视觉信息。由此可见,上述两种研究方法均涉及了原始音频信号的研究,它是情感识别方向的关键所在,也是大量研究人员的工作重心。

在SER的研究中,许多成熟的语音分析和分类技术用于提取有效的情感信息。随着深度学习技术的发展,利用神经网络模型解决SER问题已成为流行的情感识别解决方案。常见的深度学习模型[3]有卷积神经网络(Convolutional neural network, CNN)[4-5]、循环神经网络(Recurrent neural network, RNN)[6]等,此外,多通道技术和注意力机制也常用来实现SER。Yao等[7]对比3种不同的深度学习模型的性能,指出深度融合显著性区域的重要性,但仅针对于传统的空气介质环境下的语料库。

RNN作为常见的深层神经网络之一,被广泛应用于自然语言处理和时间序列有关的任务中[8-9]。例如,Tzinis和Potamianos[10]利用RNN研究局部特征和全局特征,并对比在不同音素级别下的SER性能。

由于RNN存在梯度消失和梯度爆炸问题,研究人员常使用长短期记忆网络(Long-short term memory, LSTM)[11]、门控循环单元(Gated recurrent unit, GRU)[12]等RNN变种模型设计SER模型。Xie等[13]融合LSTM和注意力机制,以找到与情绪识别相关的重要时间段。Tang等[14]设计了基于GRU的情感监测模型,以监测显著性区域内的连续语音情感。就现有的端到端SER模型,为提升识别精度,融合注意力机制,寻找显著性区域是一种常见的方法。

此外,现有的RNN类模型往往通过搭建人工神经网络,实现大脑生物神经网络的模拟,这是一种被动的处理信号和信息的方式。当传输介质发生突变或环境发生改变时,RNN需要重新训练网络模型参数,以适应最佳的识别效果,此时工作效率较低。因此,部分研究人员将含有介质信息的声波物理模型应用于相关领域中[15]。Hughes等[16]利用物理波动系统的概念,实现元音的高精度识别。目前这种物理系统被证明针对语音识别任务是有效的,但是尚未应用于情感识别任务中。Li等[15]设计了多维变速度声波方程用于多均匀介质中的波形模拟。Zhang等[17]采用纯声波方程(Pure acoustic wave equations, PAWEs)实现在各向异性介质中的模拟传播。基于上述研究可以发现,声学波动方程可用于跨介质的波形模拟,然而鲜有研究人员将其应用于跨介质的SER任务中。

从2020年起新型冠状病毒肆虐全球,人们出行经常佩戴口罩或其他防护设置,此时语音情感的表达将由传统的空气介质切换为口罩佩戴的环境,使用原有的传统介质设计的模型的识别精度受到较大影响。

本文的目标是解决SER任务在不同介质下的识别精度问题。基于此,本文设计了一种融合图像显

著性和门控循环的声波动方程情感识别模型(Image saliency gated recurrent acoustic wave equation emotion recognition, ISGR-AWEER),它由图像显著性提取和基于门控循环的声波动模型构成。前者用于模拟注意力机制,提取语音中的有效区域,后者设计了一个声波动情感识别模型模拟RNN的流程,该模型可以有效的提升跨介质下SER的精度,同时可快速的实现跨介质下的模型迁移。

1 融合图像显著性和门控循环的声波动情感识别模型

1.1 ISGR-AWEER模型整体结构

声波动物理模型的数据来源是原始声学信号,声学信号中蕴含海量的特征,并非全部信号均为有效数据,这为识别的精度提升带来了很大的困难。在声波传输介质发生变化时,现有的模型无法实现有效的迁移。针对上述问题,本文设计了融合图像显著性和门控循环的声波动方程情感识别模型,将其分为2个阶段:音频信号的图像显著性提取和基于门控循环的声波动模型。ISGR-AWEER模型整体结构如图1所示。

显著性音频信号提取时,注重有效信息的计算和统计,仅提取感情表达最强烈的部分音频。基于门控循环的声波动模型设计了一种类似GRU的模型,可以用连续声波动力学的自动演化过程来执行计算,通过物理本身的时间动力学,自然的体现前后帧之间的关联。通过推导证明,此模型与GRU网络存在异曲同工之处,同时在跨媒介的语料库中具有最优的识别效果。下文将分别介绍这两个阶段的具体设计思路。

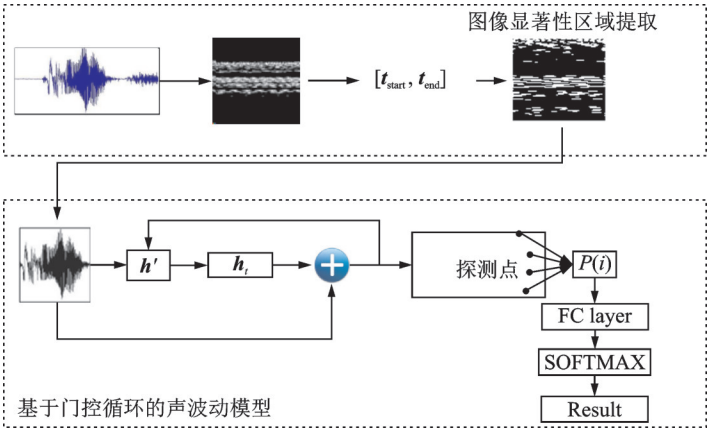

图1 ISGR-AWEER模型整体结构

Fig.1 Model structure of ISGR-AWEER

1.2 显著性音频信号提取

声波动情感识别模型以原始音频信号作为输入,基于现有的采样率,输入数据的维度较高,散度较大,而且存在大量的干扰信号,例如背景噪声等,这些导致了语音表达的情感在时间和空间上难以衡量,直接影响情感识别精度和效率。虽然流行的语谱图可通过变换得到的时频域信息提取语音中的显著性信息,但是语谱图对原始音频信号进行短时傅里叶变换的预处理破坏了语音信号中的波动特性,进而影响波动模型对于SER的精度。

因此,提出一种基于自定义音频图象的显著性音频信号的提取方法,用于快速锁定一段音频中的显著性区域。由于一个完整的情感表达的最短时间仅需2~5个连续语音帧,如果能够找到这段区间,

就可以用其替代整段音频,这段区间通常被称为显著性区域,而剩余的音频区间均可以作为背景区域处理。提取过程分为2个阶段:音频图像重构和显著性区域提取。下文将详细介绍具体提取流程。

1.2.1　音频图像重构

作为显著性音频信号提取的第一个阶段,音频图像重构用于将音频转换为$\sqrt{N}*\sqrt{N}$的图像表达。生成规则为

$$\boldsymbol{t}_{i,j}=\boldsymbol{p}_{i*\sqrt{N}+j}\quad i\in[0,\sqrt{N}\,],j\in[0,\sqrt{N}\,] \tag{1}$$

式中:N为单个语音作为原始音频的维度;$\boldsymbol{p}_{i*\sqrt{N}+j}$为原始语音信号中的第$i*\sqrt{N}+j$维数据;$\boldsymbol{t}_{i,j}$为原始语音信号的矩阵表达,对其进行归一化后,可生成一个像素矩阵$\boldsymbol{t}'_{i,j}$,即当前语音时域的图像表达。图2分别是angry和happy类别情感生成的图像,可以观察到这2种情感的表达方式存在明显差异。

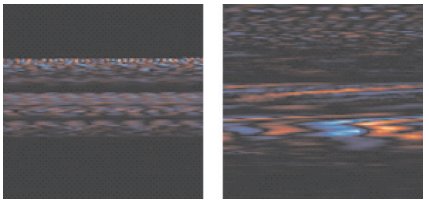

图2　Angry和happy类别的图像表达

Fig.2　Image expression of angry and happy class

1.2.2　显著性信号提取

图像表达中的数据是语音的原始声学信号,该信号的特点是具有时序性。信号发生突变的时间区间,在语音中会出现同步的波动。因此图像形式的音频表达很直观的显示出当前区域是否存在信号突变或者维持光滑平缓。当图像局部出现突变时,往往是音频出现波动的起始或终止阶段。当图像局部显示为平坦区域时,可以认为此次未有大范围的波动,即维持原有的信号强度不变。基于此,可以利用突变出现的2个时机判断音频出现大规模能量变化的区间。不同的情绪的显著性区域表达存在差异化特征,因此对于显著性区域的锁定有助于快速的识别某种特殊的情绪。

寻找情感表达显著性区域方法如下:

(1)准备一个宽度为$\sqrt{N}$滑动窗口,高度取值区间[$\sqrt[4]{N}/2,\sqrt[4]{N}$],可将整张图像切分为[$2*\sqrt[4]{N},\sqrt[4]{N},$]个区域。

(2)此窗口从图像的起始位置开始,依次扫描每个块的全部区域,进行区块间的差分,具体方法为

$$\Delta\boldsymbol{t}_{i,j}=\boldsymbol{t}'_{i+\sqrt[4]{N}/2,j}-\boldsymbol{t}'_{i,j}\quad i\in[0,\sqrt{N}\,],j\in[0,\sqrt{N}\,] \tag{2}$$

$$\Delta\Delta\boldsymbol{t}_{i,j}=\Delta\boldsymbol{t}_{i+1,j}-\Delta\boldsymbol{t}_{i,j}\quad i\in[0,\sqrt{N}\,],j\in[0,\sqrt{N}\,] \tag{3}$$

式中:$\boldsymbol{t}_{i,j}$为式(1)得到的像素矩阵;$\Delta\boldsymbol{t}_{i,j}$为差分后的矩阵,显著性区域即此矩阵中的有效数据,二次差分$\Delta\Delta\boldsymbol{t}_{i,j}$则用于寻找显著性区域的起始和结束位置。

(3)添加约束条件用于区分平坦和波动区域,获得的$\boldsymbol{t}_{\text{start}}$和$\boldsymbol{t}_{\text{end}}$即为最终显著性区域的具体位置,如式(4~5)所示。

$$\boldsymbol{t}_{\text{start}}=\boldsymbol{i}_1,\Delta\Delta\boldsymbol{t}_{i1,j}>\frac{1}{N}\sum_{i,j=0}^{\sqrt{N}}\Delta\Delta\boldsymbol{t}_{i,j},\Delta\Delta\boldsymbol{t}_{i1,j}^2>\frac{1}{N^2}\sum_{i,j=0}^{\sqrt{N}}\Delta\Delta\boldsymbol{t}_{i,j}^2 \tag{4}$$

$$\boldsymbol{t}_{\text{end}}=\boldsymbol{i}_2,\Delta\Delta\boldsymbol{t}_{i2,j}>\frac{1}{N}\sum_{i,j=0}^{\sqrt{N}}\Delta\Delta\boldsymbol{t}_{i,j},\Delta\Delta\boldsymbol{t}_{i2j}^2>\frac{1}{N^2}\sum_{i,j=0}^{\sqrt{N}}\Delta\Delta\boldsymbol{t}_{i,j}^2 \tag{5}$$

(4)将[$\boldsymbol{t}_{\text{start}},\boldsymbol{t}_{\text{end}}$]区间内的图像信息提取出,作为新的音频图像表达。图3分别为angry、happy类别的显著性区域图像,可以观察出4种情感的显著性区域表达有各自的特征。

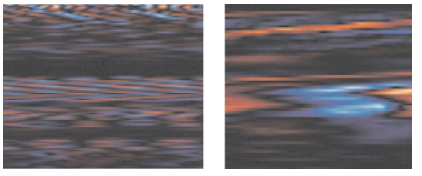

图3　Angry和happy类别的显著性信号区域表达

Fig.3　Expression of significant regions in angry and happy class

将显著性区域的图像表达转化为原始声学信号,然后输入

基于门控循环的声波动模型中，具体模型结构见1.3节。

1.3 基于门控循环的声波动模型

1.3.1 基于波动方程的物理模型

当声波在空气等介质中传播时，原始音频信号的处理方式由声波物理系统自适应学习得到。与RNN相比，这种声波传输机制并非刻意地实现信号的处理和反馈，而是使用物理本身的动力学原理自然的呈现声音序列的递归关系。基于波的物理系统的动力学，标量波场分布 $\boldsymbol{u}(x,y,z)$ 的动力学由二阶偏微分方程控制，即

$$\frac{\partial^2 \boldsymbol{u}}{\partial t^2}-\boldsymbol{c}^2\cdot\nabla^2\boldsymbol{u}=\boldsymbol{f} \tag{6}$$

式中 $\nabla^2=\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}$ 为拉普拉斯算子。$\boldsymbol{c}=\boldsymbol{c}(x,y,z)$ 为波速的空间分布，非线性材料的波速取决于波幅，$\boldsymbol{x}=\boldsymbol{x}(x,y,z,t)$ 是源项。可以使用时间步长为 Δt 的中心有限差分进行时间离散，如式(7)所示。

$$\frac{\boldsymbol{u}_{t+1}-2\boldsymbol{u}_t+\boldsymbol{u}_{t-1}}{\Delta t^2}-\boldsymbol{c}^2\cdot\nabla^2\boldsymbol{u}_t=\boldsymbol{x}_t \tag{7}$$

式中：t 为给定时间步的标量场的值。矩阵表达形式为

$$\begin{bmatrix}\boldsymbol{u}_{t+1}\\ \boldsymbol{u}_t\end{bmatrix}=\begin{bmatrix}2+\Delta t^2\cdot\boldsymbol{c}^2\cdot\nabla^2 & -1\\ 1 & 0\end{bmatrix}\cdot\begin{bmatrix}\boldsymbol{u}_t\\ \boldsymbol{u}_{t-1}\end{bmatrix}+\Delta t^2\cdot\begin{bmatrix}\boldsymbol{x}_t\\ 0\end{bmatrix} \tag{8}$$

1.3.2 基于门控循环的声波动模型

在对1.3.1节提到的波动方程进行改进后，本文提出了一种适用于跨介质的基于门控循环的声波动模型，传统的GRU常用于SER任务，可提升空气介质下的识别精度，但是无法应用于跨介质或跨语料库等环境迁移场景中，本节设计的声波动模型不仅保留了传统GRU的优势，可以解决传统RNN中的模型迁移问题。

由于声波动模型的输入是原始声学信号，所以需要将显著性区域转化为原始声学信号后再执行输入。图4描述了声波动模型的实现过程，此模型分为3个阶段：正向扩展阶段、探测点观察阶段和材料物理设置阶段。在正向扩展阶段设计了声波场动力学模型，使用波动的动力学模拟声音的传播与演化过程。在探测点观察阶段中，从多个探测点中观察声波传递到当前位置的特点。在材料物理设置阶段，动态调整步长，设置材料参数，实现跨媒介场景的自动模拟。

(1)正向扩展阶段

正向扩展的区域是在 X-Y 平面上的二维区域，它沿着 Z 轴的方向无限延伸。$\boldsymbol{x}_t$ 表示每个声波场由域左侧的一个单元输入，发射出的声波，由于传输媒介可以更换，正向扩展区域传播信号时，波速的分布和介质参数均可训练。在式(8)的基础上进行变换，可以得到式(9～10)。

$$\begin{bmatrix}\boldsymbol{u}_t\\ \boldsymbol{u}_{t-1}\end{bmatrix}=\frac{\boldsymbol{A}z_o(t)}{z_o(t)}\cdot\begin{bmatrix}\boldsymbol{u}_{t-1}\\ \boldsymbol{u}_{t-2}\end{bmatrix}+\frac{\Delta t^2 z_o(t)}{z_o(t)}\cdot\begin{bmatrix}\boldsymbol{x}_t\\ 0\end{bmatrix} \tag{9}$$

$$\boldsymbol{A}=\begin{bmatrix}\frac{2+\Delta t^2(\boldsymbol{c}+\varepsilon)^2\nabla^2}{1+\Delta tb} & \frac{-1+\Delta tb}{1+\Delta tb}\\ 1 & 0\end{bmatrix} \tag{10}$$

式中：b 为自适应的阻尼系数；ε 为不同媒介中的波速变化值。

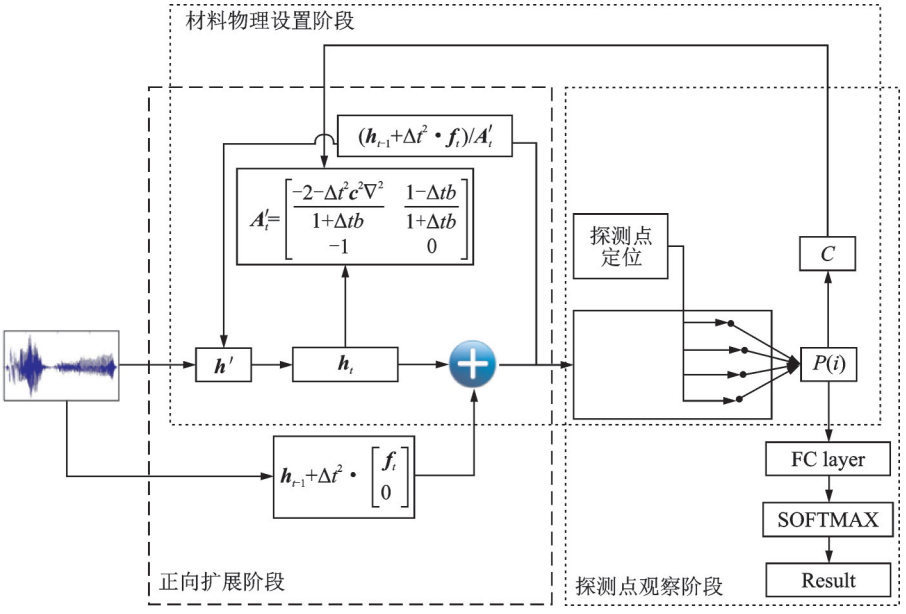

图4　声波动模型结构

Fig.4　Structure of acoustic wave equation model

设置 $\boldsymbol{h}_t=[\boldsymbol{u}_t,\boldsymbol{u}_{t-1}]^{-1}$,$\boldsymbol{A}'_t=-\boldsymbol{A}_t$,于是有

$$\boldsymbol{h}_t=(1-\boldsymbol{A}'_t)\boldsymbol{h}_{t-1}+\boldsymbol{h}_{t-1}+\Delta t^2\cdot\begin{bmatrix}\boldsymbol{f}_t\\0\end{bmatrix} \tag{11}$$

可以发现,式(11)将输入序列转化为具有时序关系的输出序列,之前的每一步操作都被编码成隐藏状态,在每一步中都得到了更新。GRU的结构如式(12~13)所示。可以看出,当前模型与GRU的形式一致

$$\boldsymbol{h}_t=(1-z)\odot\boldsymbol{h}_{t-1}+z\odot\boldsymbol{h}' \tag{12}$$

$$\boldsymbol{h}'=\tanh\left(\boldsymbol{w}\cdot\frac{\boldsymbol{x}_t}{\boldsymbol{h}'_{t-1}}\right) \tag{13}$$

基于式(12,13),可得到式(14),并获得其中的 $\boldsymbol{w}$ 的表达,如式(15)所示。

$$\boldsymbol{h}'=\tanh\left(\boldsymbol{w}\cdot\frac{\boldsymbol{x}_t}{\boldsymbol{h}'_{t-1}}\right)=(\boldsymbol{h}_{t-1}+\Delta t^2\cdot\boldsymbol{f}_t)/\boldsymbol{A}'_t \tag{14}$$

$$\boldsymbol{w}=\arctan h\left(\frac{\boldsymbol{h}_{t-1}+\Delta t^2\cdot\boldsymbol{x}_t}{\boldsymbol{A}'}\right)\cdot\boldsymbol{h}'_{t-1}\cdot\boldsymbol{x}_t{}^{-1} \tag{15}$$

此时,已经完成了正向扩展阶段。可以发现,最终的模型训练过程与GRU模型相似,它的优势是通过动态调整波速和阻尼系数等因子,来确保在不同媒介下的声音传播有效性。

(2)探测点观察阶段

本阶段利用探测点输出信息。在此区域定位了若干个探测点,默认探测点的数量与分类数量相同。每个探测点吸收的数据是声源传输至当前点位置的信号 $\boldsymbol{u}_t$,输出的是当前点位置对应的每个分类的概率值。因此,每个探测点的坐标不完全相同,这样才可以尽量保留分类的完整性。将所有观测点的输出值拼接为一个非负向量,该向量即为当前音频在物理系统中的特征表达。

探测点位置的设计思路如下，第i个探测点的边界坐标为(P_{ix},P_{iy})，探测点的数量是N，则每个探测点的坐标为

$$\begin{cases} P_{ix}=\dfrac{(1+i)*P_x}{2N} \quad i\in[0,N-1] \\ P_{iy}=\begin{cases} \dfrac{\mathrm{e}^{-i}*P_y}{1+\mathrm{e}^{-i}} \quad i\in[0,N-1]\text{且}\,i\%2==0 \\ \dfrac{P_y}{1+\mathrm{e}^{-i}} \quad i\in[0,N-1]\text{且}\,i\%2!=0 \end{cases} \end{cases} \tag{16}$$

(3)材料物理设置阶段

考虑传输声波的材质和波速对此模型产生较大的作用，此区域存在的目标是通过当前模型进行反向传播，依次估计每次输出信息的梯度、一阶矩阵、二阶矩阵，通过不断微调，对一阶矩阵和二阶矩阵进行校正，减少偏置的影响，然后开始执行随机梯度下降，最终达到模型的结果收敛。此过程与Adam相似，在本阶段通过以下的依赖关系得出波速的更新分布c'，$c'=c+\Delta^2\varepsilon$。其中，c为原有的波速，ε为材料区域中的非线性关系。由式(16)可知，波速c的反向传播是可行的，通过微调获得当前介质下的波速相当值。

上述3个阶段组成了跨媒介的声波动模型，它是基于波场动力学进行设计的，此模型模拟声波在实际物场中的传播过程，其本质与GRU相似，却同时具备较强的跨媒介迁移能力。

2 实验与结果分析

2.1 数据集

为了测试SER模型的有效性和多种介质下模型的迁移能力，本文使用自建的多介质情感语音语料库和交互情感二元动作捕捉(Interactive emotional dyadic motion capture, IEMOCAP)情感语料库[18]进行实验验证。

2.1.1 多介质情感语音语料库

基于流行的SER方向，研究人员设计了大量的多模态语料库，但是极少数语料库设计多介质下的音频数据，例如空气、液体、遮挡物等。因此，设计了一个多介质的短语音汉语语料库。

为了确保情感语料的覆盖面和规模，主要在不同媒介中采集诱导情感语音，目标是设计一个规模大、年龄层覆盖面广、情感类别平衡、语音质量高、情感表达基本正确的情感语音数据库。目前，此数据库中收录的情感包括高兴、愤怒、平静和悲伤4种情感。每条语音使用空气传播和遮挡物(佩戴口罩)的方式使用4种不同的情感朗读相同的文字内容。

在选取语料时，表达内容不可过长，力求使单个语音成为可独立表达情感、含有效发音的最小单元。为设计有效的语料，在主题设定的前提下，准备了40条相关的精短汉语语料信息。它们多为单人语料，每条语料的文字不超过5个字。语料的文本内容多数存在情感分歧，即情感的表达与语义无关，而且具备浓重的语音信息，要求受试者在融入特定环境后，以多种方式恰当的表达特定的感情。

成人自然情感语料库现有21 000余条有效语音，包含空气介质和佩戴口罩介质，音频数量占比为7∶3。

语音标注时采用多级别刻度的标注方式，每种情感分为5个等级，取值范围为[1,5]。等级1的情感表达最弱，等级5的表达最强，每个语音均需标识4类情感的等级。仅保留超过2/3专家的标注结果一致的数据。

2.1.2　IEMOCAP数据集

IEMOCAP数据集是使用动作、音频、视频录制的具有10个主题的5个二元会话中收集的,共有10 039个标准语音。每个会话由一位男性和一位女性演员执行脚本,并参与通过情感场景提示引发的自发的即兴对话。

本文仅使用IEMOCAP中的4类情感数据:happy类(与excited类合并)、sad类、angry类和neutral类。其余类别的样本数据均被丢弃。此种分类方法一共保留5 531个样本,每类样本的数据量为angry类1 103个,happy类1 636个,neutral类1 708个,sad类1 084个。除了angry和sad类别的样本量偏少之外,其他类别的情绪样本数据量较均衡。

上述两个数据集分别使用五折交叉验证方法进行实验。80%数据用于训练深度神经网络,剩余的数据被用于验证和准确性测试。

2.2　网络参数与评价标准

使用Pytorch框架进行声波动情感识别模型的搭建,设置学习速率为0.000 1。传输介质的波速c=1.0,完全匹配层(Perfectly matched layer, PML)多项式的阶是4.0,PML厚度是3,波速强度设置为1,边界点中含有的网格单元为2,空间网格步长是1.43。

采用加权精度(Weighted accuracy, WA)和未加权精度(Unweighted accuracy, UA)作为识别精度的评价指标。WA用于监测模型的整体性能,它是预测正确与样本总数之商,WA的计算完全依赖于正例的计算,尚未考虑数据倾斜时带来的负面影响。为了解决样本分布不均衡的问题,引入UA综合判定各个类别的SER精度,UA是全部类别识别精度的均值,对于不平衡的数据集,UA是一个相关性更强的特征。

2.3　实验设计与结果分析

2.3.1　*声波动情感识别模型有效性实验*

针对SER问题,设计如下实验验证当前模型的有效性,此时利用自建成人自然情感语料库和IEMOCAP数据集分别进行验证,在本实验中,以传统RNN模型[6]识别结果作为基线,同时对比以下几个模型:模型1为注意力机制+CNN模型[5];模型2为使用eGeMAPS[19]+3层双向LSTM情感识别模型;模型3为仅使用传统波动方程的情感识别模型;模型4为仅使用门控循环的声波动情感识别模型;模型5为当前模型(ISGR-AWEER)。

表1为上述模型在两个语料库中的识别精度,其中"*"表示文献中未含有对应的实验结果。

表1　两个语料库中的情感识别实验结果

Table 1　Experimental results of speech emotion recognition in two emotional speech corpus

模 型	Self-built corpus UA	Self-built corpus WA	IEMOCAP UA	IEMOCAP WA
基线	*	*	0.54	*
模型1	*	*	0.56	*
模型2	0.50	0.49	0.53	0.52
模型3	0.65	0.66	0.59	0.58
模型4	0.70	0.69	0.63	0.62
模型5	0.76	0.75	0.68	0.69

由表1可知，对比两组文献在IEMOCAP中的识别效果，当前模型的表现最佳，拥有最优的WA和UA。针对多介质情感语音语料库，也可获得相似的结构。由此可以确定，当前的模型可以最大限度地提升情感识别的平均准确率。

图5和图6分别是自建成人自然情感语料库和IEMOCAP数据集的情感识别混淆矩阵。尽管angry类别数据量较少，其识别准确度较高，反之，neutral类别数据量较多，其识别准确率较低。不难发现，当前模型的识别效果与数据是否倾斜无关。

为了验证当前模型的有效性，与流行的SER精度相对比，结果如表2所示。模型6[20]为使用自定义的特征和RNN模型；模型7[21]为提取常用的声学特征和语言学特征，融合RNN与联结主义时间分类(Connectionist temporal classification, CTC)进行情感识别；模型8[22]为提供eGeMAPS特征集与Attentive CNN模型融合的解决方案；模型9[23]为使用频谱图像，结合GRU模型进行处理；模型10为当前模型。通过实验，上述模型的对比结果如表2所示。

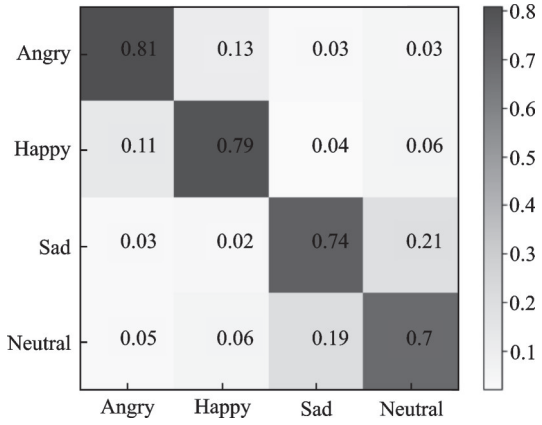

图5 自建语料库的情感识别混淆矩阵

Fig.5 Emotion recognition confusion matrix based on self-built corpus

图6 IEMOCAP情感识别混淆矩阵

Fig.6 Emotion recognition confusion matrix based on IEMOCAP

对比表2中的实验结果可以发现，与其他SER模型相比，当前模型的识别精度较高，优于一些先进的SER模型，这说明了本文设计的模型的有效性。

表2 流行SER模型的UA对比

Table 2 UA of popular SER

模型	IEMOCAP UA
模型6[20]	0.620
模型7[21]	0.480
模型8[22]	0.560
模型9[23]	0.594
模型10(当前模型)	0.680

2.3.2 跨介质有效性实验

针对跨媒介的SER模型的迁移问题，设计相关的实验验证当前模型在不同介质下迁移的有效性，此时利用多介质情感语音语料库进行验证，在本实验中，以模型2(详见表1)作为基线，将整个语料库视为一个整体数据集。分别对比在MASK(佩戴口罩)和UMASK(未佩戴口罩，空气)介质中，以下模型的有效性：模型11为使用eGeMAPS[19]+3层双向LSTM，每个介质单独训练模型；模型12为使用eGeMAPS[19]+3层双向LSTM，迁移MASK介质模型至UMASK中；模型13为使用

ISGR-AWEER,每个介质单独训练模型;模型14为使用ISGR-AWEER,迁移MASK介质模型至UMASK中。

表3为经过实验验证后,仅在多介质情感语音语料库下的识别效果。由表3可知,针对多介质情感语音语料库,采用流行的LSTM模型时,每个介质单独设计模型比迁移模型的精度提升9.6%,但整体识别效果不佳。利用本文提出的模型可规避跨介质的识别问题,单独设计模型与迁移模型结果相近,而且精度改善达到25%。通过实验证明,当前模型在跨介质SER中的有效性。

表3　多介质语料库中的情感识别实验结果

Table 3　Experimental results of speech emotion recognition in multi-media emotional speech corpus

模型	UMASK UA	UMASK WA	MASK UA	MASK WA
模型2(基线)	0.50	0.49	0.50	0.49
模型11	0.57	0.56	0.55	0.54
模型12	0.52	0.51	0.55	0.54
模型13	0.77	0.76	0.75	0.76
模型14	0.76	0.75	0.76	0.75

3　结束语

本文设计了一种融合图像显著性和门控循环的声波动方程情感识别模型,用于解决跨介质下的SER问题,该模型包含显著性区域提取和基于门控循环的声波动方程情感识别模型,可分别模拟注意力机制和RNN循环。通过在两种不同的语料库上验证,此模型可有效的实现跨介质下的情感识别,与传统的RNN模型相比,识别精度具有25%的改善。此外,该模型的迁移能力较强,适用于不同介质下的混合情感识别。

在未来的研究将进一步扩充多介质情感语音语料库的数据,同时添加文字、视频等模型数据,通过设计一种多模态的网络结构,结合显著性特征,实现对于情感识别精度的进一步提升。

参考文献:

[1] RAJASEKHAR B, KAMARAJU M, SUMALATHA V. Glowworm swarm based fuzzy classifier with dual features for speech emotion recognition[J]. Evolutionary Intelligence, 2019(1): 1-9.

[2] MA X, ZHANG T, XU C. Deep multi-modality adversarial networks for unsupervised domain adaptation[J]. IEEE Transactions on Multimedia, 2019, 21(9): 2419-2431.

[3] SARMA M, GHAHREMANI P, POVEY D, et al. Emotion Identification from Raw Speech Signals Using DNNs[C]// Proceedings of Interspeech 2018. Hyderabad, India: ISCA, 2018.

[4] KIM J, TRUONG K P, ENGLEBIENNE G, et al. Learning spectro-temporal features with 3D CNNs for speech emotion recognition[C]//Proceedings of the 7th International Conference on Affective Computing and Intelligent Interaction (ACII). [S.l.]: IEEE Computer Society, 2017: 383-388.

[5] NEUMANN M, VU N T. Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech[EB/OL]. (2017-06-02). [2020-10-23]. https://arxiv.org/abs/1706.00612.

[6] CHERNYKH V, STERLING G, PRIHODKO P. Emotion recognition from speech with recurrent neural networks[EB/OL]. (2017-01-27) [2020-10-25]. https://arxiv.org/abs/1701.08071.

[7] YAO Z, WANG Z, LIU W, et al. Speech emotion recognition using fusion of three multi-task learning-based classifiers: HSF-DNN, MS-CNN and LLD-RNN[J]. Speech Communication, 2020, 120: 11-19.

[8] SWIETOJANSKI P, RENALS S. Differentiable pooling for unsupervised acoustic model adaptation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016, 24(10): 1773-1784.

[9] LEE Jinkyu, TASHEV I. High-level feature representation using recurrent neural network for speech emotion recognition[C]//Proceedings of Interspeech. Dresden, Germany: ISCA, 2015: 1-4.

[10] TZINIS E, POTAMIANOS A. Segment-based speech emotion recognition using recurrent neural networks[C]//Proceedings of Affective Computing and Intelligent Interaction (ACII), 2017 IEEE International Conference. [S.l.]: IEEE, 2018: 190-195.

[11] KARIM F, MAJUMDAR S, DARABI H, et al. LSTM fully convolutional networks for time series classification[J]. IEEE Access, 2018, 6(99): 1662-1669.

[12] CHUNG J, GULCEHRE C, CHO K H, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling [EB/OL]. (2014-12-11) [2020-11-12]. https://arxiv.org/abs/1412.3555.

[13] XIE Y, LIANG R, LIANG Z, et al. Speech emotion classification using attention-based LSTM[J]. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 2019(99): 1-4.

[14] TANG Y, WU Z, MENG H, et al. Analysis on gated recurrent unit based question detection approach[C]//Proceedings of Interspeech 2016. Sam Francisco, USA: [s.n.], 2016: 735-739.

[15] LI K, LIAO W, LIN Y. A compact high order alternating direction implicit method for three-dimensional acoustic wave equation with variable coefficient[J]. Journal of Computational and Applied Mathematics, 2019, 361(1): 113-129.

[16] HUGHES T W, WILLIAMSON I, MINKOV M, et al. Wave physics as an analog recurrent neural network[EB/OL]. (2019-04-29) [2020-10-07]. https://arxiv.org/abs/1904.12831.

[17] ZHANG Y, LIU Y, XU S. Efficient modelling of optimised pure acoustic wave equation in 3D transversely isotropic and orthorhombic anisotropic media[J]. Exploration Geophysics, 2019, 50(5): 561-574.

[18] 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.

[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): 10-18.

[20] LEE J, TASHEV I. High-level feature representation using recurrent neural network for speech emotion recognition[C]//Proceedings of Interspeech. [S.l.]: ISCA, 2015: 1-4.

[21] PANDEY S K, SHEKHAWAT H S, PRASANNA S. Emotion recognition from raw speech using wavenet[C]//Proceedings of IEEE TENCON 2019. [S.l.]: IEEE, 2019: 1292-1297.

[22] NEUMANN M, VU N T. Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech[C]//Proceedings of Interspeech 2017. [S.l.]: ISCA, 2017.

[23] ZHANG L, WANG L, DANG J, et al. Convolutional neural network with spectrogram and perceptual features for speech emotion recognition[C]//Proceedings of International Conference on Neural Information Processing. [S.l.]: Springer, 2018.

作者简介:

贾宁(1985-),**通信作者**,女,副教授,研究方向:语音情感识别、语音合成、人工智能等,E-mail:jianing@neusoft.edu.cn。

郑纯军(1976-),男,教授,研究方向:多模态情感识别、语音合成、人工智能等,E-mail: zhengchunjun@neusoft.edu.cn。

(编辑:张彤)