ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 34,No. 4,Jul. 2019,pp. 625－631
DOI:10. 16337/j. 1004-9037. 2019. 04. 006

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

利用语音与文本特征融合改善语音情感识别

冯亚琴　沈凌洁　胡婷婷　王　蔚

(南京师范大学教育科学学院,南京,210097)

摘　要:情感识别在人机交互中具有重要意义,为了提高情感识别准确率,将语音与文本特征融合。语音特征采用了声学特征和韵律特征,文本特征采用了基于情感词典的词袋特征(Bag-of-words,BoW)和*N*-gram模型。将语音与文本特征分别进行特征层融合与决策层融合,比较它们在IEMOCAP四类情感识别的效果。实验表明,语音与文本特征融合比单一特征在情感识别中表现更好;决策层融合比在特征层融合识别效果好。且基于卷积神经网络(Convolutional neural network,CNN)分类器,语音与文本特征在决策层融合中不加权平均召回率(Unweighted average recall,UAR)达到了68.98%,超过了此前在IEMOCAP数据集上的最好结果。

关键词:情感识别;声学特征;韵律特征;文本特征;特征融合

中图分类号:TP183　　**文献标志码**:A

Using Speech and Text Features Fusion to Improve Speech Emotion Recognition

Feng Yaqin, Shen Lingjie, Hu Tingting, Wang Wei

(School of Education Science, Nanjing Normal University, Nanjing, 210097, China)

Abstract: Emotion recognition has an important significance in human-computer interaction. The purpose of this study was to improve the accuracy of emotion recognition by fusing speech and text features. Speech features were acoustic features and phonological features, and the text features were the traditional Bag-of-Words (BoW) features based on emotion dictionary and *N*-gram model. We used these features to emotion recognition and compared their performance on the IEMOCAP data-sets. We also compared the effects of different features fusion methods, including feature-layer fusion and decision-layer fusion. Experiment results show that the performance of the fusion of speech and text features is better than that of single features; the performance of the decision-layer fusion of speech and text features is better than that of feature-layer fusion. At the same time, based on the CNN classifier, UAR of the decision-layer fusion with three features reaches 68.98%, surpassing the previous best results on the IEMOCAP data sets.

Key words: emotion recognition; acoustic features; phonological features; text features; feature fusion

引　　言

情感是人际交互的天然组成部分,具有重要作用,因而也将成为人机交互过程中不可或缺的关键要素。1997年,美国MIT实验室的Pichard在其具有里程碑意义的专著《Affective Computing》中提出

基金项目:国家社会科学基金(BCA150054)资助项目。
收稿日期:2018-01-21;**修订日期**:2018-04-04

“情感计算”这一术语,其根本宗旨就是要建立能够主动识别和理解人类情感,并能对人类情感进行正确反馈的人机交互环境。情感识别作为情感计算的基础引起了广泛关注。

一般来说,人的情感主要通过面部表情、姿态表情和言语表情来表现。而语音作为人类最直接的交流手段,其本身能传递丰富的情感信息,已被成功用于情感的自动识别中[1]。一部分研究致力于寻找更加鲁棒的声学特征去进行跨库语音情感识别[2-3]。此外,另一部分研究选择通过特征融合的方式来改进语音情感识别[1,4]。语音信息表达了大部分情感,但文本信息也能传递说话人的情感。语音与文本特征融合应用于情感识别是一个重要的研究方向。

语音情感识别就是对输入的学习者的情感化语音信号进行预处理(如降噪)后,分析和提取与学习者情感表达密切相关的语音特征参数,然后采用模式识别分类器分别进行训练和测试,最后输出学习者的情感类型,得到识别结果。针对语音情感识别,金琴等[5]在研究中采用了LLDs以及应用于LLDs上的统计函数、码字转换的声学特征、高斯超向量特征等进行情感识别,分类器为支持向量机(Support vector machine, SVM)。Lim等[6]仅采用了时频特征图在深度卷积神经网络和循环神经网络中进行识别。金琴等的研究着重于寻找合适的语音特征表示;Zhang等[7]的研究则注重改进分类器结构,他们的研究都在一定程度上提高了情感识别的准确率。

文本情感分析是指用自然语言处理,文本挖掘以及计算机语言学等方法来识别和提取原素材中的主观信息。对于文本特征,基于情感词典的稀疏特征在当前的研究中占主导地位。基于机器学习的*N*-gram的文本特征也经常用于情感识别。皇甫璐雯等[8]采用了情感词典以及情感的认知结构模型进行识别。Gamage等[9]关注文本中的动词,运用随机森林进行情感识别[7]。现有研究提出了一些改进方法,但主要还集中于浅层的语义建模,并未考虑深层次的语义联系。深度学习开始应用在自然语言处理过程后,许多深度学习方法的工作都通过神经语言模型学习词向量表示并对学习到的词向量进行语义合成用于分类[10]。李华等[11]运用词向量与神经网络相结合获取文本的情感信息。

语音或文本情感识别都只是利用了情感表达的一种方式,并未包含全部的情感信息。已有一些研究将语音与文本融合用于情感识别。Ye等[4]选用了韵律特征以及统计函数和基于词典的词袋文本特征分别在SVM和贝叶斯分类器上进行识别任务,融合方式为决策层融合。陈鹏展等[1]运用序列浮动选择算法选取的声学特征和词袋特征在高斯混合模型分类器中进行识别,融合方式为决策层融合。从已有研究可看出,语音与文本特征的表示方式多样,并未形成一个统一的标准,值得进行更多的探索,寻求更合适的特征。同时,对于语音与文本特征的融合方式,已有研究也未进行效果的比较。因此本文选择了声学特征、韵律特征、词袋特征和和*N*-gram模型,比较不同特征融合方式的识别效果。

本文从语音和文本形式提取特征,探究多特征融合的情感识别。语音特征的第1种表示是基于帧的低层次声学特征(Low level descriptors, LLDs)和应用于LLDs上的统计函数;第2种是ToBI提取到的韵律特征。文本特征采用了基于手工提取情感词典的BoW特征和*N*-gram模型。本研究应用这些特征进行情感识别,比较其在IEMOCAP数据集上4类情感识别的性能;此外还比较了不同特征融合方式在情感识别任务中的效果,包括特征层融合和决策层融合。

1 特征提取与融合方式

1.1 语音特征

1.1.1 声学特征

声学特征是从语音信号中提取的,包括韵律特征(如音高、能量共振峰等)、光谱特征(如谱截止频率,相关密度和Mel频率能量等),和倒谱特征(如线性预测倒谱系数(LPCC)、Mel频率倒谱系数(MFCC),和感知线性预测(PLP)等),已被广泛应用于情感识别任务中。

本文首先对每个语音句子提取了LLDs,在基础声学特征上应用了21个不同的统计函数,将每个句子的一组时长不等的基础声学特征转化为等长的静态特征。

使用openSMILE工具包将音频分割为帧,计算LLD,最后应用全局统计函数。本文参考了Interspeech2010年泛语言学挑战赛(Paralinguistic challenge)[12]中广泛使用的特征提取配置文件"emobose2010.conf"。它包含了38个低层次的声学特征(如MFCC,音量等),21个全局统计函数应用于低层次的声学特征和它们相应的系数。这些统计函数包括最大最小值、均值、时长、方差等,如表1所示。因此,声学特征向量的维数是1 582。

表1 帧级的低层次声学特征(LLDs)及统计函数[12]

Tab. 1 LLDs and statistical functions[12]

LLDs	统计函数
Loudness	position- max./min.
MFCC(Melfrequencycepstralcoefficients)[0-14]	arith. mean, std. deviation
LogMelfrequencyband[0-7]	skewness, kurtosis
LSPfrequency[0-7]	lin. regression coeff.
F0 by Sub-Harmonic Sum	lin. regression error Q/A
F0 envelope	Quartile
Voicing probability	quartile range
Jitter local	Percentile
Jitter DDP	quartile range
Shimmer local	Up-level time

1.1.2 韵律特征

韵律是指语音中凌驾于语义符号之上的音高、音长、快慢和轻重等方面的变化,其存在与否决定着一句话是否听起来自然顺耳、抑扬顿挫[13]。其中音高、能量、时长等特征常用于情感识别任务中。

本文使用ToBI生成韵律特征。ToBI标注主要是感知语音信号中音高的高(H)、低(L)与词与词之间的韵律间隔(0-4,从弱到强)。音高重音包含单一的音调指标(H,!H*,L*)和双音调组合(L+H*,L*+H,H+!H*)。!H代表从H向下降的目标。"*"对应于与重音音节对齐的音调。对于韵律间隔,0和1对应正常流利的单词转换,2对应没有显著的音调标记的间隔,3标记与H-,L-,或!H -相关的中间短语。

ToBI提取一句话中每个韵律表征的次数组成固定长度的特征表示。韵律特征向量的维数是260。

1.2 文本特征

1.2.1 词袋(BoW)特征

词袋(BoW)模型在文本处理中有着广泛的应用。它在不考虑词序、语义结构或语法的情况下处理文本。词汇通常使用频率-逆文档频率(Term frequency-inverse document frequency,TF-IDF)理论来选择。TF-IDF 的主要思想:如果某个词或短语在一篇文章中出现的频率高,且在其他文章中很少出现,则认为此词或者短语具有很好的类别区分能力,适合用来分类。因此,TF-IDF倾向于过滤掉常见的词语,保留重要的词语。

本文首先移除停止词,然后使用TF-IDF来选择前k词。分别从4个情感类中选出400个左右的词,去除重复词,并将它们合并成基本词汇表。词袋特征表示每个维度的只包含0或1。对于一个单词,其

中1表示发生,0代表不发生。BoW特征向量的维数是955。

1.2.2 *N*-gram模型

N-gram模型利用了上下文中相邻词间的搭配信息,获取了短时的语序信息。Bag-of-*N*-gram可看作BoW的延伸,而BoW只考虑单个词,*N*-gram考虑了连续单词。为了移除没有提供情感信息的*N*-gram,每个*N*-gram的TF-IDF被计算,选取2~50范围内的作为特征。同样,特征表示每个维度的只包含0或1。对于一个*N*-gram,其中1表示发生,0表示不发生。本文预先选取了2-gram以及3-gram在SVM中进行识别,结果发现2-gram识别率高于3-gram。因此实验选用2-gram模型。2-gram模型的特征维数为2 850。

1.3 融合方式

在情感识别任务中,融合方式一般有"早"融合和"晚"融合2种。"早"融合,也就是特征层的融合(Feature-level fusion, FL),指不同的特征在识别之前融合,如图1所示;而"晚"融合,即决策层的融合(Decision-level fusion, DL),指不同特征在识别后进行自适应加权融合,如图2所示。本文分别将语音特征与文本特征进行特征层融合与决策层融合,比较不同融合方式对情感识别的效果。

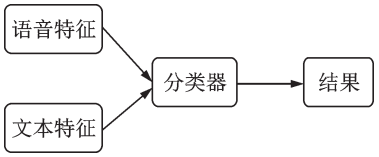

图1 语音特征与文本特征特征层融合

Fig.1 Feature-Level fusion of speech features and text features

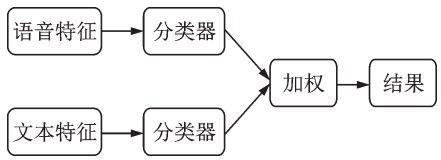

图2 语音特征与文本特征决策层融合

Fig.2 Decision-Level fusion of speech features and text features

2 实验和结果

在自动情感识别领域,实验的评测标准是UAR和准确率(Accuracy score, ACC)。本文中的所有实验都采用10组留一交叉验证模式,即用9个说话人的数据做为训练集,1个说话人的数据做为验证集。

2.1 情感数据集

本文是对Interactive Emotional Motion Capture (IEMOCAP)数据集进行4类离散情感识别,分别为高兴、悲伤、生气和中性。IEMOCAP是由南加利福利亚大学录制的情感数据库,包含约12 h的视听数据,即视频、音频和语音文本、面部表情[14]。10名演员(5段对话,每段对话1男1女)在有台词或即兴的场景下,引导出情感表达。之后,人工将每一段对话切分成单句,每句话至少由3个标注员进行类别标注。为了平衡不同情感类别的数据,将高兴(Happy)和兴奋(Exciting)合并成高兴类别。由高兴、生气、悲伤和中性最终构成了4类情感识别数据库,总共5 531个句子,如表2所示。

表2 IEMOCAP数据集中每个情感类别语句的数量

Tab. 2 The number of statements per emotion category of IEMOCAP

情感	生气	高兴	悲伤	中性	总计
样本数	1 103	1 636	1 084	1 708	5 531

2.2 分类器的选择

除了特征,情感识别的表现还取决于分类器功能。应用最广泛的机器学习算法已被用于情感识别。例如,隐马尔可夫模型(Hidden Markov model, HMM)、*K*-近邻(*K*-nearest neighbor, KNN)、人工神经网络(Artificial neural network, ANN)、支持向量机(Support vector machine, SVM)等,其中支持向

量机被认为是对不同的模式识别问题可以得到比其他的传统分类技术更好、更泛化的性能的方法。与浅学习算法相比,深度学习模型最近提高了情感识别性能。深度学习模式的网络结构允许特性表示的自动抽象。在不同的深度学习模型中,CNN和长短时记忆循环神经网络(Long short term memory,LSTM)受到了广泛关注。因此本文分别采用SVM,CNN和LSTM作为分类器。

对于CNN模型中具体参数设置,第1层使用一维的卷积层,卷积核数目采用32个,第2层卷积层采用64个卷积核,卷积核的窗长度都为10,卷积步长为1,补零策略采用"same",保留边界处的卷积结果。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",激活函数使用"ReLu"。最后连接到全连接层,通过softmax激活层后得到4类预测结果。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。使用"Adam"优化器,损失函数使用交叉熵。每10个样本计算1次梯度下降,更新一次权重。对所有训练样本循环15轮。

对于LSTM模型,采用2层LSTM,第1层的输出数目为64个,第2层的输出数目为32个。最后连接到全连接层,通过softmax激活层后得到4类预测结果。为防止过拟合,在训练过程中LSTM内部和层之间更新参数时按0.2的概率随机断开输入神经元。使用"rmsprop"优化器,损失函数使用交叉熵。每32个样本计算一次梯度下降,更新一次权重。对所有训练样本循环15轮。

2.3 语音与文本多特征情感识别

本实验分别分析了BoW特征、2-gram特征、LLDs和韵律特征(ToBi)用于情感识别的UAR和ACC。同时,将语音特征与文本特征分别特征层融合应用于识别,比较在情感识别任务中的效果。

表3列出了BoW,2-gram,LLDs,ToBi以及语音与文本特征的特征层融合在IEMOCAP数据集上的识别结果。从表3可以看出语音与文本特征的特征层融合的UAR和ACC大多高于单一特征。并且最好的结果是基于CNN的BoW,2-gram,声学和韵律特征的特征层融合,UAR为61.19%。由此可知语音与文本特征的特征层融合比单一特征在情感识别任务中表现更好。

2.4 语音与文本特征融合方式

情感识别中融合方式一般有特征层融合和决策层融合2种。本实验分别将语音特征与文本特征的

表3 语音与文本多特征在IEMOCAP上的识别结果

Tab. 3 Recognition results of speech and text features on IEMOCAP

特征	分类器					
	SVM		CNN		LSTM	
	UAR/%	ACC/%	UAR/%	ACC/%	UAR/%	ACC/%
BoW	55.09	56.28	56.08	57.42	41.23	43.69
2-gram	51.09	53.26	60.21	59.35	35.04	38.25
LLDs	52.34	50.24	59.53	58.01	56.35	54.07
ToBi	38.47	39.75	34.86	38.22	36.38	38.47
BoW+LLDs	**59.78**	**58.52**	60.96	60.07	58.23	56.68
BoW+ToBi	57.62	58.79	48.89	50.96	44.86	45.72
2-gram+LLDs	57.35	55.93	61.24	60.27	59.05	57.22
2-gram+ToBi	54.15	54.99	46.78	48.79	39.69	40.69
BoW+2-gram+LLDs+ToBi	57.64	56.48	**63.39**	**62.64**	**58.39**	**56.46**

注:+:特征层的融合;⊕:决策层的融合。

决策层融合应用于情感识别,实验的特征组合方式与上实验的组合相同。如此设置的目的是对比分析特征的决策层融合和特征层融合在情感识别中的效果。

表4列出了语音特征与文本特征的决策层融合在IEMOCAP数据集上的识别结果。从表4可以看出语音与文本特征的决策层融合的UAR和ACC都高于特征的特征层融合。基于CNN的词袋、2-gram、声学和韵律特征的决策层融合取得了最好的结果,UAR为68.98%。相较于语音与文本特征的特征层融合最好的结果提高了7.79%。由此可证明语音与文本特征的决策层融合比特征层融合在情感识别任务中表现更好。同时基于CNN分类器的情感识别取得了最好的UAR为68.98%,超过了此前在IEMOCAP数据集上的最好结果。

表4 语音与文本特征决策层融合在IEMOCAP上的识别结果

Tab. 4 Recognition results of decision-level fusion of speech and text features

特征	分类器					
	SVM		CNN		LSTM	
	UAR/%	ACC/%	UAR/%	ACC/%	UAR/%	ACC/%
BoW⊕LLDs	63.33	63.13	67.36	66.72	62.88	61.63
BoW⊕ToBi	57.71	59.12	57.16	58.77	48.54	50.65
2-gram⊕LLDs	60.93	60.65	66.80	66.35	61.16	59.79
2-gram⊕ToBi	54.51	56.22	53.01	55.17	43.69	45.83
BoW⊕2-gram⊕LLDs⊕ToBi	**64.21**	**64.61**	**68.98**	**68.36**	**65.38**	**64.56**

注:+:特征层的融合;⊕:决策层的融合。

3 结束语

结合了语音与文本信息进行情感识别,其中语音特征采用了基于帧的低层次声学特征(LLDs)以及应用于LLDs上的统计函数和ToBI提取到的韵律特征;文本特征采用了基于手工提取情感词典的词袋特征和*N*-gram模型。结果证明了语音与文本特征融合比单一特征在情感识别任务中表现更好;语音与文本特征的决策层融合比语音与文本特征的特征层融合表现更好。同时基于CNN分类器,语音语文本特征的决策层融合UAR达到了68.98%,超过了此前在IEMOCAP数据集上的最好结果。

在语音情感识别中,结合文本信息提高了情感识别的准确率。但传统的词袋特征由于参数空间的爆炸式增长,丢失了语序信息。此外,并未考虑词与词之间的内在联系性。因此未来的研究将对文本信息进行深层语义分析,建立长期的语序信息,期望提高准确率。

参考文献:

[1] 陈鹏展,张欣,徐芳萍.基于语音信号与文本信息的双模态情感识别[J].华东交通大学学报,2017,34(2):100-104.
Chen Pengzhan, Zhang Xin, Xu Fangping. Bi-modal emotion recognition based on speech signal and text information[J].Journal of East China Jiaotong University,2017,34(2):100-104.

[2] Tahon M, Devillers L. Towards a small set of robust acoustic features for emotion recognition:Challenges[J]. IEEE/ACM Transactions on Audio Speech & Language Processing, 2016, 24(1):16-28.

[3] Mohammed A, Carlos B. Domain adversarial for acoustic emotion recognition[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018:1-10.

[4] Ye Weilin, Fan Xinghua. Bimodal emotion recognition from speech and text[J]. International Journal of Advanced Computer Science and Application, 2014,5(2):26-29.

[5] 金琴,陈师哲,李锡荣,等.基于声学特征的语言情感识别[J].计算机科学,2015,42(9):24-28.

Jin Qin, Chen Shizhe, Li Xirong, et al. Speech emotion recognition based on acoustic characteristics[J]. Computer Science, 2015,42(9): 24-28.

[6] Lim W, Jang D, Lee T. Speech emotion recognition using convolutional and recurrent neural networks[C]//Signal & Information Processing Association Summit & Conference. Jeju, South Korea: IEEE,2017: 1-4.

[7] 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, 2017: 1576-1590.

[8] 皇甫璐雯,毛文吉.一种基于OCC模型的文本情感挖掘方法[J].智能系统学报,2017,12(5): 645-652.
Huangfu Luwen, Mao Wenji. A text emotion mining method based on OCC model[J].Journal of Intelligent Systems, 2017, 12 (5): 645-652.

[9] Gamage K W, Sethu V, Ambikairajah E . Salience based lexical features for emotion recognition[C]// IEEE International Conference on Acoustics. New Orleans, LA, USA: IEEE, 2017: 5830-5834.

[10] Zhou Chunting, Sun Chonglin, Liu Zhiyuan, et al. A C-LSTM neural network for text classification[J]. Computer Science, 2015,1(4): 39-44.

[11] 李华,屈丹,张文林,等.结合全局词向量特征的循环神经网络语言模型[J].信号处理,2016,32(6): 715-723.
Li Hua, Qu Dan, Zhang Wenlin, et al. A recurrent neural network language model combining global word vector characteristics [J]. Signal Processing,2016,32(6): 715-723.

[12] Schuller B, Batliner A, Steidl S, et al. The interspeech 2010 paralinguistic challenge[C]//Proceedings of Interspeech. Makuhari, Japan: ISCA, 2010: 2794-2797.

[13] 韩文静,李海峰,阮华斌,等.语音情感识别研究进展综述[J].软件学报,2014,25(1): 37-50.
Han Wenjing, Li Haifeng, Ruan Huabin, et al. Review of research progress in speech emotion recognition[J]. Journal of Software,2014,25(1): 37-50.

[14] Busso C, Bulut M, Lee CC, et al. IEMOCAP: Interactive emotional dyadic motion capture database[J]. Journal of Language Resources and Evaluation, 2008,42(4): 335-359.

作者简介:

冯亚琴(1994-),女,硕士研究生,研究方向:语音情感识别、文本情感识别,E-mail: 1181292141@qq.com。

沈凌洁(1993-),女,硕士研究生,研究方向:语音情感识别,E-mail: 602249910-@qq.com。

胡婷婷(1994-),女,硕士研究生,研究方向:语音情感识别,E-mail: 1090561350-@qq.com。

王蔚(1966-),女,博士,教授,研究方向:智能信息处理、生物信息挖掘,E-mail: wangwei5@njnu.edu.cn。

(编辑:张彤)