Speech Bullying Vocabulary Recognition Algorithm in Artificial Intelligent Child Protecting System

Tong Liu[1,2,3], Liang Ye[1,4(✉)], Tian Han[2,4], Yue Li[3,5], and Esko Alasaarela[4]

[1] Department of Information and Communication Engineering, Harbin Institute of Technology, Harbin 150080, China
yeliang@hit.edu.cn
[2] School of Software and Micro Electronics, Harbin University of Science and Technology, Harbin 150080, China
[3] Key Laboratory of Police Wireless Digital Communication, Ministry of Public Security, Harbin 150080, China
[4] Health and Wellness Measurement Research Group, OPEM Unit, University of Oulu, 90014 Oulu, Finland
[5] Electrical Engineering School, Heilongjiang University, Harbin 150080, China

Abstract. With the continuous breakthrough of various technologies, speech recognition technology has become a research hotspot. It is a way to find out the phenomenon of bullying in time by detecting whether the voice contains campus bullying vocabulary. In practical applications, an infinite network is established through sensors to transmit information, and the occurrence of campus bullying events is prevented in time. This paper studies the theory of support vector machine and its application in speech recognition. In order to identify bullying vocabulary, this paper firstly built a voice library with 250 voice audios, including 125 campus bullying word audios and 125 non-bullying audios. The first sub-frame of the speech signal was used for endpoint detection. Then mode decomposition and Fourier transform were applied. The maximum value of the primary frequency spectrum was extracted as the acoustic feature. Finally, 200 audios in the database were used for training, and 50 audios were used for speech recognition testing. The average recognition accuracy was 94%, indicating that the support vector machine theory showed a good advantage in the case of small samples for speech recognition.

Keywords: Support vector machine · Empirical mode decomposition · Intrinsic mode function component

1 Introduction

With the continuous breakthrough of various technologies, speech recognition technology has become a research hotspot, and the methods of speech recognition technology are also very rich [1]. In this paper, the application of support vector machine theory in speech recognition is studied. Firstly, the speech signal is preprocessed, and

S. Han et al. (Eds.): AICON 2019, LNICST 287, pp. 158–164, 2019.
https://doi.org/10.1007/978-3-030-22971-9_14

then the acoustic features are extracted. The support vector machine hyper-plane and kernel function are optimized for speech recognition. The voice features are used for the training of the model, and finally the speech recognition system was identified and tested.

The remainder of this paper is organized as follow: Sect. 2 describes the procedures of 2 speech signal preprocessing and feature extraction; Sect. 3 analyzes the Support Vector Machine; Sect. 4 shows the simulation results; Sect. 5 draws a conclusion.

2 Speech Signal Preprocessing and Feature Extraction

2.1 Voice Signal Preprocessing

Firstly, the speech signal is pre-emphasized by a high-pass filter to improve the high-frequency part of the speech, reduce the amplitude range of the gene line, and reduce the dynamic range of the spectrum.

Secondly, the Hanning window is used to frame and window the speech signal. Finally, endpoint detection is performed on the speech signal. In order to solve the problem that the energy curve and the zero-crossing rate curve will fluctuate in the non-speech area, the data are subjected to median smoothing in the endpoint detection. The number of wild spots in the processed speech waveform is significantly reduced (Fig. 1).

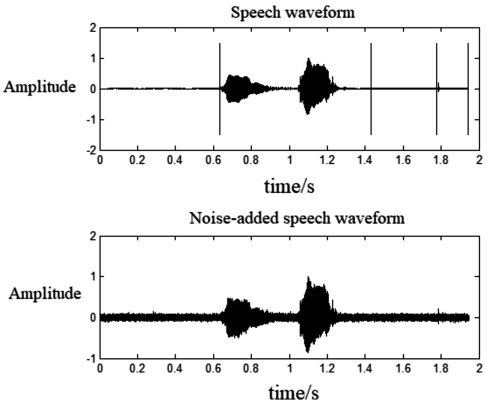

Fig. 1. Endpoint detection map for the "stupid" audio.

2.2 Speech Signal Feature Extraction

The Basic Principle of EMD. Treat the original signal $x(t)$ as a signal to be processed. Firstly, determine all the extreme points of the signal, and connect all the maxima and all the minima points with a cubic spline curve. The upper and lower envelopes of the original signal are obtained. The mean of the upper and lower envelopes is called as $m(t)$.

Subtract the upper and lower envelope mean values $m(t)$ from the pending signal $x(t)$:

$$h_1(t) = x(t) - m(t) \tag{1}$$

Verify that it is a basic mode component. If the two basic conditions are not met, it repeats the above operation as a signal to be processed until it is a basic mode weight.

$$c_1(t) = h_1(t) \tag{2}$$

After the first basic mode component is decomposed from the original signal, Subtract $c_1(t)$ from $x(t)$. The sequence of residual values is as follows:

$$r_1(t) = x(t) - c_1(t) \tag{3}$$

The $r_1(t)$ acts as a new raw signal. Repeat the above process for $r_1(t)$, and get n basic mode weights. Thus, the original signal is decomposed into the sum of several basic mode components and a remainder.

$$x(t) = \sum_{i=1}^{n} c_1(t) + r_n(t) \tag{4}$$

If the last basic mode component $c_n(t)$ is less than the threshold, the procedure will stop. In addition, the procedure also stops when the remaining component becomes a monotonic function.

Feature Selection. In this scheme, the empirical mode decomposition and the Fourier transform are first performed on each audio to obtain the spectrum of the intrinsic modal function component, and the main spectral values are taken for training and testing. Each frame of a voice audio is decomposed by empirical mode to resolve six basic mode components. In this paper, the first 7 frames in each audio are selected, and the basic mode weight in each frame is Fourier transformed. The maximum value of each mode component in each frame is taken as the main spectrum value of the natural mode function component.

3 Support Vector Machine

3.1 Support Vector Machine and Kernel Function

In the state where the training samples are linearly separable, a classification hyperplane can be found. The convex quadratic programming at this time can be transformed by (5) [2]:

$$\begin{cases} \left\{ \max\limits_{\alpha} \sum\limits_{i=1}^{n} \alpha_i - \frac{1}{2} \sum\limits_{i,j=1}^{n} \alpha_i \alpha_j x_i x_j \langle x_i, x_j \rangle \right. \\ s.t. \\ \sum\limits_{i=1}^{n} \alpha_i y_i = 0, \alpha_i \geq 0, i = 1, 2, \ldots, n \end{cases} \tag{5}$$

where α_i is the Lagrange multiplier.

$$f(x, w, b) = sign(< w, x > + b) = sign\left(\sum_{i=1}^{n} \alpha_i y_i \langle x_i, x \rangle + b \right) \tag{6}$$

where x_i is called Support Vector.

Some sample data are linearly separable in the input space, but some are not. After the dimensionally indistinguishable sample data is subjected to dimensionality reduction, it can be transformed into linearly separable sample data [3]. However, dimensionality reduction also has some shortcomings, such as losing useful information. Compared with the dimension reduction method, the kernel function method is much better. The kernel function is to map these linearly inseparable sample data into a high-dimensional space through a kernel function. In this high-dimensional space, these linearly inseparable samples are converted into linearly separable samples.

The main types of kernel functions are: linear kernel function, polynomial kernel function, S-type growth curve kernel function, and tensor product kernel function [4].

3.2 Sequence Minimum Optimization Algorithm

In the support vector machine theory, the problem to be optimized is [5]:

$$\begin{cases} \max\limits_{\alpha} W(\alpha) = \sum\limits_{i}^{m} \alpha_i - \frac{1}{2} \sum\limits_{i,j=1}^{m} y^i y^j \alpha_i \alpha_j \langle x^i, x^j \rangle \\ s.t.\ 0 \leq \alpha_i \leq C, i = 1, \ldots, m \\ \sum\limits_{i=1}^{m} \alpha_i y^i = 0 \end{cases} \tag{7}$$

Constraints can be obtained from Eq. (7):

$$\alpha_1 y^1 = -\sum_{i=2}^{m} \alpha_i y^i \tag{8}$$

Because $y \in \{-1, 1\}$, the formula (8) can be written as:

$$\alpha_1 = -y^1 \sum_{i=2}^{m} \alpha_i y^i \tag{9}$$

It can be known from Eq. (9) that α_1 and $\alpha_2, \ldots, \alpha_m$ are associated, so at least two variables must be selected at a time to satisfy the constraint [6]. Choose the best pair of

combinations α_i, α_j based on experience, then all parameters except the α_i, α_j are fixed and optimized. Therefore, the efficiency of the SMO (Sequence Minimum Optimization) algorithm is relatively high [7].

Figure 2 shows the value of α_i, α_j and the relationship between them.

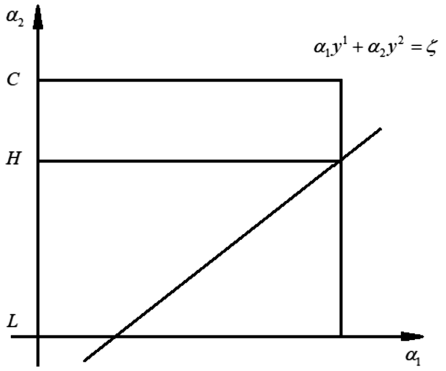

Fig. 2. The relationship of α_i, α_j

4 Classification Result

4.1 Influence of Different Kernel Functions on Training Accuracy

K-fold cross-validation solves the problem of generating more test samples [8]. Using the K-fold cross-validation method, the original data set was equally divided into 5 parts, each with 50 audio samples, and a total of 5 rounds of model training and testing were performed.

The effect of using three different kernel functions on training is shown in Table 1.

Table 1. Influence of different kernel functions on training accuracy

Kernel function type	1	2	3	4	5	Training accuracy
Polynomial kernel function	75.3%	76.2%	80.5%	78.2%	75.5%	77.1%
Two-layer neural network kernel function	84.4%	85.3%	86.2%	87%	89.4%	86.5%
Radial kernel function	93.8%	96%	94.7%	93%	94.3%	94.4%

It can be seen from Table 1 that the training accuracy with the radial kernel function is significantly higher than those with the other two kernel functions.

4.2 Simulation Results

The training samples still use the data in Table 1. The testing accuracies of the training samples and the testing samples are shown in Table 2.

Table 2. Simulation results

Data set	1	2	3	4	5	Testing accuracy
Training samples	96.2%	95.9%	97.3%	95.2%	96%	96.1%
Testing samples	93.8%	96%	94.7%	93%	94.3%	94.4%

As can be seen from Table 2, the accuracy of the testing samples reached 94.4%. This is a good illustration of the fact that in the case of small samples, the use of support vector machine theory to solve the two-class problem is a very suitable method. The support vector machine successfully found the optimal classification hyper-plane to classify the samples.

The testing accuracy after SMO optimization is shown in Table 3.

Table 3. Simulation results

	1	2	3	4	5	Accuracy
Before optimization	93.8%	96%	94.7%	93%	94.3%	94.4%
After optimization	94.5%	96.2%	95.2%	94.1%	95.0%	95.0%

It can be seen from Table 3 that after parameter optimization, the accuracy of speech recognition is increased by 0.6%.

5 Conclusion

This paper proposed a speech bullying vocabulary recognition algorithm in artificial intelligent child protecting system. After preprocessing the speech signal, the empirical mode decomposition method is used to extract the IMF of each order mode component, and then the main frequency value of the IMF is separated by Fourier transformation as the feature vector.

Through the K-fold cross-validation method, more training data were generated in the case of a small sample, and a total of five recognition tests were performed. The effects of three different kernel functions on the recognition accuracy rate were tested. The recognition rate of the radial kernel function was the highest, and the recognition accuracy reached 94.4%.

Finally, the model is optimized and the sequence minimum optimization algorithm is adopted. The SMO optimization algorithm not only improves the efficiency, but also improves the accuracy of the recognition test by 0.6%. The optimized recognition accuracy reached 95%.

Acknowledgement. This work was supported by the National Natural Science Foundation of China (61602127), the Basic scientific research project of Heilongjiang Province (KJCXZD201704), the Key Laboratory of Police Wireless Digital Communication, Ministry of Public Security (2018JYWXTX01), and partly by the Harbin research found for technological innovation (2013RFQXJ104) national education and the science program during the twelfth five-year plan (FCB150518). The authors would like to thank all the people who participated in the project.

References

1. Weidang, L., Yi, G., Xin, L., Jiaying, W., Hong, P.: Collaborative energy and information transfer in green wireless sensor networks for smart cities. IEEE Trans. Ind. Inform. **14**(4), 1585–1593 (2017)
2. Zhiyuan, T., Lantian, L., Dong, W., Ravichan-der, V.: Collaborative joint training with multitask recurrent model for speech and speaker recognition. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(3), 493–504 (2017)
3. Shu-sen, Z., Qing-cai, C., Xiao-long, W.: Convolutional deep networks for visual data classification. Neural Process. Lett. **38**(20), 17–27 (2013)
4. Mrazova, I., Kukacka, M.: Image classification with growing neural networks. Int. J. Comput. Theory Eng. **5**(3), 422–427 (2013)
5. Han, B., Davis, L.S.: Intelligent video surveillance systems and technology, pp. 79–103 (2010)
6. Geoffery, E.H., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science **313**(5786), 504–507 (2006)
7. Vapnik V.N.: The Nature of Statistical Learning Theory, pp. 131–145. Springer (2000)
8. Larochelle, H., Mandel, M., Pascanu, R., et al.: Learning algorithm for the classification restricted Boltzmann machine. J. Mach. Learn. Res. **13**, 643–669 (2012)