DE GRUYTER OPEN

BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **15**, No 4

Sofia • 2015

Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.1515/cait-2015-0054

High-Order Markov Random Fields and Their Applications in Cross-Language Speech Recognition

Jiang Zhipeng[1], *Huang Chengwei*[2]

[1]*School of Electronics and Information Engineering, Jinling Institute of Technology, Nanjing, China*
[2]*College of Physics, Optoelectronics and Energy, Soochow University, Suzhou, China*
Emails: jzp@jit.edu.cn cwhuang@suda.edu.cn

Abstract: *In this paper we study the cross-language speech emotion recognition using high-order Markov random fields, especially the application in Vietnamese speech emotion recognition. First, we extract the basic speech features including pitch frequency, formant frequency and short-term intensity. Based on the low level descriptor we further construct the statistic features including maximum, minimum, mean and standard deviation. Second, we adopt the high-order Markov random fields* (*MRF*) *to optimize the cross-language speech emotion model. The dimensional restrictions may be modeled by MRF. Third, based on the Vietnamese and Chinese database we analyze the efficiency of our emotion recognition system. We adopt the dimensional emotion model* (*arousal-valence*) *to verify the efficiency of MRF configuration method. The experimental results show that the high-order Markov random fields can improve the dimensional emotion recognition in the cross-language experiments, and the configuration method shows promising robustness over different languages.*

Keywords: *High-order Markov random fields, speech emotion recognition, cross-database recognition, dimensional emotion model.*

1. Introduction

Speech emotion recognition is an important topic in natural human-computer interaction. We can improve the user satisfaction in a harmonic environment and improve the efficiency by teaching computers to response appropriately according to user's emotional states. Nwe, Foo and De Silva [1] applied a hidden Markov model in speech emotion recognition. Steidl [2] built a naturalistic database for emotional speech analysis. Huang et al. [3], and Zou, Huang et al. [4], studied several emotion-related states for practical applications. Wu, Falk and Chan [5] proposed a new type of emotional features in speech, namely the modulation spectral features. Although promising results have been reported, dealing with continuous emotional speech in the real world is still a challenging problem.

Many of the previous studies have considered the context information in Automatic Speech Recognition (ASR). Ferreiros and Pardo [6] studied semi-continuous speech recognition in Spanish language, and they developed a contextual model. Their experiments show that the modeling of pausing in the speech data helps to improve the overall performance. Mohamedet and Nair [7] developed an automatic speech recognition system for continuous speech signals by combining Hidden Markov Models (HMMs) and Artificial Neural Networks (ANNs). GMM is also a popular model that has brought promising results in speech emotion recognition [8-10]. Yeh et al. [11] studied the context information in continuous speech. In their study, an emotion radar chart was proposed for detecting emotions from continuous speech in Chinese.

In this paper we study the continuous speech emotion recognition in Chinese and Vietnamese. We first analyze the emotional features of both languages using utterance level static features. We then improve the final results by a Markov Random Fields based configuration method [12].

2. The database

The local databases we use for verifying the emotion recognition system are collected in our lab environment, which include the ChiNese emotional speech DataBase (CNDB) and the VietNamese emotional speech DataBase (VNDB). The speech utterances are recorded during an eliciting experiment using imaging technique, noise eliciting, video clips watching and computer games. The quality of the speech recording is checked by a large number of listeners. University students, who are native speakers of Chinese and Vietnamese language are selected to participate in the experiments. The vocal data are recorded in a quiet room with limited reverberations and noise. The microphone is single channel and close to the speaker. The statistics of the two databases are shown in Table 1.

Table 1. A summary of the experimental databases

Database	Speaker number	Language type	Text types	Emotion types
CNDB	51	Chinese	Inf.	9
VNDM	6	Vietnamese	20	6

Feature analysis of the emotional data is shown in the following figures. We studied the pitch frequency, the first formant frequency, the second formant frequency, the third formant frequency and the short-time energy of speech signal. As shown in Fig. 1 we can see that the extracted features can classify negative and positive emotional regions.

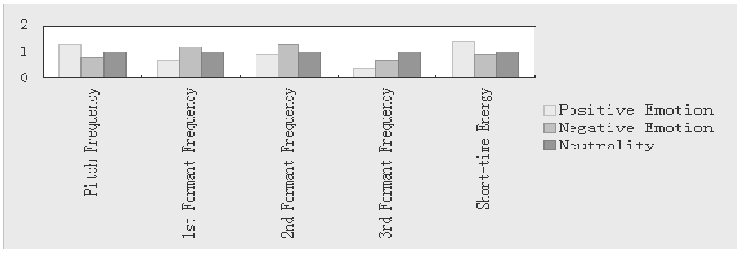

Fig. 1. Normalized feature distribution over positive and negative dimensions in a valence level

3. Methodology

3.1. Dimensional model

We adopt a dimensional model [13] for cross-language speech emotion recognition for different types of emotions. The input is an unknown speech sample. The output is the recognized arousal and valence region that can be used to verify the cross-language speech emotion recognition performance. A pre-learned model is trained using SVM-KNN (Support Vector Machine – K Nearest Neighbor) algorithm based on the annotated samples with arousal and valence labels. The dimensional space can be classified into different regions. The first region corresponds to the positive arousal dimension and the positive valence dimension, and we denote it as positive-positive to ease the notation. The second region corresponds to positive-negative, the third region corresponds to negative-negative, and the forth region corresponds to negative-positive. The functional layout is shown in Fig. 2.

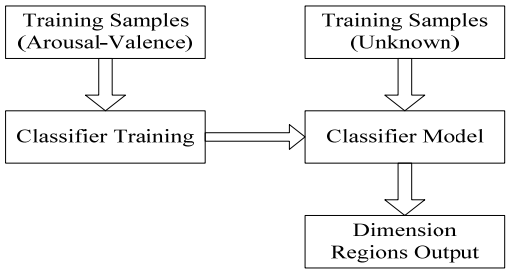

Fig. 2. Flowchart of the dimensional recognition system

3.2. The classification algorithm

In this paper we use SVM classifier to recognize the emotions based on the parameter optimization of Improved Shuffled Frog Leaping Algorithm (Im-SFLA), proposed by Z h a n g et al. [14].

In traditional SVM training, parameter optimization is based on empirical experiments. The performance of the corresponding classifier is not satisfactory. The experimental results show that using Cross Validation (CV) for the parameter selection in SVM training is better than using randomly selected parameters. The most commonly used CV methods include: K-fold Cross Validation, Hold-out Method and Leave-one-out Method.

Using some of the traditional searching algorithms, such as grid search, we may find the best classification rate. But the algorithm needs to search all the parameter points in the grid. When the search scale is very large, the traditional grid searching may cost a huge amount of time. However, using heuristic searching methods may decrease the computational cost and give the global optimal. We adopt the Im-SFLA for SVM parameter optimization.

We choose the inverse function as the fitness function and Radial Basis Function (RBF) as the kernel function of SVM. The penalty ratio C' and the width σ of the RBF need to be optimized. C' and σ form a 2-dimension frog individual, and the optimization methods [14] are:

Step a. The frog population and the value of chaotic mapping are initialized randomly. The frog individual corresponds to a value (C', σ). C' and σ are initialized separately as random variables from $[0, \theta]$ and $[0, \delta]$, where θ and δ are non-negative. The group number K in K-CV is then set to a default value and Im-SFLA algorithm is configured with initial parameters.

Step b. Search the optimal individual using Im-SFLA, and the fitness function of all frogsis calculated.

Step c. When the stopping criteria is met output the global optimal frog (C'_g, σ_g), where C'_g and σ_g are the optimized penalty ratio and width parameter of RBF kernel. Stop the iterative process when the criterion is met, otherwise go to Step b.

After finishing the parameter optimization, we may establish an Im-SFLA-SVM recognition system, its system flowchart being shown in Fig. 3.

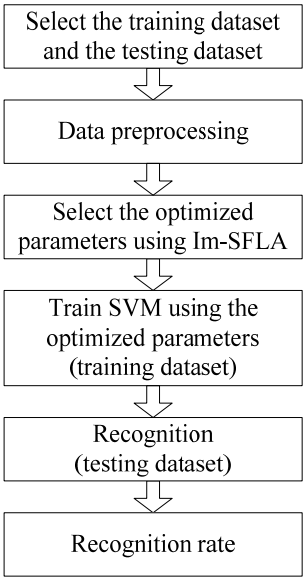

Fig. 3. Recognition model based on Im-SFLA-SVM

3.3. The configuration algorithm

In this paper we use a post configuration algorithm based on High-Order MRF configuration [15, 16], and apply it in the cross language speech emotion recognition.

Markov random field is an important machine learning algorithm. A b e n d, H a r l e y and K a n a l [17] first introduced MRF to image processing. With the development of Hammersley-Clifford theorem [18], the application of MRF in computer vision is increasingly popular. In this paper we apply the high-order MRF to the configuration in speech emotion recognition. The traditional solution to the high-order problem leads to a heavy computational cost. In this paper, we adopt the Quadratic Pseudo-Boolean Optimizer (QPBO) proposed by [16] to efficienty optimize the speech emotion recognition results.

QPBO is used for solving the optimization problem modeled by Markov random fields. The multi-labeling problem is converted into a binary labeling problem. QPBO[16] solves the linear programming relaxation of the target energy function by replacing binary integer constraints from {0, 1} with linear constraints from domain [0, 1]. The energy is defined by term e_p, which may represent the different order of energy, i.e. unary energy, pairwise energy. The QPBO algorithm concatenates all the energy values into a single vector ***e***, and parameterize it into a normal form [16]. A directed weighted graph *G* is constructed, and for each non-zero element e_p two directed arcs are added to the graph. Then a minimum *s*-*t* cut (*S*, *T*) in *G* is computed from a maximum flow, and this cut gives the final optimal solution to the relaxation. The partial labeling is [16]:

$$l_p = \begin{cases} 0 \text{ if } p \in S,\ \bar{p} \in T, \\ 1 \text{ if } p \in T,\ \bar{p} \in S, \\ \varnothing \text{ otherwise.} \end{cases} \tag{1}$$

The random field can be defined as $\boldsymbol{S} = \{s_1, s_2, \cdots, s_N\}$. Each random variable s_i can get a value in label set ***E***.

If and only if a random field satisfies the following two equations, it is called a Markov random field [19]:

$$P(s) > 0 \ \ \forall s \in \boldsymbol{S}, \tag{2}$$

$$P(s_i \mid s_1, s_2, \cdots s_{i-1}, s_{i+1}, \cdots, s_N) = P(s_i \mid \{s_j\} \ \ \forall j \in \eta_i), \tag{3}$$

where η_i is the neighboring area of s_i.

Markov random field has a close relation to the Gibbs random field [19]. The Gibbs random field is represented as:

$$P(S = s) = \frac{1}{Z} e^{-U(s)/T}, \tag{4}$$

where $U(s)$ is the energy function, *T* is the temperature and *Z* is a constant:

$$Z = \sum_s e^{-U(s)/T}. \tag{5}$$

Based on the Gibbs distribution, we can calculate the conditional probability of Markov random fields.

In the cross-language emotion recognition many factors may influence the final recognition rate and the dependency between neighboring segments may be beneficial. Since emotions can be seen as continuous variables in the dimensional space, the changes in emotions should follow a certain probability distribution. In this paper we apply the high-order MRF to the configuration of speech emotion recognition results. The context information in the continuous speech signal is considered.

First, we analyze the emotional content in the segmented level. All of the segments form the nodes (sites) of Markov Random Fields, and the possible emotion categories corresponding to a segment are the labels to this site.

During a labeling problem, a label from the label set ***E*** is assigned to each of the sites in ***S***. Each site *s* in set ***S*** corresponds to a segment in and the label set

consists of a number of emotions for each segment. The objective of the label assignment is to find a mapping from sites set ***S*** to emotion set ***E***. The configuration model of the emotion labels can be denoted as ***F*** which contains all the possible labeling.

We then adopt a high-order MRF formulation to define the configuration model. The speech emotion configuration model is fitted to the continuous speech emotion by maximizing the posterior probability of the model for the detected emotions. The energy function of the model is composed of two terms that represent the output of the classification model and distance regularization in emotion dimension model.

The distance of the labels is defined as an Euclidian distance in the arousal-valence dimensional space.

$$d_t = \left| e_i - e_j \right|_{L2} = \sqrt{(\alpha_i - \alpha_j)^2 + (\nu_i - \nu_j)^2}, \tag{6}$$

where *t* denotes the index of an edge between two nodes, *e* is the emotion label, *a* and *v* are the coordinates of the arousal-valence dimensional space.

In the continuous speech, the emotional states are overlapped with one and another, their emotion labels should be similar. We define the energy function as:

$$E(e) = \sum_{i=1}^{N} \mu_i + \alpha \sum_{t1} d_{ij} + b \sum_{t2} d_{ij}, \tag{7}$$

where *e* is the emotion label, *i* and *j* are the indexes of the speech segments, and *N* is the total number of segments; μ denotes the averaged membership function of each segments; *t*1 is the lower constraint and *t*2 is the higher constraint of distances in the dimensional space. The weights of the function are denoted by *a* and *b*.

4. Experimental results

We carry out cross-language experiments to verify our speech emotion recognition systems on both Chinese and Vietnamese. The baseline results are shown in Tables 2 and 3. The recognition rate on Vietnamese language using Im-SFLA SVM reaches 96.5% for neutrality and has dropped to 84.1% for surprise. The features used for classifying a Vietnamese emotion are efficient. Since the data is collected in an acted way the classification task is relatively simple. We can see from Table 3 that the emotion recognition rate on the Chinese language reaches 80.5% at highest and drops to 65.4% at the lowest. This database is collected from naturalistic emotional speech, and it is relatively more difficult to classify.

We further carried out cross-language experiments using the dimensional emotional model. We mix the Vietnamese and the Chinese utterances and classify the emotions into positive and negative regions. The results are shown in Table 4. We can see that after using the high-order MRF configuration framework we may improve the recognition results. The context information in the continuous speech provides extra information for emotion detection and the configuration method is suitable for both Vietnamese and Chinese language. It is a generalized optimization framework that is not dependent on the specific language or a speaker.

Table 2. Baseline results of Vietnamese speech emotion recognition test

Test sample	Recognition rate (%)					
	Happiness	Neutrality	Sadness	Surprise	Anger	Fear
Happiness	86.3	0	1.9	3.2	4.7	3.8
Neutrality	1.1	96.5	0	2.5	0	0
Sadness	5.4	2.9	77.4	6.7	4.1	3.5
Surprise	5.5	4.0	2.8	84.1	1.4	1.2
Anger	4.1	1.2	2.6	2.7	87.6	1.8
Fear	2.7	4.9	3.3	1.2	2.9	84.9

Table 3. Baseline results of Chinese speech emotion recognition test

Test sample	Fidget	Happy	Confident	Tired	Neutral	Angry	Sad	Surprise	Fear
Fidget	**76.1**	1.1	3.0	3.1	1.4	2.9	5.8	2.5	4.1
Happy	0.2	**74.1**	3.5	2.6	5.8	3.6	1.8	3.2	5.2
Confident	3.8	4.8	**68.8**	2.5	2.8	8.4	4.5	3.0	1.4
Tired	5.5	2.0	4.5	**65.4**	9.1	2.1	4.7	1.1	5.6
Neutral	1.5	6.6	5.1	7.9	**69.7**	1.2	2.5	2.9	2.6
Angry	4.8	2.0	4.9	1.3	0.9	**80.5**	0	3.2	2.4
Sad	2.1	1.2	2.4	5.3	3.9	0	**80.2**	0.4	4.5
Surprise	2.1	8.3	3.0	1.1	5.9	3.8	1.4	**71.3**	3.1
Fear	3.5	0	1.4	8.8	3.2	3.3	9.0	4.9	**65.9**

Table 4. Cross language test results with multi-segment MRF configuration

Algorithm	Recognition rate (%)			
	Positive-Positive	Positive-Negative	Negative-Negative	Negative-Positive
Baseline	68.2	71.5	59.4	79.0
Configuration	74.9	78.5	73.2	87.5
Improvement	6.7	7.0	14.2	8.5

5. Conclusions

In this paper we study the cross-language speech emotion recognition problem. We use both Chinese and Vietnamese speech utterance to verify our recognition system. We first analyze the basic acoustic parameters. We then adopt an improved SVM classifier to classify both emotional speeches. Finally we apply the high-order MRF configuration method to improve the recognition rate in the cross-language speech emotion recognition test. The results show that we may improve the recognition rate in a cross-language test which includes different utterances in languages. The high-order MRF optimization algorithm is robust against language changes. We will further explore the global constrains in the emotion dimensional space later.

***Acknowledgment*:** The work in this paper is funded by project of Jinling Institute of Technology 2015JYJG22 and JITN201523.

References

1. Nwe, T. L., S. W. Foo, L. C. De Silva. Speech Emotion Recognition Using Hidden Markov Models. – Speech Communication, Vol. **41**, 2003, No 4, pp. 603-623.
2. Steidl, S. Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech. Ph. D. Thesis, FAU Erlangen-Nuremberg, Logos Verlag, Berlin Germany, 2012.

3. H u a n g, C., Y. Z h a o, Y. J i n, Y. Y u, L. Z h a o. A Study on Feature Analysis and Recognition for Practical Speech Emotion. – Journal of Electronics & Information Technology, Vol. **33**, 2011, No 1, pp. 112-116.
4. Z o u, C., C. H u a n g, D. H a n, L. Z h a o. Detecting Practical Speech Emotion in a Cognitive Task. – In: Proc. of 20th Computer Communications and Networks, Maui, HI, USA, 2011.
5. W u, S., T. H. F a l k, W. Y. C h a n. Automatic Speech Emotion Recognition Using Modulation Spectral Features. – Speech Communication, Vol. **53**, 2011, pp. 768-785.
6. F e r r e i r o s, J., J. M. P a r d o. Improving Continuous Speech Recognition in Spanish by Phone-Class Semi-Continuous HMMs with Pausing and Multiple Pronunciations. – Speech Communication, Vol. **29**, 1999, No 1, pp. 65-76.
7. M o h a m e d, A., K. N. N a i r. HMM/ANN Hybrid Model for Continuous Malayalam Speech Recognition. – Procedia Engineering, Vol. **30**, 2012, pp. 616-622.
8. Y o u, C. H., K. A. L e e, H. L i. GMM-SVM Kernel with a Bhattacharyya-Based Distance for Speaker Recognition. – IEEE Transactions on Audio, Speech, and Language Processing, Vol. **18**, 2010, No 6, pp. 1300-1312.
9. W u, C. H., Y. H. C h i u, C. J. S h i a. Automatic Segmentation and Identification of Mixed-Language Speech Using Delta-BIC and LSA-Based GMMs. – IEEE Transactions on Audio, Speech, and Language Processing, Vol. **14**, 2006, No 1, pp. 266-276.
10. K o c k m a n n, M., L. B u r g e t, J. H. C e r n o c k y. Application of Speaker and Language Identification State-of-the-Art Techniques for Emotion Recognition. – Speech Communication, Vol. **53**, 2011, No 9, pp. 1172-1185.
11. Y e h, J.-H., T.-L. P a o, C.-Y. L i n, Y.-W. T s a i, Y.-T. C h e n. Segment-Based Emotion Recognition from Continuous Mandarin Chinese Speech. – Computers in Human Behavior, 2011, No 27, pp. 1545-1552.
12. H u a n g, C., B. A. E f r a t y, U. K u r k u r e. Facial Landmark Configuration for Improved Detection. – In: Proc. of IEEE International Workshop on Information Forensics and Security, Tenerife, Spain, 2012, pp. 13-18.
13. B a r r e t t, L. F. Discrete Emotions or Dimensions, the Role of Valence Focus and Arousal Focus – Cognition & Emotion, Vol. **12**, 1998, No 4, pp. 579-599.
14. Z h a n g, X., C. H u a n g, L. Z h a o, C. Z o u. Recognition of Practical Speech Emotion Using Improved Shuffled Frog Leaping Algorithm. – Acta Acustica, Vol. **39**, 2014, No 2, pp. 271-280.
15. I s h i k a w a, H. Transformation of General Binary MRF Minimization to the First-Order Case. – IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, No 33, pp. 1234-1249.
16. R o t h e r, C., V. K o l m o g o r o v, V. L e m p i t s k y, M. S z u m m e r. Optimizing Binary MRFs via Extended Roof Duality. – In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, 2007, pp. 1-8.
17. A b e n d, K., T. J. H a r l e y, L. N. K a n a l. Classification of Binary Random Patterns. – IEEE Transactions on Information Theory, Vol. **11**, 1965, pp. 538-544.
18. G e m a n, S., D. G e m a n. Stochastic Relaxation Gibbs Distribution and the Bayesian Restoration of Images. – IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. **16**,1984, pp. 721-741.