ICIC Express Letters
Part B: Applications
Volume **7**, Number **1**, January **2016**

ICIC International ©2016 ISSN 2185-2766
pp. **71–76**

DIMENSIONAL SPEECH EMOTION RECOGNITION USING A GENERALIZED GAUSSIAN MIXTURE MODEL BASED CONFIGURATION FRAMEWORK

Feng Zhou[1], Yan Zhao[2], Chengwei Huang[2] and Li Zhao[2]

[1]College of Information Engineering
Yancheng Institute of Technology
No. 1, Xiwang Middle Road, Yancheng 224051, P. R. China
zfycit@163.com

[2]School of Information Science and Engineering
Southeast University
No. 2, Sipailou, Nanjing 210096, P. R. China
huangcwx@126.com; zhaoli@seu.edu.cn

Received June 2015; accepted September 2015

Abstract. *In this paper we study the dimensional speech emotion recognition using a generalized configuration framework. In many practical applications, the direct estimation of emotion dimension values is challenging, and a post process of the results may be beneficial. First, we extract and analyze the basic acoustic features for the purpose of emotion classification. Second, we make use of a dimension region recognition method to initialize the system. The regions of dimensional space are classified, and accurate regression is carried out for estimating the arousal value and the valence value. Third, we propose a generalized Gaussian Mixture Model based approach to optimize the dimension values. Based on the context information between neighboring utterances, we may correct the regression results using a pre-learned statistic model. The experimental results show a promising improvement over various testing conditions.*

Keywords: Emotion recognition, Arousal-valence model, Gaussian Mixture Model

1. Introduction.

Speech emotion recognition is an important technology for natural human-computer interaction [1, 2, 3, 4]. Several past researches have addressed the class-based emotion recognition [5, 6, 7]. However, there is still a limitation in the emotion model. The traditional emotion classifiers are mostly designed for discrete emotion models, in which we have to make the assumption on the number and the types of the target emotions [8]. This model is usually adopted in the lab environment, and it is not designed for practical applications. It is very difficult in general to estimate what types of emotions may occur in real world applications, and we have to cover a very wide range of complicated human emotions.

Previous studies on dimensional emotion recognition have shown that the direct estimation of arousal and valence coefficients is still an unsolved challenging problem [1, 9, 10]. Giannakopoulos et al. [1], adopted an emotion wheel to predict the location of speech segment. In their work, the regression was based on discrete utterances and further modelling for the dependencies between segments might be beneficial for improving the location prediction. Borchert and Dusterhoft [9], proposed to use voice quality features to recognize the valence dimension of emotion. The quality features were proved efficient in classifying different valence levels with the same arousal levels. Further investigation could be done to the continuous arousal-valence space without the precondition of fixed arousal levels. Nicolaou et al. [10], used RVM (relevance vector machine) regression to predict the dimensional values. The output patterns were learned with the input patterns. In their work, the continuous prediction was designed with a predefined temporal window.

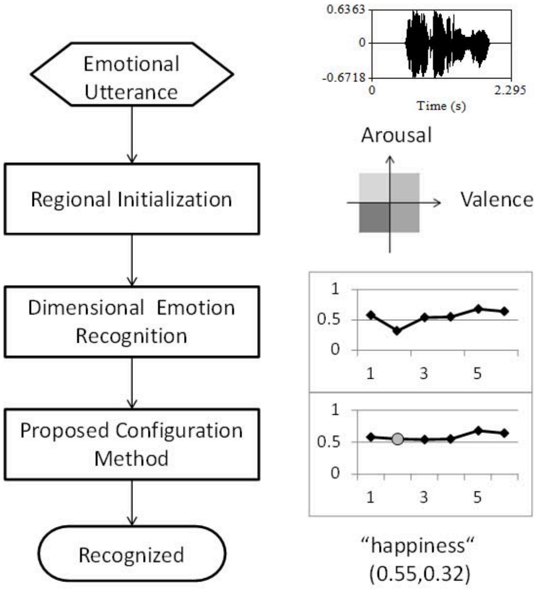

Figure 1. Configruation framework for dimensional emotion recognition

However, a pre-learned comprehensive model might be more suitable for specific speaker. Considering the above existing algorithms, a post-configuration method making use of the long time dependencies in emotional speech may improve the dimensional regression results.

In this paper we propose to use a Gaussian Mixture Model based configuration for improving the dimension regression results, as shown in Figure 1. The region recognition is used for initialization and the successive dimension regression provides the arousal value and valence value for the input utterance. In the context of continuous emotional speech signal, we may model the dependency in the dimensional space. Several configuration frameworks exist [11], in which Markov Random Fields (MRF) and Multivariate Gaussian Model are used. Preliminary experimental results are shown on image landmark configuration. However, emotion dependencies are not based on geometric shapes; therefore, they are not directly applicable to the continuous emotion configuration in speech. In this paper we proposed to use Gaussian Mixture Model to configure the context information in dimensional speech emotion. Based on the conditional probability of the model, we may improve the regression results. The constraint between neighbouring emotional utterance is considered in the configuration model. Gross errors in the continuous emotion recognition results are reduced using global optimization.

The remaining of the paper is organized as follows: in Section 2, we briefly describe the emotion features used in this paper; in Section 3, the initial stage of dimension region recognition is described; in Section 4, we give the configuration framework used for emotion recognition; in Section 5, the experimental results are provided and discussed; in Section 6, we give the conclusions of this paper.

2. **Emotional Feature Analysis.** In this section we describe the utterance level speech features that are constructed for emotion recognition.

First, basic acoustic features are extracted, including short-time energy, zero-cross rate, pitch frequency, 4 formants frequencies, and 12 Mel-frequency cepstrum coefficients. A total of 19 basic speech features are achieved.

Second, the statistical functional is applied to the basic features [2], including:

(i) maximum, minimum, mean, and standard deviation of the basic speech features;

(ii) maximum, minimum, mean, and standard deviation of the first-order difference of the basic speech features;

(iii) maximum, minimum, mean, and standard deviation of the second-order difference of the basic speech features.

A total of 228 emotion features are constructed.

Third, Principle Component Analysis (PCA) is used for reducing the feature vector dimensions. The original 228 emotion features are reduced by taking the first d PCA dimensions. d is set empirically to 9 in our experiments, which gives the best recognition rates and a low computational cost. The mean values of the first and the second PCA dimensions are shown in Figure 2 and Figure 3. We can see that the extracted features can reflect the differences among various emotion types. A total of 9 PCA dimensions are used in this paper to achieve accurate recognition of emotions and to maintain an acceptable computational cost at the same time.

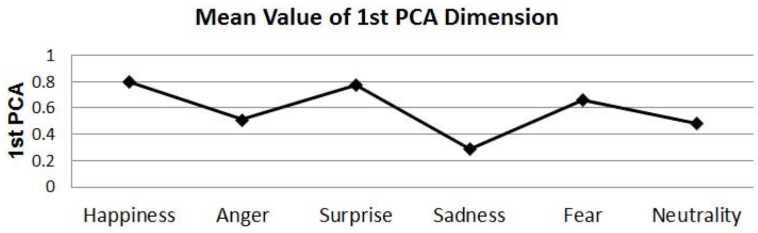

Figure 2. Distribution of various emotions on the first PCA dimension

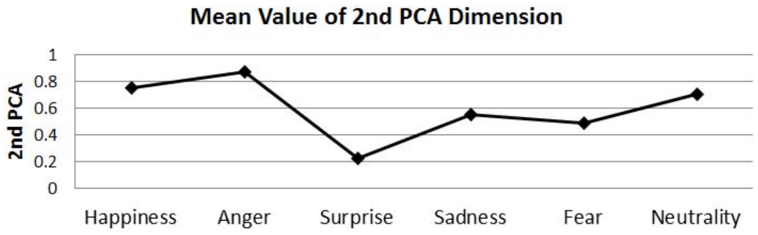

Figure 3. Distribution of various emotions on the second PCA dimension

3. **Emotion Region Recognition.** In this section we give a brief introduction to the dimension region recognition module that we used for initialization for accurate dimension regression. The input is an unknown speech sample. The output is the recognized arousal and valence region. A pre-learned model is trained using SVM-KNN (Support Vector Machine – K Nearest Neighbour) algorithm based on the annotated samples with arousal and valence labels. The dimensional space can be classified into different regions, and we use four different regions in this paper. The first region corresponds to the positive arousal dimension and the positive valence dimension, and we denote it as positive-positive for ease the notation. The second region corresponds to positive-negative, the third region corresponds to negative-negative, and the fourth region corresponds to negative-positive. The functional layout is shown in Figure 4, and detailed implementation can be found in our previous work [12].

4. **Dimensional Regression and Configuration.**

4.1. **Emotion regression.** Relevance Vector Machine (RVM) is introduced to dimensional emotion recognition by Nicolaou et al. [10]. It is very widely used in machine learning and signal processing. Emotional speech features are constructed and sent into the RVM module for emotion regression. The output vectors are then optimized using a two stage configuration framework, which will be discussed in detail in this section. The

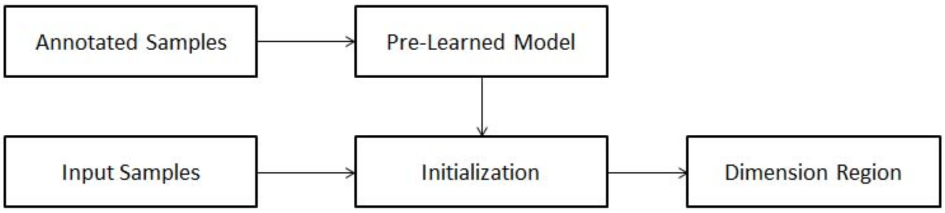

Figure 4. Flowchart of dimension region initialization

output of the RVM regression may be used for emotion prediction, and with the configuration method proposed in this paper, the errors in the prediction can be corrected using the context information based on the continuity assumption in dimensional emotion space.

The RVM regression is a Bayesian based regression algorithm. Given a multi-dimensional regression problem, the training set can be presented as: t_i, $\mathbf{x}_i$, where $\mathbf{x}_i$ is the feature vector, and the t_i is the regression output value. In the Bayesian framework, the target functional is written as [10]:

$$f = t_i - \mathbf{w}^T \phi(\mathbf{x}_i) + \epsilon_i \tag{1}$$

We assume that ϵ_i follows the normal distribution. ϕ is a non-linear projection, and $\mathbf{w}$ is the weight coefficient that can be estimated from the training dataset.

4.2. **Optimization framework.**

We propose to use a Gaussian Mixture Model based configuration method to optimize the regression result.

The emotion vector on an arousal-valence dimensional space is represented as an array of (x, y) coordinates, i.e., an n-point shape can be represented as an $n \times 2$ matrix or as a column vector $\mathbf{x} \in \mathbb{R}^{2n}$. Typically, (x, y) coordinates are normalized with respect to arousal-valence dimensional space.

We combine the Gaussian Mixture Model with a search strategy [11] for continuous optimization on the emotion vector. In each iteration, we fuse a predicted distribution based on the conditional distribution from the emotion vector model, we can predict any emotion vector distribution, given a current emotion vector.

Based on the ranking of emotion vectors, m top reliable emotion vector points are fixed according to the corresponding outputs. These landmarks are fixed in each iteration to predict the distribution of the other emotion vectors by the pre-trained Gaussian Mixture

Algorithm 1 Configuration optimization based on Gaussian Mixture Model

```
Require: Gaussian Mixture Model parameters $(a_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, Emotion vector, $\boldsymbol{e}$
Ensure: Emotion vector after configuration, $\boldsymbol{e}^*$
Rank the emotion regression result $f(x_j, y_j)$
for all $t = 1, \ldots, T$ do
    Fix $\lfloor m \rfloor$ points using rankings $f(x_j, y_j)$
    Predict the conditional emotion vector values $P(\boldsymbol{e}_k|\boldsymbol{e}_j)$ using a conditional Gaussian Mixture Model.
    Fuse prediction and the original result:
    $P(\boldsymbol{e}_k) = w \times P(\boldsymbol{e}_k|\boldsymbol{e}_j) + (1 - w) \times C(\boldsymbol{e}_k)$, where $w$ is the empirical weight, and $C(\boldsymbol{e}_k)$ is the confidence score.
    The emotion vectors are selected by maximizing the fused distribution: $\boldsymbol{e}_k = \arg\max\{P(\boldsymbol{e}_k)\}$.
    Update the fixed emotion points and the fusion weight $w$: $m = m+\delta_m$, $w = w+\delta_w$, where $\delta_m$ is the update step of emotion points, and $\delta_w$ is the update step of fusion weight.
end for
```

Model. After the prediction we fuse two channels of information: the prediction from the global trace information and the confidence score from the local emotion detector. The weighted-sum rule is adopted for fusion. At the end of each iteration, the emotion vectors are selected by maximizing the fused distribution: $\boldsymbol{e}_k = \arg\max\{P(\boldsymbol{e}_k)\}$. By gradually making better predictions, the uncertainty is reduced using the Gaussian Mixture Model, as shown in Algorithm 1.

5. **Experimental Result.** In order to verify the effectiveness of the proposed configuration algorithm based on Gaussian Mixture Model, we compared it with another two methods. The Naive Bayes method (NB) [13] and Markov Random Fields (MRF) [11] are adopted for comparison. We use them to learn the dependency between neighboring utterances. The dimension regression model is improved in an empirical way. The MRF based configuration tries to find the global optimal using energy based model, which is defined on the neighboring algorithms. Both of NB and MRF are context dependent techniques for speech emotion detection.

The dataset used in the experiments consists of Chinese emotional speech collected locally in our lab. A subset of 2,020 utterances, which are manually annotated with arousal values and valence values, are used for the emotion detection and configuration. The verification experiments are designed for the basic emotion region recognition, the accurate dimension regression, and the improved configuration.

In our previous study [12], the emotion region recognition is implemented for the classification of arousal and valence dimensional space. We adopt this method for the initialization in emotion dimension regression. The regression result of a testing utterance is considered successful if the mean error is smaller than a certain percentage of the ground-truth that is manually annotated. The success rates of the dimension regression are shown in Table 1. We can see that as the defined percentage of error distance increases, the corresponding successful rate also increases. The training-testing ratio corresponds to the size of the training set and the size of the testing set, which also impacts the final results.

Further experiments are carried out using the configuration algorithm (GMM) proposed in this paper, and the other two existing algorithms (NB, MRF) for comparison. Results

Table 1. Dimensional recognition results without configuration method

Error percentage (%)	Training-testing ratio	Defined success rate (%)		
		Arousal	Valence	Average rate
5	3:1	51.1	49.3	50.5
10	3:1	61.2	57.7	59.5
15	3:1	65.5	62.7	64.1
5	5:1	57.7	53.4	55.6
10	5:1	64.9	61.1	63.0
15	5:1	69.3	65.7	67.5

Table 2. Dimensional recognition results using various configuration methods

Error percentage (%)	Training-testing ratio	Defined success rate (%)		
		NB	MRF	GMM
5	3:1	53.1	54.2	55.6
10	3:1	62.9	61.8	63.2
15	3:1	65.2	66.4	68.9
5	5:1	57.3	58.2	59.3
10	5:1	64.5	65.5	67.7
15	5:1	69.0	70.2	72.1

are shown in Table 2, and we can see that the three configuration algorithms improve the regression results constantly, compared with the results in Table 1. The MRF method is slightly better than NB, since the MRF can model higher order context dependents. The GMM-based configuration method is able to model the various error statistic data without specific assumption on the distribution, and the proposed algorithm gives the best improvements.

6. **Conclusions.** In this paper we analyze the challenges in dimensional recognition of speech emotion. The context dependencies may contribute to the accurate estimation of dimensional values. Past researches have investigated various regression methods and a separate configuration method may further improve the results. We propose to adopt a generalized configuration framework to improve the existing recognition system. The experimental results show that after applying the GMM-based configuration model the final recognition rates are improved. In future work, we will explore the possibilities of extending the configuration framework to cross-database emotion recognition.

Acknowledgement. This work is partially supported by National Natural Science Foundation of China (No. 61231002; No. 61273266; No. 51075068) and Doctoral Fund of Ministry of Education of China (No. 20110092130004).

REFERENCES

[1] T. Giannakopoulos, A. Pikrakis and S. Theodoridis, A dimensional approach to emotion recognition of speech from movies, *Proc. of IEEE Interantional Conference on Acoustics, Speech and Signal Processing*, pp.65-68, 2009.

[2] C. Huang, Y. Zhao, Y. Jin, Y. Yu and L. Zhao, A study on feature analysis and recognition for practical speech emotion, *Journal of Electronics & Information Technology*, vol.33, no.1, pp.112-116, 2011.

[3] C. Huang, R. Liang, Q. Wang, J. Xi, C. Zha and L. Zhao, Practical speech emotion recognition based on online learning: From acted data to elicited data, *Mathematical Problems in Engineering*, pp.1-8, 2013.

[4] J. Wang, B. Wu, C. Huang, H. Qin, C. Zha and L. Zhao, Segment-based static feature analysis and recognition of emotional speech for manned space mission, *ICIC Express Letters*, vol.8, no.6, pp.1541-1546, 2014.

[5] T. L. Nwe, S. W. Foo and L. C. De Silva, Speech emotion recognition using hidden Markov models, *Speech Communication*, vol.41, no.4, pp.603-623, 2003.

[6] C. Huang, D. Han, Y. Bao, H. Yu and L. Zhao, Cross-language speech emotion recognition in German and Chinese, *ICIC Express Letters*, vol.6, no.8, pp.2141-2146, 2012.

[7] S. Wu, T. H. Falk and W. Y. Chan, Automatic speech emotion recognition using modulation spectral features, *Speech Communication*, vol.53, pp.768-785, 2011.

[8] S. Steidl, *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech*, Ph.D. Thesis, FAU Erlangen-Nuremberg, Logos Verlag, Berlin, Germany, 2009.

[9] M. Borchert and A. Dusterhoft, Emotions in speech-experiments with prosody and quality features in speech for use in categorical and dimensional emotion recognition environments, *Proc. of IEEE International Conference on Natural Language Processing and Knowledge Engineering*, pp.147-151, 2005.

[10] M. A. Nicolaou, H. Gunes and M. Pantic, Output-associative rvm regression for dimensional and continuous emotion prediction, *Image and Vision Computing*, vol.30, no.3, pp.186-196, 2012.

[11] C. Huang, B. Efraty, U. Kurkure, M. Papadakis, S. K. Shah and I. A. Kakadiaris, Facial landmark configuration for improved detection, *IEEE International Workshop on Information Forensics and Security*, pp.13-18, 2012.

[12] F. Zhou, C. Huang and L. Zhao, Dimensional speech emotion region recognition based on decomposition of feature space, *ICIC Express Letters, Part B: Applications*, vol.6, no.11, pp.3029-3034, 2015.

[13] I. Rish, An empirical study of the naive Bayes classifier, *International Joint Conference on Artificial Intelligence, Workshop on Empirical Methods in Artificial Intelligence*, vol.3, no.22, pp.41-46, 2001.