Speech Emotion Recognition based on Optimized Support Vector Machine

Bo Yu[1, 2]
[1]School of Computer Science and Technology/Harbin Institute of Technology, Harbin, China
[2]Software College/Harbin University of Science and Technology, Harbin, China
Email: hrbust_yubo1981@163.com

Haifeng Li and Chunying Fang
School of Computer Science and Technology/Harbin Institute of Technology, Harbin, China
Email: lihaifeng@hit.edu.cn, fcy3333@163.com

***Abstract*—Speech emotion recognition is a very important speech technology. In this paper, Mel Frequency Cepstral Coefficients (MFCC) has been used to represent speech signal as emotional features. MFCCs plus energy of an utterance are used as the input for Support Vector Machine. Support Vector Machine (SVM) has been profoundly successful in the area of pattern recognition. In the recent years there has been use of SVM for speech recognition. Many kinds of kernel functions are available for SVM to map an input space problem to high dimensional spaces. We lack guidelines on choosing a better kernel with optimized parameters of SVM. Some kernels are better for some questions, but worse for other questions. Which is better is unknown for speech emotion recognition, thus the thesis studies the SVM classifier and proposes methods used to select a better kernel with optimized parameters. The new method we proposed in this paper can more efficiently gain optimized parameters than common methods. In order to improve recognition accuracy rate of the speech emotion recognition system, a speech emotion recognition based on optimized support vector machine is proposed. Experimental studies are performed over the HIT Emotional Speech Database established by Speech Processing Lab in School of Computer Science and Technology at HIT. The experiment result shows that the speech emotion recognition based on optimized SVM can improve the performance of the emotion recognition system effectively.**

***Index Terms*—speech emotion recognition, MFCC optimized SVM, kernel function**

I. Introduction

Recognizing emotional state of a person from the speech signal has been increasingly important, especially in natural human-computer interaction[1]. Speech emotion recognition can be used in wide range of applications, such as remote call customer services center[2-5], speech emotion network communication system[6], improving the robustness of speech recognition[7], emotion states detection in voice mail messages[8], speech emotion recognition system in web learning[9], interactive movies[10], monitoring a driver's emotion and ensuring save drive[11] and so on. Therefore speech emotion recognition has important research value and great potential for development.

Global statistical prosodic and voice quality features have been broadly used in speech emotion recognition and gained great success. Besides the global statistical prosodic and voice quality features, spectral features are also useful features for describing speech emotion signal, such as Mel Frequency Cepstral Coefficients (MFCC). MFCC of speech signal have already been successfully used for the features of speech signals for emotion recognition. Therefore, we use MFCC plus energy with their delta and acceleration as speech emotion features.

Hidden Markov model (HMM) and Gaussian mixture model (GMM) using MFCC have achieved valuable results on speech emotion recognition[12]. However, there is a problem when using GMM to recognize speech emotion states. Effective training of GMM requires a great deal of data, while collecting emotional speech utterances will cost a lot and therefore the available training data is usually scant.

SVM has a better classification performance on a small amount of training samples. But we are lacking in guidelines on choosing a better kernel with optimized parameters of SVM. Some kernels are better for some questions, but worse for other questions. There is no uniform pattern used to the choice of SVM with its parameters and kernel function with its parameters. The paper proposed methods about selecting optimized parameters and kernel function of SVM.

The paper is organized as follows. In section 2, we give a brief description about the process of speech emotion feature extraction. Section 3 includes three works: establishing emotion recognition model based on optimized SVM, studying support vector machine classification, proposing the method for optimizing SVM. Some experimental data and results with our analysis are shown in section 4. Finally the conclusion and future work are given in section 5.

II. Experiment Data Acquisition And Feature Extraction

doi:10.4304/jsw.7.12.2726-2733

A. Speech Emotional Database Description

Referring to the domestic and foreign research, this paper divides emotion into four categories——anger, happiness, sadness, and surprise, and tries to include all kinds of feelings in them. In order to obtain experiment utterances, some non-professionals have been invited to record their emotions, thus creating an emotional database. The design of the experiment is speaker-independent and gender-independent, thus students who took part in the experiment aged about 20 include 5 males and 9 females. Recorded with Cool Edit Pro 2.0, all the data are with the technology of sampling rate of 16 kHz, a single channel audio tape recorder, 16-bit quantization, and are recorded in PC with the form of Wave. Besides, we also invite another two groups of people who were not engaged in the recording to make distinguishing experiment. Each group involved several people. The first group's task was to distinguish the emotions, thus getting rid of the unqualified recording and selecting 1256 sentences as emotion utterances stored for later usage in the experiment. Another group of people tries to tell the differences in the 1256 sentences, which can make a subjunctive evaluation of the emotion utterances in the experiment. Four types of emotions (anger, happiness, sadness, surprise) are respectively labeled by 1,2,3,4. Fig. 1 shows the emotion class distribution of 1256 samples. The 1256 items consists of 636 sentences from the males, and 620 sentences from the females. Each of the four main emotions has 300 sentences accordingly and has even distribution.

B. Speech Emotional Features Extraction

Speech emotional features extraction is a process extracting a small number of parameters from the speech signal that can be later used to represent each utterance. Speech extraction techniques include temporal analysis and spectral analysis techniques. The waveform of speech signal is used for analysis in temporal. The spectral form of speech signal is used for analysis in spectral analysis. Mel Frequency Cepstral Coefficients is a spectral analysis technique. In the recent years, MFCC feature has been widely used for not only the speaker but also speech recognition.

Mel Frequency Cepstral Coefficients are set of features reported to be robust in various kinds of pattern classification tasks in speech signal. It has been proven that human being perception of the frequency contents of sounds for speech signals does not follow a linear scale in the psychology studies. Therefore for each tone with an actual frequency f measured in Hz, a subjective pitch is measured on a scale called the Mel scale [14]. The Mel frequency scale is in the form of a linear frequency scale below 1000 Hz while a logarithmic scale above 1000 Hz. Consequently we can take advantage of the following formula to compute the Mel for a given frequency f (Hz).

$$f_{Mel} = 2595 \times \log(1 + f / 700). \quad (1)$$

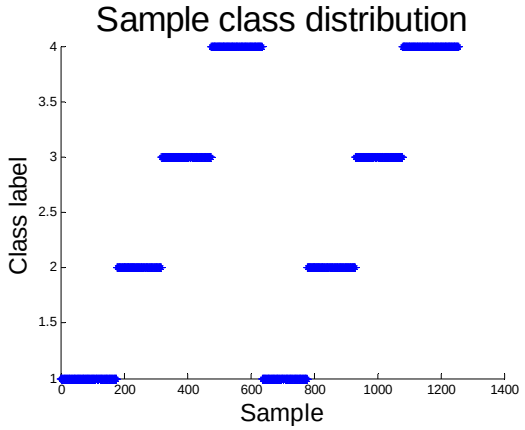

Figure 1. The distribution of four kinds of emotion.

Fig. 2 shows the MFCC extraction Algorithm. An MFCC process converts linear spectrum into nonlinear Mel-spectrum.

1). Pre-emphasis

In our system, each of the utterances is sampled by 16 kHz. Pre-emphasizing the sampled speech signals with filter is the first process in feature extraction. The purpose of pre-emphasis is to spectrally flatten the signal. The z-transform of the filter is

$$H(z) = 1 - \mu z^{-1}, \quad 0.94 < \mu < 0.97. \quad (2)$$

2). Frame-blocking

The pre-emphasized speech signal is then blocked into frames of N sample points with adjacent frames being separated by M (lower than N). The first frame is composed of the first N sample points. The second frames begin the Mth sample points after first frame and overlaps it by N-M sample points and so on. This process continues till all are accommodated within one or more frames. In our work, the frame length N = 256(16ms). There is 8ms (N-M=128) overlap between two adjacent frames to ensure stationary between frames.

3). Windowing

Every frame is multiplied by Hamming window function with N sample points. This process is to window every single frame in order to minimize the spectral distortion and to reduce edge effect at the beginning and end of every frame. This minimizes spectral distortion to taper the signal to zero. The mathematical expression of the Hamming window is

$$w(n) = 0.54 - 0.46\cos(\frac{2\pi n}{N-1}), 0 \le n \le N-1. \quad (3)$$

N represents the number of samples in each frame.

4). FFT

The next step is Fast Fourier Transform(FFT) converting each frame of N samples from data sequence to frequency spectrum. FFT is a fast algorithm to implement DFT(Discrete Fourier Transform) which is defined on the set of N samples. FFT can be calculated as

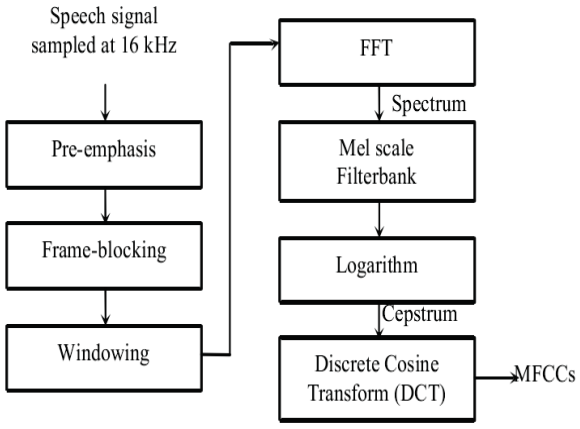

Figure 2. MFCC extraction algorithm.

$$X_a(k) = \sum_{n=0}^{N-1} x(n) e^{-j2\pi nk/N} \quad 0 \le k < N \quad (4)$$

5). Mel scale filterbanks

After the FFT processing, the frequency spectrum of each frame is filtered by a group of filters, and the power of each filter band is computed. A filter bank which spaced uniformly on the Mel scale is used to simulate the subjective spectrum. Filter banks filter the magnitude spectrum into a number of bands. Low frequencies are given more weight than high frequencies using triangular overlapping windows and sum the frequency contents of each band. The process reflects the selectivity of human ear [15].

6). Logarithm

This operation simulates the perception of loudness. We can calculate the Mel Frequency Cepstral Coefficients from the output power of the filter bank using logarithm arithmetic. The operation is mapping the logarithmic amplitudes of the spectrum obtained the Mel scale as we mention above.

7). Discrete Cosine Transform

Discrete Cosine Transform (DCT) can convert log-power spectrums to the time domain, for the Mel Frequency Cepstral coefficients are real numbers. After the DCT operation, we get a featured vector with 12 dimensional MFCC.

After computing MFCC, we calculate Logarithmic energy of each frame as one of the coefficients. Up to now we have got 13 dimensional coefficients vector consisting of 12 cepstral coefficients and one energy. To enhance the performance of the speech emotion recognition, the delta and acceleration coefficients are computed. After all the calculations, the total number of MFCC of one frame is 39. In order to predict emotion label of one sentence, we calculate the mean value of all frames of one sentence. Therefore, we get 39-dimensional feature vector of one sample.

III. Speech Emotion Recognition Based On Optimized SVM

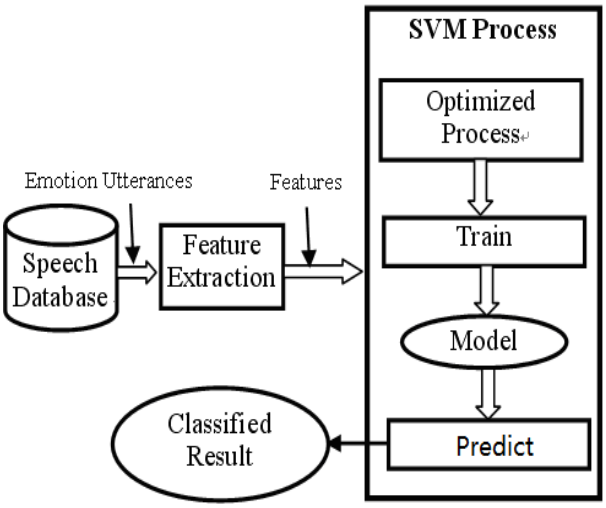

Figure 3. System model architecture.

A. System Model Architecture

Fig. 3 shows the speech emotion recognition system model architecture based on optimized SVM. The process of the system is as follows:

STEP1: Extracting speech emotion feature from 1256 utterances, the extracting features method referred section II B.

STEP2: The main task in optimized process is to improve the classification accuracy rate of the SVM. The main method and Algorithm is studied in section III C.

STEP3: After optimizing process, the system trains an optimized model used to classify.

STEP4: The system gives a classification result (class label or recognition rate) about test samples.

The principle of SVM Method is studied in next part.

B. Support Vector Machine Classification Method

Instead of using empirical risk minimization (ERM), which is commonly used in statistical learning, SVM is established on structural risk minimization (SRM). ERM only minimises an upper bound on the generalization error. Thus SVM generalize well. The major principle of SVM is to establish a hyperplane as the decision surface maximizing the margin of separation between negative and positive samples. Thus SVM is designed for two-class pattern classification. Multiple pattern classification problems can be solved using a combination of binary support vector machines.

1). Linear classification

Fig. 4 gives the idea of an optimal hyperplane for linearly classification. Triangular and square points represent the two types of training sample. H represents the hyperplane. B_1 and B_2 respectively go through the points which are the closest point of H, and parallel to H. The distance between B_1 and B_2 is called margin, which is a measurement of the expected generalization ability.

There are many linear hyperplanes that separate the samples. However only one of these achieves maximum margin. An optimal hyperplane tries to achieve maximum margin between the classes. Using a large margin to

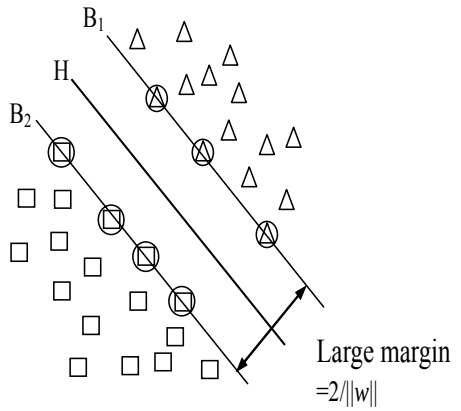

Figure 4. The idea of an optimal hyperplane for linearly classification.

separate the classes minimizes the boundary of expected generalization error.

Given a linearly Separable set of training samples (x_i, y_i), i =1, 2…N, where $x_i \in R^d$ (d is the dimension of sample space) is the real world data instances and $y_i \in \{1,-1\}$ represent the class label of x_i. A hyperplane can be represented as

$$f(x) = w \cdot x + b = 0 . \quad (1)$$

where x represents an input vector, w represents an adjustable weight vector, and b represents a bias. Given a point x, if f(x)>0, the point belongs to class 1, if f(x) < 0 the points belongs to class 2, that is f(x,w,b) = sign(w · x+b). Let the f(x) normalized so that all samples meet $|f(x)| \geq 1$. The closest vector to H that satisfy the requirement |f(x)| = 1 is called support vector. Triangular and square points with a circle represent the support vectors for classification. The closest vector to H meet |f(x)| = 1. Thus the margin equals to 2/||w||. It suggests that maximizing the margin of separation between classes is equal to minimizing the Euclidean norm of the weight vector w. Thus the classification problem can be transformed to a constrained optimization problem.

$$\begin{aligned} &\min \quad \Phi(w) = \frac{1}{2} \| w \|^2 \\ &\textit{subject to} \quad y_i(w \cdot x_i + b) - 1 \geq 0, i = 1, 2, ...N \end{aligned} \quad (2)$$

We can use the method of Lagrange function to settle the primal problem. The corresponding Lagrange function is as follows.

$$J(w, b, \partial) = \frac{1}{2} w \cdot w - \sum_{i=1}^{N} \partial_i \{ y_i (w \cdot x_i + b) - 1 \} \quad (3)$$

Where ∂_i is Lagrange multiplier. Saddle points decide the solution. After a series of transformation, we can get the following dual problem.

$$\begin{aligned} &\max \quad Q(\partial) = \sum_{i=1}^{N} \partial_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \partial_i \partial_j y_i y_j x_i \cdot x_j \\ &\text{where} \quad \sum_{i}^{n} \begin{cases} \partial_i y_i = 0 \\ \partial_i \geq 0 \end{cases}, \qquad i = 1, 2, ..., N \end{aligned} \quad (4)$$

Equation 5 can be solved by standard quadratic programming method. We can get the optimum Lagrange multipliers which is denoted by ∂_i^*. Finally, we can compute the optimum weight vector w^*.

$$w^* = \sum_{i=1}^{N} \partial_i^* y_i x_i \quad (5)$$

To compute the optimum bias b^*, we may use the w^* and a positive support vector x_i (y_i =1).

$$b^* = 1 - w^* \cdot x_i \quad (6)$$

Consequently, the optimal hyperplane, standing for a multidimensional linear decision surface in the input space, is defined by

$$w^* \cdot x + b^* = 0 . \quad (7)$$

Accordingly, the decision function f(x) = sign($w^* \cdot x+b^*$) can decide the label of the new sample. We will study the Optimal Hyperplane for nonseparable patterns in the next part.

2). Optimal Hyperplane for nonseparable patterns

The case we discuss above is hard-margin classification, that is, no sample points are allowed to be mislabeled. Actually, we cannot exclude the situation where exist some noisy points. That is, not all the points satisfy the constraints of the hyperplane. We solve this problem by introducing slack variables. The corresponding optimization problem is as follows.

$$\begin{aligned} &\min \quad \Phi(w, \xi) = \frac{1}{2} \| w \|^2 + c \sum_{i=1}^{N} \xi_i \\ &\text{subject to} \quad \begin{cases} y_i(w \cdot x_i + b) - 1 + \xi_i \geq 0 \\ \xi_i \geq 0, \qquad i = 1, 2, ...N \end{cases} \end{aligned} \quad (8)$$

Where ξ_i is a slack variable. C stands for the penalty imposed on mislabeling the point. The bigger the value of C is, the less likely the SVM model mislabels the point. However, C with high value will lead to overfitting study. The parameter C controls the tradeoff between complexity of the machine and the number of nonseparable points. The parameter C is determined experimentally via the standard use of a training/test set [16].

3). Kernel function

Not all training data is linearly separable. In order to handle the situation, kernel is introduced. SVM performs a non-linear mapping from a low-dimensional space to a high-dimensional space through a kernel. As a result, the training samples are not linearly separable in a low-dimensional space while the training samples are linearly separable in the feature space.

The main principle of the kernel function is to let operations be performed in the low-dimensional space instead of the potentially high dimensional feature space. Thus the inner product need not be computed in the feature space. According the Mercer's theory, there

always exists a kernel function K which satisfies the requirement below.

$$K(x_i, x) = \varphi(x) \cdot \varphi(x_i). \quad (9)$$

The vector $\varphi(x)$ represents the "image" induced in the feature space due to the input vector x. Therefore the inner product can be evaluated using the kernel function. In this way, we are able to reduce the scale of the problem. The dual problem is redefined as

$$\max \quad Q(\partial) = \sum_{i=1}^{N} \partial_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \partial_i \partial_j y_i y_j k(x_i, x_j)$$
$$\text{where } \sum_{i}^{n} \partial_i y_i = 0 \text{ and } \partial_i \geq 0 \quad i = 1, 2, ..., N \quad (10)$$

We compute the optimum weight vector w^* and the bias.

$$w^* = \sum_{i=1}^{N} \partial_i^* y_i \varphi(x_i)$$
$$b^* = 1 - w^* \cdot x_i \quad (11)$$

Finally, a classifier based on the support vectors will be

$$f(x) = sign(\sum_{i=1}^{N} \alpha_i y_i k(x_i, x) + b). \quad (12)$$

In our system, we select the following four kernel functions as learning machines:

(1) Linear function

It has the simplest formula form, which is given by the inner product of two vectors in low-dimensional space plus an optional constant COEF.

$$K(x_i, x_j) = x_i \cdot x_j + COEF \quad (13)$$

(2) Polynomial function

It is a unstable kernel. It is fit for problems in which all the training samples are normalized. Adjustable parameters are the slope γ, the constant term COEF and the polynomial degree d.

$$K(x_i, x_j) = (\gamma x_i \cdot x_j + COEF)^d, \quad \gamma > 0 \quad (14)$$

(3) Radial basis function (RBF)

Its common form is a Gaussian form. The adjustable parameterγplays a major role in the performance of the kernel, and should be carefully adjusted to the specific problem.

$$K(x_i, x_j) = \exp\left[-\gamma \left\| x_i - x_j \right\|^2 \right], \gamma > 0 \quad (15)$$

(4) Sigmoid function

It is also known as Multi-layer Perceptron (MLP) kernel. A SVM using a sigmoid function as kernel is equal to a two-layer perceptron neural network. The adjustable parameters in the sigmoid kernel are the slope γ and the intercept constant COEF. A common value for γ is 1/N, where N is the data dimension.

$$K(x_i, x_j) = \tanh(\gamma x_i \cdot x_j + COEF). \quad (16)$$

C. Optimizing SVM Method

In our work, we propose the method of optimzed SVM including selection of a kernel function and kernel parameters. We will compare different typical kernels with different parameters and select a better one to do the speech emotion recognition job. The system uses K-fold Cross Validation (K-CV) Algorithm to select kernel function parameters. The main idea of K-CV Algorithm with grid search in pseudo code form to decide two optimized parameters is as follows. The process of deciding more than two parameters is similar.

FOR(parameter1 = begin1; parameter1<end1; parameterr1 += step1)
BEGIN
FOR(parameter2 = begin2; parameter2<end2; parameterr2 += step2)
BEGIN
STEP1: Disorder randomly the training set to lose their order;
STEP2: Divide the disordered training set into K subsets evenly;
STEP3:
FOR i = 1 to K
BEGIN
Each subset does as a test data set;
The rest of K-1 subsets of data as training set respectively;
It would receive the K model according to the values of parameter1 and parameter2;
Every model will give a classification label of respective test subset;
END.
STEP4: The average of classification accuracy rate of k subset is the performance of the classifier;
END.
END.

In our work, we firstly use the algorithm with the parameters varying in a bigger scope and step. And then we can get a smaller scope in which parameters center to gain high recognition accuracy rate. Using the algorithm again with the parameters varying in a smaller scope and step will get another scope. Following the same step will get optimized parameters for SVM. To cover the search scope, the next search step is half of the current step. The bound value of next search scope is

$$Bound_{next} = Bound_{current} \pm n * Step_{current}$$
$$n = 1, 2, 3, 4..... \quad (17)$$

The bound of current plus or minus $n*step_{current}$ and the value of n according to the current smaller scope which parameters center to gain high recognition accuracy rate. The result of experiments will be given in section IV.

IV. Speech Emotion Recognition Experiments And Results Analysis

We perform experiments on the HIT speech database referred in section II by Matlab 7.11. The objective of these experiments is to find an optimized SVM to improve the speech emotion recognition accuracy rate. Firstly 1256 utterances' emotion features are extracted. The spectral feature of each utterance is 12 MFCC and 1 energy of a frame, together with their delta and acceleration. We use LibSVM v3.1 [17] as the implementation of support vector machine classification.

A. Selection of Optimized Parameters

We use RBF kernel with different parameters to perform a selection of optimized parameters experiments. A SVM model with RBF usually has two adjustable parameters — g (γin Gaussian function) and C (penalty parameter) . The scope of g and C is from 2^{-10} to 2^{10}, and the step of g and C is 1.2. 5 fold cross validation is performed for parameters selection. Fig. 5 and Fig. 6 shows the 5-VC parameter selection result 3-Dimensional view and in contour [18]. Observing that parameter g and c varying in a smaller scope have higher recognition accuracy rate, we can narrow the scope of grid search and the search step. Fig. 7 shows parameter selection result in contour form with smaller search scope. Parameters C and g centers in smaller scope, but they have a high recognition accuracy rate. If several groups of g and c correspond to the same recognition accuracy rate, the method will choose the group with smaller c. The reason for that c with high value will lead to overfitting study. Table I shows optimized c and g in different scope grid search and the search step and gives the recognition accuracy rate of 80 test utterances in corresponding parameters. From the results experiment (NO.1~4) using our methods referred in section III C, we can see that with the decrease of search scope and search step, the recognition accuracy rate of train set is improving. The recognition accuracy rate of test set is high. Through four experiments we gain the optimized parameters of C and g with the highest recognition accuracy rate of train set, that is C=2.3784 and g=0.0057191. We compare our methods with common methods having a bigger search scope and lower search step. Experiment NO 5* shows the result of common method. We can also gain the optimized parameters same to the result of NO 4. However, the total runtime of NO 5* is 38027.12s far greater than the runtime of NO.1~4 (606.93 +333.18+ 300.16+ 331.76 = 1572.03s). Experiments show that the new method can avoid unnecessary process of optimizing parameters, reduce sharply the time complexity of optimizing parameters, and improve the speed of training and classification. The method is especially useful for training of very large samples. It also maintains the recognition accuracy rate at the same time.

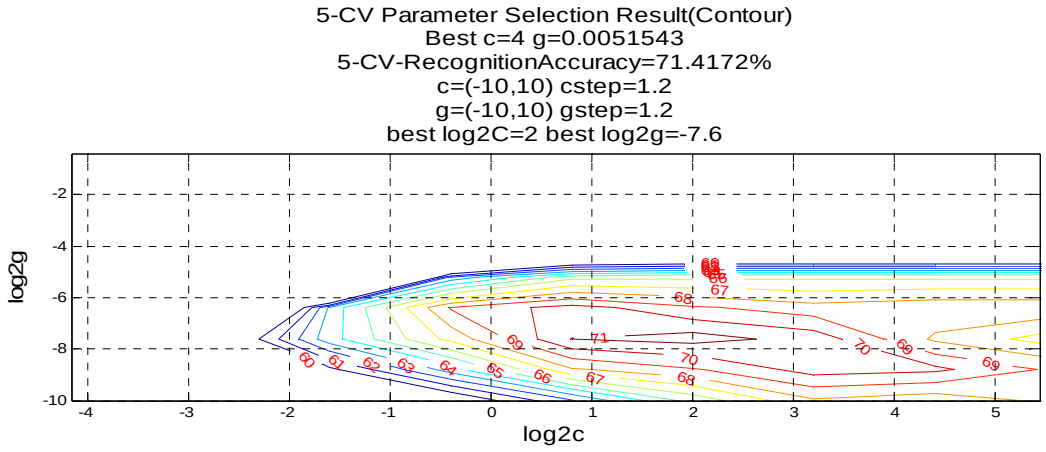

Figure 5. Parameter selection result in contour form.

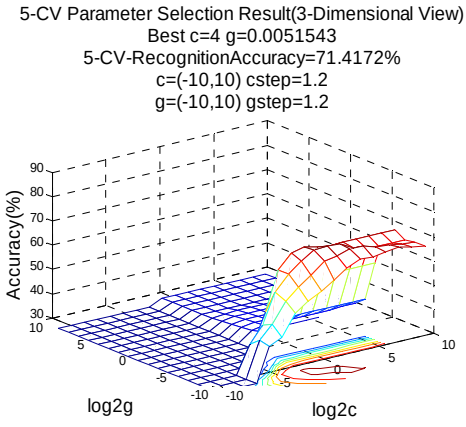

Figure 6. Parameter selection result in 3-Dimensional view.

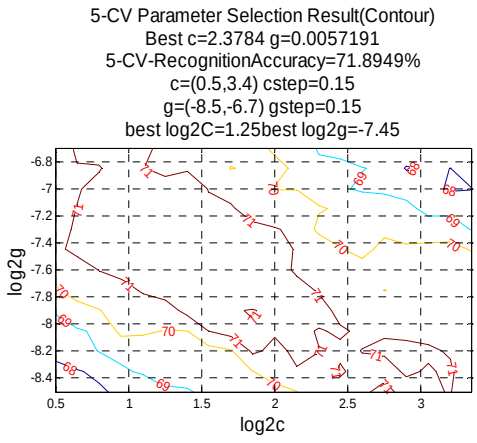

Figure 7. Contour with smaller search scope.

TABLE I.
OPTIMIZED PARAMETERS AND RECOGNITION ACCURACY

NO.	Parameter	Scope	Step	Optimized Values	Recognition Accuracy Rate		Runtime(second)
					Train set	Test set	
1	C	2^{-10}~2^{10}	1.2	4	71.4172%	92.5%	606.93
	g	2^{-10}~2^{10}	1.2	0.0051543			
2	C	$2^{-2.8}$~$2^{8.8}$	0.6	2.639	71.7357 %	95%	333.18
	g	2^{-10}~2^{-4}	0.6	0.0078125			
3	C	$2^{-0.4}$~$2^{4.6}$	0.3	2.1435	71.7357%	90%	300.16
	g	$2^{-9.4}$~$2^{-5.8}$	0.3	0.0063457			
4	C	$2^{0.5}$~$2^{3.4}$	0.15	2.3784	71.8949 %	88.75%	331.76
	g	$2^{-8.5}$~$2^{-6.7}$	0.15	0.0057191			
5*	C	2^{-10}~2^{10}	0.15	2.3784	71.8949 %	88.75%	38027.12
	g	2^{-10}~2^{10}	0.15	0.0057191			

B. Selection of Kernel Function

Table II shows the accuracy rate of speech emotion recognition based on optimized SVM using four different kernels referred in section III B. The method of selection optimized parameters of other kernels is same to RBF. It can be seen that the RBF kernel has the best performance in recognition accuracy of train set and test set.

TABLE II.
RECOGNITION ACCURACY OF SPEECH EMOTION USING DIFFERENT KERNELS

Kernel	Optimized Parameters	Recognition Accuracy	
		Train set	Test set
Linear	COEF = 0 C = 2.3784	61.305%	66.25%
Polynomial	COEF = 0 C = 0.056328 d = 3 g = 0.0046453 (γ in function)	68.1529%	82.5%
RBF	C = 2.3784 g = 0.0057191	71.8949 %	88.75%
Sigmoid	COEF = 0 C = 0.4278 g = 0.00097656 (γ in function)	48.8854%	48.75%

V. CONCLUSIONS

In this paper, we propose methods about selecting optimized parameters and kernel function of SVM and establish a speech emotion model based on optimized SVM. Support machine vector is studied as emotion classification. Mel Frequency Cepstral Coefficients plus energy with their delta and acceleration as speech emotion features are utilized as the input of SVM. Experiments show that the method of selecting optimized parameters not only sharply reduces the time complexity compared with common method, but also maintains the recognition accuracy rate at the same time. Four kernels contrast experiments show that RBF kernel has better performance in speech emotion recognition than other kernels. Based on selection optimized parameters and RBF kernels, we gain an optimized SVM improving the performance of the emotion recognition system effectively. Other methods to optimize SVM will be studied in future work.

ACKNOWLEDGMENT

We thank Speech Processing Lab of HIT for supporting this work. The work presented in this paper is supported by the National Natural Science Foundation of China under Grant No. 61171186, 60772076 and Project (HIT.KLOF.2009015) Supported by Key Laboratory Opening Funding of MOE-Microsoft Key Laboratory of Natural Language Processing and Speech. The authors are grateful for the anonymous reviewers who made constructive comments.

REFERENCES

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias and et al, "Emotion Recognition in Human-Computer Interaction," *IEEE Signal Processing Magazine. America*, vol. 18, pp. 32-80, January 2001.

[2] L. Devillers, C. Vaudable, and C. Chastagnol, "Real-life emotion-related states detection in call centers: a cross-corpora study," *in Proc. INTERSPEECH 2010. Chiba*, pp. 2350-2353, September 2010.

[3] D. Morrison and R. Wang, LC, "De Silva. Ensemble methods for spoken emotion recognition in call-centers. Speech Communication," *Speech Communication. Amsterdam*, vol. 49, pp. 98-112, February 2007.

[4] A. Batliner, K. Fischer, R. Huber, J. Spilker and E. Noth, "How to find trouble in communication," *Speech Communication. Amsterdam*, vol. 40, pp. 117-143, April 2003.

[5] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Trans. Speech and Audio Processing. America*, vol. 13, pp. 293-303, March 2005.

[6] Scherer KR, "Vocal communication of emotion: A review of research paradigms," *Speech Communication. Amsterdam*, vol. 40, pp. 227-256, April 2003.

[7] L. Ten Bosch, "Emotion, speech and the ASR framework," *Speech Communication. Amsterdam*, vol. 40, pp. 213-225, April 2003.

[8] Z. Inanoglu and R. Caneel, "Emotive alert: HMM-based emotion detection in voicemail messages," *in Proc.*

Intelligent User Interfaces. San Diego, pp. 251-253, January 2005.
[9] K. Chen, GX Yue, F. Yu, Y. Shen and A.Q. Zhu, "Research on speech emotion recognition system in e-learning," *Lecture Notes in Computer Science*. *Berlin*, vol. 4489, pp. 555-558, 2007.
[10] R. Nakats, N. Tosa and T. Ochi, "Construction of an interactive movie system for multi-person participation," *in Proc. Multimedia Computing and Systems*. *Austin*, pp. 228-232, 1998.
[11] C. M. Jones and I. M. Jonsson. Performance analysis of acoustic emotion recognition for in-car conversational interfaces," *Lecture Notes in Computer Science*. *Berlin*, vol. 4555, pp. 411-420, 2007.
[12] I. Luengo, E. Navasm, I. Hernaez and J. Sanchez, "Automatic Emotion Recognition using Prosodic Parameters," *in Proc. INTERSPEECH 2005*. *Lisbon*, pp. 493-496, 2005.
[13] H. Hao, X. Ming-Xing and W. Wei, "GMM Supervector Based SVM with Spectral Features for Speech Emotion Recognition, in Acoustics, Speech and Signal Processing," *in Conf. ICASSP 2007*. *Honolulu*, pp. 413-420, 2007.
[14] S. S. Stevens, J. Volkmann and E. B. Newman, "A scale for the measurement of the psychological magnitude pitches," *Journal of the Acoustical Society of American*, pp.185-190, August 1937.
[15] N. Kamaruddin and A. Wahab, "Speech Emotion Verification System (SEVS) based on MFCC for real time applications," *in Conf. Intelligent Environments 2008*. *Seattle*, July 2008.
[16] H. Simon, *Neural Networks A Comprehensive Foundation*, 2nd ed., Perason Education, 1999, pp.340–372.
[17] Ch. Chang, Ch. Lin, *LIBSVM: a Library for Support Vector Machines*, http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2005.
[18] MATLAB Chinese forum, *Thirty Neural Networks Examples Analysis in MATLAB*, Publication of Beijing University of Aeronautics and Astronautics, pp.110-pp.127, 2010.

Bo Yu received his B.E. in Computer Science and Technology from Harbin University of Science and Technology, Harbin, Heilongjiang Province, China in 2004 and received his M.E. in Technology of Computer Application from Harbin University of Science and Technology, China in 2007. He is currently working towards the PhD degree in Harbin Institute of Technology. His major fields are: Speech Recognition, Pattern Recognition& Software development.

He started his teaching career in the College of Software in 2007 in Harbin University of Science and Technology and promoted as a lecturer in 2008. He is the Deputy Secretary of Software Engineering Department. He used to teach Computer Theory Test in C Language in 2005 in Harbin University of Science and Technology and was the programmer in Hua Ze Digital Company from Oct. 2005 to Mar. 2006. He had his field work in Neusoft Group from Sep. 2007 to Jan. 2008. One of his published articles is Milti-agent Web Texts Mining Based on Galois Lattice.2010 International Forum on Information Technology and Applications (IFITA 2010). His current research Interests are Speech Emotion, Pattern Recognition & Software development. His previous research interests are Web text mining.

Mr.Yu received Certificate of Achievement as coach for his students getting Second Prize in 2010 The Fifth Heilongjiang Province Programming Contest, Asia Provincial-National Contests in May 16, 2010.

Haifeng Li Doctoral Supervisor, Director of Speech Processing Lab in School of Computer Science and Technology at HIT and IEEE member. He is the Dean of Honors School now. He got his Doctor's Degree from Electro-Magnetical Measuring Technique&Instrumentation from HIT in 1997 and Doctor's Degree from Computer, Communication and Electronic Science from University of Paris VI, France in 2002. He started the teaching career in 1994 in HIT, promoted as lecturer in 1995 and professor in 2003. From 1997 to 2002, he is engaged in the post-doctoral research at University of Paris VI, and presided the project of Speech Noise Reduction Research for France Telecom. In August 2004, he became the Assistant Dean of School of Software. His research fields are Audio Information Retrieval & Processing, Artificial Neural Networks. He undertakes many projects of National Natural Science Foundation, Provincial and Ministry Science Foundation and has published over 30 papers in journals and conferences at home and abroad.

Chunying Fang was born in 1978. She is currently working toward the PhD degree at the Computer Science Department, Harbin Institute of technology (HIT), Harbin, China. Her present research interests include speech recognition.