# 多模式情感识别特征参数融合算法研究

韩志艳,王 健
(渤海大学 工学院,辽宁 锦州 121000)

**摘 要**:为了克服单模式情感识别存在的局限性,文中以语音信号和面部表情信号为研究对象,提出了一种新型的多模式情感识别算法。首先,将提取的语音信号和面部表情信号特征进行融合,然后通过有放回地抽样获得各训练样本集,并利用 Adaboost 算法训练获得各子分类器。再采用双误差异性选择策略来度量两两分类器之间的差异性。最后运用多数投票原则进行投票,得到最终识别结果。实验结果表明,该方法充分发挥了决策层融合与特征层融合的优点,使整个情感信息的融合过程更加接近人类情感识别,情感识别率达 91.2%。

**关键词**:多模式;情感识别;语音信号;面部表情信号
**中图分类号**:TP391.4 **文献标识码**:A **文章编号**:1673-629X(2016)05-0027-04
doi:10.3969/j.issn.1673-629X.2016.05.006

# Research on Feature Fusion Algorithm for Multimodal Emotion Recognition

HAN Zhi-yan, WANG Jian
(College of Engineering, Bohai University, Jinzhou 121000, China)

**Abstract**: In order to overcome the limitation of single mode emotion recognition, a novel multimodal emotion recognition algorithm is proposed, taking speech signal and facial expression signal as the research subjects. First, the speech signal feature and facial expression signal feature is fused, and sample sets by putting back sampling are obtained, and then sub-classifiers are acquired by Adaboost algorithm. Second, the difference is measured between two classifiers by double error difference selection strategy. Finally, the recognition result is obtained by the majority voting rule. Experiments show the method improves the accuracy of emotion recognition by giving full play to the advantages of decision level fusion and feature level fusion, and makes the whole fusion process close to human emotion recognition more, with a recognition rate 91.2%.

**Key words**: multimodal; emotion recognition; speech signal; facial expression signal

## 1 概 述

近年来,情感识别的研究工作在人机交互领域中已经成为一个热点问题。国内外情感识别的研究主要有两大类:一类是单模式情感识别;另一类是多模式情感识别。所谓单模式情感识别,为只从单一信息通道中获得当前对象的情感状态,如从语音信号、面部表情信号或生理信号(血压、体温、脉搏、心电、脑电、皮肤电阻等)等。

对于语音情感识别,1990 年麻省理工大学多媒体实验室构造了一个“情感编辑器”对外界各种情感信号进行采样来识别各种情感,并让机器对各种情感做出适当的反应[1]。颜永红等[2]采用非均匀子带滤波器来挖掘对语音情感有益的信息,加大了各类情感之间的鉴别性,提高了情感识别的性能。毛峡等[3]通过用相关密度和分形维数作为情感特征参数来进行语音情感识别,获得了较好的性能。邹采荣等[4]提出了一种基于改进模糊矢量量化的语音情感识别方法,有效地改善了现有模糊矢量量化方法的情感识别率。Attabi 等[5]将锚模型的思想应用到了语音情感识别中,改进了识别系统的性能。Zheng 等[6]通过对传统的最小二乘回归算法进行改进,提出了不完稀疏最小二乘回归算法,能同时对标记和未标记语音数据进行情感识别。Mao 等[7]通过使用卷积神经网络来选择对情感有显著影响的特征,取得了很好的效果。

对于面部表情识别,Ekman 等[8]于 1978 年开发了

**收稿日期**:2015-08-14 **修回日期**:2015-11-22 **网络出版时间**:2016-04-
**基金项目**:国家自然科学基金资助项目(61403042,61503038);辽宁省教育厅项目(L2013423)
**作者简介**:韩志艳(1982-),女,博士,副教授,研究方向为情感识别、语音识别。
**网络出版地址**:http://www.cnki.net/kcms/detail/61.1450.TP.20160505.0828.080.html

面部动作编码系统(Facial Action Coding System, FACS)来检测面部表情的细微变化。Essa 等[9]于 1997 年提出基于视频的动态表情描述方法—FACS+,解决了 FACS 中没有时间描述信息的问题。Rahulamathavan 等[10]利用局部 Fisher 判别分析对加密面部表情信号进行了识别研究。文沁等[11]提出一种基于三维数据的人脸情感识别方法,给出了基于三维特征的眼角和嘴角新的提取算法。Zheng 等[12]提出了基于组稀疏降秩回归的多视角面部表情识别方法,能够从多尺度子域中自动选择出对情感识别贡献最大的子域。对于生理信号情感识别,Petrantonakis 等[13]采用高阶过零技术(Higher Order Crossing,HOC)提取脑电波信号中的情感信息来进行情感识别。刘光远等[14]从呼吸信号中提取特征参数进行情感识别。Zacharatos 等[15]分析研究了身体姿势和动作对情感识别的重要性。

虽然单一地依靠语音信号、面部表情信号和生理参数来进行情感识别的研究取得了一定的成果,但却存在着很多局限性,因为人类是通过多模式的方式表达情感信息的,它具有表达的复杂性和文化的相对性[16]。比如,在噪声环境下,当某一个通道的特征受到干扰或缺失时,多模式方法能在某种程度上产生互补的效应,弥补了单模式的不足,所以研究多模式情感识别的方法十分必要。例如,Kim 等[17]融合了肌动电流、心电、皮肤电导和呼吸 4 个通道的生理参数,并采用听音乐的方式来诱发情感,实现了对积极和消极两大类情感的高效识别。黄程韦等[18]通过融合语音信号与心电信号进行了多模式情感识别,获得较高的融合识别率。但是上述方法均为与生理信号相融合,而生理信号的测量必须与身体接触,因此对于此通道的信号获取有一定的困难,所以语音和面部表情作为两种最为主要的表征情感的方式,得到了广泛研究。例如,Busso 等[19]分析了单一的语音情感识别与人脸表情识别在识别性能上的互补性。Hoch 等[20]通过融合语音与表情信息,在车载环境下进行了正面(愉快)、负面(愤怒)与平静共 3 种情感状态的识别。Sayedelahl 等[21]通过加权线性组合的方式在决策层对音视频信息中的情感特征进行融合识别。

从一定意义上说,不同信道信息的融合是多模式情感识别研究的瓶颈问题,它直接关系到情感识别的准确性。因此,文中以语音信号和面部表情信号为基础,提出了一种多模式情感识别算法,对高兴、愤怒、惊奇、悲伤和恐惧五种人类基本情感进行识别。

## 2 系统结构框架

系统结构框架如图 1 所示。首先对情感数据进行一系列预处理,然后提取语音情感特征和面部表情特征,最后进行融合识别。

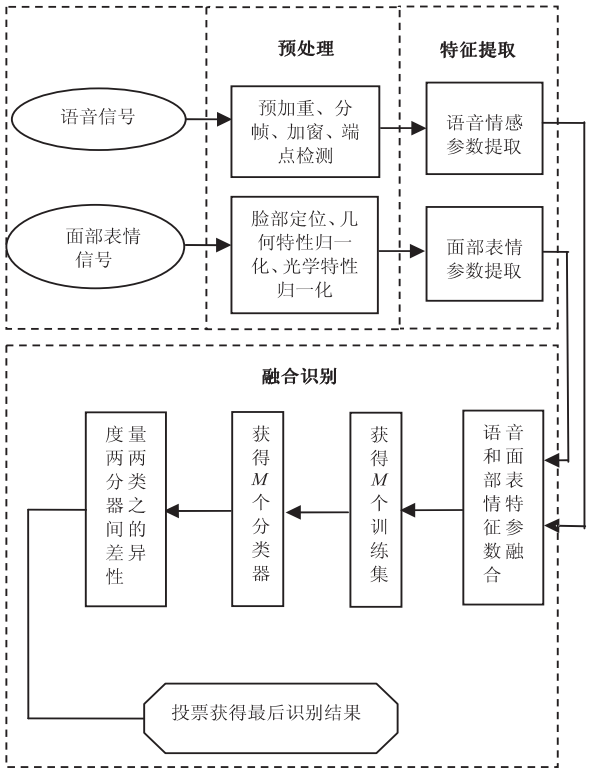


图 1 系统结构框架

## 3 参数提取

### 3.1 语音情感参数提取

以往对情感特征参数的有效提取主要以韵律特征为主,然而近年来通过深入研究发现,音质特征和韵律特征相互结合才能更准确地识别情感。Tato 等[22]研究发现,音质类特征对于区分激活维接近的情感有较好的效果,证实了共振峰等音质类特征与效价维度的相关性较强。

为了尽可能地利用语音信号中所包含的有关情感方面的信息,文中选取了语句发音持续时间与相应的平静语句持续时间的比值、基音频率平均值、基音频率最大值、基音频率平均值与相应平静语句的基音频率平均值的差值、基音频率最大值与相应平静语句的基音频率最大值的差值、振幅平均能量、振幅能量的动态范围、振幅平均能量与相应平静语句的振幅平均能量的差值、振幅能量动态范围与相应平静语句的振幅能量动态范围的差值、第一共振峰频率的平均值、第二共振峰频率的平均值、第三共振峰频率的平均值、谐波噪声比的均值、谐波噪声比的最大值、谐波噪声比的最小值、谐波噪声比的方差,作为情感识别用的特征参数。

### 3.2 面部表情参数提取

目前面部表情特征的提取根据图像性质的不同可分为静态图像特征提取和序列图像特征提取,静态图

像中提取的是表情的形变特征,而序列图像特征是运动特征。文中以静态图像为研究对象,采用 Gabor 小波变换来提取面部表情参数。具体过程如下:

(1)将预处理后的人脸图像网格化为 25×25 像素,所以每张脸共有 4 行 3 列共 12 个网格。

(2)用 Gabor 小波和网格化后的图像进行卷积,公式如下:

$$r(x,y)=\iint I(\varepsilon,\eta)g(x-\varepsilon,y-\eta)\mathrm{d}\varepsilon\mathrm{d}\eta \qquad (1)$$

式中:$I(\varepsilon,\eta)$ 为对应像素坐标 $(\varepsilon,\eta)$ 的像素值;$g(x,y)=\frac{k^2}{\sigma^2}\exp(-\frac{k^2(x^2+y^2)}{2\sigma^2})(\exp(\mathrm{i}\boldsymbol{k}\cdot\begin{pmatrix}x\\y\end{pmatrix})-\exp(-\frac{\sigma^2}{2}))$。其中,$\sigma$ 是与小波频率带宽有关的常数,取值为 $\sqrt{2}\pi$;$\boldsymbol{k}=\begin{pmatrix}k_v\cos\varphi_u\\k_v\sin\varphi_u\end{pmatrix}$,$k_v=2^{-\frac{v+2}{2}}$,$\varphi_u=u\frac{\pi}{k}$,$v$ 的取值决定了 Gabor 滤波的波长,取值为 0,1,2,$u$ 的取值表示 Gabor 核函数的方向,取值为 1,2,…,6;$k$ 表示总的方向数,取值为 6。

(3)取卷积结果的模 $\|r(x,y)\|$ 的均值和方差作为面部表情参数。

(4)用主成分分析法(PCA)对上述特征进行降维处理,获得的面部表情特征参数作为特征融合的特征参数。

## 4　算法描述

具体实施步骤如下:

第一步:通过噪声刺激和观看影视片段等诱发方式,采集相应情感状态下的语音信号和面部表情信号,并将二者绑定存储。对于语音数据,在提取特征之前要进行一阶数字预加重、分帧、加汉明窗和端点检测等预处理。对于面部表情数据,在提取特征之前要首先用肤色模型进行脸部定位,然后进行图像几何特性归一化处理和图像光学特性的归一化处理。其中,图像几何特性归一化主要以两眼位置为依据,而图像光学特性的归一化处理包括先用直方图均衡化方法对图像灰度做拉伸,以改善图像的对比度,然后对图像像素灰度值进行归一化处理,使标准人脸图像的像素灰度值为 0,方差为 1,如此可以部分消除光照对识别结果的影响。

第二步:根据 3.1 节和 3.2 节的方法提取语音情感参数和面部表情参数。

第三步:将提取的语音情感参数和面部表情参数顺序组合起来,获得多模式特征向量 $u_1$,以此类推,获得了原始训练样本集中所有的多模式特征向量 $u_2$,…,$u_r$,…,$u_W$。其中,$r=1,2,\cdots,W$,$W$ 为原始训练样本集中语音信号样本数,即面部表情信号样本数。

第四步:通过对多模式特征向量集有放回地抽样 $N$ 次获得训练样本集 $S_1$,然后依次继续抽取样本获得训练样本集 $S_2,\cdots,S_M$,即获得 $M$ 个训练样本集。

第五步:利用 Adaboost 算法对上述 $M$ 个训练样本集 $S_k,k=1,2,\cdots,M$,分别进行训练,获得每个训练样本集上的强分类器。其中,以三层 BP 神经网络作为弱分类器。

第六步:采用双误差异性选择策略来度量两两强分类器之间的差异性,并挑选出大于平均差异性的强分类器作为识别分类器,其强分类器 $H_i$ 和 $H_j(i\neq j)$ 之间的差异性公式如下:

$$\mathrm{Div}(i,j)=\frac{\mathrm{num}^{00}}{\mathrm{num}^{00}+\mathrm{num}^{01}+\mathrm{num}^{10}+\mathrm{num}^{11}} \qquad (2)$$

其中,$\mathrm{num}^{ab}$ 表示两两强分类器分类正确/错误的样本数,$a=1$ 和 $a=0$ 分别表示强分类器 $H_i$ 分类正确和错误,$b=1$ 和 $b=0$ 分别表示强分类器 $H_j$ 分类正确和错误。

第七步:运用多数优先投票原则进行投票,得到最终识别结果。

## 5　仿真实验及结果分析

为证明文中方法的识别效果,将单模式条件下与多模式条件下的识别结果进行对比。原始训练样本集包含每种情感的 200 条语音数据样本与 200 条面部表情数据样本,测试集包含每种情感的 100 条语音数据样本和 100 条面部表情数据样本。

在单模式条件下,仅通过语音信号进行识别的情感识别正确率如表 1 所示,仅通过面部表情信号进行识别的情感识别正确率如表 2 所示。

表 1　仅通过语音信号进行识别的正确率　%

| 情感类别 | 高兴 | 愤怒 | 惊奇 | 悲伤 | 恐惧 |
|---|---|---|---|---|---|
| 高兴 | 86 | 0 | 12 | 2 | 0 |
| 愤怒 | 4 | 81 | 0 | 7 | 8 |
| 惊奇 | 20 | 1 | 77 | 2 | 0 |
| 悲伤 | 5 | 4 | 0 | 88 | 3 |
| 恐惧 | 5 | 10 | 6 | 4 | 75 |

表 2　仅通过面部表情信号进行识别的正确率　%

| 情感类别 | 高兴 | 愤怒 | 惊奇 | 悲伤 | 恐惧 |
|---|---|---|---|---|---|
| 高兴 | 85 | 2 | 13 | 0 | 0 |
| 愤怒 | 0 | 79 | 7 | 10 | 4 |
| 惊奇 | 0 | 0 | 81 | 9 | 10 |
| 悲伤 | 0 | 20 | 4 | 66 | 10 |
| 恐惧 | 3 | 8 | 2 | 9 | 78 |

由表 1 和表 2 可知,仅通过语音信号进行识别的

平均识别正确率是 81.4%；仅通过面部表情信号进行识别的平均识别正确率是 77.8%。因此，单纯依靠语音信号或面部表情信号进行识别在实际应用中会遇到一定的困难，因为人类是通过多模式的方式表达情感信息的，所以研究多模式情感识别的方法十分必要。

在多模式条件下，通过简单组合文中的语音信号和面部表情信号特征进行识别的情感识别正确率如表 3 所示，通过文中方法进行识别的情感识别正确率如表 4 所示。

注：表中第 $i$ 行第 $j$ 列的元素表示真实情感状态是 $i$ 的样本被判别成 $j$ 的比例。

表 3 通过简单组合语音信号和面部表情信号进行识别的正确率 %

| 情感类别 | 高兴 | 愤怒 | 惊奇 | 悲伤 | 恐惧 |
|---|---|---|---|---|---|
| 高兴 | 92 | 1 | 6 | 1 | 0 |
| 愤怒 | 1 | 88 | 2 | 7 | 2 |
| 惊奇 | 4 | 0 | 90 | 4 | 2 |
| 悲伤 | 5 | 8 | 2 | 85 | 0 |
| 恐惧 | 1 | 3 | 2 | 6 | 88 |

表 4 文中方法进行识别的情感识别正确率 %

| 情感类别 | 高兴 | 愤怒 | 惊奇 | 悲伤 | 恐惧 |
|---|---|---|---|---|---|
| 高兴 | 96 | 2 | 2 | 0 | 0 |
| 愤怒 | 3 | 89 | 1 | 7 | 0 |
| 惊奇 | 0 | 1 | 93 | 6 | 0 |
| 悲伤 | 2 | 10 | 1 | 87 | 0 |
| 恐惧 | 0 | 5 | 1 | 3 | 91 |

从表 3 可以看出，通过简单组合语音信号和面部表情信号进行识别的正确率有所提高，但是提高的并不太明显，因此不同模式信息的融合是多模式情感识别研究的瓶颈问题，它直接关系到情感识别的准确性。从表 4 可以看出，情感识别的平均正确率达到了 91.2%，因此文中方法充分发挥了决策层融合与特征层融合的优点，使整个融合过程更加接近人类情感识别，从而提高了情感识别的平均正确率。

## 6 结束语

文中充分发挥了决策层融合与特征层融合的优点，提出了一种新型的多模式情感识别算法，从而提高了情感识别的正确率。但是文中只是针对特定文本的语音情感进行识别，要达到实用的程度尚需一定距离，所以非特定文本的语音情感识别将是下一步的研究方向。

**参考文献：**


[1] 余伶俐，蔡自兴，陈明义. 语音信号的情感特征分析与识别研究综述[J]. 电路与系统学报，2007，12(4)：76-84.

[2] 颜永红，周 瑜，孙艳庆，等. 一种用于语音情感识别的语音情感特征提取方法：中国，2010102729713[P]. 2010.

[3] Mao X，Chen L J. Speech emotion recognition based on parametric filter and fractal dimension[J]. IEICE Trans on Information and Systems，2010，93(8)：2324-2326.

[4] 邹采荣，赵 力. 一种基于改进模糊矢量量化的语音情感识别方法：中国，2008101228062[P]. 2008.

[5] Attabi Y，Dumouchel P. Anchor models for emotion recognition from speech[J]. IEEE Trans on Affective Computing，2013，4(3)：280-290.

[6] Zheng W M，Xin M H，Wang X L，et al. A novel speech emotion recognition method via incomplete sparse least square regression[J]. IEEE Signal Processing Letters，2014，21(5)：569-572.

[7] Mao Q R，Dong M，Huang Z W，et al. Learning salient features for speech emotion recognition using convolutional neural networks[J]. IEEE Trans on Multimedia，2014，16(8)：2203-2213.

[8] Ekman P，Friesen W. Facial action coding system：a technique for the measurement of facial movement[M]. Palo Alto：Consulting Psychologists Press，1978.

[9] 梁路宏，艾海舟，徐光祐，等. 人脸检测研究综述[J]. 计算机学报，2002，25(5)：449-458.

[10] Rahulamathavan Y，Phan R C W，Chambers J A，et al. Facial expression recognition in the encrypted domain based on local fisherdiscriminant analysis[J]. IEEE Trans on Affective Computing，2013，4(1)：83-92.

[11] 文 沁，汪增福. 基于三维数据的人脸表情识别[J]. 计算机仿真，2005，22(7)：99-103.

[12] Zheng W M. Multi-view facial expression recognition based on group sparse reduced-rank regression[J]. IEEE Trans on Affective Computing，2014，5(1)：71-85.

[13] Petrantonakis P C，Hadjileontiadis L J. Emotion recognition from EEG using higher order crossings[J]. IEEE Trans on Information Technology in Biomedicine，2010，14(2)：186-197.

[14] 林时来，刘光远，张慧玲. 蚁群算法在呼吸信号情感识别中的应用研究[J]. 计算机工程与应用，2011，47(2)：169-172.

[15] Zacharatos H，Gatzoulis C，Chrysanthou Y L. Automatic emotion recognition based on body movement analysis：a survey[J]. IEEE Computer Graphics and Applications，2014，34(6)：35-45.

[16] Zeng Z，Pantic M，Roisman G I，et al. A survey of affect recognition methods：audio，visual，and spontaneous expressions[J]. IEEE Trans on Pattern Analysis and Machine Intelligence，2009，31(1)：39-58.

[17] Kim J，Andre E. Emotion recognition based on physiological changes in music listening[J]. IEEE Trans on Pattern Analysis and Machine Intelligence，2008，30(12)：2067-2083.


(下转第 35 页)

次数最多的前几项为:http://www. arris. com 有 204 次,http://mikrotik. com 有 72 次,http://www. mikrotik. com/有 8 次。这些都是相关厂商的页面。

## 5　结束语

文中提出一种基于网络协议报文和 Web 页面特征在互联网中发现物理设备的方法,并通过多种手段扩充了设备的信息,对设备进行了物理、信息和社会多域描述。实验还存在一些不足之处,比如在 Web 页面分析中,某些页面需要根据脚本或者 location 字段进行二次跳转,对这些页面进一步分析会扩充发现的物理设备的数目。通过该文,可以认识到互联中存在很多没有高级安全防护措施的设备,主要是小型化家用网络设备,这其中潜在着较大的网络安全隐患。

**参考文献:**


[1] 于海宁,张宏莉,方滨兴,等. 物联网中物理实体搜索服务的研究[J]. 电信科学,2012,28(10):111-119.

[2] 华为技术有限公司. 安全预警-涉及华为家庭网关产品的多个 RomPager 漏洞[EB/OL]. 2014-12-19. http://www.huawei. com/cn/security/psirt/security - bulletins/security - advisories/hw-407667. htm.

[3] 中兴通讯公司. 中兴通讯家庭网关产品受多个 RomPager 漏洞影响[EB/OL]. 2015-01-09. http://support. zte. com. cn/support/news/LoopholeInfoDetail. aspx? newsId = 1006322.

[4] 红黑联盟. 多个 TP-Link 路由器 RomPager 拒绝服务漏洞[EB/OL]. 2014-06-22. http://www. 2cto. com/Article/201406/310905. html.

[5] 张　庆,宋　芬,沈国良. 网络设备安全措施分析与研究[J]. 网络安全技术与应用,2008(8):33-34.

[6] 武传坤. 物联网安全关键技术与挑战[J]. 密码学报,2015(1):40-53.

[7] 张友春,魏　强,刘增良,等. 信息系统漏洞挖掘技术体系研究[J]. 通信学报,2011,32(2):42-47.

[8] Wang H, Tan C C, Li Q. Snoogle: a search engine for pervasive environments[J]. IEEE Transactions on Parallel and Distributed Systems, 2010, 21(8): 1188-1202.

[9] Tan C C, Sheng B, Wang H, et al. Microsearch: when search engines meet small devices[C]//Proceedings of the 6th international conference on pervasive computing. Sydney, Australia: [s. n.], 2008: 93-110.

[10] Yap K K, Srinivasan V, Motani M. MAX: human - centric search of the physical world[C]//Proceedings of 3rd conference on embedded networked sensor systems. San Diego: [s. n.], 2005: 166-179.

[11] Frank C, Bolliger P, Mattern F, et al. The sensor internet at work: locating everyday items using mobile phones[J]. Pervasive and Mobile Computing, 2008, 4(3): 421-447.

[12] Ostermaier B, Romer K, Mattern F, et al. A real-time search engine for the web of things[C]//Proceedings of internet of things. Tokyo, Japan: [s. n.], 2010: 1-8.

[13] Krämer B J. Evolution of cyber-physical systems: a brief review[M]. New York: Springer, 2014.

[14] Horváth I. What the design theory of social-cyber-physical systems must describe, explain and predict[M]//An anthology of theories and models of design. London: Springer, 2014: 99-120.

[15] Sheth A, Anantharam P, Henson C. Physical-cyber-social computing: an early 21st century approach[J]. IEEE Intelligent Systems, 2013, 28(1): 78-82.


(上接第 30 页)


[18] 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.

[19] Busso C, Deng Z, Yildirim S, et al. Analysis of emotion recognition using facial expressions, speech and multimodal information[C]//Proc of the sixth international conference on multimodal interfaces. USA: IEEE, 2004: 205-211.

[20] Hoch S, Althoff F, Mcglaun G, et al. Bimodal fusion of emotional data in an automotive environment[C]//Proc of IEEE international conference on acoustics, speech, and signal processing. USA: IEEE, 2005: 1085-1088.

[21] Sayedelahl A, Araujo R, Kamel M S. Audio-visual feature-decision level fusion for spontaneous emotion estimation in speech conversations[C]//Proc of 2013 IEEE international conference on multimedia and expo workshops. USA: IEEE, 2013: 1-6.

[22] Tato R, Santos R, Kompe R, et al. Emotion space improves emotion recognition[C]//Proceedings of the 2002 international conference on speech and language processing. USA: IEEE, 2002: 2029-2032.