第 55 卷　第 5 期
2019 年 9 月

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

Vol. 55, No. 5
Sept., 2019

DOI:10.13232/j.cnki.jnju.2019.05.007

基于语料库的语音情感识别的性别差异研究

曹欣怡,李　鹤,王　蔚*

(南京师范大学教育科学学院教育技术系机器学习与认知实验室,南京,210097)

摘　要:性别是语音情感识别中重要的影响因素之一.用机器学习方法和情感语音数据库对语音情感识别的性别差异进行探究,并进一步从声学特征的角度分析了性别影响因素.在两个英文情感数据集以及它们的融合数据集上进行实验,分别用三种分类器对男女语音情感进行识别,并用注意力机制挑选出在男女语音情感识别中的重要特征并比较其差异.结果表明,女性语音的情感识别率高于男性.梅尔倒谱系数、振幅微扰、频谱斜率等频谱特征在男女语音的情感识别中的重要性差异较大.

关键词:机器学习,性　别,情感识别,语音情感,注意力机制

中图分类号:H107　　**文献标识码:**A

A study on gender differences in speech emotion recognition based on corpus

Cao Xinyi, Li He, Wang Wei*

(MLC Lab, Department of Educational Technology, School of Educational Science, Nanjing Normal University, Nanjing, 210097, China)

Abstract: Gender is one of the important factors in speech emotion recognition. This study aims to explore the gender differences in speech emotion recognition by using machine learning and emotional speech database, and further explore the gender differences from the perspective of acoustic features. The research was experiment on two English emotional datasets and their fusion dataset, respectively. We employed three kinds of classifiers for men's and women's speech emotion recognition. Besides, attention mechanism was used to select the important features in speech emotion recognition and compare the differences of males and females. The results show that the recognition rate of female speech emotion is higher than that of male. The importance of spectrum features such as Mel Frequency Cepstral Coefficient, Shimmer and Spectral slope in speech emotion recognition varies greatly between men and women.

Key words: machine learning, gender, emotion recognition, speech emotion, attention mechanism

语音情感识别是当前人机交互的一个研究热点,目的是从一个人的声音中自动识别这个人当前的情感状态.在语音情感自动识别领域,性别被认为是影响语音情感识别准确率的重要因素[1-3].当男性和女性的语音数据同时存在时,一些情感特征的类间方差可能很大,而

基金项目:国家哲学社会科学基金(BCA150054)

收稿日期:2019－08－14

*通讯联系人,E-mail:wangwei5@njnu.edu.cn

群间方差可能很小,这导致了语音情感识别的困难[4],所以相关的研究多为考虑性别因素以提高情感识别率,而没有对语音情感识别的性别差异进行探究.情感心理和行为学的大量研究和实验表明,男女在语音情感识别上存在差异[5-7].相关研究主要包括两个方面:一是说话者性别对语音情感识别的影响;二是听者性别对语音情感识别的影响.本文旨在用自动语音情感识别的方法进行实验,仅考虑说话者性别对语音情感识别的影响.

情感语音中可以提取多种声学特征来反映说话人的情感行为的特点[8].有声学研究表明语音特征的性别差异是影响语音情感识别性能的主要因素之一,不同性别的人在语音的长度、强度、音高以及音质上存在显著差异[4].本文在公用情感数据库上采用自动语音情感识别方法,探究说话者性别对语音情感识别结果的影响,并从声学特征的角度进一步对性别差异进行分析,主要工作如下:

(1)用自动语音情感识别对男女语音数据的情感进行分类,相对客观的从男女语音情感识别的召回率上分析差异.

(2)用注意力机制挑选出对男女语音情感分类贡献较高的特征,并根据特征所具有的物理意义,进一步从特征角度分析语音情感识别中的性别差异.

1　相关工作

性别对语音情感识别的影响的研究主要来自两个领域:一是在情感心理与行为领域,探究性别对人类语音中情感表达和感知的区别;二是语音情感自动识别领域,利用性别提高语音情感自动识别的性能.

在语音情感自动识别领域,有关性别和情感的研究多为利用性别以提高情感识别率.如Vogt and André[9]通过加入自动性别检测提高情感分类的准确率,验证了性别和情感识别系统优于性别独立的系统.Shahin[10]提出一个将性别识别、情感识别和说话人识别组合的三阶段识别模型,并验证了使用性别和情感线索的说话人识别模型优于仅使用说话人线索的识别模型.Ladde and Deshmukh[11]提出将性别识别和混合分类器情感识别结合的方法,用隐马尔科夫模型(Hidden Markov Model,HMM)进行训练,用支持向量机(Support Vector Machine,SVM)进行分类,减少了情感识别的时间还获得了更高的准确率.以上研究说明性别对于语音情感识别存在明显的影响,但没有对性别差异做进一步的分析.

在情感心理与行为研究领域,Belin et al[12]测试了基于听者和表演者性别的识别准确性差异,参与者被要求从激活-效价-强度三个维度来评估情感.实验结果显示,女性表演者的语音情感在激活和强度维度的平均识别率都高于男性.Lausen and Schacht[7]在对声音的情感识别是否与听者和说话者的性别有关的实验中发现,总的来说,女性在解码声音情感时比男性更准确,而且说话者的性别对听者如何从声音中判断情感有显著的影响.在具体的情感类型上,有研究表明男性说话者的愤怒和恐惧具有更高的识别率,而女性说话者的厌恶更容易被识别[13].这些研究通过测试实验分析了说话者性别在语音情感识别中的差异,但未从声学特征角度进一步分析.

2　研究方法

为了探究语音情感识别中的性别差异,考虑在语境因素中社会文化的性别差异可能会影响情感识别的性别差异[14],在两个英文语音情感识别的公用数据集IEMOCAP和eNTERFACE′05上进行了实验.首先将情感数据集中的男女语音数据分开,分别提取男女语音的声学特征;然后将提取到的声学特征分别输入支持向量机(SVM)、卷积神经网络(Convolution Neural Network,CNN)、长短时记忆网络(Long-Short Term Machine,LSTM)中进行情感识别,实验采用的衡量指标为机器学习常用的不加权平均召回率(Unweighted Average Recall,

UAR);最后,用基于注意力机制的LSTM从特征的角度对性别差异进行分析.

2.1 数据集 IEMOCAP数据集是南加利福利亚大学录制的用于情感识别的英文数据集,由十名专业演员(五男五女)在录音室录制,每个句子都由标注者进行离散情感和维度情感的标注.离散情感包括愤怒、悲伤、开心、厌恶、恐惧、惊讶、沮丧、激动、中性,共九类情感.由于在之前的研究中,激动和开心在情感聚类时的表现相似,区分不明显,因此合并为开心[15].本研究选取了四种最具代表性的情感进行预测:中性、愤怒、开心和悲伤,共5531个样本.

eNTERFACE′05数据集是由44个分别来自14个不同国家的人录制的英语数据集,描述了六种离散情感:愤怒、开心、悲伤、厌恶、恐惧以及惊讶.实验只选择了愤怒、开心、悲伤三种情感,共630个样本.

2.2 特征选择 实验中不同的情感特征对情感最终识别效果存在重要影响,提取和选择能有效反映情感变化的语音特征是目前语音情感识别领域最重要的问题之一.为消除特征集可能造成的影响,实验选用eGeMAPS(Extended Version Of Geneva Minimalistic Acoustic Parameters Set)和Emobase2010两种声学特征集进行实验.

首先利用开源工具包openSMILE进行帧水平的低层次声学特征提取,然后再应用全局统计函数得到全局特征.eGeMAPS由25个低水平描述符(Low Level Descriptors,LLDs)组成,包含频率、能量、频谱相关的参数,与传统的高维特征集相比仅有88维特征,但是对语音情感建模问题表现了更高的鲁棒性[16].Emobase2010是Interspeech2010年泛语言学挑战赛中广泛使用的特征集,包含34个LLDs,共1582维特征.

2.3 分类器模型 为了消除分类器对识别结果可能产生的影响,本研究分别采用SVM,CNN和LSTM建立情感识别模型.

SVM是机器学习的经典算法,在情感识别中已得到广泛使用,并且具有良好的分类效果.本实验使用线性SVM建立分类模型.CNN和LSTM是深度学习的代表算法,在自然语音处理如语音识别、语言翻译等领域取得重大成功,在语音情感识别中也有最优表现,本实验采用这两个分类器进行对比分析.CNN模型由两个卷积池化层与一个全连接层构成,卷积核的窗长度为10,卷积步长为1,激活函数使用"Relu",池化层采用最大池化的方式,最后经过softmax激活层后得到预测结果.优化器选择"Adam",损失函数为交叉熵.为了防止过拟合,在训练过程中每次变更参数时按0.2的概率随机断开输入神经元.LSTM作为常见的设置,优化器选择"Adam",固定的学习速率设置为0.001,损失函数为交叉熵.为了防止过拟合,在训练过程中每次变更参数时按0.2的概率随机断开输入神经元,最后经过softmax激活层后得到预测结果.在训练和测试的过程中,都采用十折交叉验证法,将数据集中的90%作为训练数据,10%作为测试数据.

基于注意力机制的LSTM模型是将注意力机制与按上述参数设置的LSTM模型相结合,将注意力机制接入LSTM的输出,得到注意力矩阵X;之后将注意力特征矩阵X与原LSTM的输出A融合,进行内乘运算后得到矩阵Z,再接一个全连接层,得到注意力权重系数.

3 实验结果与分析

3.1 两个数据集以及融合数据集的识别结果 提取IEMOCAP和ENTERFACE′05数据集中男女语音的两种声学特征分别输入2.3所述参数设置的三种分类器中。两个数据集男女语音情感的识别率分别如表1和表2所示.

由表1和表2可知,两个特征集上三种分类模型对女性语音情感的识别率都高于男性.为了验证二者识别率是否存在差异,对两个数据集上不同分类器和不同特征集的识别结果做独立样本T检验,结果如表3所示.T检验的显著

表1　IEMOCAP数据集上,在不同特征集和不同分类器模型中的召回率

Table 1　*UAR* of different classifiers with different feature sets on IEMOCAP dataset

性别	特征集	SVM	CNN	LSTM
Male	eGeMAPS	0.601	0.565	0.6145
	Emobase 2010	0.5445	0.6365	0.6475
Female	eGeMAPS	0.6575	0.5785	0.6455
	Emobase 2010	0.584	0.661	0.6775

表2　eNTERFACE′05数据集上,在不同特征集和不同分类器模型中的召回率

Table 2　*UAR* of different classifiers with different feature sets on eNTERFACE′05 dataset

性别	特征集	SVM	CNN	LSTM
Male	eGeMAPS	0.6533	0.4867	0.7067
	Emobase 2010	0.8667	0.6733	0.8
Female	eGeMAPS	0.7267	0.5467	0.7933
	Emobase 2010	0.88	0.6867	0.8267

性差异(p)都小于0.05(大部分为0.000),证明在两个数据集中语音情感识别确实存在性别差异,不同分类器模型对女性语音的情感识别率都更高,这说明该模型更容易识别女性声音中所包含的情感.这与许多心理学和社会学研究得出的女性比男性在语音情感的表达上更加情绪化、更易识别的结论是一致的.

为了消除两个数据集在数据采集和语言设计上的差异,本文将两个数据集融合进行实验,男女样本在不同分类器的召回率如表4所示.相同地,对融合数据集中男女的识别率结果进行T检验,在不同分类器和不同特征集下显著

表3　两个数据集上,在不同分类器和不同特征集中男女情感识别率T检验的sig值

Table 3　The significance values of *T* test for *UAR* of men and women with different classifiers and feature sets on two datasets

数据集	特征集	SVM	CNN	LSTM
IEMOCAP	eGeMAPS	0.000	0.000	0.000
	Emobase 2010	0.000	0.000	0.000
eNTERFACE′05	eGeMAPS	0.000	0.000	0.000
	Emobase 2010	0.0139	0.000	0.000

表4　融合数据集上,在不同特征集和不同分类器模型中的召回率

Table 4　*UAR* of different classifiers with different feature sets on the fusion dataset

性别	特征集	SVM	CNN	LSTM
Male	eGeMAPS	0.5871	0.5426	0.5992
	Emobase 2010	0.5308	0.6292	0.6654
Female	eGeMAPS	0.6388	0.5770	0.6483
	Emobase 2010	0.5846	0.6833	0.6829

性差异(p)都为0.000,小于0.05(数据未列出).这表明在融合特征集上语音情感的识别存在性别差异,三种分类模型对女性语音情感的识别率都高于男性,与在单一数据集上的结果一致,女性所传达的语音情感更易被识别.

3.2　基于注意力机制的特征分析结果　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[17].由于Emobase2010特征集应用的全局统计函数较多,其物理意义过于抽象;而eGeMAPS每一维特征具有更好的解

释性,同时,在特征选择过程中,特征权重的分布更加明显,所以实验采用eGeMAPS特征集. eGeMAPS中包含88维特征,是在25个LLDs上做统计函数产生的. 这些LLDs主要有F0基频、频率微扰、共振峰频率、共振峰带宽、响度、梅尔倒谱系数1~4等等. 此外,还有六个时间特征包括响度峰值的比率、连续有声区域、无声区域的平均长度和标准差以及每秒连续发音区域的数量.

本文在IEMOCAP,eNTERFACE以及它们的融合数据集上分别进行实验,用注意力机制挑选出在男女语音情感识别中较为重要的特征,并将其按照LLDs和时间特征的类型归类,比较在这三个数据集中对于女性语音情感识别共同重要的特征与识别男性语音情感共同重要的特征存在何种差异,结果如表5所示.

当前,用于语音情感识别的声学特征大致可分为韵律特征、基于谱的相关特征和音质特征这三种类型. 韵律特征主要是用来体现语音的语调、停顿、节奏,常用的韵律特征包括音高、声强、时长等. 基于谱的相关特征被认为是声道形状变化和发声运动之间相关性的体现[8],常用的谱特征包括梅尔倒谱系数(Mel Frequency Cepstral Coefficient,MFCC)、振幅等. 音质特征主要是用于衡量语音是否清晰、纯净[18],常用的音质特征包括共振峰频率、带宽等等. 据此,分析表5可发现,谱特征如MFCC4、振幅微扰、频谱斜率、谐波差异等在男女语音情感识别中重要性相差较大;其次是F1带宽、F2带宽、F1频率等音质特征;再后是每秒连续发音区域的数量、有声区域的平均长度等时间特征,属于韵律特征.

表5 三个数据集中,在男女语音情感识别中的特征重要性的差异排序

Table 5 The features in the men´s and women´s speech emotion recognition of the importance of differences on three datasets

重要性相差大的前15个特征名称	语音情感识别中的重要性排名(女性/男性)
梅尔倒谱系数4(MFCC4)	9/24
振幅微扰(Shimmer)	20/6
频谱斜率(Spectral Slope0-500 Hz)	13/26
谐波差异H1-A3(Harmonic difference)	29/19
F3相关能量	22/32
F1带宽(F1 bandwidth)	32/22
F2带宽(F2 bandwidth)	31/23
F1频率(F1 frequency)	16/8
每秒连续发音区域的数量	10/18
Hammarbeg指数	8/15
有声区域的平均长度	11/4
无声区域的平均长度	21/14
频谱流量(Spectral flux)	7/1
频率微扰(jitter)	23/17
等效音级(equivalent Sound Level_dBp)	4/10

4 结 论

本文用机器学习方法构建语音情感识别模型,对男女语音情感数据进行情感识别,目的是用与已有研究不同的方法来探究语音识别中的性别差异,消除性别刻板印象可能带来的影响;用注意力机制挑选特征,按注意力权重排序得到对男女语音情感分类较为重要的特征,进一步从特征角度进行分析. 实验结果表明,说话者性别对语音识别结果存在影响,女性的语音情感的识别率高于男性. 这一差异一般是根据社会和文化背景来解释的[19-20]. 根据性别角色理论,女性比男性更强烈的情绪强度源于男性和女性社会角色所产生的对性别差异的规范性期望. 女性要想在各自的社会角色中取得成功,就必须具有情感表达能力;而男性被期望情绪稳定,更能够控制自己的情感[21]. 所以,男性可能会隐藏自己的情感,而女性可能会更自由地表达自己的情感[14]. 这使得在进行语音情感识别的时候,女性表达的语音情感更易被识

别.在声学属性层面,男女语音的情感识别在频谱特征如MFCC4、振幅微扰、频谱斜率,音质特征如第一共振峰频率及带宽、第二共振峰带宽以及一些时长特征,如有声区域的平均长度和无声区域的平均长度上差异较大.

本文的研究仍存在不足,例如,只选择了四种常用的代表性的离散情感,但这四种情感可能并不是性别差异的典型情感.在后续的实验中将从多种离散情感或从维度情感的角度进行实验,更深入地分析情感识别的性别差异.

参考文献

[1] Dehghan A, Ortiz E G, Shu G, et al. DAGER: deep age, gender and emotion recognition using convolutional neural network. arXiv: 1702.04280, 2017.

[2] Kim J, Englebienne G, Truong K P, et al. Towards speech emotion recognition "in the wild" using aggregated corpora and deep multi-task learning. arXiv: 1708.03920, 2017.

[3] Wang Z Q, Tashev I. Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks//2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New Orleans, LA, USA: IEEE, 2017, doi: 10.1109/ICASSP.2017.7953138.

[4] Fu L Q, Wang C J, Zhang Y M. A study on influence of gender on speech emotion classification//2nd International Conference on Signal Processing System. Dalian, China: IEEE, 2010, DOI: 10.1109/ICSPS.2010.5555556.

[5] Brody L R, Hall J A. Gender, emotion, and socialization//Chrisler J C, McCreary D R. Handbook of Gender Research in Psychology. Springer Berlin Heidelberg, 2010: 429－454.

[6] Chaplin T M, Aldao A. Gender differences in emotion expression in children: a meta-analytic review. Psychological Bulletin, 2013, 139(4): 735－765.

[7] Lausen A, Schacht A. Gender differences in the recognition of vocal emotions. Frontiers in Psychology, 2018, 9: 882.

[8] 赵力,黄程韦.实用语音情感识别中的若干关键技术.数据采集与处理,2014,29(2):157－170. (Zhao L, Huang C W. Key technologies in practical speech emotion recognition. Data Acquisition and Processing, 2014, 29(2): 157－170.)

[9] Vogt T, André E. Improving automatic emotion recognition from speech via gender differentiation//Proceeding of Language Resources and Evaluation Conference. Genoa, Italy: LREC, 2006: 1－4.

[10] Shahin I. Speaker identification in emotional talking environments using both gender and emotion cues//International Conference on Communications, Signal Processing, and their Applications (ICCSPA). Sharjah, United Arab Emirates: IEEE, 2013: 1652－1659.

[11] Ladde P P, Deshmukh V S. Use of multiple classifier system for gender driven speech emotion recognition//International Conference on Computational Intelligence and Communication Networks. Jabalpur, India: IEEE, 2015, DOI: 10.1109/CICN.2015.145.

[12] Belin P, Fillion B S, Gosselin F. The Montreal affective voices: a validated set of nonverbal affect bursts for research on auditory affective processing. Behavior Research Methods, 2008, 40: 531－539.

[13] Collignon O, Girard S, Gosselin F, et al. Women process multisensory emotion expressions more efficiently than men. Neuropsychologia, 2010, 48(1): 220－225.

[14] Gong X M, Wong N, Wang D H. Are gender differences in emotion culturally universal? Comparison of emotional intensity between Chinese and German samples. Journal of Cross-Cultural Psychology, 2018, 46(8): 993－1005, doi: 10.1177/0022022118768434.

[15] Metallinou A, Wollmer M, Katsamanis A, et al. Context-sensitive learning for enhanced audio-visual emotion classification. IEEE Transactions on Affective Computing, 2012, 3(2): 184－198.

[16] Eyben F, Scherer K R, Schuller B W, et al. The geneva minimalistic acoustic parameter set

(GEMAPS) for voice research and affective 0computing. IEEE Transactions on Affective Computing,2016,7(2):190－202.

[17] Mirsamadi S, Barsoum E, Zhang C. Automatic speech emotion recognition using recurrent neural networks with local attention//IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New Orleans, LA, USA: IEEE,2017:2227－2231.

[18] Gobl C,Chasaide A N. The role of voice quality in communicating emotion, mood and attitude. Speech Communication, 2003, 40(1－2): 189－212.

[19] Jansz J. Masculine identity and restrictive emotionality//Fischer A H. Gender and Emotion: Social Psychological Perspectives. Cambridge, England: Cambridge University Press, 2000: 166－188.

[20] Shields S A. Speaking from the heart: gender and the social meaning of emotion. Cambridge, England:Cambridge University Press,2002,230.

[21] Wood W,Eagly A H. A cross-cultural analysis of the behavior of women and men: implications for the origins of sex differences. Psychological Bulletin,2002,128(5):699－727.

(责任编辑 杨可盛)