Speech Emotion Recognition Based on PSO-optimized SVM
Chuan-zhen LI¹, Fang-kun LIU¹, Yu-tian WANG², Hui WANG²,* and Qin ZHANG²

¹Information Engineering School, Communication University of China, Beijing, China
²Key Laboratory of Media Audio & Video (Communication University of China) Ministry of Education, Beijing, China
*Corresponding author

Keywords: Speech emotion recognition, Fourier parameter, Particle Swarm Optimization, Support Vector Machine, Frame length.

Abstract. Speech emotion recognition, as a significant part of man-machine interaction, has been widely used in manufacture, information industry, criminal investigation and security protection. However, recent research shows that it is hard to get a high recognition rate. This paper intends to enhance precision of recognition based on Support Vector Machine and Particle Swarm Optimization algorithm. Particle Swarm Optimization algorithm is applied to SVM's parameter selection and optimization. The experimental results on German EMO-DB database and Chinese CASIA database show that the accuracy rate using with PSO-optimized SVM reaches 90.09% and 73.5% respectively.

Introduction
Speech emotion recognition (SER) has become one of the major issues in the study of artificial intelligence. The recognition results can be used to analyze the emotion characteristics, judge and simulate the speaker's emotions. Consequently, the above applications can be widely applied to information industry, entertainment, criminal, security areas, as well as medical fields.

To recognize speech emotion, Renée Van Bezooijen etc.[1] created the precedent of classifying emotions using acoustic statistical characteristics. After that, R. Picard etc.[2] proposed the concept of Affective Computing, and stated that "Affective Computing is computing that relates to, arises from, or deliberately influences emotion or other affective phenomena." Moriyama etc.[3] proposed a linear correlation model between voice and emotion, which conducted the emotion recognition through linear system with the relation information. The above studies established theoretical basis for speech emotion recognition.

Recently, different techniques and different features have been generally applied to speech emotion recognition. Yelin Kim[4] used Deep Belief Network (DBN) to get the complex non-linear feature interactions in multimodal data, which was effective for emotion recognition. It even could be used for unsupervised data. Hari Krishna Vydana[5] adopted Gaussian Mixture Modeling with a universal background model (GMM-UBM) to speech emotion recognition. In this way, researchers could improve the recognition rate by studying the presence of emotion information of smaller segments of speech. Reza Lotfian and Carlos Busso[6] adopted test to speech (TTS) technology to synthesize normal voice to contrast with emotional voice, which can be used to recognize speech emotion in non-label mode. The Fourier parameters were used as the features of speech emotion recognition by Kunxia Wang etc[7]. The recognition rate reached 88.88% for the six types of emotions in the EMO-DB database and 79% for the six types of emotions in the CASIA database respectively. Yu Gu[8] extracted the voiced segments of a piece of audio through the spectrograms. Features were extracted from the voiced segments in order to get better recognition rate. In [7] and [8], researchers used SVM as the classifier for speech emotion recognition.

This paper adopts Fourier parameters[7] as the features of speech signals. And PSO-optimized SVM is used to improve the recognition rate through selecting optimized combination of parameters of SVM. Simultaneously, the recognition rates with different frame length are presented in this paper.
The rest of the paper is organized as follows. Section 2 describes the system design, signal preprocessing, feature extracting and the procedure of optimize the SVM parameters. Section 3 presents the experimental results with different database and different frame length. Finally, our conclusions are given in Section 4.

**Design of Speech Emotion Recognition System**

To improve the recognition accuracy, we design a speech emotion recognition system based on PSO-optimized SVM. The framework of this recognition system is shown as Fig.1.

![Figure 1. Speech emotion recognition system.](image)

The system includes several parts: preprocessing of speech signals, the features extraction, SVM training and parameters optimization, SVM classifier modeling, and speech emotion recognition. The system will be described in detail as below.

**Preprocessing of Speech Signal**

This paper adopts Fourier parameters as emotion features of the speech. Fourier parameters can be obtained by Fourier transform. However, since the speech signal is non-stationary, we could not get the spectral features of the signal at specific point in time through global Fourier transform. It is necessary to divide the speech signal into frames and add window before conducting discrete-time Fourier transform (DTFT). It should be noted that, in the preprocessing, the frame length is a significant factor which affects the final recognition rate. If the frame length is too long, the Fourier transform is meaningless because the signal is in a non-stationary state; if the frame length is too short, the amount of information contained in each frame is too small, and the spectrum will result in the low final recognition rate.

After dividing the signal into frames, we add window on each frame. There are three major windows: rectangular window, Hamming window and Hanning window. Hamming window is chosen because it has the largest side-lobe attenuation and spectrum leakage could be minimized.

**Feature Extraction**

Each speech signal frame could be regarded as a stationary signal, and discrete-time Fourier transform can be conducted. The Fourier parametric model[7] of each frame is shown as Eq.1 and Eq.2.

\[
x(n) = \sum_{k=0}^{N} H_k^l(n)(\cos(2\pi\frac{f_k^l}{F_s}n) + \phi_k^l)
\]  

\[
H_k^l(n) = \sum_{m=0}^{N-1} x(n)e^{-j \frac{2\pi km}{N}}
\]

where \(l\) represents the number of frames of one piece of speech signal. \(H_k^l\) and \(\phi_k^l\) respectively represent the amplitude and phase of the \(k\)th frequency components. \(F_s\) represents the sampling rate of the piece of speech and \(f\) represents the frequency of the \(k\)th frequency components. \(N\) represents the number of frequency components contained in the frame.
PSO-optimized SVM

This paper applies PSO to SVM classifier to optimize SVM parameters. Then optimized parameters are used for speech emotion recognition.

Support Vector Machine. SVM[9] has unique advantages in solving small sample and nonlinear pattern of recognition. In general, if the data are linearly separable, the solution is to find a hyperplane that satisfies the classification requirements and to make the train-to-focus points far from the classification surface as much as possible. However, the feature data for speech emotion classification are not linearly separable. In this case, it is required to convert the low-dimensional, nonlinear problem to a high-dimensional, linearly separable problem. In order to avoid the curse of dimensionality, we have to use kernel function to project all sample points to the high-dimensional space[10]. Currently there are three kernel functions of common use:

Polynomial kernel function:

\[ K(x_i, x_j) = (r + \gamma x_i^T x_j)^d \]  \hspace{1cm} (3)

Gaussian radial basis function:

\[ K(x_i, x_j) = \exp(-\gamma \| x_i - x_j \|^2) \]  \hspace{1cm} (4)

Sigmoid kernel function:

\[ K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r) \]  \hspace{1cm} (5)

where \( d \) represents the terms' degree of the polynomial kernel function. \( \gamma \) represents the width value of the kernel function. \( r \) represents the bias coefficient of the kernel function. We have to notice that \( d > 0, \gamma > 0, r > 0. \)

We use Gaussian radial basis function as the kernel function of SVM in this paper, which is widely used in nonlinear classification.

Particle Swarm Optimization. PSO[11] simulates the bird flock’s behavior of foraging, in which the flock coordinate and cooperate to achieve the optimal goal. It is a global search algorithm with simple procedure, brief parameters, and uncomplicated adjustment, so it is very suitable for solving optimization problems. In the standard particle swarm algorithm, the velocity and position update of the particle in the search space are determined according to Eq.(6) and Eq.(7) [12].

\[ v_{ij}(t + 1) = w(t) \times v_{ij}(t) + c_1 \times \text{rand}() \times (p_{bj}(t) - x_{ij}(t)) + c_2 \times \text{rand}() \times (g_{bj}(t) - x_{ij}(t)) \]  \hspace{1cm} (6)

\[ x_{ij}(t + 1) = x_{ij}(t) + v_{ij}(t + 1) \]  \hspace{1cm} (7)

where \( w(t) \) represents inertia weight. \( c_1 \) and \( c_2 \) are acceleration factors. \( \text{rand}() \) is any number that is evenly distributed in interval \([0,1] \). \( p_{bj} \) is the best fled position of the present particle. \( g_{bj} \) is the best position of the entire population corresponding to the particle and it is the best position that current iteration has experienced. \( x_{ij}(t) \) and \( v_{ij}(t) \) is the position and velocity at the time \( t \).

SVM parameter optimization by PSO. This optimized SVM algorithm is based on Lin Zhiren's Lib-SVM, and its parameters \( \gamma \) and \( c \) are optimized to improve the accuracy of the speech emotion recognition rate. \( \gamma \) is the parameter of the kernel function in SVM, \( c \) is the penalty factor.

Each particle represents the combination of \((c, \gamma)\). Parameters in SVM model are modified according to the particle velocity and position shown in eq.(6) and eq.(7), so that the exhaustive artificial experiment can be avoided.

As shown in Fig.2, velocity and position of the new particles are updated constantly to obtain current global position of the particle, which contains the best kernel function parameter \( \gamma \) and the penalty factor \( c \) corresponding to the SVM classifier.
Experimental Results and Analysis

EMO-DB German Corpus

The EMO-DB German corpus includes 7 categories of emotions performed by 10 actors and 5 actresses and each emotion induces 10 sentences. The emotions contain anger, joy, neutral, sadness, anxiety, disgust, and boredom. A total of 535 sentences were collected. The distribution of the emotions in the corpus is as follows: 127 for anger, 81 for boredom, 71 for joy, 64 for anxiety, 51 for disgust, 62 for sadness, and 79 for neutral.

In this experiment, five categories of emotions in the EMO-DB corpus are selected: anger, boredom, joy, sadness and neutral. There are 355 training sentences and 55 test sentences. The results of all experiments are averaged over five cross validation experiments. And we compare our results with those of [7] and [13], as shown in Table 1.

Table 1. Speech Emotion Recognition Rate Based on SVM and Speech Emotion Recognition Rate of SVM Based on Particle Swarm Optimization Algorithm (%).

<table>
<thead>
<tr>
<th>Emotion Type</th>
<th>Anger</th>
<th>Boredom</th>
<th>Joy</th>
<th>Sadness</th>
<th>Neutral</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>72.7</td>
<td>81.8</td>
<td>90.9</td>
<td>72.7</td>
<td>54.6</td>
<td>74.55</td>
</tr>
<tr>
<td>SVM Based on PSO</td>
<td>100.0</td>
<td>72.73</td>
<td>100.0</td>
<td>100</td>
<td>81.82</td>
<td>90.09</td>
</tr>
<tr>
<td>Result in [7]</td>
<td>98.29</td>
<td>71.48</td>
<td>92.92</td>
<td>91.21</td>
<td>87.64</td>
<td>88.31</td>
</tr>
<tr>
<td>Result in [13]</td>
<td>86.88</td>
<td>92.02</td>
<td>81.24</td>
<td>90.68</td>
<td>93.80</td>
<td>88.92</td>
</tr>
</tbody>
</table>

According to the results in Table 1, for the EMO-DB corpus, the average result (90.09%) obtained by the optimized SVM classifier is higher than the average result obtained by the general SVM classifier (74.55%) under the same experimental conditions. The improvement is due to the optimized combination of parameter $\gamma$ and penalty factor $c$. With optimized SVM classifier, the recognition rates are 100% for anger, joy and sadness, and 72.73% for boredom. This is because the spectrum of anger, joy and sadness are different from each other while the spectrum of boredom is similar to anger and neutral, so it is easier to recognize anger, joy, and sadness.
The recognition rates with different frame length are shown in Table 2. It can be seen that, with frame length changes from 256 to 2048, the average emotion recognition rate changes from 90.09% to 80%. It is because the short frame length contains less confusing information.

<table>
<thead>
<tr>
<th>Emotion Type</th>
<th>Anger</th>
<th>Boredom</th>
<th>Joy</th>
<th>Sadness</th>
<th>Neutral</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>N=256</td>
<td>100.0</td>
<td>72.73</td>
<td>100.0</td>
<td>100</td>
<td>81.82</td>
<td>90.09</td>
</tr>
<tr>
<td>N=512</td>
<td>100.0</td>
<td>81.8</td>
<td>100.0</td>
<td>72.7</td>
<td>90.9</td>
<td>89.09</td>
</tr>
<tr>
<td>N=1024</td>
<td>72.73</td>
<td>72.73</td>
<td>100.0</td>
<td>81.82</td>
<td>72.73</td>
<td>80</td>
</tr>
<tr>
<td>N=2048</td>
<td>63.64</td>
<td>45.45</td>
<td>100.0</td>
<td>100</td>
<td>90.91</td>
<td>80</td>
</tr>
</tbody>
</table>

CASIA Chinese Corpus

The CASIA Chinese Corpus includes 6 categories of emotions performed by 2 actors and 2 actresses. The emotions contain anger, joy, surprise, neutral, sadness and fear. A total of 9600 sentences were collected.

In this experiment, all the categories of emotions are selected. 6000 training sentences and 1200 test sentences are selected from the database. The results of all the experiments are averaged over five cross validation experiments. And we compare our results with those of [7], as shown in Table 3.

<table>
<thead>
<tr>
<th>Emotion Type</th>
<th>Anger</th>
<th>Surprise</th>
<th>Fear</th>
<th>Joy</th>
<th>Sadness</th>
<th>Neutral</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>78.5</td>
<td>65.5</td>
<td>73.9</td>
<td>68.5</td>
<td>63.3</td>
<td>83.0</td>
<td>72.1</td>
</tr>
<tr>
<td>SVM Based on PSO</td>
<td>83.5</td>
<td>60.5</td>
<td>74.4</td>
<td>69.5</td>
<td>67.8</td>
<td>85.0</td>
<td>73.5</td>
</tr>
<tr>
<td>Result in [7]</td>
<td>86</td>
<td>84</td>
<td>81</td>
<td>81</td>
<td>67</td>
<td>75</td>
<td>79</td>
</tr>
</tbody>
</table>

For the CASIA corpus, we could see from Table.3 that the average recognition rate (73.5%) of PSO-SVM classifier is higher than that (72.1%) of general SVM classifier. The pattern is similar to the results of experiments on EMO-DB corpus. The improvement is also because of the optimized combination of parameter $\gamma$ and penalty factor $c$.

The recognition rate with different frame length are shown in Table 4. The pattern is also similar to the results of the experiments on EMO-DB. It means that the frame length leads to the difference in the final recognition results directly. If the frame is too long, the Fourier transform is meaningless and the signal is in a non-stationary state; even if the signal is in a stationary state, it would contain more confusing information.

<table>
<thead>
<tr>
<th>Emotion Type</th>
<th>Anger</th>
<th>Surprise</th>
<th>Fear</th>
<th>Joy</th>
<th>Sadness</th>
<th>Neutral</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>N=256</td>
<td>83.5</td>
<td>60.5</td>
<td>74.4</td>
<td>69.5</td>
<td>67.8</td>
<td>85.0</td>
<td>73.5</td>
</tr>
<tr>
<td>N=512</td>
<td>71.0</td>
<td>65.5</td>
<td>72.9</td>
<td>69.0</td>
<td>61.30</td>
<td>74.5</td>
<td>69.0</td>
</tr>
<tr>
<td>N=1024</td>
<td>71.0</td>
<td>44.0</td>
<td>66.8</td>
<td>63.5</td>
<td>62.3</td>
<td>71.0</td>
<td>63.1</td>
</tr>
<tr>
<td>N=2048</td>
<td>69.5</td>
<td>40.0</td>
<td>68.8</td>
<td>54</td>
<td>62.3</td>
<td>72.5</td>
<td>61.2</td>
</tr>
</tbody>
</table>

Conclusion

In this paper, PSO is used to optimize the combination of the two parameters: penalty factor $c$ and the kernel function parameter $\gamma$ of SVM. Then, PSO-optimized SVM is used to recognize emotion from speech sentence. The average recognition rate for 5 categories of emotions in the EMO-DB database is 90.09% and the lowest recognition rate is 73.5% for boredom. The average recognition rate for 6 categories of emotions in the CASIA database is 73.5% and the lowest recognition rate is 60.5% for surprise. Simultaneously, this paper shows the effect on emotion recognition caused by frame length. In the future, we will try to find the more suitable features and better methods to obtain better emotion recognition rate.
Acknowledgments

This research is supported by the National Natural Science Foundation of China (Grant No. 61231015, 61501410), The State Administration of Press, Publication, Radio, Film and Television of the People's Republic of China scientific research project (Grant No. 2015-53) and the Engineering Planning Project of Communication University of China (Grant No.3132014XNG1425).

References


