Tone Recognition of Continuous Mandarin Speech Based on Binary-Class SVMs

Shui-ping Wang[1,2], Zhen-ming Tang[1], Ying-nan Zhao[2], Sai Ji[2]
1. Department of Computer Science & Technology,
Nanjing University of Science & Technology, Nanjing, China
2. Department of Computer Science & Technology,
Nanjing University of Information Science & Technology, Nanjing, China
shuipingw@126.com, tang.zm@mail.njust.edu.cn, yingnanzhao@163.com,jisai@nuist.edu.cn

***Abstract*—Tone is an essential component for word formation in all tone languages. It plays a very important role in the transmission of information in speech communication. In this paper, we look at using support vector machines (SVMs) for automatic tone recognition in continuously spoken Mandarin. Wavelet transform and Teager Energy Operation (TEO) are used to detect the voiced segments. Considerable improvement has been achieved by adopting binary-SVMs scheme in a speaker-independent Mandarin tone recognition system.**

Keywords-tone language; tone recognition; fundamental frequency; support vector machines

I. INTRODUCTION

Mandarin Chinese is a tonal and syllabic language, in which, different characters may have the same base-syllable but with different tones and with different information. Speech recognition of tone languages depends not only on the articulatory composition but also on the tone patterns. Mandarin contains five tones, characterized by fundamental frequency (F_0) contour pattern: 'high-level', 'high-rising', 'low-dipping', 'high-falling' and 'neutral'. Different tones have different fundamental F_0 pattern. In continuous speech, the neutral tone which is not considered in this paper always occurs in word-end or sentence-end and dose not have a stable F_0 contour.During the last decades, many approaches have been proposed for tone recognition. Hidden Markov Models (HMM) [1][2], Neural Networks [3], Decision-tree Classification[4], Support Vector Machine (SVM) [5] have been applied to recognize tones in tone languages, such as Mandarin, Cantonese and Thai. Among the approaches, Support Vector Machine is an excellent pattern classifier, which is proposed in resent years. It has unique advantages in solving non-linear classification problems of small samples [5]. SVM is a static classifier, the dimension of model in which should be consistent, however speech signals are of typical dynamic models, and the dimension of each syllable feature vectors is variable. To solve this problem, a new tone feature homogeneous method based on curve fitting is proposed in this paper. Finally, experiments are carried out based on SVM and BP Neural Network.

II. SYLLABLE SEGMENTATION

As tone information is mostly focused on the part of the voiced speech, syllable segmentation is carried out before tone feature extraction. In this paper, the voiced segments are detected based on wavelet transform. Specific processes are as follows:

(1) Wavelet Transform

Four-level wavelet decompositions are done for speech signal *f(n)*, *W(j, k)* is the forth wavelet coefficients of db 2.

(2) TEO nonlinear operation

Teager Energy Operation (TEO) was a nonlinear operation, proposed by Kaiser [6], which was applied in speech signal procession successfully [7].

e (j, k) is the answer of *W(j, k)* under TEO operator.

$$e(j,k) = T[W(j,k)] \tag{1}$$

(3) Amplitude contour extraction

Acquiring the modulus of *e (j, k)*, through low-pass filter *h (k)*, amplitude contour *P (j, k)* is obtained.

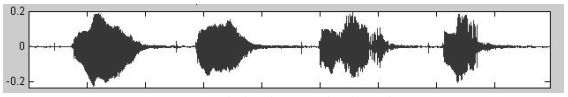
$$P(j,k) = abs[e(j,k)] * h(k), \quad k = 1,...N/2^4 \tag{2}$$

(4) Voiced segments detection

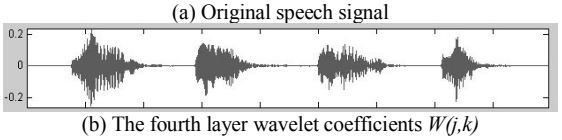
As the amplitude of voiced segments is obviously larger than the non-voiced sections, the relationship between dullness segments and threshold th is as follow. In the voiced segments, *X (j, k)* is 1.

$$X(j,k) = \begin{cases} 1, P(j,k) \geq th \\ 0, P(j,k) < th \end{cases} \tag{3}$$

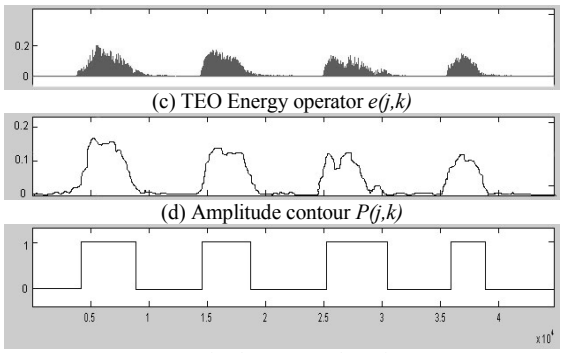
The processing results of each step are shown as Fig.1.



(a) Original speech signal



(b) The fourth layer wavelet coefficients *W(j,k)*



(c) TEO Energy operator *e(j,k)*

(d) Amplitude contour *P(j,k)*

(e) Voiced segments detection

Figure 1. Syllable segmentation processing

III. HOMOGENEOUS TONE FEATURE EXTRACTION

A. Characteristic parameters calculating

The tone of a syllable is mainly determined by its F_0 contour, and the duration and energy are also related to the tone. The different trajectory of fundamental frequency represent different tone model [8]. Fig.2 plots the spectrograms with different tones of Mandarin Chinese monosyllable shu spoken by a female adult.

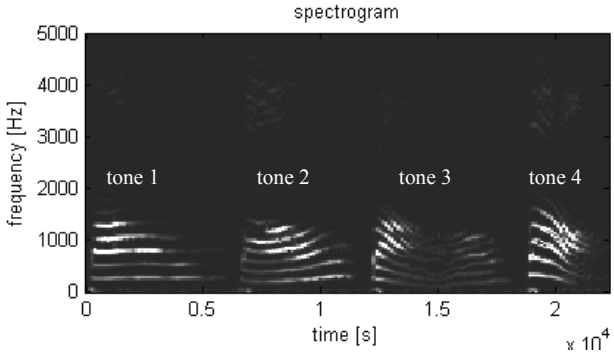


Figure 2. Spectrogram of monosyllable bao

The average magnitude difference function is useful for F_0 estimation and offers less computational complexity than the autocorrelation function. Suppose that a function *x (t)* is periodic with period *T*, then a function

$$d(n) = x(n) - x(n-k) \tag{4}$$

will be zero for *k=0, ±T, ±2T, ⋯*. For short segments of voiced speech one would expect that *d (n)* would be nearly zero, although not exactly so, at multiples of the period. The short time average magnitude of *d (n)* as a function of *k* will be small whenever *k* is close to the period. Thus, the short time average magnitude difference function (AMDF) is defined as

$$AMDF(k) = \sum_{m=-\infty}^{\infty} | x(n+m)w(m) - x(n+m-k)w(m-k) | \tag{5}$$

AMDF requires fewer multiplications than the autocorrelation function. Consequently, it has been used in real-time speech processing systems.

After detecting the F_0 contour, we use Median filter to deal with the random errors on the trajectory points. Then A group of characteristic parameters of fundamental frequency:

$$F_0 = \{f_0^1, f_0^2, f_0^3, \cdots, f_0^N\} \tag{6}$$

are detected. Where N is the total number of a voice paragraph frame. The differential F_0 parameters are expressed as:

$$\Delta F_0 = \{\Delta f_0^1, \Delta f_0^2, \Delta f_0^3, \cdots, \Delta f_0^{N-1}\} \tag{7}$$

The energy parameters are expressed as:

$$E = \{e_1, e_2, e_3, \cdots, e_N\} \tag{8}$$

Where $$e_i = \sum_{n=0}^{M-1} | x(n) |^2 \tag{9}$$

Furthermore, we use the log-scale transformation of energy parameters to match the ear characteristics of auditory perception.

$$\log_{10} E = \{\log_{10} e_1, \log_{10} e_2, \log_{10} e_3, \cdots, \log_{10} e_N\} \tag{10}$$

The differential energy parameters are expressed as:

$$\Delta \log_{10} E = \{\Delta \log_{10} e_1, \Delta \log_{10} e_2, \Delta \log_{10} e_3, \cdots, \Delta \log_{10} e_{N-1}\} \tag{11}$$

A frame-related feature vector consists of 4 parameters: $\{f_0^i, \Delta f_0^i, \log_{10} e_i, \Delta \log_{10} e_i\}$, where $1 \leqslant i \leqslant$ N-1. Associating the frame-related feature vectors, we can get a tone-related feature parameter matrix *X*, which is expressed as (12).

$$X = \begin{bmatrix} f_0^1 & \Delta f_0^1 & \log_{10} e_1 & \Delta \log_{10} e_1 \\ f_0^2 & \Delta f_0^2 & \log_{10} e_2 & \Delta \log_{10} e_2 \\ \vdots & \vdots & \vdots & \vdots \\ f_0^{N-1} & \Delta f_0^{N-1} & \log_{10} e_{N-1} & \Delta \log_{10} e_{N-1} \end{bmatrix} \tag{12}$$

B. Homogeneous

As the duration of every syllable is different from each other, and then the frame number N of each syllable is distinct, the dimension of every feature vector is certainly not the same. But, SVM can only deal with the same-dimension pattern classification problems. Therefore, we should convert the feature vectors to the new one with the same dimension. This procedure is named as feature homogeneous. In this paper, we use curve fitting method to normalize the feature vectors.

Feature parameters of each row in matrix *X* are supposed to be the sampling points on one fitting curve, which are used to calculate the fitting coefficients. Composing the fitting coefficients, a same-dimension feature vector is generated. In this paper, we use the least square method based on the Legendre Polynomials basis functions. The top 6 orders Legendre Polynomials curves are plotted as follow.

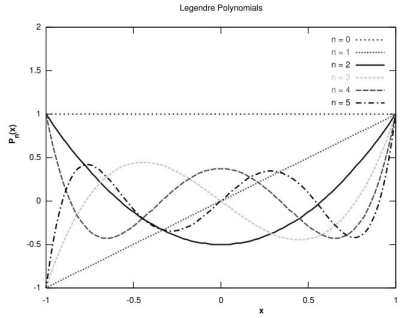


Figure 3. The top 6 orders Legendre Polynomials curves

Let $g_0^i(x) = \sum_{j=1}^{M} c_j^i e_j(x)$ be the curve constituted by the voice feature parameters of column i in matrix X. Where, $e_j(x)$ is the j order Legendre Polynomial, c_j^i is the curve fitting coefficient, and M, which is 4 here, is the order of Legendre Polynomial. Using the least square method, fitting coefficient c_j^i is obtained. A new feature matrix C composed of the normalized voice feature parameters is expressed as (13).

$$C = \begin{bmatrix} c_1^1 & c_1^2 & c_1^3 & c_1^4 \\ c_2^1 & c_2^2 & c_2^3 & c_2^4 \\ c_3^1 & c_3^2 & c_3^3 & c_3^4 \\ c_4^1 & c_4^2 & c_4^3 & c_4^4 \end{bmatrix} \tag{13}$$

IV. CLASSIFIER DESIGN

Alongside with normalization, tone recognition is also a problem of classification. In this section, SVM-based binary-class classifiers will be introduced. SVMs are a set of related supervised learning methods used for classification and regression, and possess the well-known ability of being universal approximators of any multivariate function to any desired degree of accuracy. It has been found that SVMs show better or comparable results than the outcomes estimated by neural networks and other statistical models [8]. As shown in Fig. 4, the training examples are divided into two classes, in which the empty circles belong to class Ⅰ and the black solid circles belong to class Ⅱ. All the hyperplanes, C_0, C_1 and C_2, have zero empirical error, but C_0 is the optimal hyperplane because it maximizes the distance between the H_1 and H_2, thereby offering better generalization.

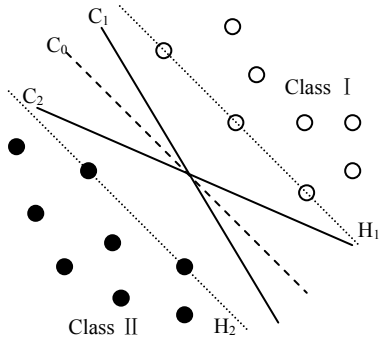


Figure 4. Binary classification problem

The binary-class SVM classifier can be extended to perform multi-class SVM classification. Six (C_4^2) binary class classifiers are trained, each of them has the structure shown as Fig 5.

Where, $K(x_i, x)$ is the inner product function. There are 3 kinds of commonly used product functions: (1) Polynomial Function, (2) Radial Basis Function (RBF), and (3) Sigmoid Nuclear Function. In this paper, we use RBF as the inner production function.

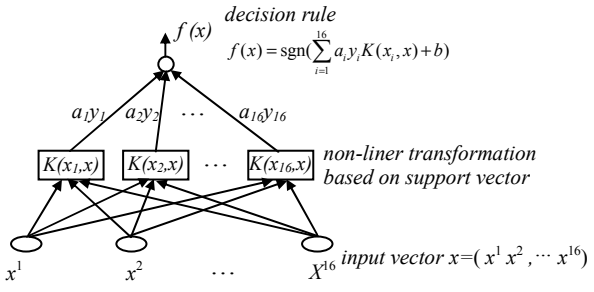


Figure 5. Support vector machine map

Tone recognition modulate matrix R is designed as Tab.1, where 1/2 represents Tone 1 vs. Tone 2, 1/3 represents Tone 1 vs. Tone 3, etc. Tone recognition modulate matrix R

TABLE I. TONE RECOGNITION MODULATE MATRIX R

One-versus-one						
	1/2	*1/3*	*1/4*	*2/3*	*2/4*	*3/4*
Tone 1	1	1	1	0	0	0
Tone 2	-1	0	0	1	1	0
Tone 3	0	-1	0	-1	0	1
Tone 4	0	0	-1	0	-1	-1

For a given token x, the 6 binary-class classifiers are applied to produce 6 hypotheses, h_1, h_2, … , h_6. Then the class label of token x can be predicted by choosing the rth row of the matrix R which is closest to (h_1, h_2, … , h_6).

V. EXPRIMENTS

A. Voice database establish

Speech samples were obtained from 12 native Mandarin Chinese speaking undergraduates (21-25 years of age), which is composed of 6 male students and 6 female students. All recordings were conducted with a sampling frequency of 22.05 kHz and 16-bit resolution. The speech samples were recordings of spontaneous productions of the four Mandarin tone pattern of 45 sets of Mandarin syllables *(a, ai, bao, bo, can, chi, du, duo, fa, fu, ge, hu, ji, jie, ke, la, ma, na, pao, pi, qi, qie, sha, shi, shu, tu, tuo, wan, wen, wu, xia, xian, xu, ya, yan, yang, yao, yi, ying, yu, yuan, yun, zan, zhi, zi)*, then ruslted in 2160 tokens(12 subjects × 45 syllables × 4 tones). Half of them (spoken by 3 males and 3 females) were chosen to be the training set, and the rest to be the testing set.

B. Experimental results

Six combinations of tone confusion include Tone 1-Tone 2, Tone 1-Tone 3, Tone 1-Tone 4, Tone 2-Tone 3, Tone 2-Tone 4, and Tone 3-Tone 4. The results on test set are expressed as

Tab.2. In the test set, each tone has 270 samples, and the total accuracy is 93.52%.

TABLE II. CONFUSION MATRIX OF TONE RECOGNITION

		heard				
		Tone 1	*Tone 2*	*Tone 3*	*Tone 4*	*Accuracy(%)*
spoken	**Tone 1**	252	6	3	9	93.33%
	Tone 2	9	250	4	7	92.59%
	Tone 3	2	3	259	6	95.93%
	Tone 4	10	6	5	249	92.22%
Overall						93.52%

Neural Network is also a non-liner classifier with wide rang of applications, which is one of the few tone recognition methods with good results [9]. In this section, comparative experiments on SVM and BP Neural Network are performed not only on the test set, but also on the training set. The results are shown in Fig.6.

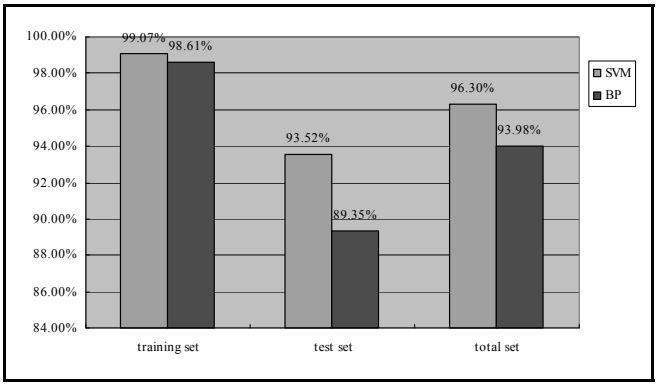


Figure 6. Different classifier results

VI. CONCLUSION AND DISCUSSION

In this paper, we have explored tone recognition in Mandarin continuous speech. Using voiced segments extraction method based on wavelet transform, speech syllables were detected correctly. Curve fitting method based on the Legendre Polynomials basis functions was used to normalize the feature vectors, then a 4×4 – dimension feature matrix was obtained. Six binary-class SVM classifiers were trained, and were used to recognize the 4 tones of Mandarin speech. From the experiments, we can draw the conclusion that the recognition rate of SVM classifier is higher than it of BP Neural Network classifier, and that SVM has stronger generalization ability than BP network.

VII. ACKNOWLEDGEMENTS

We are grateful to Xiao-ping Zhao, Feng Jiao, and Rui Ma for their technical support. This study is supported in part by National Natural Science Foundation of China (60702076).

REFERENCES

[1] Chen,X.-X.,Cai,C.-N.,Guo,P.,Sun,Y.,1987.A Hidden Mardov model applied to Chinese four-tone recognition. In: Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp.797-800.

[2] Yang,w.-J.,1988. Hidden Markov Model for Mandarin lexical tone recognition. IEEE Trans. Acoust. Speech Signal Process. 36,988-992.

[3] Emonts,M.,Lonsdale,D.,2003. A memory-based approach to Cantonese tone recognition. In: Proc. 8th European Conference on Speech Communication and Technology (EUROSPEECH), pp.2305-2308.

[4] Cao Yang et al. “Tone Recognition in Mandarin using Focus”, INTERSPEECH ,2005, 3301-3304.

[5] Gang Peng, William S.-Y. Wang, 2005. Tone recognition of continuous Cantonese speech based on support vector machines. Speech Communication 45, 2005, pp. 49-62.

[6] Kaiser J F. On a simple algorithm to calculate the energy of a signal. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP’90), Albuquerque, USA, 1900:381-384.

[7] Bahoura M and Rouat J. Wavelet speech enhancement based on the Teager energy operator, IEEE. Signal Processing Letters, 2001, 8(1): 10-12.

[8] Kecman, V. (2005) Support Vector Machines – An Introduction. In: Support Vector Machines: Theory and Applications, ed. By L. Wang, Springer-Verlag Berlin Heidelbe rg, New York, pp.1-48.

[9] Xu Li, Zhang Wen-le, Zhou Ning, etc. Mandarin Chinese Tone recognition with an artificial neural network. Journal of Otology, 2006, 1(1): 30-34.