a a

Fuhai Li[1], Jinwen Ma[1,**], and Dezhi Huang[2]

1 , A ,
g s , g 100871,
lfh@water.pku.edu.cn, jwma@math.pku.edu.cn
2 g , s g s ,
g 100084,

Abstract. g s s s s - s s g s g. , s s s . s , s g s s s ( ) - - s (s) s 's s. s s g g s s s g s ( ) s s 's s.

I

Recently, automatic speech processing becomes very important and significant since it can contribute to the natural language recognition and understanding as well as the applications to robot and automation. In fact, speech processing technology has been developed very quickly in many fields, especially in the Internet, telecom and security. Speech recognition, understanding and synthesis are the main contents of the speech processing [1, 2, 3, 4], in which the vowel recognition generally plays an important role.

Chinese language (Mandarin) is the largest natural language in the world. So it is very important and significant to study the Chinese speech processing. Actually, Chinese has five major vowels, i.e., /a/, /o/, /e/, /i/, /u/. Clearly, the recognition of the five Chinese vowels is very useful for Chinese speech recognition and understanding. However, there has not been any e cient method to do such a di cult task, to our best knowledge. For the purpose of Chinese vowel recognition, there are two tasks. The one is to extract a set of features from each Chinese vowel such that should be

In signal processing, cepstral analysis has turned out to be an e cient tool for speech analysis [4]. Actually, the Mel-Frequency Cepstral Coe cients (MFCCs) [2, 4] can be considered as a favorable choice of the features of a vowel. Moreover, since the number of features for a vowel is large, the Supporter Vector Machine (SVM) [5, 6, 7, 8, 9] can be used to classify the vowels with the selected features.

In this paper, we will use the MFCCs as the vowel's features and the SVM as the classifier for Chinese vowel recognition. It is shown by the experiments that the MFCCs and SVM based Chinese vowel recognition system can reach good recognition accuracy on the given data and outperform the SVM with the Linear Prediction Coding Cepstral (LPCC) coe cients as the vowel's features.

q y a

We begin to introduce the MFCCs as the features of the vowels. Actually, the MFCCs is one of the cepstral analysis technology that has been applied to speech analysis successfully. The Mel frequency describes that the human's apperception of the frequency is nonlinear. The following formula shows the relationship between the Mel frequency and the real frequency:

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

where f is the real frequency and the unit is Hz.

The extraction process of MFCCs is given as follows.

(1). **Pretreatment**

A. Pre-emphasis. The aim of pre-emphasis is to elevate the high frequencies of the speech and make the spectrum smooth. The used first-order digital filter and it's output are given as follows:

$$H(z) = 1 - \mu z^{-1}. \tag{2}$$

$$y(n) = x(n) - \mu x(n-1). \tag{3}$$

where $\mu = 0.9375$ and $x(n)$is the original signal.

B. Windowing. We divide each speech signal into certain frames by Hamming sliding window $h_j(n)$ whose length is 512 and o set is 128.

$$h\ (n) = \begin{cases} 0.54 - 0.46 \quad \text{s } 2n\pi/(N-1), & 128(j-1) \le n \le N-1+128(j-1); \\ 0, & \text{s .} \end{cases} \tag{4}$$

$$s\ (n) = y(n)h\ (n). \tag{5}$$

where $h_j(n)$ is the $j-th$ window, $s_j(n)$ is the $j-th$ frame signal (We neglect the zero signals outside the Hamming window), and $j = 1, 2, ..., 13$.

(2). **The Spectrum Calculation**

We then calculate the amplitude spectrum [4] of each frame $|X_j(k)|$ through the Discrete Fourier Transformation as follows.

$$X_j(k) = \sum_{n=0}^{N-1} s_j(n) e^{-j\frac{2\pi}{}}. \tag{6}$$

(3). $m(l)$ **Calculation**
We further calculate the Mel-scaled triangular filter's output $m(l)$. The triangular filters are equal interval in the mel-frequency scale.

$$m(l) = \sum_{k=o(l)}^{k(l)} W_l(k)|X_j(k)|, l = 1, 2, ..., L. \tag{7}$$

where

$$W_l(q) = \begin{cases} \frac{k-o(l)}{c(l)-o(l)}, o(l) \leq k \leq c(l); \\ \\ \frac{h(l)-k}{h(l)-c(l)}, c(l) \leq k \leq h(l). \end{cases} \tag{8}$$

and $o(l)$,$c(l)$,$h(l)$ are the triangular filter's lower limit, central, and upper limit respectively. L is the number of the triangular filters and we select L = 16.

(4). **MFCCs Calculation**
Finally, the MFCCs can be got via the Discrete Cosine Transformation.

$$c_{\text{mfcc}}(i) = \sqrt{2/L}\sum_{l=1}^{L} \log m(l) \cos[(l - 1/2)i\pi/L]. \tag{9}$$

. . . . . y. . . a

In this section, we introduce the SVM and then construct a Chinese vowel recognition system via SVM.

3.1 Support Vector Machine (SVM)

The basic ideas of the SVM is to find the optimal hyperplane of two linear separation classes. Figure 1 shows an intuitional example in the two dimension case, where the white and black balls represent two type of samples, respectively. The line H separates the two types of balls without any mixture. H1 and H2 drill through balls which are the nearest to H and are parallel to H. The word of "margin" means the distance between H1 and H2. The so-called optimal hyperplane (Here we means the optimal classification line) is to maximize the margin. In this case, the samples on the lines H1 and H2 are called support vectors which actually support the optimal hyperplane.

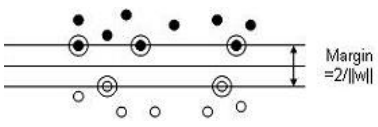


Fig. 1. ss f s - ss s

Mathematically, we consider two linearly separated classes C1 and C2 with the samples set $\{(x_i, y_i), i = 1, 2, ..., n, x \in \mathbb{R}^m, y \in \{+1, -1\}\}$. And the linear classification function takes the general form of

$$g(x) = \omega \cdot x + b. \tag{10}$$

Via certain normalization, all the samples in the two classes satisfy $|g(x)| \geq 1$, i.e., the samples which are the nearest to the hyperplane H satisfy $|g(x)| = 1$. So the margin is equal to $2/\parallel \omega \parallel$. In this setting, the problem of solving the optimal hyperplane can be described as the following optimization problem:

$$\text{minimize} \qquad \phi(\omega) = \frac{1}{2} \parallel w \parallel^2 = \frac{1}{2}(\omega \cdot \omega); \tag{11}$$

$$\text{subject to} \qquad y_i[(w \cdot x_i) + b] - 1 \geq 0. \tag{12}$$

By the method of Lagrange Multipliers, We can get the following dual optimization problem:

$$\text{maximize} \quad Q(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i x_j); \tag{13}$$

$$\text{subject to} \quad \sum_{i=1}^{n} \alpha_i y_i = 0, \ \ and \ \ \alpha_i \geq 0, \ \ i = 1, 2, ..., n. \tag{14}$$

In this way, the optimization problem turns to find a quadratic function's extremum with a linear equation and positive constraints. Actually, it has a unique solution. If the optimal solution is α_i^*, we have

$$w^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i. \tag{15}$$

According to the Kühn-Tucker condition, generally there are only small part of α_i^* are nonzero which correspond to the support vectors.

Finally, we have the discriminant function:

$$g(x) = \text{sgn}\{(\omega^* \cdot x) + b^*\} = \text{sgn}\{\sum_{i=1}^{n} \alpha_i^* y_i (x_i \cdot x) + b^*\}. \tag{16}$$

where b^* is the threshold value and can be calculated by any support vector via the inequality (12).

From the function (16), we notice that the classification function just has relation with the inner products of the new sample and the support vectors. For the nonlinear case, we can project the original feature space into a higher dimension space in which the samples are linearly separable. But for SVM, we just need to find an appropriate inner product $K(x_i, x_j)$, to replace (x_i, x_j), i.e., we do nothing for the space transformation. In the new feature space, the objective function and the discriminant function are:

$$Q(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j). \tag{17}$$

$$g(x) = \text{sgn}\{(\omega^* \cdot x) + b^*\} = \text{sgn}\{\sum_{i=1}^{n} \alpha_i^* y_i K(x_i, x) + b^*\}. \tag{18}$$

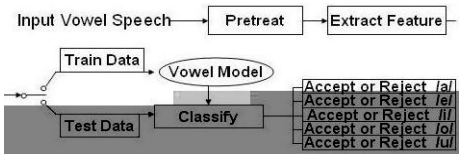
There are three commonly used kernel functions:

(1). linear: $K(x_i, x_j) = x_i^T x_j$;
(2). polynomial: $K(x_i, x_j) = (x_i^T x_j + 1)^q$;
(3). radial basis function (RBF): $K(x_i, x_j) = \exp\{-\frac{|x - x|^2}{\sigma^2}\}$.

For the multi-class problem, we can combine some two-class SVMs together. There are two possible ways to do such a task: “one vs. all” and “one vs. one”, which are given and compared in detail in [9]. Actually, there are many softwares of SVM available on the web and we will use a typical software of SVM given at a website.

3.2 The Chinese Vowel Recognition System

Based on the MFCCs and SVM, we construct the Chinese vowel recognition system. Figure 2 gives the frame of the Chinese vowel recognition system. The sample data are divided into 2 groups: train data and test data. We firstly use the train data to construct the vowel model via the SVMs in the combination of “one vs. one”, and use the test data to check the vowel model. The output of the system is to give the class of an input vowel speech signal.



4.1 Experimental Results

For comparison, we use the three kinds of SVMs for the Chinese vowel recognition system: Radial basis function SVM (RBF Kernel), 3-poly SVM (cubic polynomial kernel), and Linear SVM (no kernel). We get the SVM software from website: http://www.csie.ntu.edu.tw/ cjlin/libsvmtools/. We use the 4-fold cross-validation test method. We divide the total 1567 vowel samples into 5 classes according to their class labels and divide each class into 4 folds averagely. In each experiment, we choose one fold of each class as the test data set and the remainders as the train data set. Then, we conduct 4 times experiments. At last, we use the average accuracy of the 4 times experiments as the final recognition or classification result. Tables 1-3 give the average accuracies of the vowel recognition system with the di erent kinds of SVMs.

Table 1. [illegible] s (s [illegible] s) s [illegible] s

	/ /	/ /	/■/	/ /	/ /
/ /	96.91%	0.00%	0.00%	3.09%	0.00%
/ /	0.27%	96.98%	1.92%	0.82%	0.00%
/■/	0.00%	0.60%	94.50%	0.89%	4.46%
/ /	3.63%	2.07%	2.07%	85.75%	6.48%
/ /	0.00%	0.00%	12.80%	28.80%	58.40%

Table 2. [illegible] s (s [illegible] s) s [illegible] 3- [illegible] s

	/ /	/ /	/■/	/ /	/ /
/ /	99.16%	0.00%	0.00%	0.84%	0.00%
/ /	0.00%	98.35%	0.55%	1.10%	0.00%
/■/	0.00%	0.00%	97.32%	0.60%	2.08%
/ /	1.55%	1.30%	0.52%	94.04%	2.59%
/ /	0.00%	0.00%	16.00%	12.00%	72.00%

Table 3. [illegible] s (s [illegible] s) s [illegible] s

	/ /	/ /	/■/	/ /	/ /
/ /	98.60%	0.00%	0.00%	1.40%	0.00%
/ /	0.00%	98.90%	0.27%	0.82%	0.00%
/■/	0.00%	0.30%	98.51%	0.30%	0.89%
/ /	1.04%	1.04%	0.00%	96.11%	1.81%
/ /	0.00%	0.00%	10.40%	12.80%	76.80%

From tables 1-3, we can find that the MFCCs and SVM based Chinese vowel recognition system is very e cient to recognize the vowels in Chinese speech. Actually, the recognition accuracy on each vowel is very high. Comparing the three kinds of SVMs, we can find that on the average, the Chinese vowel recognition system with the RBF SVMs is slightly better than that with the linear or 3-poly SVMs. However, the system with the 3-poly SVMs get the highest recognition accuracy 99.16% for the vowel /a/.

4.2 Comparisons with the LPCC Based System

We further compare our Chinese vowel recognition system with the same SVM system under the LPCC coe cients as the features of a vowel. Actually, the LPCC coe cients are also widely used in the speech signal processing and have certain favorable properties on the vowel recognition. Here, each vowel also is

Table 4. g g s (s s) -
s g s s

	/ /	/ /	/ /	/ /	/ /
/ /	90.73%	0.00%	0.00%	9.27%	0.00%
/ /	1.10%	96.43%	0.27%	2.20%	0.00%
/ /	0.00%	0.89%	91.37%	2.38%	5.36%
/ /	7.77%	4.15%	1.55%	81.61%	4.92%
/ /	0.80%	0.00%	30.4%	20.80%	48%

Table 5. g g s (s s) -
s g s s 3-

	/ /	/ /	/ /	/ /	/ /
/ /	95.51%	0.00%	0.00%	4.49%	0.00%
/ /	0.82%	97.53%	0.00%	1.65%	0.00%
/ /	0.00%	0.60%	94.35%	1.49%	3.57%
/ /	5.70%	4.40%	1.30%	84.72%	3.89%
/ /	0.00%	0.00%	26.40%	25.60%	48.00%

Table 6. g g s (s s) -
s g s s

	/ /	/ /	/ /	/ /	/ /
/ /	96.35%	0.00%	0.00%	3.65%	0.00%
/ /	0.00%	98.90%	0.27%	0.82%	0.00%
/ /	0.00%	0.00%	94.35%	1.49%	4.17%
/ /	3.89%	4.66%	0.26%	88.86%	2.33%
/ /	0.00%	0.00%	23.2%	24.00%	52.80%

Table 7. s g s s s s g s s

	/ /	/ /	/ /	/ /	/ /
	6.18%	0.55%	2.68%	4.14%	10.40%
3-	3.65%	0.82%	2.97%	9.32%	24.00%
	2.25%	0.00%	4.16%	7.25%	24.00%

represented by a 208 (13 ∗ 16) LPCC coe cients. Tables 4-6 show the average accuracy of Chinese vowel recognition system with three kinds of SVMs. We can see that the Chinese vowel recognition system based on the LPCC coe cients can also lead to a high recognition accuracy, but its recognition accuracy is considerably lower than that of the MFCCs based system, which is given more accurately in Table 7.

We have investigated the recognition of Chinese vowels. The MFCCs of a vowel signal are extracted as the features of the vowel and we then use the SVMs to construct the Chinese vowel recognition system. It is shown by the experiments that this MFCCs and SVM based system can reach a high recognition accuracy on the given vowels database and outperform the SVM with the Linear Prediction Coding cepstral (LPCC) coe cients as the features of each vowel.

1. , , g, .: s g . , g - s, (1993)
2. , ., , ., , .: g s s g . g A s (s), 5 (2002) 127 128
3. g, ., , ., .: g ss g g s (s). s - g s s g (1995)
4. , g ss g (s). ss, g (2003)
5. , . ., , . ., , . .: ss . 2 (s). ss s g s , g (2003)
6. , ., g, ., .: g (s). s g s ss, g (2000)
7. , .: s g (1998)
8. g, ., , . .: s A - g s A s g . g , 36 (2003) 2429 2439
9. s, . ., , . .: A s s - ss s. s. s, 13 (2002) 415 425