Adaptive algorithm for feature selection of speech emotion recognition based on genetic algorithm and SVM

Hong Chen¹, Bo Zhang²,*

¹School of Computer Science, Wuhan Donghu University, Wuhan 430212, China
²School of Computer Science, Wuhan Donghu University, Hubei Key Laboratory of Big Data in Science and Technology (Wuhan Library of Chinese Academy of Science), Wuhan 430212, China
*Corresponding author’s e-mail: zhangbo@wdu.edu.cn

Abstract. As a key carrier for people to obtain information in daily life, speech not only provides text signals, but also contains emotional signals. As an important factor of semantics, emotion signals are the focus of improving the intelligent level of human-computer interaction. Therefore, it is important to improving the accuracy of speech emotion recognition. This paper proposes to incorporate derivative feature parameters into the voice emotion recognition feature selection, and to improve the feature selection algorithm based on genetic algorithm and SVM. This article may provide technical support to promote the development of artificial intelligence.

1. Introduction
In recent years, with the rapid development of artificial intelligence, in order to achieve more intelligent human-computer interaction, human-computer interaction products must be able to fully understand human emotions. Although researchers from all over the world have made many research results in the field of speech emotion research, the efficiency and accuracy of the entire speech emotion information processing needs to be further improved. Therefore, this paper proposes to incorporate derivative feature parameters into the feature selection of speech emotion recognition, and improves the feature selection algorithm based on genetic algorithm and SVM to provide technical support for improving the accuracy of speech emotion recognition.

2. Extraction of emotional feature parameters of speech signal

2.1. Amplitude energy parameter extraction
Usually, the frame length of the speech signal is 20ms. A frame of signal $s(n)$ that has been taken out is subjected to windowing processing, and a certain window function $w(n)$ is used to multiply $s(n)$ to form a windowed speech $s_w(n)$.

\[ s_w(n) = s(n) \times w(n) \]

The window function uses Hamming window, and its calculation formula is as follows (where $N$ is the frame length):

\[
\begin{align*}
   w(n) &= \begin{cases} 
   0.54 - 0.46 \cos \left( \frac{2\pi n}{N-1} \right), & 0 \leq n \leq N - 1 \\
   0, & \text{else}
   \end{cases}
\end{align*}
\]
The short-term frame energy is the weighted sum of squares of the sampling point values of a frame. The short-time frame energy $E_n$ is represented by the following formula:

$$E_n = \sum_{m=0}^{\infty} [x(m)w(n-m)^2] = \sum_{m=-N+1}^{n} [x(m)w(n-m)^2]$$

Among them, $s(n)$ is the time series of discrete speech signals, and $w(n)$ is the Hamming window function. Here, the physical meaning of the window function $w(n)$ squared is a filter with an impulse response of $w(n)^2$.

Similarly, the short-term average amplitude calculation formula of the signal is as follows:

$$M_n = \sum_{m=-N+1}^{n} |x(m)w(n-m)|$$

2.2. Time-related parameter extraction

Many time-related parameters, short-term zero-crossing analysis of the speech time domain analysis is the most common and effective one. For continuous speech signal, which can be examined by the case of time domain waveform through time axis.

The short-term average zero-crossing number of the speech signal $x(n)$ is defined as:

$$Z_n = \sum_{m=-\infty}^{\infty} [sgn[x(m)] - sgn[x(m-1)]]w(n-m) = |sgn[x(n)] - sgn[x(n-1)]|^w(n)$$

In the formula, $sgn[]$ is a sign function, when $x(n)$ is greater than or equal to zero, take positive. Otherwise, it is negative. $w(n)$ is the window sequence.

2.3. Pitch related parameter extraction

Voiced signal is a quasi-periodic signal. The fundamental tone refers to the periodicity caused by the vibration of the vocal cords when voiced, and the fundamental period is reciprocal of vocal cord vibration frequency. Since it is quasi-periodic, it can only be estimated by the short-time average method. In the aspect of pitch detection, this paper uses the cepstrum method to calculate. The speech signal $s(n)$ is excited by the glottal pulse $e(n)$ and filtered by the vocal tract response $v(n)$.

$$s(n) = e(n) \times v(n)$$

The cepstrum of these three are $\hat{s}(n)$, $\hat{e}(n)$, $\hat{v}(n)$ respectively, so

$$\hat{s}(n) = \hat{e}(n) + \hat{v}(n)$$

A simple inverse filtering method can be used to separate and recover $e(n)$ and $v(n)$. According to the nature of the excitation $e(n)$ and its cepstrum, the pitch period can be obtained. Through linear predictive coding (LPC) analysis, the speech signal $s$ (force) can be expressed as:

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + G e(n)$$

Among them, $a_i$ is the prediction coefficient, $P$ is the prediction order, $e(n)$ is the excitation signal, and $G$ is the amplitude factor. Perform LPC analysis on the input speech signal to obtain the prediction coefficient $a_i$, which constitutes the inverse filter $A(z)$:

$$A(z) = \sum_{i=0}^{p} a_i z^{-i}, \quad a_0 = 1$$

The original speech is inversely filtered through the filter $A(z)$ to obtain the prediction margin signal $\varepsilon(n)$. $\varepsilon(n)$ does not contain channel response information, but contains complete excitation information. Cepstrum analysis of the prediction margin signal $\varepsilon(n)$ can obtain clearer and more accurate pitch information.

2.4. Derivative feature extraction

The research results of many scholars show that the characteristics of prosody can effectively describe speech emotion. At the same time, scholars have unanimously recognized the importance of changes to emotions also affects the changes in emotional characteristic parameters. If we directly describe this change, theoretically it is closer to the essence of expressing emotion. The mathematical description of change is the derivative. Considering the digital discreteness and short-term stability of speech processing, it is necessary to convert this discrete value into a continuous function. This paper adopts the method of curve fitting, the fitting function adopts cubic spline curve; on the basis of the obtained fitting curve, the derivative is calculated, and these derivatives and their derivative values are used as new characteristic parameters. The derivative is directly obtained, after the discrete sequence of the
extracted prosody feature parameters is fitted by the cubic spline curve. In this article, MATLAB's spline function is used for curve fitting, and then the mmspder function is used for derivative evaluation.

3. Method for selecting voice emotion features based on genetic algorithm and SVM

3.1. Overall structure of feature selection methods
In this system, the structure of genetic algorithm selection features is shown in Figure 1. The basic flow of the algorithm is as follows:
- The training feature data enters the SVM classifier after data selection;
- The training feature selection method follows the rules:
  - For the first time, follow the initial selection sequence of GA survival;
  - The rest generate the selection sequence according to the evaluation function of the GA algorithm;
- After SVM recognition, the result data generated is processed to generate an evaluation function, and the evaluation function value knows the feature selection of the next round;
- The evaluation function guides the new feature selection sequence, and then enters the SVM classifier;
- Repeat this until the required accuracy;
- The feature selection sequence finally obtained is the result of this feature selection.

![Figure 1. The structure of feature selection algorithm based on genetic algorithm](image)

3.2. Feature selection method algorithm
The main purpose of using genetic algorithm is to select the subset of features that can distinguish voice emotions from the many extracted speech features. For the convenience of description, we define the following variables. \( E_i = [e_{i1}, e_{i2}, ..., e_{in}], i=1,2, ..., 6 \): represents all feature sets, in our experiment \( n=23 \), a total of 23 features are selected; \( i \) is 1 to 6 respectively representing 6 kinds of emotions; \{\( E_1, E_2, ..., E_6 \)\} represents all the emotional feature sets; \( X=[x_1, x_2, ..., x_n] \), \( x_i=0 \) or \( x_i=1 \): Represents a selection sequence consisting of 01, 1 means the feature is selected, 0 means not selected; Each \( X \) sequence represents an individual. When GA is initialized, \( N \) \( X \) sequences will be generated as the initial individual set; for the convenience of description, \( X_i \) is used below to denote the set of individuals whose fitness is calculated for the \( i \)-th time in the whole process of a genetic algorithm. The final output of the algorithm is a 01 sequence subset of \( X \), which represents the result of feature selection.

The coding rules adopt binary coding, and binary coding is used for individual coding. The precision is set to \( e^6 \), \( X'=[x_1', x_2', ..., x_n'] \), \( x_i' \in [0,1] \) is the individual code string, \( x_i \) is a vector of decimals between 0 and 1. To facilitate feature selection, we do the following conversion: Let \( X=[x_1', x_2', ..., x_n'] = round(X')=[round(x_1'), round(x_2'), ..., round(x_n')] \), where \( round(x_i') \), If \( |x_i' -
1|is 0, otherwise the value is 1. This represents the integer with the shortest distance from xi, that is, xᵢ=0 or 1, 1 means the feature is selected, and 0 means not.

According to its genetic algorithm code value, X can be used to represent the selected subset of features, and corresponding feature parameters can be selected for training and recognition according to the feature subset X.

### 3.3. Result analysis of feature selection method

We use the above genetic algorithm to divide all data into two parts: training data and test data, and use OSU_SVM3.0 mat lab library and GA library GAOT to do the following algorithm test on matlab7. Firstly, training the SVM with the training data to get a maximum recognition rate \( X_1 \), and then testing the test data under \( X_1 \) to get the test recognition rate. According to the results of preliminary experiments, the default number of evolutions in this study is 100 generations. From the speech database, we choose 60 sentences for each emotion and recognize them. The recognition results are as follows:

| Table 1. Comparison of SVM one-to-one and SVM one-to-many recognition results |
|-----------------------------|--------------------------|--------------------------|---------------------------|--------------------------|--------------------------|-----------------------------|
|                             | happiness | sadness | astonishment | anger | fear | disgust | Total average             |
| SVM one-to-one               | 73.6      | 79.9    | 78.6         | 80.2  | 76.4 | 82.8    | 78.58                     |
| SVM one-to-many              | 72.5      | 82.1    | 75.3         | 77.3  | 79.5 | 81.9    | 78.1                      |

From the above table, the actual recognition rate of the two has little difference.

### 4. Speech emotion recognition method based on adaptive genetic algorithm

#### 4.1. Adaptive improvement of algorithm

According to the above analysis, various speech emotions should be distinguished from other emotions with different feature subsets. Therefore, it is necessary to select individual characteristics based on emotion in the selection algorithm. In the following adaptation, there are two modifications. According to the one-to-one method of SVM, each emotion is selected as the positive side to be recognized, and all the remaining emotion samples are used as the negative side. The genetic algorithm is used to select the most suitable feature subset to make the recognition rate of the emotion highest. In this way, to identify all 6 kinds of emotions, it must be recognized 5 times, and the result will produce a binary classification tree, the tree nodes from top to bottom, representing the optimal classification of emotion selection order. The value of the adaptive function represents the emotion recognition rate that is distinguished from all other emotions. Specifically, it refers to the maximum recognition rate of each emotion that is distinguished from other emotions in the case of a specific feature selection subset. For the first recognition, using genetic algorithm, it is necessary to find a feature subset \( X_i \) and an emotion \( E_j \) so that emotion \( E_j \) uses \( X_i \) feature subset to distinguish the other five emotions with the highest recognition rate. That is, the feature subset \( X_i \) is the most essential feature of the emotion \( E_j \) to distinguish all other emotions; while the effect of \( X_i \) for certain emotions other than \( E_j \) to distinguish all other emotions is not very good. In this way, the fitness function for the first recognition can be expressed as follows:

Let \( E_j \) represent various emotions, then\( \{E_1, E_2, \ldots, E_6\} \) represents a set of 6 emotions, and \( Svm(E_j, X_i) \) represents the recognition rate of \( E_j \) under the feature subset \( X_i \). \( EVAL_3 \) represents the evaluation function of identifying one of the six emotions in a certain generation selection (that is, the emotion is recognized as the positive side, and the other 5 emotions are recognized as the negative side), then

\[
EVAL_3^{GA} \leftarrow \max (Svm(E_1, X_i), Svm(E_2, X_i), \ldots, Svm(E_6, X_i))
\]

Genetic algorithm is bound to find a \( X_{i_{\text{max}}} \) and \( E_{j_{\text{max}}} \), making \( EVAL_3 \) maximum.
4.2. Adaptive optimal recognition classifier construction algorithm

The entire algorithm flow of classifier training based on feature selection and adaptive optimization of recognition is as follows:

Step 1: Set \(i = 1\);

Step 2: randomly generate feature selection subset \(X_0\), and generate training data and test data according to the feature selection of \(X_0\). One of the six emotions is selected as the positive party, and the others are recognized as the negative party. The maximum recognition rate of the six recognitions is used as the fitness function of this operation. Assume that the recognition rate of \(E_0\) emotion is the highest.

Step 3: Use the fitness function that was obtained in the first step to guide the feature selection subset \(X_i\). According to the best search features of genetic algorithm, \(E_0\) still has the greatest probability to produce the greatest recognition rate. The recognition rate is used as the fitness function of this operation. Step 4: As shown in the second step, the genetic algorithm continues to perform the operations of all the following generations, and finally an optimal feature selection subset \(X_{max}\) and \(E_{max}\) (the greatest probability is \(E_0\)) will be found. Its meaning is that \(X_{max}\) is most conducive to distinguish \(E_{max}\) from all other emotions.

Step 5: \(i = i + 1\), if \(i > 6\), go to step 6, otherwise go to step 2.

Step 6: End.

The above steps will generate the first emotionally selected node on the binary decision tree and the corresponding feature selection subset. The following steps are still used for the remaining 5 other emotions, and finally a complete decision tree and selected subsets for each node will be generated.

![Figure 2. Adaptive genetic algorithm decision tree](image)

4.3. Experimental results of the improved algorithm

All experiments of the adaptive algorithm are implemented on matlab7.0, using OSU-SVM3.0 and GAT toolkit, and the parameter selection of genetic algorithm remains unchanged. Since the search result of genetic algorithm is close to optimal, it is not optimal. Therefore, it can be run several times, and the results may be different each time. In the process of generating a decision tree result, each result has to be run 5 times. We have done 22 experiments and obtained 22 sets of results. Among the 22 sets of data, one of the experiments in which the largest recognition result was obtained, the statistics of various emotion selection orders (ie, binary tree root node sequence) are shown in Table 2:

|                | happiness | sadness | astonishment | anger | fear | disgust |
|----------------|-----------|---------|--------------|-------|------|---------|
| the first time | 0         | 31.7    | 0            | 0     | 0    | 71.1    |
| the second time| 0         | 70.9    | 0            | 0     | 0    | 32.4    |
| the third time | 0         | 0       | 29.3         | 61.2  | 8.2  | 0       |
| the fourth time| 0         | 0       | 39.0         | 16.4  | 46.1 | 0       |
| the fifth time | 78.4      | 0       | 7.5          | 17.9  | 0    | 0       |
| the sixth time | 26.8      | 0       | 47.8         | 0     | 29.2 | 0       |
As described in Table 2, the order of emotion selection is disgust -> sadness -> anger -> fear -> happy -> surprise. Among the 22 experiments, 71.1% chose disgust, which means that disgust is the easiest to distinguish from the other 5 emotions. In order to compare the final recognition results, we compared the results obtained by the neural network contribution analysis method and the adaptive genetic algorithm. The results are as follows:

| Table 3. Comparison of various algorithms recognition results |
|-------------------------------------------------------------|
|                | happiness | sadness | astonishment | anger | fear | disgust | Total average recognition rate |
| Neural network contribution                        | 69.8      | 83.2     | 83.7         | 79.4  | 81.7 | 71.2     | 78.17                       |
| Adaptive GA                                    | 89.6      | 81.6     | 90.2         | 78.7  | 78.4 | 79.6     | 83.02                       |

In this article, derivative features are used to describe emotional features. The adaptive algorithm finally obtains a subset of features, which represents the individual characteristics of emotions. In the experiment to obtain the maximum recognition rate, the distribution of feature selection is shown in Table 4.

| Table 4. Distribution of feature selection (number) |
|------------------------------|
| Short-term amplitude       |
| Short-term energy          |
| Fundamental frequency      |
| Fundamental frequency derivative |
| Energy derivative          |
| Amplitude derivative       |
| number                     |
| disgust                    | 2 | 1 | 3 | 1 | 1 | 3 | 11 |
| sadness                   | 2 | 1 | 2 | 4 | 3 | 2 | 14 |
| anger                     | 2 | 0 | 3 | 2 | 2 | 2 | 11 |
| fear                      | 0 | 2 | 1 | 1 | 4 | 1 | 9  |
| happiness                 | 0 | 1 | 3 | 3 | 1 | 0 | 8  |

This paper proposes to incorporate derivative feature parameters into the voice emotion recognition feature selection, and to improve the feature selection algorithm based on genetic algorithm and SVM. The results show that the adaptive genetic algorithm proposed in this paper can effectively improve the accuracy of speech emotion recognition.

Acknowledgment
This paper was co-supported by Hubei Key Laboratory of Big Data in Science and Technology in 2019, "Research and application of regional science and technology based on big data of hot spring industry in Dawan District of Guangdong, Hong Kong and Macao" under grant 20KF011014.

References
[1] Ye, J.X, Tu, Q.Y, Chen, Y.T. (2019) Multi-level random forest speech emotional recognition based on importance score. Journal of Changsha University of Science and Technology (Natural Science Edition), 16: 77-83.
[2] Li, H, Ran, N, Wang, W. (2020) Research and Analysis of Speech Emotion Recognition Based on Knowledge Graph. Computer Technology and Development, 278: 141-146.
[3] Sun, X.H, Li, H.J. (2020) Overview of speech emotion recognition. Computer Engineering and Applications, 954: 6-14.
[4] Zhuo, G, Bian, W.D, Jiang, J. (2020) Research on Tibetan speech endpoint detection based on short-term average energy and short-term zero-crossing rate. Computer Knowledge and Technology, 112: 3-7.
[5] Sun, J. (2015) Research on Time-Frequency Analysis of Speech Signal Based on MATLAB. Computer Knowledge and Technology: Academic Exchange, 15: 11-13.