基于注意力机制的 LSTM 语音情感主要特征选择

胡婷婷，冯亚琴，沈凌洁，王 蔚
(南京师范大学教育科学学院机器学习与认知实验室，江苏南京 210097)

摘要：传统的语音情感识别方式采用的语音特征具有数据量大且无关特征多的特点，因此选择出与情感相关的语音特征具有重要意义。通过提出将注意力机制结合长短时记忆网络(Long Short Term Memory, LSTM)，根据注意力权重进行特征选择，在两个数据集上进行了实验。结果发现：(1) 基于注意力机制的 LSTM 相比于单独的 LSTM 模型，识别率提高了 5.4%，可见此算法有效提高了模型的识别效果；(2) 注意力机制是一种有效的特征选择方法。采用注意力机制选择出了具有实际物理意义的声学特征子集，此特征集相比于原有公用特征集在降低了维数的情况下，提高了识别准确率；(3) 根据选择结果对声学特征进行分析，发现有声片段长度特征、无声片段长度特征、梅尔倒谱系数(Mel-Frequency Cepstral Coefficient, MFCC)、F0 基频等特征与情感识别具有较大相关性。

关键词：特征选择；语音情感识别；深度学习；注意力机制
中图分类号：H107 **文献标识码**：A **文章编号**：1000-3630(2019)-04-0414-08
DOI 编码：10.16300/j.cnki.1000-3630.2019.04.010

The salient feature selection by attention mechanism based LSTM in speech emotion recognition

HU Ting-ting, FENG Ya-qin, SHEN Ling-jie, WANG Wei
(*Machine learning and cognition lab, School of Education Science, Nanjing Normal University, Nanjing* 210097, *Jiangsu*, *China*)

Abstract: The traditional approaches to speech emotion recognition use the acoustic features characterized by large amount of data and redundancy. So, it is of great significance to choose the important phonetic features related to emotion. In this study, the attention mechanism is combined with Long Short Term Memory (LSTM) to conduct feature selection according to the attention parameters. The results show that: (1) the recognition rate of the attention mechanism based LSTM is increased by 5.4% compared with the single LSTM model, so this algorithm effectively improves the recognition accuracy; (2) the attention mechanism is an effective feature selection method, by which, the subsets of acoustic features with practical physical significance can be selected to improve the recognition accuracy and reduce the dimension compared with the original common feature set; (3) according to the selection results, the acoustic features are analyzed, and it is found that the emotion recognition is correlated with the features of voiced segment length, unvoiced segment length, fundamental frequency F0 and Mel-frequency cepstral coefficients.

Key words: feature selection; speech emotion recognition; deep learning; attention mechanism

0 引 言

情感计算是人工智能一个重要研究领域，在人机交互中情感交互具有重要意义。语音情感识别是情感计算的一个主要研究课题。在语音情感识别中，选择与情感相关的语音特征语音是情感识别中重要的工作环节。在情感识别中，研究者们通过各种特征选择方法去选择合适的语音情感特征，迄今为止，如何选择出最好的特征集，仍然没有一致清晰的意见。

声学特征是语音识别中最常用的一类特征，语音识别与语音情感识别之间有着不可分割的关联。因此，从众多语音声学特征中寻找与情感相关的特征具有重要研究意义。常用的声学特征包括音高、音强等韵律特征，频谱特征以及声音质量特征。语音特征采用开源工具 openSMILE(open-Source Media Interpretation by Large Feature-space Extraction)进行提取，关于具体提取方式与算法详见文献[1]。由于语音提取工具的标准化以及语音识别研究的逐步深入，提取的语音特征数量也越来越多。从 INTERSPEECH 2009 Emotion Challenge 中的声学特征集的 384 维[2]，到 INTERSPEECH 2010 Paralinguistic Challenge 中声学特征集 1 582 维[3]，到 INTERSPEECH 2014 Computational Paralinguistics

收稿日期：2018-08-09；**修回日期**：2018-09-03
基金项目：中国国家社会科学基金会项目(BCA150054)
作者简介：胡婷婷(1994－)，女，安徽芜湖人，硕士研究生，研究方向为器学习与深度学习，语音情感识别。
通讯作者：王蔚，E-mail: 769370106@qq.com

ChallengE 中的声学特征集已达到 6 373 维[4]。尽管这些特征集在情感识别中取得了不错的效果，但因其维数过大，若直接使用所有的情感特征建模，由于冗余特征与噪声数据的存在，会造成计算效率低、计算成本高、建模精度差、特征之间相互影响等问题。因此，为了得到维数较低、效果较好的特征集，需要使用特征选择算法从所有原始特征中选择出一个子集。

特征选择指从已有特征集中选取维数更小的子集，且识别效果不降低或更佳。目前常用的特征选择方法有以下几种：对原始数据进行随机的试探性的特征选择算法，如顺序前进选择法，其选择时随机性较大[5]；对原始数据进行数学变换的特征选择算法，如主成份分析(Principal Component Analysis, PCA)[6]以及线性判别分析(Linear Discriminant Analysis, LDA)等[7]，对原始特征空间进行数学变换与降维，导致无法对原始特征进行选择；还有一些基于机器学习的选择方法，对原始数据用分类器进行特征选择。CAO 等[8]采用随机森林的特征选择算法，选择出最有效的声学特征以提高识别效果。姜晓庆等[9]使用二次特征选择的方法，选择出具有情感区分性的语音特征子集。KIM[10]使用线性特征选择方法，结合高斯混合模型以选取声学特征。陶勇森等[11]提出将信息增益与和声搜索算法相结合的方法进行语音情感特征选择，以上研究中结合分类器对特征进行选择，旨在提高识别准确率。

在声学特征分析中，WU 等[12]得出梅尔倒谱系数(Mel-Frequency Cepstral Coefficient, MFCC)情感识别效果优于音高和能量特征，相比于前两种特征，持续时长特征识别效果较差。在对语音情感识别的特征重要性分析中，得出 F0 类识别效果优于持续时长特征，其中不同的应用统计函数得到的特征效果差异也较大，例如 F0 均值分类效果最佳，而 F0 最大值位置分类的效果较差[13]。在情感维度分类识别中，研究得到音质特征与情感的愉悦度有密切关系，韵律特征与情感激活度相关性较大的结论[14]。因此，选择出一致认同的，具有物理意义的，与情感具有较大关联性的声学特征，对于语音情感识别具有重要意义。

注意力机制最早提出于手写字生成，后来逐渐运用于多个领域。现今在机器翻译、图像标题生成、语音识别、自然语言处理多领域得到成功运用[15-18]。在语音识别中，注意力机制被用来选择出基于时序的帧水平的特征中，整个时间序列上一句话的某一帧或者某些帧的片段在整句话中的重要程度[19]。本研究受此启发，采用注意力机制在句子水平的全局特征中选择出具有重要作用的特征种类，将注意力机制结合长短时记忆网络(Long Short Term Memory, LSTM)作为一种特征选择方式。基于注意力矩阵参数选择出重要的声学情感特征并对其进行分析。同时，通过注意力机制改进深度学习中的 LSTM 识别算法，以提高情感识别效果。

1　基于注意力机制的LSTM情感识别模型

1.1　注意力机制

注意力机制的目的是在训练过程中，让模型知道输入数据中哪一部分信息是重要的，从而使模型高度关注这些信息[19]。在本研究中，对每一维声学特征，使用 Softmax 函数，使用注意力机制去获得在训练过程中每一维特征的注意力权重，进行求和后归一化。计算得到注意力特征矩阵 $\boldsymbol{A}(\boldsymbol{A}=[\alpha_1,\alpha_2,\cdots,\alpha_i,\cdots,\alpha_n])$，$\alpha_i$ 为特征集中第 i 个特征的注意力权重，此权重由注意力机制训练数据产生，理解成每一维特征对于情感识别的贡献率)。将 $\boldsymbol{A}$ 和 LSTM 层的输出 $\boldsymbol{B}$(本研究中 $\boldsymbol{B}$ 为 88×32 的二维向量)做内积运算得到 $\boldsymbol{Z}$(本研究中 $\boldsymbol{Z}$ 为 88×32 的二维向量)矩阵，将 $\boldsymbol{Z}$ 连接到模型中的全连接层，继续进行情感四分类训练。

在本研究的注意力机制中，经过 LSTM 层对输入的 $\{X_n\}$ 进行训练后(其中，$\{X_n\}$ 代表语音声学特征为 LSTM 层输入，将在 1.3 中详细介绍)，得到 LSTM 层的输出参数 $\boldsymbol{B}$，此输出参数作为注意力机制的输入。对于注意力输入序列 $\boldsymbol{B}(\boldsymbol{B}=[b_1,b_2,\cdots,b_i,\cdots,b_n])$ 中的每个参数 b_i (b_i 为 32 维的向量)，本研究中有 88 维声学特征，i 取值区间为 1～88。因此，序列 $\boldsymbol{B}$ 是 88×32 的二维矩阵。注意力权重 α_i 可通过式(1)计算：

$$\alpha_i=\frac{\exp(f(b_i))}{\sum_j \exp(f(b_j))} \tag{1}$$

其中，$f(b)$ 为计分函数，在本实验中，$f(b)$ 是线性函数 $f(b)=\boldsymbol{W}^{\mathrm{T}}b$，其中 $\boldsymbol{W}$ 是模型中可训练的参数。注意力机制的输出为 $\boldsymbol{Z}$，是由输入序列加权求和后得出的：

$$Z=\sum \alpha_i b_i \tag{2}$$

1.2　LSTM 模型

循环神经网络(Recursive Neural Network, RNN)是包含循环的网络，循环可以使得信息可以从当前步传递到下一步 LSTM 结构，允许信息的持久化。然而，相关信息和当前预测位置之间的间隔不断增大时，RNN 会丧失连接远距离信息的学习能

力。LSTM 由 HOCHREITER 及 SCHMIDHUBER 提出，并被 GRAVES 进行了改良和推广，是一种 RNN 特殊的类型，可以学习长期依赖信息[20]。

1.3 基于注意力机制的 LSTM

采用 LSTM 结合注意力机制的方式，去训练语音声学特征，建立情感识别模型。情感识别模型结构如下图 1 所示。

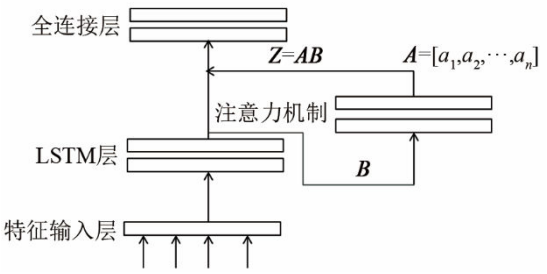

图 1 结合注意力机制的 LSTM 模型结构图

Fig.1 The structure diagram of LSTM model combined with the attention mechanism

输入序列 $\{X_n\}$ 代表语音情感特征，由 $\{X_1, X_2, \cdots, X_n\}$ 组成，在 (Geneva Minimaliastic Acoustic Parameter Set, GeMAPs)特征集中共包含 88 维声学特征，因此 n 值为 88。X_i 代表一种声学特征。时间步设为 88 步，输入维度为 1 维。将输入特征序列连接到 LSTM 层中，每个 LSTM 由 32 个神经元节点组成。将 LSTM 输出接入注意力机制，连接到一个 88 个节点的全连接层，通过一个 Softmax 进行识别，调用注意力机制计算方法，得到注意力矩阵 $\boldsymbol{A}$。在连接到全连接之前 LSTM 进行维度转置为(32,88)，以便将 88 维特征对应到每个节点上。经过全连接后再转置为(88,32)的形式，与原 LSTM 进行运算。之后将基于注意力特征矩阵 $\boldsymbol{A}$ 与原 LSTM 的输出 $\boldsymbol{B}$ 融合，进行内积乘运算后得到矩阵 $\boldsymbol{Z}$，进行压平后接入情感识别中全连接层。连接到全连接层，全连接层一设 300 个节点，激活函数使用‘ReLu’，为防过拟合，在训练过程中每次更新参数时按 0.2 的概率随机断开输入神经元。将全连接一的输出连接到全连接层二，设置四个节点，对应四种情感分类，激活函数使用‘Softmax’。使用‘Adam’优化器，计算交叉熵作为损失函数对模型进行编译。对数据循环 20 轮，采用批梯度下降更新权重，每一个 batch 大小设为 128。

2 语音特征介绍

本研究采用开源软件 openSMILE 进行帧水平的低层次基础声学特征的提取，应用全局统计函数得到句子水平全局特征[1]。比如 F0 基频特征，通过 openSMILE 软件，提取每一帧的特征，之后使用均值、方差、百分位数等函数进行全局统计，得到本研究中使用的全局特征。本研究参考之前研究中提出的 GeMAPs 特征集，提取出相关的 88 个声学特征。以下内容对 Gemaps 特征集中包含的特征做一个简单介绍，详细内容参见文献[21]。

GeMAPs 声学特征集是用于语音情感计算的常用特征集之一。采用其扩展特征集包含以下 88 个声学特征参数。特征集中包含以下 18 个低水平描述特征(Low Level Descriptors, LLDs)特征参数：

(1) 频率相关参数：F0 基频，频率微扰(jitter)，振峰频率(第一、第二、第三共振峰的中心频率)，共振峰(第一共振峰的带宽)。

(2) 能量/振幅相关参数：振幅微扰(shimmer)，响度，谐噪比(R_{HN})。

(3) 频谱(平衡)参数：Alpha 比，Hammarberg 指数，频谱斜率(0～500 Hz 和 500～1 500 Hz)，第一、第二、第三共振峰相关能量是 H1、H2、H3，第一、第二谐波差值(H1-H2)，第一、第三谐波差值(H1-H3)。

以上所有的 18 个 LLDs 都用 3 帧长对称移动平均滤波器在时间上进行平滑处理。在音高、振幅微扰和频率微扰 3 项特征上，只在有声片段进行平滑处理，对于从无声到有声片段之间的转换区域不做平滑处理。算术均值和变异系数(算术均值标准化后的标准差，变异系数)作为统计函数应用在所有的 18 个 LLDs 上，产生了 36 个特征参数。对于响度和音高额外应用了以下 8 个统计函数：20，50 和 80 的百分位数，以及 20～80 范围的百分位数，信号部分上升、下降的斜率的均值和标准差。所有的函数都应用在有声音的区域(非 0 的 F0 基频区域)，一共产生了 52 个参数。

此外，在无声片段的 Alpha 比，Hammarberg 指数，频谱斜率(0～500 Hz 和 500～1 500 Hz)的算术平均数这 4 个参数以及以下介绍的 6 个时间特征也被加入特征中，这 6 个时间特征是：

(4) 时间特征：响度峰值的比率，连续声音区域(F0>0)的平均长度和标准差，无声区域(F0=0，近似停顿)的平均长度和标准差，每秒钟连续发声区域的数目(伪音节率)。

之前的研究证明，倒谱系数在情感状态模型中具有重要作用。因此添加了以下 7 个 LLDs 成为我们扩展的特征集：

(5) 倒谱特征参数

频谱参数：梅尔频率倒谱系数 1～4，频谱流量。

频率相关参数：第二、第三共振峰的带宽。

对这 7 个 LLDs 在所有的部分(包括无声和有声部分)应用算术均值和变异系数，对共振峰带宽参数(仅在有声部分应用统计函数)，得到 14 个参数。加上频谱流量只在无声部分的算术均值，以及频谱流量和 MFCC 1-4 在有声部分的算术均值和变异系数，得到 11 个参数。此外，等效声级也被包括进来，共得到额外的 26 个参数，从而得到共 88 个参数的扩展的 eGeMAPS(Extend Geneva Minimaliastic Acoustic Parameter set)特征集。

3　情感识别与特征选择实验

3.1　数据集介绍

数据是进行研究的基础，良好的实验数据对实验结果有着直接的影响。本研究采用由美国南加州大学 SAIL 实验室收集的 IEMOCAP(interactive emotional dyadic motion capture database)公用英文数据集中语音数据进行语音情感特征选择与情感识别[22]，作为本研究的数据集一，进行模型训练与特征选择。使用 The eNTERFACE'05 Audio-Visual Emotion Database 数据集作为数据集二，用于验证我们选取的声学特征子集在情感识别中的适用性与普遍性[23]。

本研究采用 IEMOCAP 数据集中语音数据提取情感识别中的语音声学特征。IEMOCAP 数据集由 5 男 5 女在录音室进行录制，每个句子样本对应一个情感标签，情感在离散方式上标注为“愤怒”“悲伤”“开心”“厌恶”“恐惧”“惊讶”“沮丧”“激动”“中性情感”九类情感。在之前的研究中，在情感聚类识别时，由于激动和开心表现相似，区分不明显。因此将其处理为一类情感，合并为“开心”[24]。最终本研究参考一种常用情感识别方式，选取“中性”“愤怒”“开心”“悲伤”4 类情感，共 5 531 个样本进行模型训练。eNTERFACE'05 数据集被设计用于测试和评价语音与视频中情感识别任务。数据集由来自 14 个不同国家，共 44 个说话人进行录制。每个说话人根据要求录制“愤怒”“沮丧”“害怕”“开心”“悲伤”“惊讶”6 种情感的句子，每种情感包含 5 个句子。本研究选取“愤怒”“开心”“悲伤”3 种情感，共 630 个样本来验证选取的情感特征的有效性。

3.2　基于注意力机制 LSTM 的情感识别

使用数据集一中的 5 531 句声音数据，作为实验样本。根据 eGeMAPs 特征集，使用 openSMILE 工具对每句话提取出 88 维声学特征。每句话对应的手工情感标注作为训练标签。采用 1.3 节介绍的基于注意力机制的 LSTM 模型，将 88 维的声音特征作为输入序列输入到该模型中，对该模型进行训练，模型输出每句语音对应的情感的类别。采用十折交叉方式验证模型预测效果，使用样本的 9/10 进行训练，1/10 进行测试，进行 10 轮训练与预测，对 10 次的预测结果进行平均取值。在数据集一中的预测结果如表 1 所示，准确率(Accuracy, ACC)和不加权平均召回率(Unweighted Average Recall, UAR)分别达到了 0.570 和 0.582。没有注意力机制的 LSTM 分类结果 ACC 和 UAR 分别为 0.516 和 0.529。因此通过添加注意力机制，ACC 和 UAR 分别提高了 5.4%和 5.3%，证明通过注意力机制改进的情感识别模型，有效提高了情感识别准确率。

表 1　基于注意力机制 LSTM 与 LSTM 模型识别准确率对比
Table 1　Comparison of recognition accuracy between LSTM and attention mechanism based LSTM

特征集	分类器模型	ACC	UAR
eGeMAPS88	基于注意力机制 LSTM	0.570	0.582
eGeMAPS88	LSTM	0.516	0.529

在之前的基于 IEMOCAP 数据集的研究中，使用四类情感 5531 个样本，采用不同的分类器、特征集、样本得到不同的识别结果[25-28]，如表 2 所示。与之前的实验结果相比，本研究的实验结果得到了较高的识别准确率。可见，本研究实验结果表现较好。

表 2　基于 IEMOCAP 数据集研究的识别率
Table 2　Recognition rate based on IEMOCAP data set

特征集	分类器	样本标签	UAR/%	时间/年
Interspeech2011	SVM	愤怒、开心、悲伤、中性	54.23	2013
Interspeech 2011	SVM	愤怒、开心、悲伤、中性	50.64	2013
Interspeech2009	SVM	愤怒、开心、悲伤、中性	56.75	2015
eGeMAPS	CNN	愤怒、开心、悲伤、中性	54.73	2017

3.3　基于注意力机制的特征选择

特征选择一直是机器学习中至关重要的一个步骤，算法改进可以提高识别率，特征的好坏决定了准确率的高低。因此在语音情感识别中选取对情感识别影响力大的特征具有重要意义。选择具有实际可以解释的、具有物理意义的声学特征对特征选择起到至关重要的作用。选择出重要的特征后，使得后续的研究者们可以参考与借鉴。本研究采用注意力机制进行特征选择。

在注意力机制中，得到注意力参数矩阵，对所有参数进行求和后进行标准化(标准化是数据处理中，类似于归一化的预处理方式，将数据处理为均值为 0，标准差为 1 的一组数据)，得到每个特征的在情感识别模型中的概率。本研究使用 IEMOCAP 中的 5 531 个样本，提取出 88 个声学特征，对识别模型训练进行特征选择，使用十折交叉验证的方式对模型进行评估，根据注意力矩阵中每个特征对应的注意力参数，选择出对情感识别作用较大的特征。根据阈值选择出的特征数与识别率如表 3 所示，根据特征注意力参数，选择出参数大于 0.08 的特征有 81 个，大于 0.01 的有 51 个，大于 0.16 的只有 7 个。

表 3 根据阈值选择出的特征数与识别率

Table 3 The feature numbers and recognition rates based on different threshold values

阈值	特征数	UAR	ACC
0.006	88	0.585	0.573
0.008	81	0.585	0.571
0.009	68	0.584	0.570
0.010	51	0.590	0.575
0.011	35	0.585	0.570
0.012	22	0.580	0.568
0.013	19	0.571	0.560
0.014	12	0.547	0.534
0.016	7	0.536	0.522
0.017	5	0.510	0.502
0.018	4	0.507	0.499
0.019	4	0.507	0.499
0.020	4	0.507	0.499
0.025	2	0.428	0.443
0.026	1	0.377	0.403

图 2 为不同数量特征分类的结果。由图 2 可知，在选择阈值设置为 0.01 时选择出的 51 个特征取得了较高的识别效果，因此选取前 51 个特征作为本次研究的情感特征的子特征集。

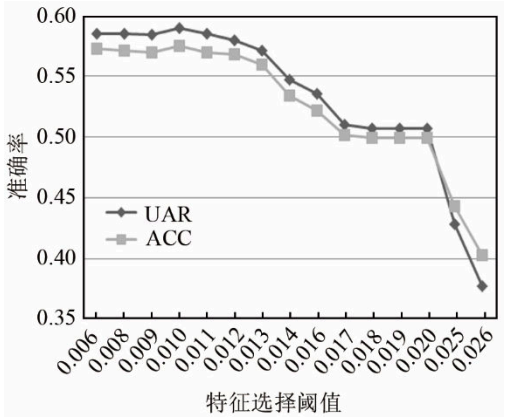

图 2 不同特征数的分类结果

Fig.2 Classification results of different numbers of features

为了对此子特征集的有效性进行验证，在 eNTERFACE’05 中使用此子特征集进行验证。由于 IEMOCAP 数据样本数多，eNTERFACE’05 样本数相比较少，因此 IEMOCAP 作为对模型进行训练与特征选择的主要实验数据集，eNTERFACE’05 作为验证数据集。在 eNTERFACE’05 数据集中使用选择出的子特征集对基于注意力机制的 LSTM 模型进行训练，发现本次选取的子特征集在验证数据集上也表现良好，如表 4 所示，相比于选取之前的 88 维特征集，在降低了维数的情况下，识别准确率有小幅度提升。有效验证了选取的子特征集不仅在选取的原数据集上表现良好，在其他数据集也表现良好，证明了此子特征集的有效性。

表 4 子特征集在验证集 eNTERFACE’05 上的表现

Table 4 The performances of feature subsets on the validation set - eNTERFACE’05

特征集	注意力	ACC	UAR
eGeMAPS_88_eNTERFACE’05 原特征集	有	0.630	0.629
eGeMAPS_51_eNTERFACE’05 子特征集	有	0.640	0.639

为了更好地比较两个数据集间的异同，补充了两数据集之间迁移学习的实验。使用数据集一的样本数据与标签训练模型，使用本研究选择后的 51 维特征集，采用基于注意力机制的 LSTM 分类器，对模型进行训练与预测，并将训练好的模型进行保存后，再使用数据二的数据来进行预测，将数据集一训练好的模型直接导入使用，分析数据集一训练好的模型在情感识别中是否具有可迁移性与通用性。由于数据集二中不含中性情感样本，因此对于中性情感标签在模型导入使用时进行补 0 处理。实验结果发现，数据集二使用该模型预测的 ACC 为 0.403，UAR 为 0.403。可以分析，数据集一与数据集二在情感表达上具有一定的相似性。

3.4 声学特征重要性分析

在对特征进行重要性排序时，基于注意力机制的特征选择步骤如图 3 所示。

首先对 IEMOCAP 数据集中样本数据提取的 88 维特征使用基于注意力机制的 LSTM 进行训练，再根据注意力参数进行排序，得到每个特征的重要性排序。之后 eNTERFACE’05 数据集使用基于注意力机制的 LSTM 再进行训练，根据注意力参数对特征进行重要性排序。比较两个数据集选取出的重要情感特征是否具有一致性，验证特征在识别中的稳定性与普遍性。

表 5 列出了根据注意力机制计算出的特征重要性排序。表 5 中的第一列表示由 IEMOCAP 数据选

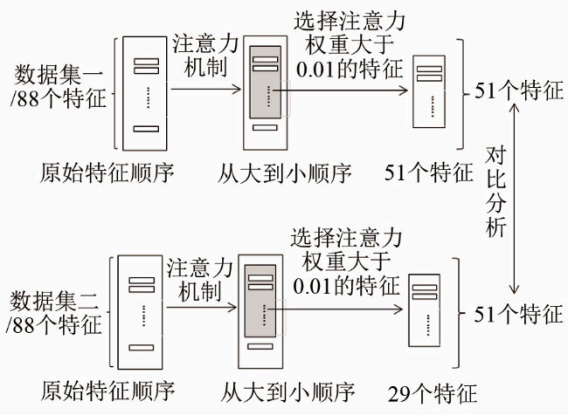

图 3　基于注意力机制的特征选择步骤

Fig.3　The feature selection procedure based on the attention mechanism

择出的重要特征，第二列表示 eNTERFACE’05 验证集数据选择出的排序靠前的特征的名称，第三列是其特征在两个数据集中的排名。由于篇幅限制只列出了前 15 个特征。

分析发现，在数据集一中，F0 排名最高，只用一个音高特征 F0_stddevNorm 进行预测时，准确率已经能到达 0.403，可见其在语音情感识别中的重要性，然而在数据集二中，该特征则表现一般。可见在不同数据集中，由于说话人、环境不同等原因会造成特征的差异。在表 5 中对两个数据集中表现差异大的特征进行了斜体标注，两个数据集中都表现良好的进行了粗体标注以方便分析。

其中，无声片段的长度(Stddev_Unvoiced Segment Length)、有声片段的长度(Stddev_Voiced Segment Length)、MFCC1 均值这 3 个特征在两个数据集中的表现均很好，而且保持稳定。基于本研究的分析中，这 3 种特征与情感之间具有很大关联，在情感识别中起较大作用。而之前研究中得出的结论为 F0 基频、响度特征优于持续时长的表现，本研究中时长特征表现良好，且在两个数据集中表现稳定。

另外，无声部分 Alpha 比表现良好，与 F0 特征两者结合在数据集一中识别准确率可达 0.443，且在数据集二中也表现良好。使用标准差统计的无声区域长度，以及响度的标准差参数在数据集一上也表现很好，以上 4 个特征已经可以达到 0.499 的准确率。其中响度的标准差参数、F1 频率均值、有声片段频谱流量、无声部分的 hammarberg 指数，MFCC2_stddev 这几个特征在两个数据集上的表现差异很大，在数据集一中表现很好，而在数据集二中表现较差。

对于特征的统计函数进行分析发现，使用算术均值和变异系数统计的特征表现优于使用百分位数或者斜率等函数统计的同类特征。更多信息我们可以从表 5 中获得，不再做详细描述。

基于选取的前 50 个声学特征可以分析出，F0 基频、Alpha 比、Hammarberg 指数、等效声级、响度斜率相关特征、MFCC 和频谱流量类的倒谱特征、jitter、shimmer、振峰频率、频谱斜率、连续声音区

表 5　根据注意力参数的特征排序

Table 5　Feature ranking by attention parameters

数据集一选择出前 15 个特征	数据集二选择出前 15 个特征	数据集一特征排名/数据集二特征排名
F0semitoneFrom27.5Hz_sma3nz_stddevNorm	mfcc1_sma3_amean	1/49
alphaRatioUV_sma3nz_amean	StddevVoicedSegmentLengthSec	2 /16
StddevUnvoicedSegmentLength	StddevUnvoicedSegmentLength	3/3
loudness_sma3_stddevNorm	F2bandwidth_sma3nz_amean	4/74
mfcc2_sma3_amean	mfcc2V_sma3nz_amean	5/37
mfcc1_sma3_amean	F2frequency_sma3nz_stddevNorm	6/1
StddevVoicedSegmentLengthSec	F0semitoneFrom27.5Hz_sma3nz_pctlrange0-2	7/2
mfcc1V_sma3nz_amean	VoicedSegmentsPerSec	8/21
F3bandwidth_sma3nz_amean	spectralFluxUV_sma3nz_amean	9/36
F1frequency_sma3nz_amean	logRelF0-H1-A3_sma3nz_amean	10/86
spectralFlux_sma3_stddevNorm	MeanUnvoicedSegmentLength	11/44
mfcc1_sma3_stddevNorm	MeanVoicedSegmentLengthSec	12/ 27
spectralFluxV_sma3nz_stddevNorm	loudness_sma3_meanFallingSlope	13/66
hammarbergIndexUV_sma3nz_amean	loudnessPeaksPerSec	14/76
mfcc2_sma3_stddevNorm	loudness_sma3_stddevFallingSlope	15/63

注：表中，amean：算术平均；stddevNorm：变异系数；sma3：三帧长对称移动平均滤波器；nz：非零 F0；V：有声；UV：无声

表中斜体标注特征表示两个数据集中差异较大，黑体标注特征表示该特征在两个数据集中均表现良好

域和无声区域的平均长度和标准差、伪音节率等特征在数据集一中表现良好。

相比以上的特征来说，共振峰带宽，第一、第二、第三共振峰的中心频率的频谱谐波峰值能量和 F0 频谱峰值能量的比、谐波差异、谐噪比，以及部分响度的参数等特征在识别中注意力参数较小，识别力较差。

4 结 论

注意力机制是通过计算特征的注意力参数，将其与深度学习模型结合训练的一种方式。本研究通过加入注意力机制，改进了 LSTM 模型，有效提高语音情感识别准确率，相比于单 LSTM 模型，准确率提高了 5.4%。

使用注意力机制进行特征选择是一种有效的特征选择方法。基于此方式选取了重要的声学特征，并且根据注意力参数，对特征进行重要性排序。本研究基于原有通用的 88 位特征集的基础上，选取了 51 维的子特征集，在降低了特征维数的情况下，取得更好的识别效果，在数据集一、二上均取得良好的结果。

对特征进行分析发现，无声片段的长度、有声片段的长度、MFCC1 均值三个特征在训练数据集与验证数据集中均表现良好，证明此 3 个特征对于情感识别的重要作用。F0、alpha 比、响度特征等与情感也具有较强关联性，在情感识别中起重要作用。算术均值与变异系数相比于其他百分位、斜率等统计函数更加具有表现力。

采用了两个数据集进行了模型的训练与特征的选择。分别使用注意力参数选择靠前的特征，发现重要的特征虽然在两次选择时，参数会有小幅波动，但是波动范围较小，说明重要的特征即使在不同数据集中，仍然保持稳定的表现，情感识别效果良好。

5 讨 论

本研究采用两个英文数据集进行情感识别与特征选择实验，由于数据集的采集方式、说话人、环境等因素不同，会对特征选择的结果产生一定程度的影响，产生不一致的结论。因此克服数据不同带来的影响，从而获得更一般性的结论至关重要。本研究为了克服数据的影响，在大样本的数据集上进行特征选择实验，在小样本的数据集上进行验证。为了消除数据产生的影响，对小样本数据集也进行了选择实验，对实验结果进行对比分析，以求获得一般性的可靠结论。但是由于数据集二中包含的样本与数据集一中有所不同，没有包含中性情感，对结果会造成一定程度的影响。在未来的工作中，希望能够发现或者制造出包含相同情感种类、相同语言并且样本数量较多的数据集以供使用。

当前语音情感识别的研究中，由于深度学习对数据量的要求增加，数据量越大模型的训练效果越可靠。但是由于单一的数据集样本量有一定限制，因此多数据集、跨数据集是研究的必然趋势。在未来的研究中，可以进行跨库、跨语言以及多语言的情感识别实验，进行更多深层次关于迁移学习在情感识别中的研究。分析不同语言、不同文化在表达情感时的共同点，分析语音中包含的信息特定情感之间关联性。

参 考 文 献

[1] EYBEN F. Opensmile: the munich versatile and fast open-source audio feature extractor[C]//Firenze, Italy: MM '10 Proceedings of the 18th ACM international conference on Multimedia, 2010: 1459-1462.

[2] SCHULLER B, STEIDL S, BATLINER A. The interspeech 2009 emotion challenge[C]//Brighton,UK:Interspeech(2009), ISCA, 2009: 312–315.

[3] SCHULLER B, STEIDL S, BATLINER A, et al. The interspeech 2010 paralinguistic challenge[C]//Chiba, Japan: Conference of the International Speech Communication Association, 2010: 2794-2797.

[4] SCHULLER B, STEIDL S, BATLINER A, et al. The interspeech 2014 computational paralinguistics challenge: cognitive & physical load[C]//Singapore:Proc. Interspeech 2014, 2014: 427-431.

[5] PÉREZ-ESPINOSA H, REYES-GARCÍA C A, VILLASEÑOR-PINEDA L. Acoustic feature selection and classification of emotions in speech using a 3D continuous emotion model[J]. Biomedical Signal Processing & Control(S1746-8094), 2012, **7**(1): 79-87.

[6] SONG P, HENGW Z, LIANG R. Speech emotion recognition based on sparse transfer learning method[J]. Ieice Transactions on Information & Systems(S1745-1361) , 2015, **98**(7): 1409-1412.

[7] ZHANG X, ZHA C, XU X, et al. Speech emotion recognition based on LDA+kernel-KNNFLC[J]. Journal of Southeast University (S1003 -7985), 2015, **45**(1): 5-11.

[8] CAO W H, XU J P, LIU Z T. Speaker-independent Speech Emotion Recognition Based on Random Forest Feature Selection Algorithm[C]//Dalian, China: Proceedings of the 36th Chinese control conference, 2017: 10995-10998.

[9] 姜晓庆, 夏克文, 林永良. 使用二次特征选择及核融合的语音情感识别[J]. 计算机工程与应用, 2017, **53**(3): 7-11.
JIANG Xiaoqing, XIA Kewen , LIN Yongliang. Speech emotion recognition using secondary feature selection and kernel fusion[J]. Computer Engineering and Applications, 2017, **53**(3): 7-11.

[10] KIM W G. Speech emotion recognition using feature selection and fusion method[J]. Transactions of the Korean Institute of Electrical Engineers(S1975-8359), 2017, **66**(8): 1265-1271.

[11] 陶勇森, 王坤侠, 杨静. 融合信息增益与和声搜索的语音情感特征选择[J]. 小型微型计算机系统, 2017, **38**(5): 1164-1168.
TAO Yongsen , WANG Kunxia , YANG Jing. Hybridizing information gain and harmony search for feature selection on speech emotion[J]. Journal of Chinese Computer Systems, 2017, **38**(5): 1164-1168.

[12] WU D, PARSONS T D, NARAYANAN S S. Acoustic feature analysis in speech emotion primitives estimation[C]//Makuhari, Chiba, Japan: Conference of the International Speech Communication Association, 2010: 785-788.

[13] TAO J, KANG Y. Features importance analysis for emotional speech classification[C]//Berlin: International Conference on Affective Computing & Intelligent Interaction, 2005, 3784: 449-457.

[14] 黄程韦, 赵艳, 金赟. 实用语音情感的特征分析与识别的研究[J]. 电子与信息学报, 2011, **33**(1): 112-116.
HUANG Chengwei , ZHAO Yan , JIN Yun. A sstudy on feature analysis and recognition of practical speech emotion[J]. Journal of Electronics & Information Technology, 2011, **33**(1): 112-116.

[15] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[J]. Computer Science, 2014, arXiv: 1409.0473.

[16] XU K, BA J, KIROS R, et al. Show, attend and tell: Neural image caption generation with visual attention[C]//ICML, 2015, 14: 77–81.

[17] CHOROWSKI J K, BAHDANAU D, SERDYUK D, et al. Attention-based models for speech recognition[J]. Computer Science (S2333-9721), 2015, **10**(4): 429-439.

[18] ADEL H, SCHUTZE H. Exploring different dimensions of attention for uncertainty detection[C]//Valencia, Spain: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 2016: 22-34.

[19] MIRSAMADI S, BARSOUM E, ZHANG C. Automatic speech emotion recognition using recurrent neural networks with local attention[C]//New Orleans, LA, USA: IEEE International Conference on Acoustics , 2017: 2227-2231.

[20] GREFF K, SRIVASTAVA R K, KOUTNIK J, et al. LSTM: a search space odyssey[J]. IEEE Transactions on Neural Networks & Learning Systems(S2162-237X), 2015, **28**(10): 2222-2232.

[21] EYBE F, SCHERER K, TRUONG K, et al. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing[J]. IEEE Transactions on Affective Computing(S 1949-3045), 2016, **7**(2): 190-202.

[22] BUSSO C, BULUT M, LEE C C. IEMOCAP: interactive emotional dyadic motion capture database[J]. LanguageResources&Evaluation(S1574-020X), 2008, **42**(4): 335-359.

[23] MARTIN O, KOTSIA I, MACQ B. The eNTERFACE'05 audio-visual emotion database[C]//Atlanta, GA, USA: Conference on Data Engineering Workshops, 2006: 8-12.

[24] METALLINOU A, WOLLMER M, EYBEN F, et al. Context-sensitive learning for enhanced audiovisual emotion classification[J]. IEEE Transactions on Affective Computing(S1949- 3045), 2012, **3**(2): 184-198.

[25] MARIOORYAD S, BUSSO C. Compensating for speaker or lexical variabilities in speech for emotion recognition[J]. Speech Communication(S0167-6393), 2014, **57**(1): 1-12.

[26] MARIOORYAD S, BUSSO C. Exploring cross-modality affective reactions for audiovisual emotion recognition[J]. IEEE Transactions on Affective Computing(S1949-3045), 2013, **4**(2): 183-196.

[27] GAMAGE K W, SETHU V, LE P N, et al. An i-vector GPLDA system for speech based emotion recognition[C]//Asia-Pacific Signal and Information Processing Association Summit and Conference. IEEE, 2015: 289-292.

[28] 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]//Stockholm, Sweden ： Interspeech, 2017: 1263-1267.