Journal of Southeast University (English Edition) Vol. 29, No. 4, pp. 366 –371 Dec. 2013 ISSN 1003—7985

Annoyance-type speech emotion detection in working environment

Wang Qingyun[1,2] Zhao Li[1] Liang Ruiyu[1] Zhang Xiaodan[1]

([1]School of Information Science and Engineering, Southeast University, Nanjing 210096, China)
([2]School of Communication Engineering, Nanjing Institute of Technology, Nanjing 211167, China)

Abstract: In order to recognize people's annoyance emotions in the working environment and evaluate emotional well-being, emotional speech in a work environment is induced to obtain adequate samples of emotional speech, and a Mandarin database with two thousands samples is built. In searching for annoyance-type emotion features, the prosodic feature and the voice quality feature parameters of the emotional statements are extracted first. Then an improved back propagation (BP) neural network based on the shuffled frog leaping algorithm (SFLA) is proposed to recognize the emotion. The recognition capability of the BP, radical basis function (RBF) and the SFLA neural networks are compared experimentally. The results show that the recognition ratio of the SFLA neural network is 4. 7% better than that of the BP neural network and 4. 3% better than that of the RBF neural network. The experimental results demonstrate that the random initial data trained by the SFLA can optimize the connection weights and thresholds of the neural network, speed up the convergence and improve the recognition rate.

Key words: speech emotion detection; annoyance type; sentence length; shuffled frog leaping algorithm
doi: 10. 3969/j. issn. 1003 –7985. 2013. 04. 003

Emotions in vocal communication between humans are very important for understanding a speaker's emotion, mood and attitude[1-3]. Unlike the linguistic information contained in words and sentences, the speech emotions carry the affective information even without the notice of the speaker. They are naturally expressed and uneasy to disguise or control. Another reason to use speech emotion to evaluate people's emotional well-being is that speaking is a very natural way of interaction. Speech-enabled interaction becomes increasingly important in various work situations and in daily life. Recently, real-world applications of the speech emotion recognition system have been studied. Jones et al. [4] studied an in-car emotion recognition system to recognize driver emotional states and respond appropriately in speech interaction between the car and the driver. In their study, boredom, happiness, surprise, anger and sadness were treated for the in-car system. Clavel et al. [5] studied the fear-type extreme emotions for audio-based surveillance systems. Fear-type speech emotions were treated as a sign of threat situations. Ang et al. [6] investigated the use of prosody for detection of frustration and annoyance in natural human-computer dialog. These studies bring the speech emotion research closer to the applications in the real world.

In this paper, research on the annoyance-type speech emotion mainly aims at providing objective cues to evaluate people's working states in long-time repeated tasks in special missions, such as long-term manned space missions. Annoyance-type negative emotions easily occur in these situations, and usually become a threat to the stability of people's performance at work. It is very important to detect the potential threat in work situations regarding people's emotional well-being. Yet, to date, little research has been done on the emotional threats in long-term special missions. Unobvious emotions in working dialogues are difficult to perceive and categorize, and this is even true to human listeners. Annoyance-type emotion manifestations in special working situations are expected to present themselves quite differently, compared with active speech emotions or natural emotions in daily life. Due to these reasons, research on the annoyance-type emotion in speech is meaningful not only in practical situations but also in literature. Previous studies on basic emotion categories (e. g. joy, anger, surprise, sadness, and disgust) [7-8] are not enough to meet the needs of the above mentioned application. For real-world application, we collect emotional speech data in long-term repeated tasks with an eliciting method and categorize the data into different practical emotion classes according to a listening perception test and a self-evaluation.

In order to recognize annoyance-type emotion in the above speech data, a new swarm intelligence algorithm, shuffled frog leaping algorithm (SFLA) neural network[9-10], is proposed in this paper. Integrating the advanced genetic algorithm and particle swarm optimiza-

Received 2013-08-10.
Biography: Wang Qingyun(1972—), female, doctor, associate professor, wangqingyun@ vip. 163. com.
Foundation items: The National Natural Science Foundation of China (No. 61375028, 61301219), China Postdoctoral Science Foundation (No. 2012M520973), the Scientific Research Funds of Nanjing Institute of Technology (No. ZKJ201202).
Citation: Wang Qingyun, Zhao Li, Liang Ruiyu, et al. Annoyance-type speech emotion detection in working environment[J]. Journal of Southeast University (English Edition), 2013, 29(4): 366 – 371. [doi: 10. 3969/j. issn. 1003 –7985. 2013. 04. 003]

tion, the SFLA has the characteristics of fresh concept, small population, fast convergence, strong global optimization capability and easy implementation. Therefore, compared with the traditional evolutionary algorithms, the SFLA can more easily obtain the balance between searching and optimization, which overcomes the shortage of slow learning speed and low recognition accuracy of the traditional neural network models. Compared with the experimental results of the BP and the RBF neural networks, the SFLA neural network for speech emotion recognition can achieve significant improvement in recognition performance. The research in this paper has practical significance. For example, in the stress environment of manned spaceflight, it is easy to generate the emotions of irritability, anxiety and tensity which will be reflected in the speech[11-13]. So the speech signal can be used to monitor the emotional states of astronauts and make psychological assessments and psychological intervention.

1 Annoyance-Type Emotion Detection

1.1 Speech emotion data collection

In order to study the annoyance-type speech emotion in special work environments, 1 344 utterances of annoyance-type emotion and 693 utterances of neutral emotion were induced and collected in an eliciting experiment. Ten native speakers, five male and five female, participated in the experiment and Mandarin speech data was collected. Subjects were given a series of simple mathematical problems to solve. Subjects were required to settle the math problem sets repeatedly and report the questions and answers vocally, and at the same time subjects were required to wear earphones and noise was used to induce annoyance. The answers and the typical communication phrases were collected as the emotion speech materials. The participating subject was alone in a separate room to better induce the negative emotion.

As the experiment went on, the subjects tended to get impatient and annoyed, and the accuracy of their vocal reports of the math problems dropped. Several abnormal vocal cues could be perceived indicating a sign of working ability shift influenced by negative emotions. Longer and unusual pauses, slip of tong and speech breath change were usually perceived in speech when annoyance was induced. Such emotionally colored behaviors in speech are common in real-life dialogue; however, these subtle emotions have been rarely studied for emotion recognition.

Near the end of the experiment the subjects tended to quit the on-going task due to the long-time boring mathematic works and the noise stimulations. The speech materials recorded near the end of the experiment are almost all self-evaluated as annoyance by the subjects. The typical emotions in speech are labeled by auditory perception experiments. Since our emotion recognition system is designed to serve the evaluation of the threat caused by negative emotions, we make a comparison with the listening perceived evaluation and the self-evaluation to assess the emotions in the elicited speech.

The speech emotion data are classified into “neutral”, “annoyance-type” and “other emotions” classes. Negative emotion especially like annoyance is considered as a threat on the long-term special missions. In our experiment we are able to assess this threat roughly by the mistakes made by the subjects in the vocal reports. During the experiment, subjects are required to solve math problems, report vocally and read materials in a long-term noise environment. Subjects show different abilities to accomplish these tasks by examining the correctness of their vocal reports. One can notice that continually making mistakes is a sign of drop in work ability. Abnormal vocal cues can also reflect the influence of negative emotion, such as long pause during the vocal report and skipping a question. Neutral emotional speech is considered as a non-threat situation. The results of threat analyses are shown in Fig. 1.

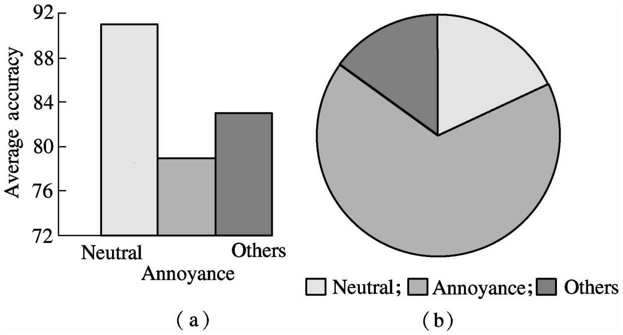

Fig. 1 Threat analysis of annoyance-type emotion. (a) Accuracy of each emotion type; (b) Mistake distribution over emotion types

1.2 Feature analysis

Speech data throughout the emotion datasets are categorized into three sentence length classes. Within each type of emotion dataset speech, utterances are evenly distributed over “short”, “middle” and “long” sentence length classes. And feature analysis is performed on these subclasses to show how time duration can influence the emotion features used in current recognition systems relationships.

The relationships between emotion and various acoustic features are studied. Prosodic features are reported mainly related to arousal dimension in the dimensional emotion model. Similarly, voice quality features are reported related to valence dimension[14-15]. Both temporal features and statistical features can be used for speech emotion recognition. However, statistical features are considered less dependent on phoneme information. In this paper, statistical features including maximum, minimum, mean, standard deviation and range are adopted to construct the emotional feature set. Acoustic features, as low-level description, including prosodic and voice quality features are extracted from the speech utterances. Totally, 372 features are generated as listed in Tab. 1.

Tab. 1 Feature extraction

Feature index	Feature description
1 to 10	Max, min, mean, std, range of pitch and diff pitch
10 to 11	Jitter, shimmer
12 to 52	Max, min, mean, std, range of F_1 to F_4, and diff of F_1 to F_4
52 to 62	Max, min, mean, std, range of intensity, and diff intensity
62 to 192	Max, min, mean, std, range of MFCC_1 to MFCC_{13}, and diff of MFCC_1 to MFCC_{13}
192 to 372	Max, min, mean, std, range of BBE_1 to BBE_{18}, and diff of BBE_1 to BBE_{18}

Notes: diff stands for differentiation; BBE stands for bark band energy; F_1 to F_4 stand for the first to the fourth formants; MFCC stands for the Mel frequency cepstrum coefficient.

In speech emotion research, different sets of features of different databases are selected. Pitch features are among the most popular and important features. They are usually extracted through the maximum of the autocorrelation function (ACF) or the average magnitude difference function (AMDF) of speech segments. In order to increase the accuracy and anti-noise of the pitch extraction, we combine the autocorrelation energy function

$$R_{ss}(k) = \left[\sum_{m=1}^{N/2} s(m)s(m+k)\right]^2 \qquad k = 1, 2, \ldots, \frac{N}{2} \tag{1}$$

and the magnitude difference energy function

$$D_{ss}(k) = \left[\sum_{m=1}^{N/2} \left| s(m) - s(m+k) \right| \right]^2 \qquad k = 1, 2, \ldots, \frac{N}{2} \tag{2}$$

The pitch extraction object function is defined as

$$F_{\text{pitch}}(k) = \frac{R_{ss}(k)}{D_{ss}(k)} \tag{3}$$

For every speech frame $s(m)$, the pitch is estimated by the maximum of the function F_{pitch}.

The pitch mean distributions of samples of the neutral emotion and annoyance-type emotion are illustrated in Fig. 2. The vertical coordinate stands for the amount of samples in each region of the feature value, which reflects the sample distribution density. Subcategorizing the emotion classes according to the sentence length type can simplify the emotion expression forms of each class, which causes the difficulty in modeling. Noticeably the distributions of the two pitch features change significantly in different sentence lengths. The distributions in a short sentence are less complicated, suggesting that different models should be used for each subcategorized dataset. The analysis of standard deviation is similar to that of pitch mean distributions.

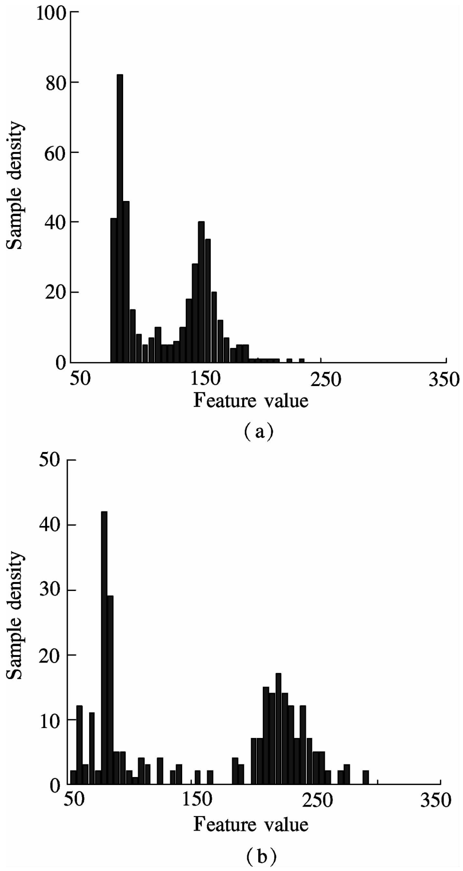

Fig. 2 Pitch mean distributions of sentences with different emotions. (a) Annoyance-type emotion of short sentence (447 samples); (b) Neutral emotion of short sentence (231 samples)

The new research in emotional feature selection is to generate large amounts of acoustic features, including both prosodic features and voice quality features, and use the features selection method to optimize the emotional features for recognition. The principal components analysis (PCA) is used to further explore the merits of subcategorizing speech utterances according to the natural sentence length. We randomly select 100 samples for each emotion, and use the PCA to project the original feature space onto an optimized feature space. The first three dimensions of the new feature space are plotted in Fig. 3. Compared with the traditional method of modeling with all types of utterances, the two types of emotions are better distinguished with our method.

1.3 Recognition methodology

The annoyance-type emotion recognition method proposed in this paper is based on the neural network. The common learning algorithm for the BP neural network is the BP algorithm. The BP algorithm is easy to get into local extreme points and has the curse-of-dimensionality problem. Therefore, during the BP neural network training time, the SFLA can be used to optimize the random

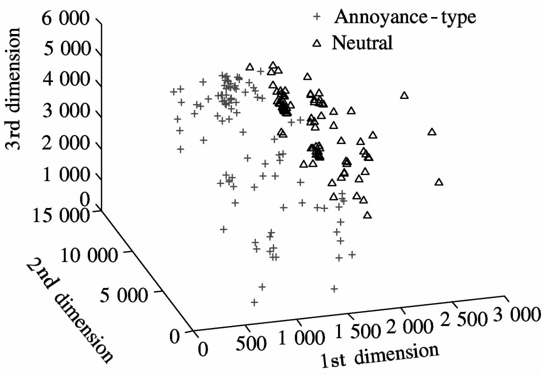

Fig. 3 PCA analysis of sentences with different emotions

initial parameters of the network, such as the connection weights between the input layer and the hidden layer, the connection weights between the hidden layer and the output layer, as well as the threshold value, in order to improve the convergence rate and learning ability[16]. The algorithm is as follows:

1) The initialized population is randomly provided in n-dimensional space, and every unit element is the pending values of weights and thresholds.

2) The fitness function value of every unit with its current status is calculated, and the fitness function is defined as

$$J(k,\, i) = \sum_{m=1}^{n} (y_{m,\,i} - \hat{y}^{k}_{m,\,i})^{2} \qquad k = 1, 2, \dots, N \quad (4)$$

where $J(k,\, i)$ is the fitness value of the i-th unit after k times of iteration; k is the number of iterations and N is the maximum number of iterations; m is the number of samples in the training set; $y_{m,\,i}$ is the network target output of the i-th unit with the m-th sample input; $\hat{y}^{k}_{m,\,i}$ is the network actual output of the i-th unit with the m-th sample input after k times of iteration.

3) The fitness values of all individuals are sorted from good to bad, and each unit is in turn divided into individual sub-populations.

4) The worst unit in current sub-populations will be updated. And this updating will continue until the number of iterations meets the requirements given in advance.

5) When all the sub-populations finish updating, then repeat steps 3 to 5 until the mixed number of iterations is reached.

Parity is a common criterion to evaluate the neural network performance. It can validate the performance of the SFLA neural network by comparison with the BP neural network. The maximum evolution generation is 100, the node function of the hidden layer is tansig, the node function of the output layer is purelin, the population size is 50 and the network error is less than 0.1×10^{-5}. Every network will simulate 20 times. Then the BP neural network gets the average convergence generation of 19.75, and the SFLA neural network gets the average convergence generation of 11.65. So the SFLA neural network has higher convergence speed and better performance.

2 Experimental Results

Experiments of relationship between average recognition rate and features dimension are conducted by using the collected speech sentences and the results are shown in Tab. 2.

Tab. 2 Average recognition rate of different methods %

Method	Characteristic dimension					
	6	7	8	9	10	11
SFLA	67.1	73.9	78.1	79.8	81.6	81.2
RBF	62.8	71.0	75.9	76.8	77.3	77.3
BP	61.5	70.4	74.0	75.8	76.0	75.8

From Tab. 2, we can see that all three methods reach the maximum average recognition rate with 10-dimensional feature space, so the original feature should be reduced to 10 dimensions. For the study of annoyance-type emotion features, the top ten best features are selected by the Fisher discriminate ratio (FDR). For each feature i, the FDR is computed as

$$\mathrm{FDR}_i = \frac{(\mu_{i,\,\text{neutral}} - \mu_{i,\,\text{annoyance}})^2}{\sigma^2_{i,\,\text{neutral}} - \sigma^2_{i,\,\text{annoyance}}} \qquad (5)$$

where $\mu_{i,\,\text{neutral}}$ and $\mu_{i,\,\text{annoyance}}$ are the class mean value of feature vector i for the neutral emotion class and the annoyance-type emotion class, respectively; $\sigma^2_{i,\,\text{neutral}}$ and $\sigma^2_{i,\,\text{annoyance}}$ are the variance values. Selected features include max_F_2, std_BBE_3, range_diff_F_1, mean_F_2, mean_F_1, mean_F_3, min_diff_F_4, range_diff_I, jitter and range_I. Here, diff stands for differentiation; I stands for intensity. For instance, “min_diff_I” stands for minimum of intensity differentiation.

According to the feature selection results, voice quality features (e. g. formants) and spectral features (e. g. MFCC) are especially important for the detection of annoyance-type emotion from speech. Since the natural speech emotions are less distinguishing in the arousal dimension than the acted speech and annoyance-type emotion is mainly located in the negative region of the valence dimension, the valence features seem more important in the study of annoyance-type emotion than traditional prosodic features which are more useful in the assessment of levels of excitement. Especially in the work environment, people usually tend to hide their negative emotions. Although the level of excitement in speech emotion is easy to control consciously, the voice quality features and the spectral features may provide important information to assess the unconsciously expressed emotion cues in speech.

In this paper, 200 utterances are randomly selected as the testing sets and the remaining utterances are used for training. For each neutral emotion class, the training set consists of 131 utterances, and for each annoyance-type emotion class, the training set consists of 347 utterances. Experimental results are shown in Fig. 4. From the figure, the average recognition rate of annoyance-type emotion with the SFLA neural network can reach 78%, which is obviously higher than that of the RBF neural network (70%) and the BP neural network (65%). Accordingly, the error rate 1 (neutral but recognized to be annoyance-type) and the error rate 2 (annoyance-type but recognized to be neutral) of the SFLA neural network are lower than those of the BP and the RBF neural networks. It means that the SFLA neural network proposed in this paper is suitable for the recognition of annoyance-type emotion. The weights obtained by traditional methods are adjusted during training, which are very sensitive to the initial value of the network. Once the value is improper, the network will surge and cannot converge. At the same time, it is easy to fall into the local extreme value and cannot obtain the best weight distribution, which will decrease learning ability when recognizing confusing emotions. So the recognition rate is relatively low. Since the SFLA neural network integrates the advantages of the genetic algorithm and particle swarm optimization, it can obviously improve the recognition rate. It also gains better results when optimizing neural network weights. Compared with the traditional evolutionary algorithms, the SFLA neural network can more easily obtain the balance between searching and optimization, which overcomes the shortage of slow learning speed and low recognition accuracy with traditional neural network models.

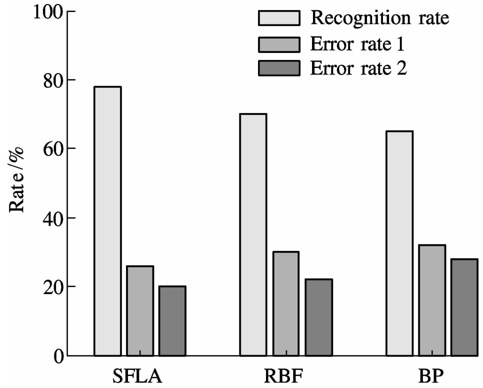

Fig. 4 Recognition results by SFLA, BRF and BP neural networks

3 Conclusion and Future Work

Annoyance-type emotion occurs in long-term repeated work situations and with the aid of noise we induced the annoyance emotional speech in an elicited experiment. Combining the SFLA with artificial neural networks, the annoyance-type emotion of speech is identified in this paper. And prosody and voice quality features are picked up from the emotional speeches. By these features, the BP neural network, the RBF neural network and the SFLA neural network are experimented to recognize annoyance-type emotion. Under the same test conditions, the average recognition rate of the SFLA neural network is higher than that of the BP neural network by 4.7%, and is higher than that of the RBF neural network by 4.3%. It means that with the SFLA the random initial parameters (learning weights and thresholds) of the network can be trained to optimize the neural networks. And then quicker convergence and better learning ability can be achieved.

The emotion detection system described in this paper has a potential to evaluate people's work abilities from the view of the psychology status of the individual. On the other hand, physiological signals will provide another angle to assess the emotional status. The fusion of the multimodal emotion recognition system is worth further studying.

References

[1] Barbara A, Spellman D, Willingham T. *Current directions in cognitive science* [M]. Beijing: Beijing Normal University Press, 2007: 1 – 5.

[2] Vinciarelli A, Pantic M, Bourlard H. Social signal processing: survey of an emerging domain[J]. *Image and Vision Computing*, 2009, **27**(12): 1743 – 1759.

[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] Jones C M, Jonsson I M. Automatic recognition of affective cues in the speech of car drivers to allow appropriate responses[C]//*Proceedings of the 17th Australia Conference on Computer-Human Interaction: Citizens Online: Considerations for Today and the Future*. Canberra, Australia, 2005: 1 – 10.

[5] Clavel C, Vasilescu I, Devillers L, et al. Fear-type emotion recognition for future audio-based surveillance systems [J]. *Speech Communication*, 2008, **50**(6): 487 – 503.

[6] Ang J, Dhillon R, Krupski A, et al. Prosody-based automatic detection of annoyance and frustration in human-computer dialog [C]//*7th International Conference on Spoken Language Processing*. Denver, CO, USA, 2002: 16 – 20.

[7] Mitsuyoshi S, Ren F, Tanaka Y, et al. Non-verbal voice emotion analysis system [J]. *International Journal of Innovative Computing, Information and Control*, 2006, **2**(4): 819 – 830.

[8] Pao T L, Chen Y T, Yeh J H, et al. Emotion recognition and evaluation from Mandarin speech signals[J]. *International Journal of Innovative Computing, Information and Control*, 2008, **4**(7): 1695 – 1709.

[9] Eusuff M, Lansey K, Pasha F. Shuffled frog-leaping algorithm: a memetic meta-heuristic for discrete optimization[J]. *Engineering Optimization*, 2006, **38**(2): 129 – 154.

[10] Amiri B, Fathian M, Maroosi A. Application of shuffled frog-leaping algorithm on clustering[J]. *The International Journal of Advanced Manufacturing Technology*, 2009, **45**(1/2): 199 – 209.

[11] Huang C W, Jin Y, Zhao Y, et al. Recognition of practi-

cal emotion from elicited speech[C]//*IEEE* 1*st International Conference on Information Science and Engineering*. Nanjing, China, 2009: 639 – 642.

[12] Huang C W, Jin Y, Wang Q Y, et al. Speech emotion recognition based on decomposition of feature space and information fusion[J]. *Signal Processing*, 2010, **26**(6): 835 – 842.

[13] Huang C W, Jin Y, Zhao Y, et al. Design and establishment of practical speech emotion database[J]. *Technical Acoustics*, 2010, **29**(1): 63 – 68.

[14] Gobl C, Ní Chasaide A. The role of voice quality in communicating emotion, mood and attitude[J]. *Speech Communication*, 2003, **40**(1): 189 – 212.

[15] Johnstone T, van Reekum C M, Hird K, et al. Affective speech elicited with a computer game[J]. *Emotion*, 2005, **5**(4): 513 – 518.

[16] Xu F Y. Adaptive shuffled frog-leaping algorithm for motion estimation[D]. Pingtung, China: National Pingtung Institute of Commerce, 2013.

工作环境中的语音烦躁情绪检测方法

王青云[1,2]　赵　力[1]　梁瑞宇[1]　张潇丹[1]

([1] 东南大学信息科学与工程学院, 南京 210096)

([2] 南京工程学院通信工程学院,南京 211167)

摘要:为了检测工作人员的烦躁情绪,实现情感状态的评价,通过在工作环境中诱发情感语音,获取了足够的测试样本,建立了2 000条样本的工作环境情感语音数据库.在检测烦躁情绪过程中,首先提取语音的韵律特征和音质特征参数,然后利用基于蛙跳算法的改进的BP神经网络进行烦躁情绪识别.实验比较了BP,RBF和SFLA神经网络的性能,结果显示SFLA神经网络的识别率比BP神经网络高4.7%,比RBF神经网络高4.3%.实验结果表明,使用蛙跳算法训练随机初始数据可以优化神经网络的连接权重和阈值,加快收敛速度,提高识别率.

关键词:语音情感检测;烦躁类型;句子长度;蛙跳算法

中图分类号:TN912.3