Journal of Southeast University (English Edition) Vol. 32, No. 2, pp. 158 - 163 June 2016 ISSN 1003—7985

A novel speech emotion recognition algorithm based on combination of emotion data field and ant colony search strategy

Zha Cheng[1,2] Tao Huawei[1] Zhang Xinran[1] Zhou Lin[1] Zhao Li[1] Yang Ping[2]

([1]Key Laboratory of Underwater Acoustic Signal Processing of Ministry of Education, Southeast University, Nanjing 210096, China)
([2]College of Big Data and Informaton Engineering, Guizhou University, Guiyang 550025, China)

Abstract: In order to effectively conduct emotion recognition from spontaneous, non-prototypical and unsegmented speech so as to create a more natural human-machine interaction; a novel speech emotion recognition algorithm based on the combination of the emotional data field (EDF) and the ant colony search (ACS) strategy, called the EDF-ACS algorithm, is proposed. More specifically, the inter-relationship among the turn-based acoustic feature vectors of different labels are established by using the potential function in the EDF. To perform the spontaneous speech emotion recognition, the artificial colony is used to mimic the turn-based acoustic feature vectors. Then, the canonical ACS strategy is used to investigate the movement direction of each artificial ant in the EDF, which is regarded as the emotional label of the corresponding turn-based acoustic feature vector. The proposed EDF-ACS algorithm is evaluated on the continueous audio/visual emotion challenge (AVEC) 2012 dataset, which contains the spontaneous, non-prototypical and unsegmented speech emotion data. The experimental results show that the proposed EDF-ACS algorithm outperforms the existing state-of-the-art algorithm in turn-based speech emotion recognition.

Key words: speech emotion recognition; emotional data field; ant colony search; human-machine interaction
doi: 10.3969/j.issn.1003-7985.2016.02.005

It is important to consider human behavior informatics and behavioral signal processing[1-2] when attempting to create an intelligent and natural human-machine interaction. As an important part of the quantitative analysis of human behavior, emotion recognition plays an important role in many applications, such as virtual agents, mobile phones, and in-car interfaces, in appropriately responding and reacting to the emotional state of the interacting users[3-4]. This results in social competence and increased acceptance among potential users.

Even though current works[5-6] report remarkable emotion recognition accuracy, when attempting to assign an emotional label to an emotionally colored speech turn, most works are not suitable for automatic speech emotion recognition in a real-life setting. One of reasons is that automatic emotion recognition (AER) algorithms are mostly evaluated on acted and prototypical data such as German "Berlin" or Danish emotional speech, which is relatively easy to assign to a set of predefined emotional labels and causes the recognition performance to overestimate[7-8]. The other reason is that AER methods fail to effectively deal with unsegmented and spontaneous speech, which even requires an immediate response to the emotional state of a user before he/she has finished speaking[9].

It is an interesting and challenging problem how to automatically recognize the user's emotional state from spontaneous, non-prototypical and unsegmented speech, since the speaker's emotion always affects the temporal dynamics of spectral, prosodic, and voice quality acoustic features[10-11]. For accurately recognizing speech emotion turn by turn, the temporal context of the preceding speech turns has to be considered when modeling emotional history. The static classifiers like SVM[12] and KNN[13] do not model the temporal dynamics context but capture this contextual information only by temporal-related features such as duration or statistical functionals over features. Though the dynamic classification methods such as hidden Markov models [14] and hidden conditional random fields[15] are the ones to reduce this limitation, these classifiers have no possibility to model the temporal dynamics context by a flexible self-learned manner[9]. Recently, a dynamic classifier called the long-term short memory (LSTM) recurrent neutral network is used to model the temporal dynamics context. Wöllmer et al. [9] used the LSTM to model the temporal dynamics context of real-life turn-based speech emotion. This method achieves state-of-the-art quality for turn-based speech emotion recognition. However, LSTM has drawbacks for modeling the context information by strictly sequential propagation[16], and ignores the prior information of emotion variation. This information may be influenced by the

Received 2015-10-09.
Biographies: Zha Cheng (1979—), male, graduate; Zhao Li (corresponding author), male, doctor, professor, zhaoli@seu.edu.cn.
Foundation items: The National Natural Science Foundation of China (No. 61231002, 61273266, 61571106), the Foundation of the Department of Science and Technology of Guizhou Province (No. [2015]7637).
Citation: Zha Cheng, Tao Huawei, Zhang Xinran, et al. A novel speech emotion recognition algorithm based on combination of emotion data field and ant colony search strategy[J]. Journal of Southeast University (English Edition), 2016, 32(2): 158 - 163. doi: 10.3969/j.issn.1003-7985.2016.02.005.

speaker's personality and cultural background, and can be captured by the labeled training data.

To address the above issues, in this paper we present a novel turn-based speech emotion recognition algorithm, which can not only model the temporal context but also learn the emotion varied prior information for each turn. The proposed speech emotion algorithm is developed by the following two steps. First, we develop the EDF by using the canonical data field (DF). Using its potential function, we can quantitatively investigate the emotional label relationship among acoustic feature vectors of different turns. Secondly, an artificial ant colony is used to mimic the acoustic feature vectors of turns, and the ant colony search (ACS) strategy is used to investigate the movement and transformation of each artificial ant in the EDF.

To evaluate the proposed emotion recognition algorithm, all simulation experiments are performed on the SEMAINE (machine-human interaction using nonverbal expression) corpus. This corpus contains the spontaneous, non-prototypical and unsegmented speech by users' spontaneous and emotionally colored conversations under the virtual agent scenario.

1 Emotional Database and Acoustic Feature Extraction

The AVEC 2012 challenge dataset is a subset of the SEMAINE corpus[17], which is continuously annotated in a two-dimensional space spanned by valence and activation. Each recording in this challenge dataset is collected by the Wizard-of-Oz SAL interface, which allows users to speak to four different virtual characters. Each of them represents one of four emotional quadrants: "Prudence" represents relaxed/serene (quadrant Ⅰ); "Poppy" represents happy/excited (quadrant Ⅱ); "Obadiah" represents sad/bored (quadrant Ⅲ); and "Spike" represents angry/anxious (quadrant Ⅳ). Although the quadrant annotation of emotion space can reflect continuous emotional change, a categorical decision has to be made for humanizing the output of the AER algorithm. Following the previous work[9], we consider the five-class emotion task: relaxed/serene, happy/excited, sad/bored, angry/anxious and neutral. Neutral is also considered in this work since it is dominant in real-world emotional categories and can avoid the ambiguous emotional categorical annotation for a given speech turn[9]. The recordings in the AVEC 2012 challenge dataset were split into three parts: training, development and test set, respectively.

As each recording in this data set contains long continuous time, a segmentation of the recording has to be performed. Pauses of more than 200 ms are used to segment a recording into turns based on the voice activity detection. Therefore, we can perform turn-based speech emotion recognition in this work. For each recording, the annotations for arousal and valence are quasi-time-continuous labels using the FEELtrace system[18]. As ground truth of each turn, the mean annotation of each turn is calculated and is mapped to the five-class emotion using the two-dimensional space spanned by the valence and activation.

In this work, we perform turn-based speech emotion recognition using the 2 268-dimensional acoustic baseline feature set, which are 32 energy and spectral-related low-level descriptors × 42 functionals, 6 voice related LLD × 32 functionals, 32 delta coefficients of the energy/spectral LLD × 19 functionals, 6 delta coefficients of the voice related LLD × 19 functionals, and 10 voiced/unvoiced durational features. Due to space limitation, details for the LLD and functionals are given in Ref. [17]. The functionals are computed over turn detected by voice activity detection. To avoid redundancy among the 2 268-dimensional feature set, we perform feature selection using the correlation-based feature selection with the linear forward search[19]. This results in the selection of 55-dimensional feature vector for the five-class emotion recognition task.

2 Algorithm

In this section, we present a novel turn-based speech emotion recognition algorithm, called ACS based on the EDF (EDF-ACS). The proposed EDF-ACS algorithm aims at recognizing the spontaneous, non-prototypical and unsegmented speech emotion by simultaneously modeling turn-based emotional history and the prior emotion variance information for each turn. The following subsections 2.1 and 2.2 describe the two steps required to build the EDF-ACS algorithm.

2.1 EDF

Gan et al.[20] proposed an effective algorithm called data field (ED) to describe the inter-relationship (e.g., interaction, transformation and movement) among different data similar to the source field of physics, and each piece of data in ED is not independent and can be associated with other data by the potential function of ED. Using this potential function, data can move along the path and direction of the largest potential function imposed by other data. This ED has been proved a powerful algorithm for data mining, such as data clustering and probabilistic density's estimation[21].

Let $\{\boldsymbol{x}_i\}_{i=1}^{N}$ be a data set and N is the number of data. The potential function between data x_i and $\{\boldsymbol{x}_j\}_{j=1,j\neq i}^{N}$ is defined as

$$F(\boldsymbol{x}_i, \{\boldsymbol{x}_j\}_{j=1,j\neq i}^{N}) = \sum_{j=1,j\neq i}^{N} \exp(-\mathrm{dis}(\boldsymbol{x}_i, \boldsymbol{x}_j)) \quad (1)$$

where $\mathrm{dis}(\cdot,\cdot)$ is the distance measure between different data. Without using the common Euclidean distance, this distance cannot catch any statistical regularities in the

data estimated from the training set with the labeled data. According to Ref. [21], we compute the potential function using the Mahanalobis distance for the kNN classifier. Using the Mahanalobis distance, Eq. (1) can be rewritten as

$$F(\boldsymbol{x}_i, \{\boldsymbol{x}_j\}_{j=1,j\neq i}^{N}) = \sum_{j=1,j\neq i}^{N} \exp\{-(\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}} \boldsymbol{M} (\boldsymbol{x}_i - \boldsymbol{x}_j)\} \tag{2}$$

where the linear transform matrix $\boldsymbol{M}$ is semi-definite, and can be solved by a defined semi-definite program problem[21]. The distance is trained to aim at the objective that the k-nearest neighbors belonging to the same class are always separated from the data of other classes by a larger margin. Therefore, the unlabeled data tends to have the largest potential value produced by the k-nearest neighbors from the same class data, other than the smallest from different classes. Correspondingly, the recognition rule for the DF is formulated as

$$\text{If } \arg\max_{p} \{V_p(\boldsymbol{x}_i) = \sum_{x_j \in \delta_p(\boldsymbol{x}_i)} F(\boldsymbol{x}_i, \boldsymbol{x}_j)\},\ p = 1, 2, \cdots, P$$
$$\text{then} \quad \boldsymbol{x}_i \in \text{the } p\text{-th class}$$

where $V_p(\boldsymbol{x}_i)$ is $\boldsymbol{x}_i$'s potential value produced by the data set $\delta_p(\boldsymbol{x}_i)$; $\delta_p(\boldsymbol{x}_i)$ is the k-nearest neighbors of $\boldsymbol{x}_i$ from the p-th class train set; $p = 1, 2, \cdots, P$ and P is the number of class. As shown in the rule, this unlabelled data $\boldsymbol{x}_i$ is assigned to the label when its correspondingly k-nearest neighbors of the p-th class training set produce the largest potential value for the data $\boldsymbol{x}_i$. Similar to the kNN recognizer, the DF also uses the nearest neighbors of an unlabelled data to recognize its label. The kNN recognizer recognizes an unlabelled data by the major label of the k-nearest neighbors in the training set. Different from the kNN, the DF exploits the comparison of potential values, which is respectively produced by the k-nearest neighbors from the training set of different classes. According to the recognition rule, the DF can quantitatively measure the label's attribute to an unlabelled data by the value of the potential function. Using DF for speech emotion recognition, each data in DF is represented with the reduced acoustic feature vector using the CFS-based feature selection[19]. Then, this DF is referred to as EDF, and each acoustic feature vector in this EDF has its own emotion label. However, the EDF cannot effectively model the emotion varied history. To address the issue, we develop the EDF-ACS algorithm in the following subsection further.

2.2 ACS based on EDF

ACS[22] is a metaheuristic algorithm used to find an optimal path in a graph with respect to its predefined functions. The artificial ants defined by the ACS move along the path of the graph by mimicking the forging behavior of its biological counterparts. Finding the optimal path for ants is a stochastic procedure based on two elements, namely heuristic value and pheromone. The heuristic value is the prior knowledge of edge selection for a single ant, without communicating with the other ants. A pheromone is the weight of edge and a way that ants communicate with other ants. The more pheromones on the one edge, the greater the possibility that the ant selects the edge. The key idea of the ACS-type algorithm contains two interwoven rules. Each ant selects the edge of the graph by considering the amount of pheromone corresponding to this edge. On the other hand, ants constantly move along the graph and an outer observer appreciates the path of each ant according to a certain quality function. The rules allow ACS to dynamically investigate the interaction, transformation and movement of acoustic feature vectors in EDF. Thus, this results in that the turn-based emotional history and the emotion varied prior information may be simultaneously investigated by the proposed EDF-ACS algorithm.

The necessary precondition of applying the ACS to any problem is to reformulate the edges and nodes of graph and give reasonable quality functions. It is easy to assume that each node corresponds to one of the emotional categories, and the edge links the possible emotion change mode. Fig. 1 plots the graph of the five-class emotion changing modes. Subsequently, we define the heuristic value and pheromone of this graph. The heuristic value reflects the prior information of emotional change, which can be calculated by the potential value produced by the labeled training data, since this potential value gives an indication of the likely emotion label and location information in the EDF for the corresponding acoustic feature vector. More specifically, let $\boldsymbol{x}(n)$ be the acoustic feature vector of the n-th speech turn, and the emotional label of the $(n-1)$-th speech turn is the c-th class. Then, the potential value $\eta_{\boldsymbol{x}(n)}(c, d)$ of $\boldsymbol{x}(n)$ produced by the d-th class training set can be written as

$$\eta_{\boldsymbol{x}(n)}(c, d) = V_d(\boldsymbol{x}(n)) = \sum_{x_j \in \delta_d(x(n))} F(\boldsymbol{x}(n), \boldsymbol{x}_j) \tag{3}$$

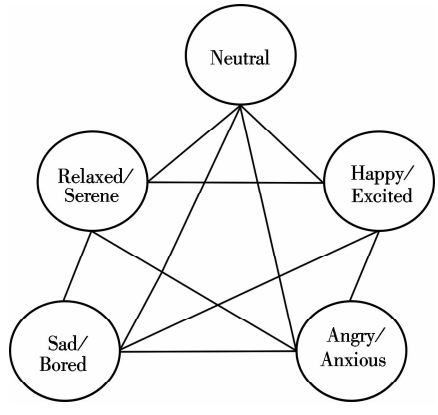

Fig. 1 The five-class emotion changing graph

where the potential value $\eta_{\boldsymbol{x}(n)}(c,d)$ represents the heuristic value of $\boldsymbol{x}(n)$ changing from the c-th class emotion to the d-th class. $\delta_d(\boldsymbol{x}(n))$ is the k-nearest neighbor of $\boldsymbol{x}(n)$ from the d-th class training set.

For the specific n-th speech turn, the pheromone represents the information of the emotional history, which is influenced by the acoustic feature vectors $\{\boldsymbol{x}(1),\boldsymbol{x}(2),\cdots,\boldsymbol{x}(n-1)\}$ of the precedent turns. Similar to $\eta_{\boldsymbol{x}(n)}(c,d)$, $\eta_{\{\boldsymbol{x}(1),\boldsymbol{x}(2),\cdots,\boldsymbol{x}(n-1)\}}(c,d)$ is defined as the pheromone of the edge from the c-th class emotion to the d-th class, which is produced by the set of acoustic feature vectors $\{\boldsymbol{x}(1),\boldsymbol{x}(2),\cdots,\boldsymbol{x}(n-1)\}$. Suppose that $\boldsymbol{x}(n-1)$ belongs to the c-th class emotion, and that there is a likely emotional label change for $\boldsymbol{x}(n)$. Using the canonical ACS strategy, the probability that $\boldsymbol{x}(n)$ selects the edge from the c-th class emotion to the d-th class can be defined as

$$p_{\boldsymbol{x}(n)}(c,d) = \frac{[\eta_{[\boldsymbol{x}(n)]}(c,d)]^{\alpha}[\eta_{\{\boldsymbol{x}(1),\boldsymbol{x}(2),\cdots,\boldsymbol{x}(n-1)\}}(c,d)]^{\beta}}{\sum_{u\in S_{\boldsymbol{x}(n)}(c)}[\eta_{[\boldsymbol{x}(n)]}(c,u)]^{\alpha}[\eta_{\{\boldsymbol{x}(1),\boldsymbol{x}(2),\cdots,\boldsymbol{x}(n-1)\}}(c,u)]^{\beta}} \tag{4}$$

where

$$\eta_{\{\boldsymbol{x}(1),\boldsymbol{x}(2),\cdots,\boldsymbol{x}(n-1)\}}(c,d) = (1-\rho)\eta_{\{\boldsymbol{x}(1),\boldsymbol{x}(2),\cdots,\boldsymbol{x}(n-2)\}}(c,d) + \eta_{\boldsymbol{x}(n-1)}(c,d) \tag{5}$$

where $S_{\boldsymbol{x}(n)}(c)$ is the allowing emotion set of $\boldsymbol{x}(n)$ changing from the c-th class emotion; α is the parameter that determines the relative importance of the heuristic value and β determines the relative importance of pheromone; ρ $(0<\rho<1)$ is the pheromone evaporating parameter, by which we can learn the previous temporal context information with the exponential decay manner. Based on the above analysis, the proposed EDF-ACS algorithm is described in Fig. 2.

Fig. 2 shows that using Eq. (4), there are two cases for recognizing the unsegmented and turn-based speech emotion. One case is that there is likely emotional change for the acoustic feature vector $\boldsymbol{x}(n)$. Although we cannot accurately estimate the necessary condition of emotion change, $\min p_y(c,d)$ ($y\in$ the c-th class training set) results in that Eq. (4) cannot ignore any likely emotion change. Therefore, we use Eq. (4) with $\beta\neq 0$ to calculate the emotional change probability for the acoustic feature vector $\boldsymbol{x}(n)$, which considers the emotion information history of precedent turns. Conversely, Eq. (4) uses $\beta=0$ to calculate the emotional change probability for $\boldsymbol{x}(n)$, which is reduced to the probability representation of the recognition rule.

3 Experiments

The parameters k, α, β and ρ described in the previous section affect the performance of the proposed EDF-ACS algorithm. In Tab. 1, we set a range for each of the four parameters and uniformly partition these ranges into eight, eleven, eleven and nine discrete values for parameters $k(k=8,9,\cdots,14,15)$, $\alpha(0\leqslant\alpha\leqslant 2)$, $\beta(0\leqslant\beta\leqslant 2)$ and $\rho(0<\rho<1)$, respectively.

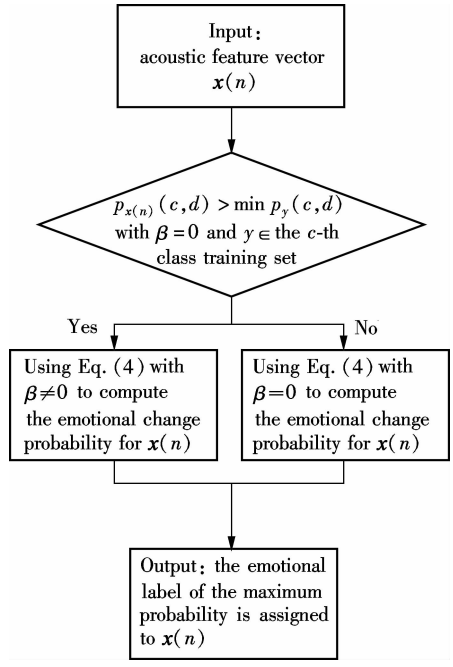

Fig. 2 Flowchart of the proposed EDF-ACS algorithm

The value ranges of parameters α, β and ρ are set to be the same as the canonical ACS algorithm. The start trial of k is 8, since the linear transform matrix $\boldsymbol{M}$ in Eq. (2) is solved only using three target neighbors and a 50/50 cross validation strategy over the training set. When k is increased, the performance of EDF-ACS appears to converge rapidly and the higher k can result in over-smoothing of the recognition boundary. Therefore, the final trial for k is set to be 15. This results in 8 712 different parameter combinations. For different k, we greedily search the combination of the parameters ρ, α and β to make the EDF-ACS algorithm achieve the highest average recognition accuracy using the development set. The experimental results are shown in Tab. 1.

Tab. 1 Average recognition accuracy over the development set

Fixed parameter k	Adjusted parameters by greedy search	Average recognition accuracy/%
8	$\rho=0.4$, $\alpha=1$, $\beta=1.2$	37.88
9	$\rho=0.4$, $\alpha=1$, $\beta=1.2$	37.96
10	$\rho=0.5$, $\alpha=0.8$, $\beta=1.4$	39.42
11	$\rho=0.5$, $\alpha=0.8$, $\beta=1.4$	40.14
12	$\rho=0.5$, $\alpha=0.8$, $\beta=1.4$	41.23
13	$\rho=0.6$, $\alpha=0.6$, $\beta=1.6$	40.61
14	$\rho=0.6$, $\alpha=0.6$, $\beta=1.6$	39.74
15	$\rho=0.6$, $\alpha=0.6$, $\beta=1.6$	38.27

As shown in Tab. 1, for smaller k ($k=8,9$), more context information is exploited in recognition task by

using the smaller ρ and larger α. On the contrary, for larger k ($k = 13, 14$ and 15), less context information is exploited in the recognition task by using the larger ρ and smaller α. This may be because the larger k can effectively describe the distribution and location information of acoustic feature vectors in the emotional data field. Then we evaluate the proposed EDF-ACS for the test set with $k = 12$, $\alpha = 0.8$ and $\beta = 1.4$, which can achieve the highest average accuracy for the development set. To fairly evaluate the proposed EDF-ACS, we compare the proposed algorithm with ED and the common algorithms of turn-based speech emotion recognition. The common algorithms are KNN, SVM and discriminatively trained LSTM. In particular, the discriminatively trained LSTM achieves state-of-the-art quality for turn-based speech emotion recognition[9]. The input layer size of LSTM is set to be 55, which is the same as the dimension of the reduced acoustic feature vector. In addition, LSTM contains a recurrent hidden layer with 50 memory blocks for each LSTM cell. The output layer size of LSTM is equal to the number of emotional categories. SVM uses the Gaussian kernel function. The kernel bandwidth and regularization parameter is also tuned by the development set.

Tab. 2 shows the recognition results of the above algorithms for the AVEC 2012 dataset. LSTM outperforms the SVM, kNN and ED, which indicates that the dynamic context information is useful in the turn-based speech emotion recognition task. In particular, when using the same number of neighbors, ED achieves better recognition accuracy than kNN. This shows that the Mahanalobis distance can better detect the distribution characteristic of speech emotion data than the Euclidean distance. Furthermore, the best average accuracy for EDF-ACS is 35.01%, which indicates that the turn-based speech emotion recognition also should consider the emotional varied prior information.

Tab. 2 Average recognition accuracy on AVEC 2012 dataset

Algorithms	Accuracy/%
kNN ($k = 12$)	31.62
SVM	30.48
LSTM	33.65
ED ($k = 12$)	32.24
Proposed EDF-ACS	35.01

4 Conclusion

In this paper, we propose a novel speech emotion recognition algorithm based on the combination of the emotional data field (EDF) and the ant colony search (ACS) strategy. Compared to the existing and the state-of-the-art algorithms, the proposed EDF-ACS algorithm can effectively recognize emotion from spontaneous, non-prototypical and unsegmented speech. In this algorithm, the EDF models the varying information of speech emotion by using its potential function to investigate the inter-relationship among the turn-based acoustic feature vectors, and the canonical ACS strategy investigates the movement direction of each artificial ant in the EDF to model the speech emotion history and prior emotion information for each speech turn.

References

[1] Jin Q, Li C, Chen S, et al. Speech emotion recognition with acoustic and lexical features[C]//*IEEE International Conference on Acoustics, Speech and Signal Processing*. Brisbane, Australia, 2015: 4749-4753.

[2] Ramakrishnan S, El Emary I M M. Speech emotion recognition approaches in human computer interaction[J]. *Telecommunication Systems*, 2013, **52**(3): 1467-1478.

[3] Lu H, Frauendorfer D, Rabbi M, et al. StressSense: Detecting stress in unconstrained acoustic environments using smartphones[C]//*Proceedings of the* 2012 *ACM Conference on Ubiquitous Computing*. Pittsburgh, PA, USA, 2012: 351-360.

[4] Lee J S, Shin D H. A study on the interaction between human and smart devices based on emotion recognition [C]//*Communications in Computer and Information Science*. Berlin: Springer, 2013: 352-356.

[5] Anagnostopoulos C N, Iliou T, Giannoukos I. Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011[J]. *Artificial Intelligence Review*, 2015, **43**(2): 155-177. DOI: 10.1007/s10462-012-9368-5.

[6] Ingale A B, Chaudhari D S. Speech emotion recognition [J]. *International Journal of Soft Computing and Engineering*, 2012, **2**(1): 235-238.

[7] Lanjewar R B, Chaudhari D S. Speech emotion recognition: a review[J]. *International Journal of Innovative Technology and Exploring Engineering*, 2013, **2**(4): 68-71.

[8] Huang C W, Wi D, Zhang X J, et al. Cascaded projection of Gaussian mixture model for emotion recognition in speech and EGG signal[J]. *Journal of Southeast University*(*English Edition*), 2015, **31**(3): 320-326.

[9] Wöllmer M, Schuller B, Eyben F, et al. Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening[J]. *IEEE Journal of Selected Topics in Signal Processing*, 2010, **4**(5): 867-881. DOI: 10.1109/jstsp.2010.2057200.

[10] Gharavian D, Sheikhan M, Nazerieh A, et al. Speech emotion recognition using FCBF feature selection method and GA-optimized fuzzy ARTMAP neural network[J]. *Neural Computing and Applications*, 2012, **21**(8): 2115-2126. DOI: 10.1007/s00521-011-0643-1.

[11] Wu D, Parsons T D, Narayanan S S. Acoustic feature analysis in speech emotion primitives estimation[C]//*INTERSPEECH* 2010, *Conference of the International Speech Communication Association*. Makuhari, Chiba, Japan, 2010: 785-788.

[12] Swain M, Sahoo S, Routray A, et al. Study of feature combination using HMM and SVM for multilingual Odiya speech emotion recognition[J]. *International Journal of*

Speech Technology, 2015, **18**(3): 387 - 393. DOI: 10.1007/s10772-015 - 9275-7.

[13] Khan M, Goskula T, Nasiruddin M, et al. Comparison between KNN and SVM method for speech emotion recognition[J]. *International Journal on Computer Science and Engineering*, 2011, **3**(2): 607 - 611.

[14] Meng H, Bianchi-Berthouze N. Naturalistic affective expression classification by a multi-stage approach based on hidden markov models[M]//*Affective Computing and Intelligent Interaction*. Berlin: Springer, 2011: 378 - 387.

[15] Ramirez G A, Baltrušaitis T, Morency L P. Modeling latent discriminative dynamic of multi-dimensional affective signals[M]//*Affective Computing and Intelligent Interaction*. Berlin: Springer, 2011: 396 - 406.

[16] Gers F A, Schraudolph N N, Schmidhuber J. Learning precise timing with LSTM recurrent networks[J]. *The Journal of Machine Learning Research*, 2003, **3**(1): 115 - 143.

[17] Schuller B, Valster M, Eyben F, et al. Avec 2012: The continuous audio/visual emotion challenge[C]//*Proceedings of the* 14*th ACM International Conference on Multimodal Interaction*. New York, USA: ACM, 2012: 449 - 456.

[18] Cowie R, Douglas-Cowie E, Savvidou S, et al. "FEELTRACE": An instrument for recording perceived emotion in real time[C]//*ITRW on Speech and Emotion*. Newcastle, Northern Ireland, UK, 2000: 19 - 24.

[19] Hall M A. Correlation-based feature selection for machine learning[D]. Hamilton, Zealand: Department of Computer Science, The University of Waikato, 1999.

[20] Gan W Y, Li D Y, Wang J M. Hierarchical clustering method based on data fields[J]. *Acta Electronica Sinica*, 2006, **34**(2): 258 - 262.

[21] Weinberger K Q, Saul L K. Distance metric learning for large margin nearest neighbor classification[J]. *The Journal of Machine Learning Research*, 2009, **10**(1): 207 - 244.

[22] Parpinelli R S, Lopes H S, Freitas A A. An ant colony based system for data mining: Applications to medical data[C]//*Proceedings of the Genetic and Evolutionary Computation Conference*. San Francisco, CA, USA, 2001: 791 - 797.

一种新的结合情感数据场和蚁群策略的语音情感识别算法

查 诚[1,2] 陶华伟[1] 张昕然[1] 周 琳[1] 赵 力[1] 杨 平[2]

([1] 东南大学水声信号处理教育部重点实验室, 南京 210096)

([2] 贵州大学大数据与信息工程学院, 贵阳 550025)

摘要:为了有效识别自发、非典型及未分割语音的情感以建立更自然的人机交互界面,提出了一种新的结合情感数据场和蚁群策略的语音情感识别算法.用情感数据场中势函数建立基于块的声学特征向量之间的内在联系.为识别自发语音情感,用人工蚁群模拟基于块的声学特征向量,然后用典型的蚁群策略研究每个人工蚂蚁在情感数据场的运动轨迹,并把该蚂蚁的运动轨迹作为对应的声学特征向量的情感标签.利用2012年连续音视频情感挑战赛中的语音数据对所提算法进行测试.实验结果表明:该算法较已有算法能更好地对基于块的语音情感进行识别.

关键词:语音情感识别;情感数据场;蚁群搜索;人机交互

中图分类号:TN912.3