文章编号:1005-0523(2017)02-0100-05

基于语音信号与文本信息的双模态情感识别

陈鹏展,张 欣,徐芳萍

(华东交通大学电气与自动化工程学院,江西 南昌 330013)

摘要:情感识别已成为人机交互不可或缺的部分,目前单模态情感识别具有识别率低、可靠性差的特点,故提出一种融合语音信号与文本信息的双模态情感识别方法。首先,采集特定情感状态下的语音信号及文本信息;然后提取语音相关特征参数以及文本情感关键词特征参数并对其进行优化;最后,对两个单模态识别器的输出结果进行加权融合获得识别结果。针对所提算法进行了相关实验研究,结果表明双模态情感识别技术具有更高识别精度。

关键词:语音信号;文本识别;参数优化;高斯混合模型

中图分类号:TP391 文献标志码:A

DOI:10.16749/j.cnki.jecjtu.2017.02.014

情感是人类交流的具体体现,在人机交互中起着重要的作用。而情感识别是情感计算的基础,能否进行情感识别直接影响情感计算的实现。语音信息作为人类最直接的交流手段,其本身能传递丰富的信息资源[1-3],但介于音频信号本身存在一些固有缺陷,如信号弱、噪声强等,从单一的模型获得正确的情感状态很难满足当前情感识别系统的需求。多模态的融合利用语音、生理信号、面部表情等多个通道的情感信息互补性提高分类器的识别性能,从而提高识别分类器的准确度。多模态融合的优势在于,当某一个通道的特征或者识别过程受到缺失或者影响时,另一个通道仍能保证较好识别率,使识别系统具有一个良好的鲁棒性。

以语音信号与文本信息为基础,研究语音信号与文本信息的相应的情感特征分析及融合算法。通过对语音识别与文本识别判决结果进行加权融合,构建基于双模态分类器,并比较了基于语音信号与文本信息的单模态分类器以及基于双模态分类器的识别率。

1 特征提取

在人机交互中,情感识别技术所面临的最大挑战之一是评价说话者的情绪。通常对于说话者情绪的判断,从音频中提取特征,而语音信号所表述的文本信息也可以用来被监测说话者的情绪。通过音频信号与文本信息的双模态融合,计算机可以识别“谁说”、“说的是什么”、“如何说”,以更正确、更自然的实现与人的互动。同时,该技术具有很高的应用价值,如呼叫中心、电子服务中心、电子学习及娱乐等。

1.1 语音信号特征提取

在语音的情感识别中,能够表示语音的情感相关的特征相对较多[4-6],除一些较为广泛认同的参数,如

收稿日期:2016-10-24

基金项目:国家自然科学基金资助项目(61164011);江西省研究生创新专项资金项目(YC2015-S242);江西省博士后科研择优资助项目(2015KY19)

作者简介:陈鹏展(1975—),男,副教授,博士,研究方向为传感网络、人机交互。

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

能量、共振峰、语速、语调、基音等，还有其他参数，如能量谱分布、线性预测倒谱系数(LPCC)、Mel 频率倒谱系数(MFCC)。

针对语音信号的时域和频域特性，经过序列浮动前向选择算法(sequential floating forward selection，SFFS)[7]对特征集进行反复实验，最终选取了 74 个全局统计特征，其中，特征 1~10 为基音及其一阶差分的均值、最大值、最小值、中值、方差，特征 11~20 为短时能量及其差分的均值、最大值、最小值、中值、方差，特征 21~25 为基因频率的均值、最大值、最小值、中值、方差，特征 26~45 为第 1~第 4 共振峰均值、最大值、最小值、中值、方差，特征 46~50 过零率均值、最大值、最小值、中值、方差，特征 51~74 为 24 阶 MFCC 均值。

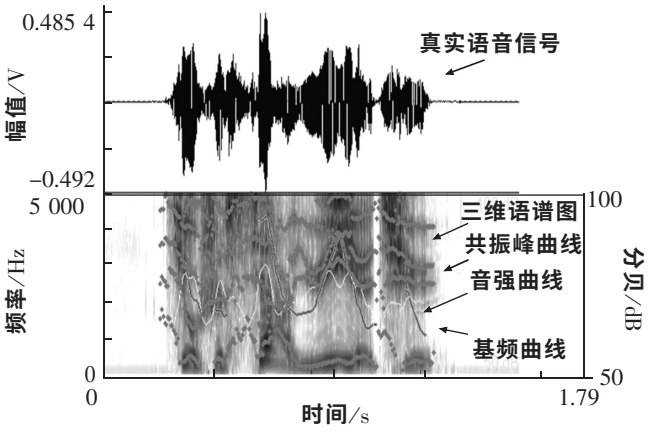

图 1　语音信号分析图

Fig.1　Analysis of speech signal

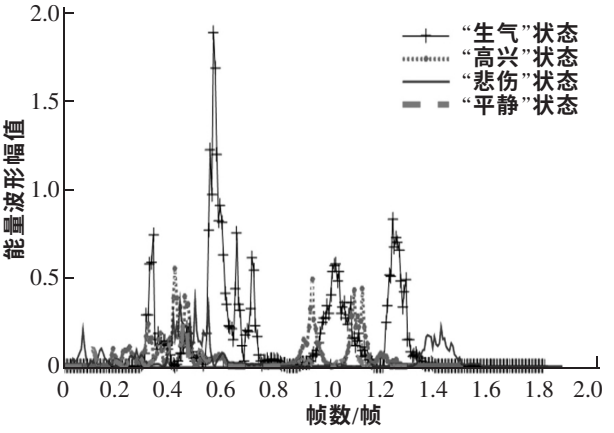

图 2　同一语句四种情感状态下能量波形对比图

Fig.2　Comparison of energy waveforms of four emotional states under the same statement

1.2　文本信息特征提取

文本信息的特征分析主要在于文本的预处理和文本向量化表述。文本预处理主要有语句拆分、简化等。句子拆分是将文本分割成一系列单独的单词文本，便于后面的测试，分词采用的是大连理工信息检索研究室整理及标注的中文情感词库。

去除停用词处理在于去除对分类没有意义的词语减少文本特征向量的维度和不必要的运算量。根据创建的停用词表使用布隆过滤器来去掉所有文本中的停用词。

特征提取采用信息增益(X)[8]，名词 Q 的 X 值定义为

$$X(Q)=-\sum_{k=1}^{n} p(A_k)\log P(A_k)+P(Q)\sum_{i=1}^{n} p(A_k)\log P(A_k|Q)+P(\overline{Q})\sum_{i=1}^{n} p(A_k|\overline{Q})\log P(A_k|\overline{Q}) \quad (1)$$

其中：A_k(其中 $k=1,\cdots,m$)表示第 k 类；$p(A_k)$是在训练样本集中是 A_k 类的概率；$p(Q)$，$p(\overline{Q})$分别是名词 Q 在训练样本集，不在训练样本集中出现的概率；$p(A_k|Q)$，$p(A_k|\overline{Q})$分别是名词 Q 出现的前提下样本是 A_k 类的概率，及名词 Q 不出现的前提下样本是 A_k 类的概率。X 值越高，对分类预测提供的信息就越多。通过设定阈值，可以将 X 值小于阈值的名词删除掉，从而降低特征空间维度。

2　分类器模型创建

2.1　单通道语音情感识别模型

音频情感识别模型创建思想是：对原始语音信号进行适当的预处理获得有效音频信号，如分帧、加窗、端点监测等，然后运用 SFFS 算法对语音信号所提取的特征进行选择获得获取最优特征子集，总共包含 74 个特征向量，再通过创建训练样本与测试样本，进行高斯混合模型分类器(gaussian mixture model,GMM)[9]进行样本比对，获得语音情感识别结果。基于单通道的语音模型分类器的识别框架如图 3 所示。

2.2　单通道文本情感识别模型

文本情感识别模型主要是通过对句子中情感关键词的锁定进行判断。通过对文本内容进行预处理、特

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

征提取及相应向量转化，然后通过 GMM 算法进行情感状态的测定。而基于单通道文本分类器识别框图如图 4 所示。

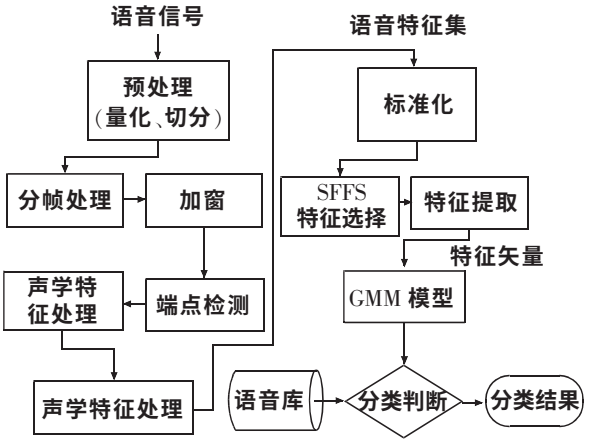

图 3 语音情感识别流程图

Fig.3 Flow chart of speech emotion recognition

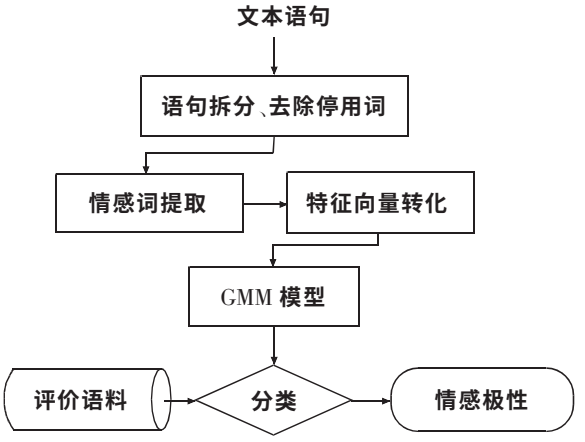

图 4 文本情感识别流程图

Fig.4 Flow chart of text emotion recognition

2.3 双模态融合识别模型创建

针对语音信号与文本信息的双模态融合识别框图如图 5 所示。该方法可使得单通道情感识别效果最大化。两个单通道识别结果作为融合的输入，通过行加权融合实现双模态情感识别分类。

本文对两种单模态分类器均采用 GMM 算法来进行生气、高兴、平静、悲伤四种情感识别。高斯混合模型是 m 个单高斯分布的加权和，表示形式如下

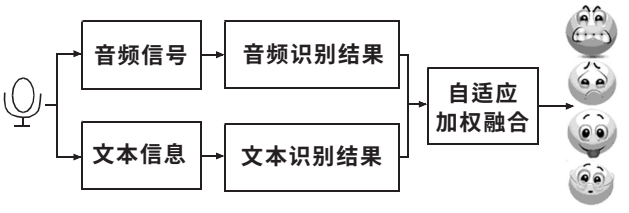

图 5 双模态情感识别系统框图

Fig.5 Block diagram of double-modal emotion recognition system

$$P(\boldsymbol{x}_t|\lambda)=\sum_{i=1}^{m} a_i p_i(\boldsymbol{x}_t;\boldsymbol{\mu}_i,\sum i) \tag{2}$$

式中：$\boldsymbol{x}_t$ 为第 t 个单高斯分布的 D 维随机向量；a_i 为第 i 个单高斯分布的权值，且 $\sum_{i=1}^{m} a_i=1$；$p_i(\boldsymbol{x}_t)(i=1,\cdots,m)$ 为单高斯分布函数，其均值矢量为 $\boldsymbol{\mu}_i$，协方差矩阵为 $\sum i$，即

$$p_i(\boldsymbol{x}_t;\boldsymbol{\mu}_i,\sum i)=\frac{1}{(2\pi)^{D/2}|\sum i|^{1/2}}\times\exp\left\{-\frac{1}{2}(\boldsymbol{x}_t-\boldsymbol{\mu}_i)^{\mathrm{T}}\sum{}^{-1}(\boldsymbol{x}_t-\boldsymbol{\mu}_i)\right\} \tag{3}$$

式中协方差矩阵可以用满矩阵，也可以用简化对角矩阵。高斯混合分布密度如公式(4)所示。其中，GMM 模型的参数估计采用 EM 算法来完成，使得 GMM 能够最佳的表示样本的分布概率。

$$\lambda_i=\{a_i,\boldsymbol{\mu}_i,\sum i\}\qquad i=1,\cdots,m \tag{4}$$

由于单通道在工作环境中存在一定干扰，本文采用自适应加权融合算法[10]实现对两个通道信息进行更新和融合，各分类器加权系数根据其对当前样本可靠性进行动态调整，置信度高的分类器所占权重更高，算法以自适应的方式找到每个分类器的最优加权因子，利用得到的加权因子实现双模态数据融合，获得最终的分类结果。对于待测样本特征 y，假设，两个子分类器均给出了四种情感类的 GMM 似然度，分别记为 $P(y|\lambda_k)$，其中 k 代表情感类别，取值为 1~4。各类别的 GMM 似然度直接决定该分类器的判决置信度的高低。子分类器融合权值表达式如公式 5 所示，其中 n 为分类器编码，取 1，2。

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

$$I_n=\frac{\sum_{1\le j<i\le 4}\left|\ln(P(y|\lambda_j))\ \ln(P(y|\lambda_i))\right|}{\left|\sum_{k=1}^{4}\ln(P(y|\lambda_k))\right|} \tag{5}$$

子分类器的判决置信度的高低与样本所处概率分布模型的非重叠区域有关，更直接的表现在分类器给出的似然度值的分散程度，似然值较为分散的分类器，其判决置信度较高，性能则较为可靠。最后，通过对两个子分类器的判决进行加权融合，获得最终的分类结果，加权融合表示形式如下

$$Y=\sum_{n=1}^{2}I_nA_n \tag{6}$$

其中：Y 为双模态分类器最终识别结果；A_n 表示子分类器分类结果，由公式(7)求得。当 $I_1>I_2$ 时，则 $Y=A_1$；同理，当 $I_2>I_1$ 时，则 $Y=A_1$。

$$A_n=\max\{P_n(y|\lambda_k)\} \tag{7}$$

3　试验结果与分析

验证通过 3 个试验结果对比来实现，分别为采用单模态语音的情感识别，采用单模态文本的情感识别以及采用双模态融合的情感识别。图 6 显示了单模态语音情感识别、单模态文本识别和基于语音与文本的双模太融合识别对情感的平均识别率。由图 6 可见，多模态的情感识别技术对每类情感的识别精度均有所提高。

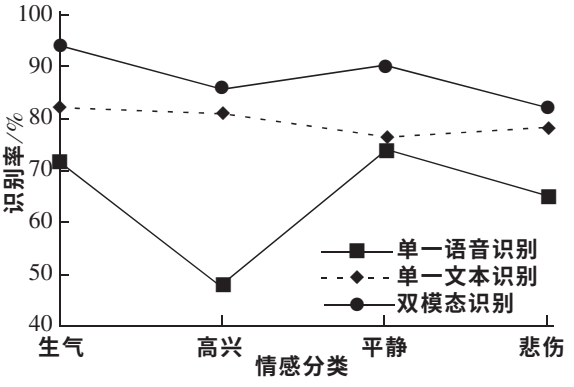

图 6　3 种方法识别率对比图

Fig.6　Comparison of the recognition rate of three methods

表 1　双模态融合算法分类情况

Tab.1　Classification of the dual mode fusion algorithm

%

分类	生气	高兴	平静	悲伤
生气	47	3	0	0
高兴	2	43	4	1
平静	0	2	45	3
悲伤	1	2	6	41

表 1 显示了双模态融合方法对每类情感的正确分类数。由表 1 可知，通过融合算法减少了每类情感的误判率，其中生气，高兴，平静，悲伤的误判率分别为 3%，7%，5%，9%。

4　结论

目前的情感识别系统多数是采用单通道情感数据进行识别研究，而本文通过加权融合方法将两种不同来源的数据的分类结果进行再次融合，实现基于语音信号与文本信息的双模态情感识别系统的研究，进行了单模态语音信号、文本信息的分类实验及双模态语音信号与文本信息融合情感识别实验。实验结果表明，基于语音信号和文本信息的双模态融合相对于单模态分类器识别率、鲁棒性均得到提高。

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

参考文献:

[1] VINCIARELLI A,PANTIC M,BOURLARD H,et al. Social signal processing survey of an emerging domain[J]. Image Vis Comput J,2009,27(12):1743-1759.

[2] CASALE S,RUSSO A,SCEBBA G,et al. Speech emotion classification using machine learning algorithms[C]// 20008 IEEE Internat ional Conference on Semantic Computing. IEEE,Cgnta Clara,CA,USA,2008:158-165.

[3] ZENG Z,PANTIC M,ROISMAN G I,et al. A survey of affect recognition methods audio, visual and spontaneous expressions[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2009,31(1):39-58.

[4] **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.

[5] ZHANG X,SUN Y,DUAN S. Progress in speech emotion recognition[J]. TENCON 2015-2015 IEEE Region 10 Conference,2015:1-6.

[6] **张跃进,刘邦桂,谢昕. 噪声背景下语音识别中的端点检测**[J]. **华东交通大学学报**,2007,24(5):135-138.

[7] OVA B N. Floating search methods in feature selection[J]. Pattern Recognition Letters,2010,15(11):1119-1125.

[8] **申红,吕宝粮,内山将夫,等. 文本分类的特征提取方法比较与改进**[J]. **计算机仿真**,2006,23(3):222-224.

[9] **黄程韦,金赟,王青云,等. 基于语音信号与心电信号的多模态情感识别**[J]. **东南大学学报:自然科学版**,2010,40(5):895-900.

[10] **叶云青,王长征,周日贵. 基于最佳指数因子的自适应权值图像融合**[J]. **华东交通大学学报**,2011,28(2):74-79.

Multimodal Emotion Recognition Based on Speech Signal and Text Information

Chen Pengzhan,Zhang Xin,Xu Fangping

(School of electrical and Automation Engineering, East China Jiaotong University, Nanchang 330013, China)

Abstract:Emotion recognition has become an indispensable part of human-computer interaction. This paper propsesa fusion method of speech signal and the text information in emotion recognition,because of the low recognition rate and poor reliability of single modal emotion recognition. First of all,collecting specific emotional state of the speech signal and text information;then extracting the speech feature parameters and keywords emotional characteristic parameters of text information and optimize it; finally, recognition results are obtained by weighted fusion of the output results of two single modal identification devices. According to the results of experimaental,it showed that the dualmodal emtoion recognition technology has higher recognition accuracy.

Key words: speech signal;text recognition;parameter optimization;gauss mixture model

(责任编辑　姜红贵)

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net