第 55 卷　第 4 期
2019 年 7 月

南京大学学报(自然科学版)
JOURNAL OF NANJING UNIVERSITY
(NATURAL SCIENCE)

Vol. 55, No. 4
July, 2019

DOI:10.13232/j.cnki.jnju.2019.04.016

基于深度学习的自然与表演语音情感识别

王　蔚*,胡婷婷,冯亚琴
(MLC实验室,南京师范大学教育科学学院教育技术系,南京,210097)

摘　要:语音是情感表达的重要途径,自然状态和表演状态下的语音所蕴含的情感信息并不完全相同.为了探索自然状态和表演状态下语音情感识别的差异,采用深度学习算法分析了IEMOCAP公用数据集,对自然状态和表演状态下的中性、愤怒、开心和悲伤等四类情绪语音数据进行实验:首先提取语音数据的声学特征(对比了emobase2010特征集和eGeMAPs特征集),然后利用卷积神经网络(Convolutional Neural Networks,CNN)对自然与表演状态下的语音情感进行识别,比较了两种状态下的情感识别率,再利用混淆矩阵分析两种状态下不同情绪之间的误分率和相似性.实验结果显示,自然状态下的情感识别率明显高于表演状态下,还发现愤怒和悲伤在两种状态下的误分率有明显区别.该现象对理解情绪的表达机制有启发意义.

关键词:情感类别,语音情感识别,深度学习,伪装语音
中图分类号:H107　　**文献标识码**:A

Speech emotion recognition in nature and scripted state based on deep learning

Wang Wei*, Hu Tingting, Feng Yaqin
(MLC Lab, Department of Educational Technology, School of Educational Science, Nanjing Normal University, Nanjing, 210097, China)

Abstract: Speech is an important way of emotional expression. The emotional information is not the same under the speech state of nature and scripted. In order to explore the difference of speech emotion recognition under the nature and the scripted state, the deep learning algorithm is used to analysis IEMOCAP public datasets. Four types of emotions, such as neutral, anger, happy and sad, are analyzed in the following experiments. Firstly, acoustic features are extracted (compared in the emobase2010 and eGeMAPs features sets). Then, Convolution Neural network (CNN) was carried out to recognize speech emotion in the nature and scripted state, respectively. Finally, confusion matrix is used to analyze the difference of the recognition accuracy of two states in every emotions. Results show that the emotion recognition accuracy in natural state was significantly higher than the one in the scripted state. There was also significant difference in the two states for angry and sad emotions. The results would be helpful for understanding the mechanism of emotional expression.

Key words: emotion categorization, speech emotion recognition, deep learning, deceptive speech

基金项目:国家哲学社会科学基金(BCA150054)
收稿日期:2019－03－05
*通讯联系人,E-mail:wangwei5@njnu.edu.cn

情感识别在人机交互中具有重要意义,在语音情感识别和合成中情绪的准确判断与语音的正确表达有直接关系.同时,语音的可控性和训练性使语音的情感表达具有很大的伪装性,语音欺骗检测是一个通过语音检测说话人真实性的问题,在许多应用领域如执法、军事和情报机构等会产生深远的影响[1-3].所以探讨自然语音情感识别对提高语音人机交互和语音合成的准确性具有重要意义.在现有的语音情感研究中,为了收集到信效度高,实验结果可解释的数据,常常通过演员表演去收集和记录情感语音数据.但是,表演状态下的数据和真实情况下自然激发的情感具有差异,使用表演语音数据集训练的模型,在实际使用中会造成识别率的下降[4],这促进了对自然语音情感表征的研究,也为本研究提供了前期数据和研究基础.

对语音的真实性研究一直是备受关注的领域,Hirschberg et al[5]建立了第一个大型欺骗语音语料库,包含约七小时的人声,发现声学韵律特征在欺骗与真实的区分中起到重要作用.Mendels et al[6]采用不同的声学特征集,使用深度学习分类器根据识别结果,去比较不同声学特征在语音欺骗检查中的效果.Fan et al[7]和Hirschberg et al[8]描述了基于语料库中对欺骗语音和真实语音进行区分的实验,并根据识别结果对特征进行了分析.这些研究发现语音在真实状态与表演状态下包含的声学特征的差异,但这些特征差异依赖于个体,较少具有稳定性和泛化性.随着MIT媒体实验室的Ekman[9]提出情感计算,研究者对语音交流中的自动情感识别产生了极大的关注.研究发现,语音的欺骗状态与害怕和高兴等情感有密切关联.借助表演中的声学情感信息去研究语音的伪装是一项值得研究的工作.

Douglas-Cowie et al[10]对现有一些情感数据集的分析发现,现有的数据集大多是说话人被要求去表现指定的情感,这样能够简化数据集收集的难度,但是收集的数据是模拟的情感语音,这种数据训练的识别模型在真实场景中运用时,识别率性能会显著的下降,可见自然和表演状态下的数据存在一些待研究的差异.Batliner et al[11]和Devillers et al[12]也发现基于表演的语音情感分类器表现不能代表实际应用中的分类准确率,实际应用中的情感识别需要通过自然状态下记录的数据去提升.Neumann and Vu[13]通过将语音情感识别使用的语音数据分为自然和表演两种状态分别进行情感识别,发现自然状态下总的情感识别率高于表演状态,且两种状态下对各种情感的识别率也有较大差异.

本研究受此启发,针对自然和表演状态下情感识别效果差异大的问题,试图去发现语音情感识别中两者语音情感表征的共同与差异.本研究通过数据挖掘去探究自然与表演状态下的语音情感的差异,在两个角度上对差异进行分析.首先是情感识别模型的水平上,根据情感识别结果,分析两种状态在不同情感中的表现.其次在误分率上,探讨在两种状态下各种情感混淆的差异性,为发现语音表达上的各情感的关联性和稳定性提供依据.

1　数据与方法

1.1　数据集与样本选择　首先分析现有的几个常用语音情感数据集,包括德语Berlin数据集、中文的CASIA数据集、英文SAVEE、enterface′05以及IEMOCAP(Interactive Emotional Dyadic Motion Capture Database)数据集.经比较发现,IEMOCAP数据集在收集时考虑到了演员表演状态与自然状态下的区别,并通过设计情境化的演员对话的交互,记录自然状态下的数据,因此选择IEMOCAP数据集进行实验.IEMOCAP数据集设计了自然和表演两种状态下的数据,将十个演员的对话记录为五个会话周期,每个周期中包含两个说话人的数据.说话人被要求表演三种选定的脚本,其中脚本内容包含明显的情感.除了脚本之外,说话人还被要求在假设的场景中去即兴开展对

话,场景通过设计去引出指定的情感(开心、愤怒、悲伤、激动、中性). 其中关于录制材料的选择、演员的选择、数据的收集与标注都是经过详细的设计的,具体可参考文献[14]. 本研究中采取实验样本情感范围包含开心、悲伤、愤怒、中性几种情感,由于激动样本与开心情感表现相似,且开心样本数量较少,因此将开心和激动合并为同一类情感,后续以开心表示[15]. 最终得到包含四类情感的5531个语音样本,其中脚本表演的语音样本,即表演状态样本数为2588句,在真实场景中即兴进行的对话,即自然状态下的样本数为2943句. 两种状态下四种情感样本分布如表1所示,自然状态下愤怒样本数较少,其余样本总体分布均衡.

表1 自然与表演状态下的样本分布

Table 1 Samples distribution under natural and scripted states

样本种类	中性	愤怒	开心	悲伤	总数
自然(improvise)	1099	289	947	608	2943
表演(scripted)	609	814	689	476	2588

1.2 声学特征 声学特征是语音情感识别中最常用的一类特征,主要包含三大类:音高音强等韵律特征、频谱特征以及声音质量特征. 本研究中声学特征采用开源软件openSMILE进行帧水平的低层次基础声学特征LLDs(Low Level Discriptors)的提取,应用全局统计函数得到全局特征. 本研究采用Interspeech 2010年泛语言学挑战赛(Paralinguistic Challenge)中广泛使用的特征提取配置文件emobase2010提取出的1582维声学特征作为特征集一[16],eGeMAPs特征集作为特征集二进行实验[17],对实验结果进行分析.

1.2.1 特征集一:emobase2010特征集 特征集包含音高、MFCC、F0等特征系数,并使用均值、最大值、最小值、斜率、百分位数等函数进行统计,最终得到包含1582维声学特征的特征集,具体参数可以参考文献[16].

1.2.2 特征集二:eGeMAPs(Extend Geneva minimaliastic acoustic parameter set)特征集 eGeMAPs声学特征集是用于语音情感计算的常用特征集之一,包含18个LLDs特征参数,如频率相关参数、能量/振幅相关参数、频谱(平衡)参数、时间特征、倒谱特征参数等. 对这些LLDs在所有的部分(包括无声和有声部分)应用算数均值和变异系数,就可以得到共88个参数的扩展eGeMAPs特征集. 更多关于特征集的介绍参见文献[17].

1.3 分类器 为消除分类器带来的影响,本研究采用支持向量机(Support Vector Machine, SVM)和卷积神经网络(Convolutional Neural Networks,CNN)两种分类器进行对比实验.

SVM是基于统计学习理论和结构风险最小化理论的,它通过建立超平面对样本进行分类,但是只能进行二分类,对于多分类问题需要构造多个分类器. 该算法在小样本领域、非线性模式识别、函数拟合等方面都具有极强的优势,是经典分类器中的典型算法.

CNN是深度学习中一种非常流行的前向神经网络,有很高的计算速度和很好的特征降维和选择能力,被广泛应用于图像和语音领域. CNN有三个特性:稀疏连接、权值共享和相等表示. 局部感受域指一个神经元只与部分邻层神经元连接,即只与一部分输入数据相连,然后在更高层将局部信息组合起来形成高级抽象表征,得到相应的全局信息. 这是一种自下而上的提取特征的过程,即层层依次接收局部的输入信息,最后进入全连接层聚合. 与每层具有相当大小的全连接网络相比,CNN能够有效降低网络模型的学习复杂度,更容易训练. 参数共享即CNN存在多个特征平面,同一特征平面的神经元共享权值,即共享卷积核参数. 有了局部感受域和参数共享这两个特点,CNN大大减少了需要训练的参数个数,同时降低了模型的复杂度,减少计算时间,使模型不易过拟合. 同时,多特征图能抽取更多特征,也能获得更好的识别率.

2　实验与结果

2.1　数据与分类　基于选取的2943个自然状态样本和2588个表演状态样本,分别使用emobase2010和eGeMAPs两组声学特征集对于自然和表演的语音情感进行分类,使用SVM和CNN分别进行实验.

实验中CNN分类器模型参数基于大量的实验和调整,具体设置如下,使用一维卷积神经网络模型,两个卷积层加上一个全连接层,输出层使用softmax激活层后得到四类预测结果.使用"Adam"优化器,损失函数使用交叉熵.每十个样本计算一次梯度下降,更新一次权重.对于模型中具体参数设置,第一层使用一维的卷积层,卷积核数目采用32个,第二层卷积层采用64个卷积核,卷积核的窗长度为10,卷积步长为1,补零策略采用"same",保留边界处的卷积结果.激活函数使用"ReLu",为防止过拟合,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.池化层采用最大值池化方式,池化窗口大小设为2,下采样因子设为2,补零策略采用"same",保留边界处的卷积结果.对所有训练样本循环20轮.

实验结果如表2所示.可以发现,自然状态下识别率显著高于表演状态下情感识别率,且在两个特征集和两种分类器上的表现一致.SVM识别率偏低,CNN识别率高.因此,后续实验均采用CNN进行.

同时,通过对两个特征集的实验结果分析发现,emobase2010特征集的识别率在自然和表演两种状态下,均高于eGeMAPs特征集的识别率.究其原因,可能是由于emobase2010声学特征集包含了更多种类的声学特征,使用了更多的统计函数对特征进行处理得到高维的特征集,包含了更多的信息,因此分类识别率更高.

表2　自然与表演状态下四类情感总的识别准确率

Table 2　The total accuracy of emotions recognition in the states of nature and scripted

数据	特征集	CNN		SVM	
		UAR	ACC	UAR	ACC
表演语音	emobase2010	0.599	0.609	0.530	0.562
自然语音	emobase2010	0.659	0.669	0.622	0.628
表演语音	eGeMAPs	0.544	0.556	0.504	0.518
自然语音	eGeMAPs	0.622	0.663	0.588	0.631

2.2　识别结果分析　图1显示了自然和表演状态下不同情绪识别准确率的对比.可以发现,自然状态下悲伤、开心和中性情感识别率均高于表演状态,愤怒识别率低于表演状态,这可能与自然状态下愤怒的样本数远少于表演状态下的愤怒样本数有关.自然状态下总体识别率优于表演状态,这是因为自然状态包含更多自然的情感信息,所以在训练模型时得到了更好的结果.然而,根据此前研究假设,表演状态下的情感表达可能更加刻意和明显,因此情感识别率可能会高于自然状态下的识别率,但是实验结果却给出了与假设相反的结论.

图2是使用混淆矩阵对识别结果的可视化图,混淆矩阵的纵向坐标为真实情感标签,横向坐标为预测情感结果,左上至右下对角线为对应情感的准确识别率,其余的为错分结果,以百分制标识.从图中可以发现,在自然状态下中性和悲伤情感的识别率最高,在表演状态下愤怒和悲伤情感的识别率更高.由于悲伤情感本身更容易被区分,所以在两种状态下识别率都较高.而愤怒情感在自然状态下识别率最低,在表演状态下识别率最高,这可能是因为自然状态下愤怒的样本数较少,所以会对识别结果产生一定的影响.同时,通过分析可以发现,自

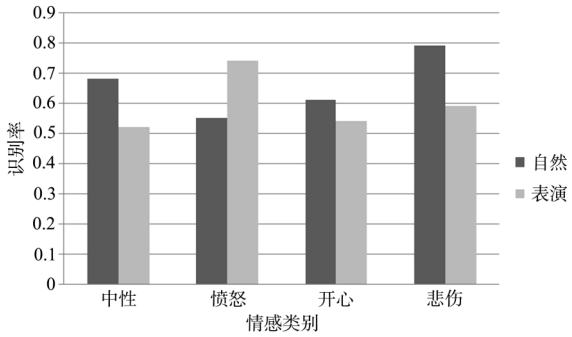

图 1 自然与表演各种情绪识别准确率结果对比

Fig. 1 The accuracy of natural and scripted emotion recognition

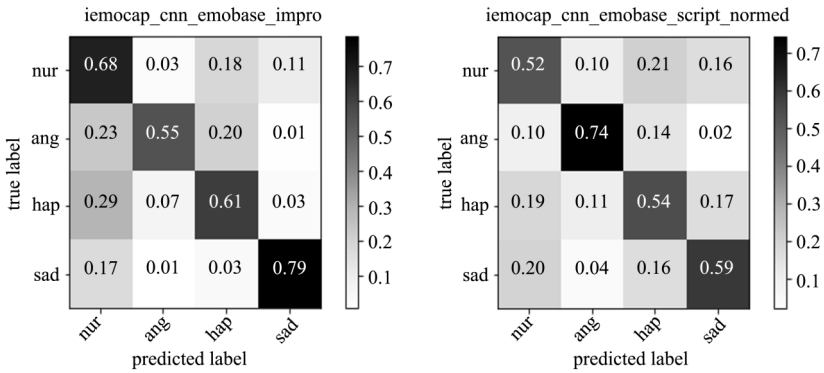

图 2 自然与表演状态下情感识别混淆矩阵

Fig. 2 Emotional recognition distribution of natural and scripted states

然状态下情感更容易被误分为中性情感,表演状态下开心和悲伤之间容易误分. 表明自然语音与表演语音在不同情感表达上存在差异.

在两组状态下四种情绪的误识率如图 3 所示,在自然状态下,愤怒误分为中性和开心比较多,表演状态下这种错识较少. 自然状态下悲伤较少被误识为开心,但在表演状态下悲伤常常被误识为开心或中性.

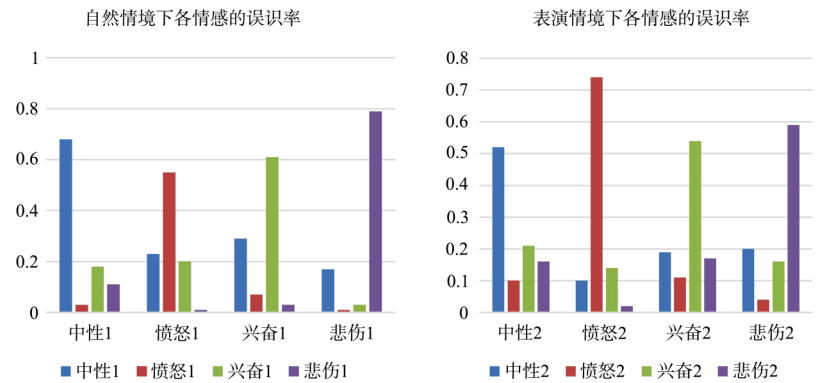

图 3 自然和表演状态下各情绪的误识率

Fig.3 Comparison of emotions misrecognition rates under natural and scripted states

3　讨论与结论

针对目前IEMOCAP数据库中的表演状态下录制的语音情感数据和诱发的自然情绪语音数据,利用深度学习算法对中性、愤怒、开心和悲伤四种情绪进行识别.实验发现,在自然状态下的语音情感识别率高于表演状态,且有的情感本身如悲伤情感在两种状态下识别率均表现良好,可能是由情感本身因素引起,导致在识别中易于区分.

在自然状态下愤怒常常被误识为开心,而表演状态下这种误识率较低.表演状态下悲伤常常被误识为开心或中性,而自然状态下较少发生.这是一个非常有意义的现象,可能暗示不同情绪的易表达性,也可能表示不同情绪之间的关联性,揭示这些现象对情感理论的研究也会有一定启发.

本研究试图通过发现这些差异特征,为后续研究中对语音表达中自然与表演的区分提供借鉴意义.然而本研究仍存在一定的局限性,由于自然和表演语音样本收集的困难性,因此研究采用的数据较少,没有采取跨数据集验证.在未来的研究中,通过使用更多的更均衡的样本,进一步分析机理.

参考文献

[1] Fan X H, Zhao H M, Chen X Q, et al. Deceptive speech detection based on sparse representation// 2016 IEEE 12th International Colloquium on Signal Processing & Its Applications. Malacca City, Malaysia: IEEE, 2016, DOI: 10.1109/CSPA.2016.7515793.

[2] Pan X Y, Zhao H M, Zhou Y. The application of fractional Mel cepstral coefficient in deceptive speech detection. PreeJ, 2015, 3: e1194.

[3] Levitan S I, An G Z, Wang M D, et al. Cross-cultural production and detection of deception from speech//Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection. New York, NY, USA: ACM, 2015, DOI: 10.1145/2823465.2823468.

[4] Devillers L, Vidrascu L, Lamel L. Challenges in real-life emotion annotation and machine learning based detection. Neural Networks, 2005, 18(4): 407－422.

[5] Hirschberg J, Benus S, Brenier J M, et al. Distinguishing deceptive from non-deceptive speech//Interspeech 2005. Lisbon, Portugal: ISCA, 2005: 1833－1836.

[6] Mendels G, Levitan S I, Lee K Z, et al. Hybrid acoustic-lexical deep learning approach for deception detection//Proceedings of Interspeech 2017. Stockholm, Sweden: ISCA, 2017, DOI: 10.21437/Interspeech.2017－1723.

[7] Fan C, Zhao H M, Chen X Q, et al. Distinguishing deception from non-deception in Chinese speech// 2015 6th International Conference on Intelligent Control and Information Processing. Wuhan, China: IEEE, 2016, DOI: 10.1109/ICICIP.2015.7388181.

[8] Hirschberg J. Deceptive speech: clues from spoken language. Chen F, Jokinen K. Speech Technology. Boston: Springer, 2010, 79－88.

[9] Ekman P. Telling lies: clues to deceit in the marketplace, politics, and marriage (Revised Edition). New York: W. W. Norton & Company, 2009, 416.

[10] Douglas-Cowie E, Campbell N, Cowie R, et al. Emotional speech: towards a new generation of databases. Speech Communication, 2003, 40(1－2): 33－60.

[11] Batliner A, Fischer K, Huber R, et al. Desperately seeking emotions or: actors, wizards and human beings//ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion. Northern Ireland, United Kingdom: Newcastle, 2000: 195－200.

[12] Chenchah F, Lachiri Z. A bio-inspired emotion recognition system under real-life conditions. Applied Acoustics, 2017, 115: 6－14.

[13] Neumann M, Vu N T. Attentive convolutional neural network based speech emotion recognition: s study on the impact of input features, signal length, and acted speech. 2017, arXiv: 1706.00612.

[14] Busso C, Bulut M, Lee C C, et al. IEMOCAP: interactive emotional dyadic motion capture database. LanguageResources and Evaluation, 2008,42(4):335－359.

[15] 胡婷婷,沈凌洁,冯亚琴等. 语音与文本情感识别中愤怒与开心误判分析. 计算机技术与发展, 2018,28(11):130－133. (Hu T T, Shen L J, Feng Y Q, et al. Research on anger and happy misclassification in speech and text emotion recognition. Computer Technology and Development, 2018,28(11):130－133.)

[16] Schuller B, Steidl S, Batliner A, et al. The INTERSPEECH 2010 paralinguistic challenge - age, gender, and affect//Interspeech 2010. Makuhari, Japan: ISCA, 2010:2794－2797.

[17] Eybe, Scherer K R, Schuller B W, et al. The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Transactions on Affective Computing, 2016, 7(2):190－202.

(责任编辑 杨可盛)