软件学报 ISSN 1000-9825, CODEN RUXUEW
*Journal of Software*,2016,27(Suppl.(2)):130−136

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

# 自闭症干预中无监督自编码的语音情感识别*

葛 磊, 强 彦, 赵涓涓

(太原理工大学 计算机科学与技术学院,山西 太原 030024)

通讯作者: 强彦, E-mail: 27420265@qq.com

**摘 要**: 语音情感识别是人机交互中重要的研究内容,儿童自闭症干预治疗中的语音情感识别系统有助于自闭症儿童的康复,但是由于目前语音信号中的情感特征多而杂,特征提取本身就是一项具有挑战性的工作,这样不利于整个系统的识别性能.针对这一问题,提出了一种语音情感特征提取算法,利用无监督自编码网络自动学习语音信号中的情感特征,通过构建一个 3 层的自编码网络提取语音情感特征,把多层编码网络学习完的高层特征作为极限学习机分类器的输入进行分类,其识别率为 84.14%,比传统的基于提取人为定义特征的识别方法有所提高.

**关键词**: 语音情感识别;极限学习机;无监督自编码;人机交互

中文引用格式: 葛磊,强彦,赵涓涓.自闭症干预中无监督自编码的语音情感识别.软件学报,2016,27(Suppl.(2)):130−136. http://www.jos.org.cn/1000-9825/16028.htm

英文引用格式: Ge L, Qiang Y, Zhao JJ. A speech emotion recognition based on unsupervised autoencoder in the intervention of autism. Ruan Jian Xue Bao/Journal of Software, 2016,27(Suppl.(2)):130−136 (in Chinese). http://www.jos.org.cn/1000- 9825/16028.htm

## A Speech Emotion Recognition Based on Unsupervised Autoencoder in the Intervention of Autism

GE Lei, QIANG Yan, ZHAO Juan-Juan

(College of Computer Science and Technology, Taiyuan University of Technology, Taiyuan 030024, China)

**Abstract**: Speech emotion recognition is an important research area in human computer interaction (HCI). The speech emotion recognition system used in the intervention therapy for autistic children is helpful for their rehabilitation. However, the variation and complexity in speech emotion features, the extraction of which itself is a challenging task, will contribute to the difficulty to improve the recognition performance of the whole system. In view of this problem, this paper proposes a new method of speech emotion feature extraction with unsupervised auto-encoding network to learn emotional feature in speech signal automatically. By constructing a 3-layer auto-encoding network to extract the speech emotional feature, the high level feature is used as the input of extreme learning machine classifier to make final recognition. The speech emotion recognition rate of the system reaches 84.14%, which is higher than the traditional method based on human defined feature extraction.

**Key words**: speech emotion recognition; extreme learning machine; unsupervised autoencoder; human computer interaction

儿童自闭症对儿童成长和发展有严重的影响,自闭症儿童在临床上主要表现为不同程度的言语发育障碍、人际交往障碍、兴趣狭窄和行为方式刻板[1].研究发现,自 1943 年美国医生 Kanner 曾首次提出自闭症概念,自闭症的患病率一直处于上升趋势[2].目前,自闭症的治疗主要依靠人为干预,包括体育舞蹈干预、康复花园园艺疗法、自然教法、感觉统合游戏、箱庭疗法、音乐疗法等[3,4].

* 基金项目: 国家自然科学基金(61540007, 61373100); 北京航空航天大学虚拟现实技术与系统国家重点实验室开放基金(BUAA-VR-15KF02, BUAA-VR-16KF13)

Foundation item: National Natural Science Foundation of China (61540007, 61373100); Virtual Reality Technology and National Key Laboratory of Open Foundation (Beihang University)(BUAA-VR-15KF02, BUAA-VR-16KF13)

收稿时间: 2016-05-01; 采用时间: 2016-11-21

目前人工干预治疗面临的问题之一在于自闭症患者群体庞大,所以,在我国人为干预自闭症是一项非常艰巨的任务[5].针对上述问题,我们可以通过借助计算机干预,以互动方式帮助自闭症儿童获得自我认识,走出自我封闭世界.计算机在与自闭症患者的互动的过程中需要及时捕获他们语言表达中情感的变化,及时正确地引导和稳定自闭症患者的情绪变化.这就需要计算机实时地对语音情感进行识别.目前,语音情感识别技术往往是基于人为定义特征的提取和选择,也就是说当输入新的语音信号时,往往要先对语音信号进行特征提取、特征选择,然后对特征子集进行识别判断.由于不同情感之间语音信号的差异性,有些特征会有缺失的现象,所以要准确地提取全部的特征是非常困难的事,因此很难对不同情感都能够准确地识别.

## 1 相关工作

近年来众多学者在语音情感识别方面做了大量研究.金琴等人[6]提出基于声学特征的语言情感识别,在不同的语音信号数据集中采用不同的声学特征进行实验,获得较好的识别效果.Tao 等人[7]提出了一种新的语图谱特征提取算法,并将其用在语音情感识别中,该算法所提特征的识别率较早期声学特征至少提高 5%.毛启容等人[8]提出基于情感上下文的情感推理算法.该算法首先利用传统语音情感特征和上下文语音情感特征分别识别待分析情感语句的情感状态,与采用传统语音情感特征的方法相比,文中提出方法平均识别率提高 12.17%.何凌等人[9]提出了一种基于声门信号特征参数及高斯混合模型的情感识别算法,该算法能够有效地识别各类情感状态,其情感判别正确率优于传统的基音频率及共振峰特征参数.叶吉祥等人[10]利用希尔伯特黄变换(HHT)对情感语音进行处理,得到情感语音的希尔伯特边际能量谱,通过分析能量谱获取新的情感特征,实验结果表明,通过该方法提取的新的情感特征具有较好的识别效果.Wu 等人[11]融合调制声谱特征和韵律特证用于语音情感识别,实验表明同时使用这两种特征时,系统的整体识别率为 91.6%.Liang 等人[12]从输入语音中提取声学和韵律信息并且结合语义标签进行情感分类,整体识别率高达 83.55%.

尽管现在不同形式的特征被用在语音情感识别系统上,但是研究人员还无法确定最适合于语音情感识别的特征子集.而且,由于语音情感和其他因素紧密相关,这些人为定义的特征能否充分有效地表征语音的情感内容也无法确定.目前,研究人员更倾向于融合多种特征进行识别,但是提取大量语音特征本身就是一项具有挑战性的工作.针对这一问题,本文提出一种基于极限学习机多层无监督编码的语音情感识别方法,该方法通过对语音信号自编码直接得到高层情感特征,避免繁琐的手动提取特征的过程.实验结果表明,本文的方法对于语音情感识别具有较好的分类性能.

## 2 方法介绍

本文提出的方法先采用基于极限学习机的无监督多层自编码,通过学习一种多层非线性网络结构,发现语音信号数据间的相关性,实现数据特征的自动提取,然后将高层的特征作为情感识别分类器的输入,进行最后的决策.本文方法的识别过程如图 1 所示.

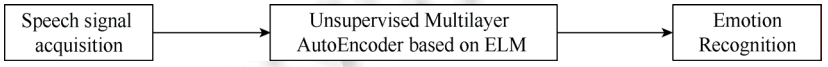



Fig.1 Speech emotion recognition process

图 1 语音情感识别过程图

### 2.1 基于极限学习机自编码

自编码神经网络是一种对输入数据 $X\in R^d$ 用某种特征表示 $H\in R^L$ 编码来重构原始输入 $X$ 的神经网络[13].也就是说,自动编码的输出理想上要等同于输入:$Y\approx X$.传统上自动编码网络通常使用反向传播的算法进行训练.图 2 显示的就是简单的单隐层的自编码网络结构图,其中隐含层神经元的个数可以大于或者小于输入数据的维度,它们都可以发现输入数据之间紧凑或者稀疏的结构.如果隐含层是线性的并且使用最小均方误差进行训练,这样自动编码的效果和 PCA 相似,隐含层通过学习去显示输入数据最主要的部分.如果隐含层是非线性的,自动编

码能够捕获输入数据分布的多方面特征[14].

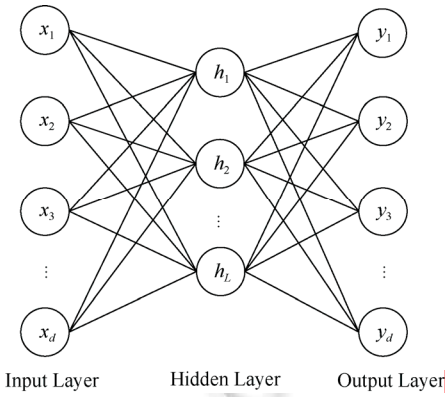



Fig.2 Network structure of one hidden layer autoencoder

图 2 单隐层自编码网络结构

与传统自编码中使用反向传播的进行训练来确定参数不同,本文中使用极限学习机(extreme learning machine,简称 ELM)作为训练算法进行无监督自编码.因为极限学习机的网络框架和单隐含层自编码框架完全一样,所以极限学习机算法可以通过设置期望输出等于输入来训练自编码网络.在本文使用的 ELM 自编码中,输入数据 $X\in R^d$ 首先被映射到 ELM 特征空间 $H\in R^L$,然后特征空间 H 通过重构矩阵 $\beta\in R^{L,d}$ 重构成原始输入数据 $X$,即 $X=H\beta$.正如图 3 所示,重构矩阵 $\beta$ 可能包含输入数据的高层信息,因此被保留下来作为重构 ELM 的输入权重.正如文献[15,16]中阐述的,输出权重 $\beta$ 是通过奇异值来表现输入数据,所以可以用来对数据进行无监督的训练,而且在解决大型非结构化数据时会取得很好的泛化性能.

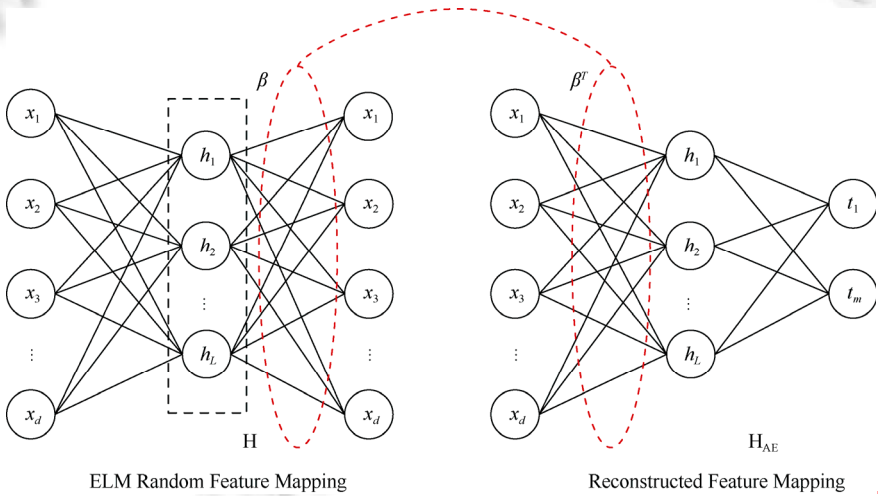



Fig.3 In ELM autoencoder, ELM is used as the training algorithm for the network, the transpose of output weight is used as the input weight of a normal ELM

图 3 在 ELM 自编码中,ELM 用来作为网络的训练算法,输出权重向量的转置作为特征重构的输入权重

在含有 $L$ 个隐含层单元的 ELM 自编码中,$N$ 个随机样本,其中 $x_i=[x_{i1},x_{i2},\ldots,x_{id}]^T\in R^d$,$t_i=[t_{i1},t_{i2},\ldots,t_{im}]^T\in R^m$,在输出层通过公式(1)被重构出来,

$$\sum_{i=1}^{N}\beta_i \cdot g(a_i \cdot x_j + b_i) = x_j, j \in [1, N] \tag{1}$$

其中,$a_i=[a_{i1},a_{i2},\ldots,a_{id}]^T$ 是随机产生的输入权重,$\beta_i=[\beta_{i1},\beta_{i2},\ldots,\beta_{id}]^T$ 是重构矩阵,$g(\cdot)$是隐含层的激励函数.式(1)可以写成矩阵的形式:

$$H\beta = X \tag{2}$$

重构矩阵 $\beta$ 可以简单地由广义逆求出 $\beta=H^{+}X$.当我们使用 ELM 编码的重构矩阵的转置作为另一个 ELM 的输入权重时,隐含层的输出 $H_{AE}$ 可以通过公式(3)获得:

$$H_{AE}=\begin{bmatrix} g(\beta_1^T \cdot X_1 + b_1) & \cdots & g(\beta_l^T \cdot X_1 + b_L) \\ \vdots & \ddots & \vdots \\ g(\beta_1^T \cdot X_N + b_1) & \cdots & g(\beta_l^T \cdot X_N + b_L) \end{bmatrix}_{N\times L} \tag{3}$$

其中,$\beta=H^{+}X$.

### 2.2 栈式自编码算法

栈式自编码网络是由多层自编码网络构成的神经网络,本文中基于 ELM 的多层自编码网络如图 4 所示.

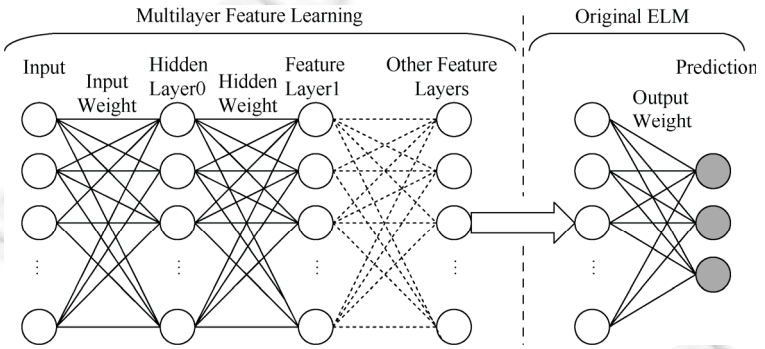



Fig.4 Multilay autoencoder network based on ELM

图 4 基于 ELM 多层自编码网络图

在进行无监督特征学习之前,先将输入的语音信号转换到 ELM 特征空间中,这样有助于挖掘训练样本中隐藏的特征信息.然后,通过 $N$ 层无监督学习最终获得高层稀疏特征.每个隐含层输出的数学表达式为

$$H_i = g(H_{i-1} \cdot \beta) \tag{4}$$

其中,$H_i$ 是第 $i$ 层的输出($i\in[1,K]$),$g(\cdot)$表示隐含层激励函数,$\beta$ 代表输出权重.这里多层编码框架中每个隐含层都是单独的特征分离器.随着层数地增加,产生的特征会变得更加紧凑.一旦前隐含层的特征被提取,当前隐含层的权重 $\beta$ 就可以通过计算隐含层的奇异值确定,不再需要微调.同时,为了获得更加紧凑的高层特征,我们将稀疏约束加入到编码中,优化模型见式(5).

$$O_\beta = \arg\min_{\beta}\{\|H\beta - X\|^2 + \|\beta\|\} \tag{5}$$

从图 4 中可以看出,在进行无监督的分层训练之后,第 $K$ 层的输出则是从输入数据中提取的最高层特征,然后将它作为有监督分类的输入得到整个网络的最终结果.

### 2.3 极限学习机分类

极限学习机 ELM 是一种单隐层前馈神经网络的学习算法[17],与传统机器学习的算法相比,ELM 网络模型具有学习速度快和泛化性能强等优势,且 ELM 模型可用来解决实际中的分类问题.因此本文使用基于 ELM 的神经网络情感语音识别,其中激励函数选择 sigmoid 函数,并且根据 Huang 等人的研究,随着隐含层节点个数 $L$ 的增加,网络模型的泛化性能越来越好,本文在实验的基础上选择 $L$=400,这样 ELM 既可以在很短的时间内完成学习过程又可以保证模型具有较好的识别效果.

## 3 实验结果

### 3.1 实验数据

本实验采用的是语音数据库是由山西省某医院所录制的自闭症儿童情感语音库,实验针对 5 类不同情感(高兴、悲哀、生气、惊吓、中性)进行识别,每种情感包含 500 个短句,对于每种情感随机选取 40%进行训练,60%进行测试.为了验证系统的稳定性和鲁棒性,按照以上规则随机选取训练集和测试集 5 次,在不同训练/测试集上分别进行实验,最终的结果取 5 次实验的平均值.

### 3.2 基于无监督自编码参数设置

本实验在构造整个自编码网络的时候需要指定两个参数,即自编码网络隐含层节点的个数和层数.接下来将从这两方面分析,为探究隐藏层节点数对识别效果的影响,我们首先构造含有一个隐含层的编码网络,令其隐含层节点的个数为 50~1 000,间隔为 50,观察整个网络对语音情感的识别效果,实验结果如图 5 所示.

从图 5 中可以看出,随着隐含层节点个数的增长,识别准确率呈现上升的趋势,但是,当隐含层节点的个数大于 500 以后,曲线趋于平稳状态.

为了探究多层自编码网络的隐藏层层数对识别效果的影响,我们在实验中逐步增加编码层数进行测试,并且每次都选用不同数据测试 5 次.由图 6 可以看出随着编码层数的增加,识别准确率有所提升,当编码层数为 3 时,识别准确率达到最高,但是以后的识别率有所下降.

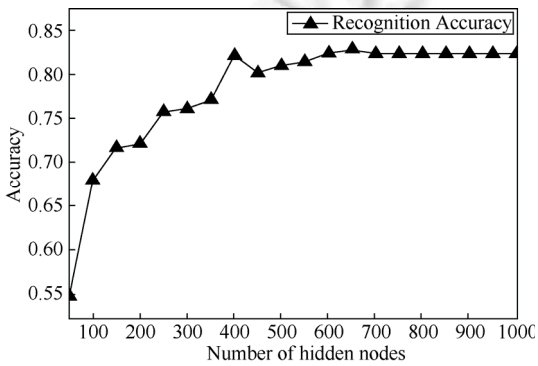



Fig.5 The recognition accuracy of multilay autoencoder versus different numbers of hidden nodes

图 5 隐含层节点个数对多层自编码网络识别性能的影响

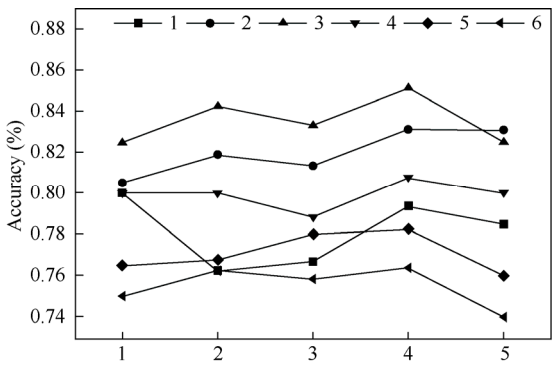



Fig.6 The recognition accuracy of multilay autoencoder versus autoencoder layers

图 6 编码层数对识别准确率的影响

### 3.3 不同方法实验分析

为了进一步验证所提方法的有效性,我们分别计算该方法对于 5 种不同情感(中性、悲哀、生气、惊吓、高兴)的识别性能,见表 1.

**Table 1** Emotion recognition accuracy of five emotion

**表 1** 5 种情感正确识别率

| 真实情感 | 识别情感类别 | | | | |
|---|---|---|---|---|---|
| | 中性 | 悲哀 | 生气 | 惊吓 | 高兴 |
| 中性 | **0.84** | 0.109 | 0.009 | 0.021 | 0.021 |
| 悲哀 | 0.005 | **0.803** | 0.121 | 0.041 | 0.03 |
| 生气 | 0.013 | 0.065 | **0.883** | 0.028 | 0.011 |
| 惊吓 | 0.124 | 0.012 | 0.027 | **0.829** | 0.008 |
| 高兴 | 0.011 | 0.024 | 0.008 | 0.105 | **0.852** |

由表 1 可以看出,高兴和生气的情感识别率较高,达到 85.2%和 88.3%,生气和悲哀之间比较容易混淆,因为这两种情感在发音时有许多相似的特征.本文方法的总体识别率为 84.14%.

图 7 中显示的是利用多层自编码网络提取语音情感特征和文献[6−10]中直接提取不同传统情感特征参数这两种方法相比,可以看出本文方法的识别准确率比文献[10]中方法的准确率要高 1.47%.

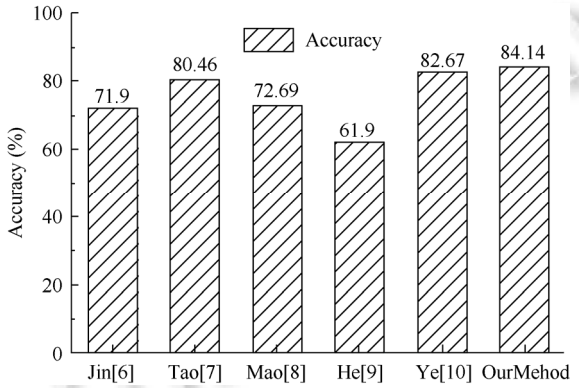



Fig.7 Comparison of recognition accuracy between different features

图 7 不同特征准确率比较

## 4 结 论

本文提出了利用多层自编码神经网络来自动提取情感语音信号中的情感特征.将多层自编码网络和极限学习机结合,提出了一种基于极限学习机的多层无监督自编码网络模型.该模型在实际训练过程中时间复杂度小,最后的识别效果比人为的选取传统情感特征参数直接提取的结果要高,明显地提高了情感语音的识别率.未来工作重点是研究大数据下的语音情感识别,通过训练海量的语音数据来有效地提高语音情感的识别率.

**References**:

[1] Muller CL, Anacker AMJ, Veenstra-VanderWeele J. The serotonin system in autism spectrum disorder: From biomarker to animal models. Neuroscience, 2015.

[2] Lovell B, Wetherell MA. The psychophysiological impact of childhood autism spectrum disorder on siblings. Research in Developmental Disabilities, 2016,49−50:226−234.

[3] Szabó MK. Patterns of play activities in autism and typical development. A case study. Procedia-Social and Behavioral Sciences, 2014,140:630−637.

[4] Rendall AR. Learning delays in a mouse model of autism spectrum disorder. Learning, 2015.

[5] Muotri AR. The human model: Changing focus on autism research. Biological Psychiatry, 2015.

[6] Jin Q, Chen SZ, Li XR, Yang G, Xu JP. Speech emotion recognition based on acoustic feature. Computer Science, 2015,42(9): 24−28 (in Chinese with Abstract English).

[7] Tao HW, Zha C, Liang RY, Zhang XR, Zhao L, Wang QY. Spectrogram feature extraction algorithm for speech emotion recognition. Journal of Southeast University (Natural Science Edition), 2015,20(9):1817−821 (in Chinese with Abstract English).

[8] Mao QR, Bai LJ, Wang L, Zhan YZ. Emotion reasoning algorithm based on emotional context of speech. Pattern Recognition and Artificial Intelligence, 2014,9:826−834 (in Chinese with Abstract English).

[9] He L, Huang H, Liu XH. Speech emotion detection based on glottal signal features. Computer Engineering and Design, 2013,34(6): 2147−2151 (in Chinese with Abstract English).

[10] Ye JX, Hu HX. Application of Hilbert marginal energy spectrum in speech emotion recognition. Computer Engineering and Applications, 2014,50(7):203−207 (in Chinese with Abstract English).

[11] Wu SQ, Falk TH, Chan WY. Automatic speech emotion recognition using modulation spectral features. Speech Communication,

2011,53(5):768−785.

[12] Wu CH, Liang WB. Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. IEEE Trans. on Affective Computing, 2011,2(1):10−21.

[13] Bengio Y. Learning deep architectures for AI. Foundations and Trends in Machine Learning, 2009,2(1):1−127.

[14] Hinton GE, Osindero S, Teh YW. A fast learning algorithm for deep belief nets. Neural Computation, 2006,18(7):1527−1554.

[15] Tang JX, Deng CW, Huang GB. Extreme learning machine for multilayer perceptron. IEEE Trans. on Neural Networks and Learning Systems, 2015.

[16] Vincent P, Larochelle H, Bengio Y, Manzagol PA. Extracting and composing robust features with denoising autoencoders. In: Proc. of the 25th Int'l Conf. on Machine Learning. ACM, 2008. 1096−1103.

[17] Daliri MR. Combining extreme learning machines using support vector machines for breast tissue classification. Computer Methods in Biomechanics and Biomedical Engineering, 2015,18(2):185−191.

**附中文参考文献**:

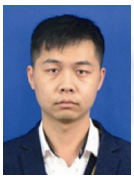
[6] 金琴,陈师哲,李锡荣,杨刚,许洁萍.基于声学特征的语言情感识别.计算机科学,2015,42(9):24−28.

[7] 陶华伟,查诚,梁瑞宇,张昕然,赵力,王青云.向语音情感识别的语谱图特征提取算法.东南大学学报(自然科学版),2015,45(5):817−821.

[8] 毛启容,白李娟,王丽,詹永照.基于情感上下文的语音情感推理算法.模式识别与人工智能,2014,27(9):826−834.

[9] 何凌,黄华,刘肖珩.基于声门特征参数的语音情感识别算法研究.计算机工程与设计,2013,34(6):2147−2151.

[10] 叶吉祥,胡海翔.Hilbert 边际能量谱在语音情感识别中的应用.计算机工程与应用,2014,50(7):203−207.



**葛磊**(1990－),男,江苏南通人,硕士,主要研究领域为医学图像处理,模式识别.

**强彦**(1969－),男,博士,教授,博士生导师,CCF 会员,主要研究领域为图像处理,物联网技术,云计算技术.

**赵涓涓**(1975－),女,博士,教授,博士生导师,CCF 会员,主要研究领域为医学图像处理,物联网技术.