APPLICATION OF SUPPORT VECTOR MACHINE BASED SPEECH RECOGNITION TECHNOLOGY IN HUMAN-COMPUTER INTERACTION TECHNOLOGY

Wangcheng Cao
School of Computer and Information Technology, Mudanjiang Normal University, No.191, Cultural Street, Aimin District, Mudanjiang, Heilongjiang 157011, China.
wangchengc80@126.com

Abstract - Speech recognition technology, an important part of human-computer interaction technology, has become a research hotspot. This study focuses on the support vector machine (SVM) algorithm based speech recognition technology. Mel frequency cepstrum coefficient (MFCC) algorithm was used to extract feature parameters. In order to make SVM algorithm more reliable, a fuzzy SVM was proposed. Then the algorithm was validated by the collected speech samples. The algorithm was found having shorter training time and higher recognition rate compared to hidden Markov model (HMM) algorithm and SVM algorithm. When the feature dimension was 16, the recognition rate was as high as 88.6%. In the study of anti-noise performance, the fuzzy SVM algorithm was also better than the other two algorithms. When the noise rate reached 50%, the algorithm still had a recognition rate more than 80%, which suggested the high reliability of the algorithm and the application prospect of the technology in human-computer interaction.

Keywords: Human-Computer Interaction, Support Vector Machine, Speech Recognition, Mel Frequency Cepstrum Coefficient.

1. Introduction

With the development of science and technology, human-computer interaction has become a reality and a part of people's daily life. Language and speech are the most effective means of transmitting information, one of the most ideal human-computer interaction means. As an important part of human-computer interaction, speech recognition technology has also been studied more and more deeply. Speech based human-computer interaction technology has broad prospects for development [1].

At present, the most common speech recognition technology is based on hidden Markov model (HMM) and neural network [2, 3]. Dahl et al. [4] proposed a deep neural network (DNN)-HMM structure and found that the recognition accuracy of this structure was significantly improved compared to the traditional Gaussian mixture model. Abdel-Hamid et al. [5] used convolutional neural network (CNN) for speech recognition and found that the recognition error rate was 6% ~ 10% lower.

Support Vector Machine (SVM) developed in recent years has been found to have a good performance in speech recognition.

Zarrouk et al. [6] combined SVM with HMM for continuous Arabic speech recognition and found that SVM-HMM algorithm has superior performance, with a recognition rate of 74.01%. Bai et al. [7] used artificial fish swarm algorithm to improve the anti-noise ability of SVM algorithm. In this study, Mel frequency cepstrum coefficient (MFCC) algorithm was used to extract the features of speech signals, and then a fuzzy SVM algorithm which combined fuzzy thought with SVM was proposed.

2. Speech Recognition

Language is the main way of information transmission and emotional communication, and it is also a unique function of human beings. Speech recognition technology can realize the communication between machines and human beings, and it is an effective way of human-computer interaction. With the development of science and technology, speech recognition has gained more and more development. At present, speech recognition technology has been applied in intelligent translators, voice phones, intelligent toys, home robots and other fields.

The basic principle of speech recognition is to preprocess the input speech signal and get the recognition result by feature extraction and comparison with the trained model. The structure of speech recognition system is shown in Figure 1.

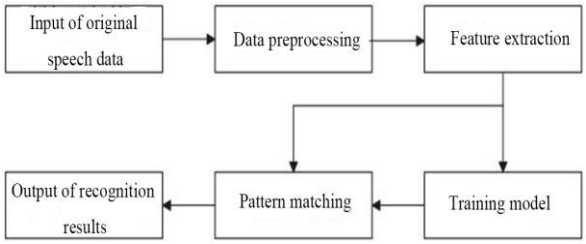

Figure 1: The structure of the speech recognition system

3. Extraction of Feature Parameters

Feature extraction of speech signal can judge unknown speech according to the characteristics of different speeches, and the extracted feature parameters should fully represent the characteristics of the speech. The characteristic parameters of speech signal are time domain and frequency domain. The frequency domain includes linear prediction coefficient, LPC cepstrum coefficient (LPCC), MFCC and so on [8, 9]. MFCC is simple and discriminative, so MFCC was chosen as the characteristic parameter in this paper.

The relationship between Mel frequency and actual frequency is:

$$f_{mel} = 2529\lg(1+f/700),$$

where f_{mel} stands for Mel frequency and f stands for actual frequency.

In MFCC algorithm, signal frequency axis needed to be converted to Mel scale, and the computational process was as follows.

Data were preprocessed by continuous subsection, and speech signals were processed by framing and windowing.

The time domain signal $x(m)$ obtained after preprocessing was filled up with 0 to be a sequence whose length was N, and discrete short time Fourier transform was performed to obtain linear frequency spectrum $X(j)$.

The square of frequency spectrum amplitude was calculated to obtain signal energy spectrum.

A triangular filter was constructed. The output spectrum of all frequency bands was obtained through band-pass filtering. Mel frequency scales was aligned evenly to obtain the middle frequency of the triangular band-pass filter. The output of the triangular filter was:

$$n(i) = \sum_{j=l(i)}^{h(i)} H_i(j)\left|X(j)\right|^2, \quad i = 1,2,\cdots,L$$

$$H_i(j) = \begin{cases} \dfrac{j-l(i)}{c(i)-l(i)} & l(i) \le j \le c(i) \\ \dfrac{h(i)-j}{h(i)-c(i)} & c(i) \le j \le h(i) \end{cases}$$

where $l(i)$, $c(i)$ and $h(i)$ stands for the lower limit, central and upper limit frequency of the filter respectively, L stands for the number of filters, and $H_i(j)$ stands for the filter.

MFCC could be calculated using

$$C_m = \sum_{i=1}^{L} \log n(i) \cos\left[\pi m(i-0.5)/L\right], m = 1,2,\cdots,\frac{L}{2}.$$

4. SVM

SVM Algorithm

If there was a training sample set $\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$, where $x_i \in X, y_i \in Y = \{1,-1\}$. If there existed $g(x) = w\cdot x + b$ and linear discrimination function was $g(x) = w\cdot x + b$, then the classification interval was:

$$d(w,b) = \min_{\{x_i|y_i=1\}} \frac{w\cdot x_i + b}{|w|} - \max_{\{x_i|y_i=-1\}} \frac{w\cdot x_i + b}{|w|} = \frac{1}{|w|} - \frac{-1}{|w|} = \frac{2}{|w|}$$

If $y_i\left[(w\cdot x_i)+b\right] \ge 0, i = 1,2,\cdots,N$ and the classification interval was the maximum, then it was the optimal classification surface.

Linearly separable SVM could be rewritten into an optimization problem.

$$\begin{cases} \min \quad \dfrac{\|w\|}{2} \\ st. \quad (y_i(w\cdot x_i)+b) - 1 \ge 0, i = 1,2,\cdots,N \end{cases},$$

where w and $b \in R$, $X_n \in R^M$ stand for characteristic vectors, and $y_n \in (-1,1)$ stands for the value of affiliated category. Lagrangian multiplier was used for solution, and the problem could be written as:

$$w(\alpha) = \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i\alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{n} \alpha_i,$$

where α_i refers to Lagrangian multiplier.

$$f(x) = \mathrm{sgn}(\sum_{i=1}^{n} \alpha_i y_i (x_i \cdot x) + b).$$

The final classification function was:

$$f(x) = \mathrm{sgn}(\sum_{i=1}^{n} \alpha_i y_i (x_i \cdot x) + b).$$

If it is linearly inseparable, then slack variable ζ was introduced. The objective function was:

$$\begin{cases} \min \quad \frac{\|w\|^2}{2} + C\sum_{i=1}^{N}\zeta_i \\ y_i[(w\cdot x_i)+b] \geq 1-\zeta \quad i=1,2,\cdots,N \\ \zeta \geq 0 \end{cases},$$

where C stands for penalty factor.

$w(\alpha) = \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i \quad x_j) \quad \sum_{i=1}^{n}\alpha_i$ was obtained through solution based on Lagrangian multiplier, where $K(x_i, x_j)$ stands for kernel function. The final classification function was

$$f(x) = \text{sgn}(\sum_{i=1}^{n}\alpha_i y_i K(x_i \quad x) + b).$$

Common kernel function

Linear kernel function was:

$$K(x, x_i) = x \cdot x_i.$$

Polynomial Kernel Function was:

$$K(x, x_i) = [(x \cdot x_i) + 1]^p,$$

where P stands for number of polynomial order.

Radial basis kernel function was:

$$K(x, x_i) = \exp(-\frac{\|x - x_i\|^2}{\sigma^2}.$$

Different kernel functions will affect the classification performance of SVM. It was found that radial basis kernel function had a good performance. Therefore radial basis kernel function was selected in this study.

Fuzzy SVM

To further improve the recognition performance of SVM, fuzzy thought was combined with SVM to be a fuzzy SVM.

For decision function $D_i = \sum_{i=1}^{n} a_i y_i K(x_i, x) + b$, input vector x was classified as:

$$x_i \in \begin{cases} Class1, D_i(x) = 1 \\ Class2, D_i(x) = -1 \end{cases}$$

If the decision function of class i and j was $D_{ij}(x) = w_{ij}^T x + b_{ij}$ and moreover $D_{ij}(x) = -D_{ji}(x)$, then

$$D_i(x) = (\sum_i sign(D_{ij}(x))),$$

vector X belonged to class $\arg\max_{i=1,2,\cdots n} D_i(x)$.

If $D_{ij}(x) \geq 1$ or $D_{ij}(x) \leq 1$, x completely belonged to class i or j. If $1 < D_{ij}(x) < 1$, then it meant degree of association between x and class i and j, and degree of membership was the linear function of $D_{ij}(x)$.

$$m_i = f(D_{ij}(x)) = \begin{cases} 1, D_{ij}(x) \geq 1 \, or \, D_{ij}(x) \leq -1 \\ |D_{ij}(x)|, -1 < D_{ij}(x) < 1 \end{cases}$$

5. Application of SVM Based Speech Recognition

Speech recognition system

In the speech recognition system designed in this study, the collected speech signals were preprocessed by pre-emphasis and endpoint detection, features of speech signals were extracted using MFCC algorithm, and the extracted feature parameters were taken as the input of fuzzy SVM for training and recognition.

Collection and preprocessing of speech samples

Voice of 6males and 6 females was recorded via computer sound card in a laboratory. Everyone read 50 words thrice, and totally 1800 speech samples were obtained, among which, 1000 samples were taken as training samples and 800 as testing samples.

Transfer function $H(z) = 1 - \mu z^{-1}$ was used in the pre-emphasis process of samples, where μ stands for pre-emphasis coefficient, 0.9375. The frame length of framing and windowing was 240 sampling points (30 ms), and the frame shift was 80 sampling points (10 ms). Hamming window was added. Then endpoint detection was performed using double threshold method through zero rate and short-time energy.

Recognition results

Firstly, SVM was used to analyze the recognition rate of MFCC algorithm. The order was set between 12 and 16. The recognition time and recognition rate are shown in Table 1.

Table 1 The recognition time and rate of MFCC under different dimensions

Feature dimension	Order				
	12	13	14	15	16
Time (s)	16.03	16.24	16.35	16.76	16.89
Recognition rate (%)	80.1	82.1	83.6	85.7	88.6

As shown in Table 1, the recognition rate of MFCC algorithm was above 80%; with the increase of dimension, the recognition time of MFCC algorithm increased, but the change was not obvious; the recognition rate of MFCC algorithm increased with the increase of dimension, but model parameters describing speech features also increased, which improved the difficulty of model establishment. Generally the dimension was set as 16.

In order to verify the reliability of this method, HMM, SVM and fuzzy support vector machine (FSVM) were used for speech recognition. The recognition results are shown in Table 2.

Table 2 The recognition results of different algorithms

	HMM	SVM	FSVM
Recognition rate (%)	92.3	94.2	97.8
Training time (min)	5.2	2.1	0.8

As shown in Table 2, the recognition rate of the proposed FSVM algorithm was the highest, up to 97.8%, which was obviously better than the HMM algorithm and SVM algorithm. Moreover, the training time of the algorithm was also significantly shorter than the other algorithms, which showed that the algorithm had better recognition performance and high reliability.

In order to verify the anti-noise ability of the algorithm, uniformly distributed noises were added to the samples. The recognition results of the three algorithms under different noises are shown in Table 3.

Table 3 The recognition results of different algorithms under different noises

Noise ratio	Recognition effect	HMM	SVM	FSVM
10%	Recognition rate (%	87.2	90.1	96.3
	Training time (min)	7.3	4.1	1.8
20%	Recognition rate (%	83.2	85.6	90.3
	Training time (min)	8.9	5.6	2.1
30%	Recognition rate (%	78.6	80.1	88.7
	Training time (min)	9.7	6.2	2.9
50%	Recognitio rate (%)	70.2	76.3	82.1
	Training time (min)	10.8	8.2	3.4

It can be noted from Table 3 that noise had a great influence on the recognition rate of the algorithm. With the increase of the noise ratio, the recognition rate of the algorithm decreased to some extent, and the training time increased. When the noise ratio was 50%, the recognition rate of HMM algorithm decreased to 70.2%, that of SVM algorithm was only 76.3%, and the recognition rate of FSVM algorithm kept above 80%. Although the training time increased, the recognition rate of HMM algorithm decreased to 70.2%. The HMM algorithm and SVM algorithm were still relatively short, indicating that FSVM algorithm remained a relatively good recognition rate even under the influence of noise.

6. Discussion

About 75% of human daily communication is achieved through language. With the development of science and technology, how to realize the communication between people and computers has got more and more attentions, and human-computer interaction technology has been continuously developed [10]. Speech recognition can achieve simple man-machine interaction [11, 12]. Speech recognition technology has great potential for development and has application values in many fields, such as smart phone, smart home and smart robot.

Traditional speech recognition methods include HMM, Neural Network, etc. [13, 14]. SVM has been found to have better generalization performance and recognition efficiency and has been widely used in speech recognition. In order to further improve the recognition performance of SVM, various optimization was carried out. In this study, a FSVM was put forward based on fuzzy thought, which could further optimize the classification surface of SVM and significantly improve the anti-noise ability.

Feature extraction is an important part of speech recognition. The most commonly used feature parameters are LPCC and MFCC. LPCC is mostly used to extract the features of different speakers' pitches. MFCC based on hearing has better performance in speech recognition. Therefore, MFCC algorithm was selected to extract the features of speech signals. The performance of the system was verified. It was found that the recognition rate of the algorithm was more than 80% and the recognition time was shorter, suggesting the high feasibility of the algorithm.

Through the collection of different speech samples, this study compared the recognition effects of HMM, SVM and FSVM. It was found that FSVM had the highest recognition rate and the shortest recognition time under the same conditions, which verified the high reliability of the algorithm.

Noise often exists in actual speech recognition, and noise will greatly affect the accuracy of speech recognition [15, 16]. In order to verify the anti-noise ability of FSVM algorithm, the recognition effect was analyzed when the noise ratio was set as 10%, 20%,

30% and 50%. The results show that the recognition effect of FSVM was the best under the same noise conditions. When the noise ratio was 50%, the recognition rate of HMM algorithm was 70.2%, that of SVM algorithm was 76.3%, and that of FSVM algorithm was 82.1%, suggesting a good anti-noise performance.

7. Conclusion

Speech recognition can realize speech communication between human and machine to provide convenience for people's life. It has a broad development prospect. In this paper, the application of SVM in speech recognition was analyzed. MFCC algorithm was used to extract the characteristic parameters of speech signal. Then SVM was improved with fuzzy thought. The experiment showed that FSVM had better recognition performance and anti-noise ability. When the noise ratio was 50%, the recognition rate of FSVM was 82.1 %, indicating the algorithm had a high reliability in speech recognition.

Acknowledgement

This study was supported by Research on the Application of SVM in Human-Computer Interaction under grant number 1351MSYYB002 and Research on Some Key Technologies of the Internet of Things Architecture and Intelligent Information Processing Theory under grant number YB2018005.

References

[1] Katore M, Bachute M R. “Speech based human machine interaction system for home automation. Bombay Section Symposium.” IEEE, 2015:1-6.

[2] Hinton G, Deng L, Yu D, et al. “Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups.” IEEE Signal Processing Magazine, 2012, 29(6):82-97.

[3] Seltzer M L, Yu D, Wang Y. “An investigation of deep neural networks for noise robust speech recognition.” IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2013:7398-7402.

[4] Dahl G E, Acero A. “Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition.” IEEE Transactions on Audio Speech & Language Processing, 2011, 20(1):30-42.

[5] Abdel-Hamid O, Mohamed A R, Jiang H, et al. “Convolutional Neural Networks for Speech Recognition.” IEEE/ACM Transactions on Audio Speech & Language Processing, 2014, 22(10):1533-1545.

[6] Zarrouk E, Ayed Y B, Gargouri F. “Hybrid continuous speech recognition systems by HMM, MLP and SVM: a comparative study.” International Journal of Speech Technology, 2014, 17(3):223-233.

[7] [7] Bai J, Yang L, Zhang X. “An anti-noise SVM parameter optimization method for speech recognition.” Journal of Central South University, 2013, 44(2):604-611.

[8] Burak Tombaloğlu, Erdem H. “Development of a MFCC-SVM Based Turkish Speech Recognition system. Signal Processing and Communication Application Conference.” IEEE, 2016:929-932.

[9] Chen T. “Study of Speech Recognition Technology Based on MFCC and SVM.” Journal of Guangxi Vocational & Technical College, 2010(5):1-4.

[10] Grudin J, Carroll J M. “From Tool to Partner:The Evolution of Human-Computer Interaction.” Extended Abstracts of the Chi Conference. Morgan & Claypool, 2017:183.

[11] Zhang X Q, Chen S W. “Speech recognition system based on DSP and SVM.” International Conference on Machine Learning and Cybernetics. IEEE, 2010:2313-2316.

[12] M. A, Ganesh B, Ratnadeep R. “Automatic Speech Recognition and Verification using LPC, MFCC and SVM.” International Journal of Computer Applications, 2015, 127(8):47-52.

[13] Frihia H, Bahi H. “HMM/SVM segmentation and labelling of Arabic speech for speech recognition applications.” International Journal of Speech Technology, 2017, 20(1):1-11.

[14] Graves A, Mohamed A R, Hinton G. “Speech recognition with deep recurrent neural networks.“ IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2013:6645-6649.

[15] Bai J, Xue P, Zhang X, et al. “Anti-noise Speech Recognition System Based on Improved MFCC Features and Wavelet Kernel SVM.” Advances in Information Sciences & Service Sciences, 2012, 4(23):599-607.

[16] Qian Y, Bi M, Tan T, et al. “Very Deep Convolutional Neural Networks for Noise Robust Speech Recognition.” IEEE/ACM Transactions on Audio Speech & Language Processing, 2016, 24(12):2263-2276.