A Pitch Smoothing Method for Mandarin Tone Recognition

Qian Liu[1,2], Jinxiang Wang[1], Mingjiang Wang[1],
Panpan Jiang[2], Xirui Yang[2] and Jiayuan Xu[2]

Microelectronics Center, Harbin Institute of Technology, Harbin, P. R. China
Department of Electronic Science and Technology, Harbin University of Science and Technology, Harbin, P. R. China

Abstract

Mandarin Chinese is known as a tonal language with four lexical tones. Tone recognition plays an important role in automatic Chinese speech recognition in that the same syllable with different tones gives quite distinct meanings. The different tone can be characterized by its pitch contour, but the pitch contours are hardly ideal smooth curves. It is because the pitch points calculated by pitch detector normally have some error points. These error pitch points can cause the erroneous classification of Mandarin four-tone recognition. It is necessary to smooth the pitch contour before tone recognition. The classic smooth algorithms can not deal with error fundamental frequencies successively. A new smoothing method proposed in this paper can deal with the error pitch point appropriately. It first checks whether the current point is a correct or error point, then the error type, and finally modifies the error point according to the error type. For different error type, the corresponding smoothing method is also different. To confirm this smoothing method, four "one vs. all" Support Vector Machine classifier are built for Mandarin Tone Recognition. The test results indicate that error rate of Mandarin Chinese four tone recognition can be reduced under the smoothing method.

Keywords: *fundamental frequency; smoothing; tone recognition*

1. Introduction

Mandarin Chinese is known as a tonal language with four lexical tones. The same syllable with different tones gives quite distinct meanings. For example, syllable "ma" with tone one means mother, with tone three it means horse; syllable "er" with tone two means son, with tone three it means ear, and with tone four it means two. Therefore, Chinese four-tone recognition plays an important role in automatic Chinese speech recognition.

The different tones can be characterized by its pitch contours. The pitch contours of different tone have different contour shapes. Fig. 1 shows the typical pitch contour of the four different tones. As shown in Fig. 1, the first tone is the high and smooth tone, the second tone is the rising tone, the third tone is the falling and rising tone, and the forth tone is the falling tone.

Normally, pitch analyses can performed in time domain or in frequency domain. The commonly used methods for pitch detection are Average Magnitude Difference Function [1], Simplified Inverse Filtering Technique [2], Autocorrelation [3] and Cepstrum [4]. But the pitch contours generated from pitch analysis are hardly ideal smooth curves. By observing the original pitch contour, two kinds of error pitch points are found: one is the half, double, or triples points; the other is the bad points with wrong fundamental frequency value. These error points can cause erroneous tone recognition results. Accordingly it is necessary to smooth the pitch contour before tone detection.

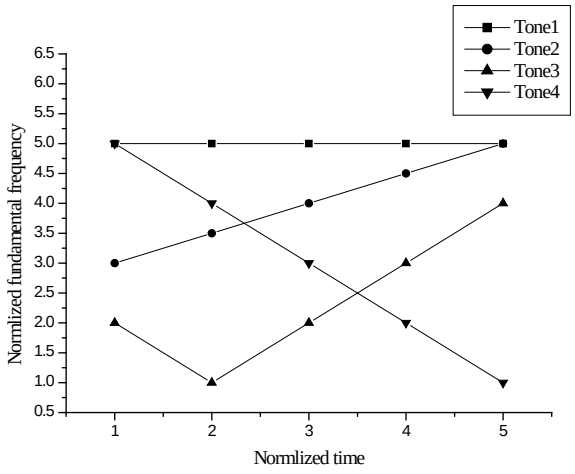

Figure 1. The Typical Pitch Contour of Chinese Four Tones

There are several approaches to smooth a curve. The simplest smoothing approach is the rectangular smoothing which simply replaces each point in the original signal with the average of N adjacent points, where N is a positive integer called smoothing width. Using rectangular smoothing method, the noise is greatly decreased but the curve shape itself is hardly changed. The other classic smooth algorithms such as curve fitting, linear smoothing and median smoothing can not deal with error fundamental frequencies successively [5]. Two specialized smoothing method [5, 6] are proposed to smooth the pitch or frequency calculated by autocorrelation method. The method in [5] can not deal with the half and triple point appropriately. The other method [6] needs the calculation of pitch not only by autocorrelation method, but also by cepstral method, which will cost much more computations. This paper proposes a new method to generate a preferable accurate pitch contour. This approach first checks the error type of the wrong fundamental frequency point, then deals with the different kinds of error points by different ways.

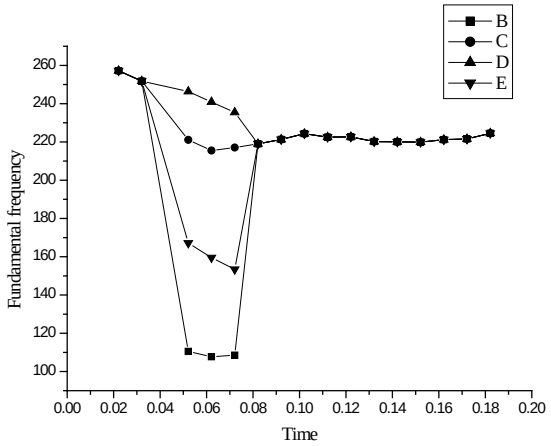

Figure 2. An Illustration of Three Smoothing Methods

Figure 2 is an illustration of rectangular smoothing, smoothing method in [5] and pitch smoothing method in this paper. Line B is the fundamental frequencies calculated from autocorrelation method. As shown in Fig2, there are three error points in it. Lines C, D, E are the smoothed fundamental frequencies with different smoothing methods. Line C is an example using the pitch smoothing method proposed in this paper; line D is an example using the pitch smoothing method proposed in [5]; and line E is an example using rectangular smoothing method, where N equals 5. It shows that modified points in line E are still jump points. Line C and D are smoother than line E. Comparing line C with line D, the line using the approach proposed in this paper is found to be smoother and more natural.

2. Pitch Detection

The pitch contour carries tone information. As shown in Figure 1, the trends of fundamental frequencies from tone 1 to tone 4 are different. For tone 1, fundamental frequencies are almost at the same value. For tone 2, the fundamental frequencies go up as time increase. For tone 3, the fundamental frequencies go down first, and then go up. For tone 4, the fundamental frequencies go down as time increase.

There are two main categories of approachs for pitch tracking, in time domain and in frequency domain [7]. The category in the time domain uses time-related features such as peak picking, Zero-Crossing Rate, and autocorrelation. The other category in the frequency domain normally applies to cepstrum and harmonic matching. Although a large number of different approaches have been proposed for detecting pitch, the pitch detector based on autocorrelation method is still one of the most robust and reliable pitch detector[5].

The essential of autocorrelation method is that it measures the similarity of the signal and its time delayed signal. If at some delay the two signals have similar waveforms, the autocorrelation is large. Assume a speech signal is $s(m)$, its time delayed signal is $s(m-k)$, the autocorrelation of $s(m)$ and $s(m-k)$ can expressd as:

$$R_s(K) = \sum_{m=-\infty}^{\infty} s(m)s(m-k) \tag{1}$$

Normally, the short-time autocorrelation function is used in pitch detection instead of autocorrelation function. The short-time autocorrelation function is obtained by windowing $s(m)$, $w(n)$ is the function of window. The short-time autocorrelation $R_n(k)$ is:

$$R_n(k) = \sum_{m=-\infty}^{\infty} s(m)w(n-m)s(m-k)w(n-m+k) \tag{2}$$

The window in formula 2 should be even symmetry, it means

$$w(n) = w(-n) \tag{3}$$

During the calculation of autocorrelation, the window is fixed, speech signal is moving. The length of the window should be at least twice larger than the pitch period. If the window length is N, $w(n)$ equals 0 when $n \le 0$ and $n \ge N-1$, then formula 2 turns

$$R_n(k) = \sum_{m=0}^{N-1-k} s(m)w(n-m)s(m-k)w(n-m+k) \tag{4}$$

In fundamental frequency calculation, $R_n(k)$ is computed for k near the estimated number of samples in a pitch period; if there is no suitable prior fundamental frequency estimate, $R_n(k)$ is estimated for k from the shortest possible period until the suitable fundamental frequency emerges. In fact, an autocorrelation pitch detector measures fundamental frequency

based on the correlation of each windowed signal with its time delayed signal. Assume the pitch period of a section of a speech signal is P, its maximum value of autocorrelation function occurs at the location of pitch periods: where the time delay equals 0, P, 2P, 3P, etc. Those time delays with maximum value are candidates of pitch period. The smallest nonzero time delay is the pitch period. Each blocked signal generates a pitch period value. The fundamental frequencies can be computed by pitch periods.

3. Pitch Smoothing

Speech signals may contain environment noise, or even be periodic with signals of different fundamental frequencies. It can also be the case that speech signal periodic with the pitch period P are also periodic with 2P, 3P, *etc*. As a result it is necessary to find the smallest pitch period or the highest fundamental frequency. If the detected fundamental frequencies have half, double, triples points or the bad points with wrong value, there is the need to smooth the fundamental frequencies by correcting the error points.

Assume the fundamental frequencies from autocorrelation are $\{f_i\}$. f_i represents the fundamental frequency of the i-th blocked signal. If there is no error point in $\{f_i\}$, the fundamental frequencies should be a smooth line with no jump point. The trend of pitch contour can tell which tone it represents. If there are error points in $\{f_i\}$, as shown in Figure 2, the trend will change a lot at the error point. The intention of pitch smoothing method is to make the pitch contour smooth by modifying the error fundamental frequency points.

The classic smoothing method can not deal with the error fundamental frequencies appropriately. As show in Figure 2, the fundamental frequencies modified by rectangular smoothing method still have jump points. Around these jump points, the trend of fundamental frequencies changes sharply. Accordingly it is necessary to smooth the fundamental frequencies by specialized algorithm.

By observing the original pitch contour, two kinds of error pitch points are found. The key problem of smoothing the fundamental frequencies is to deal with these two kinds of error fundamental frequency points appropriately.

Figure 3 shows the principle of the smoothing method in this paper. The smoothing method proposed in this paper first checks whether the current point is a correct or error point, then the error type, and finally modifies the error point according to the error type. For different error type, the corresponding smoothing method is also different. The fundamental frequency points with correct value will remain the same.

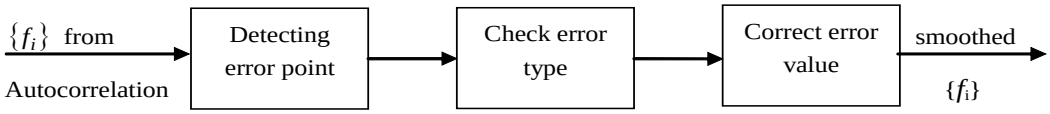

Figure 3. Principle of Smoothing Method

Assume f_i is the current fundamental frequency point; f_{i-1} is the last fundamental frequency point; f_{i+1} is the next fundamental frequency point. The fundamental frequencies calculated by autocorrelations are smoothed as follows:

if $|f_i - f_{i-1}| < TH1$ & $|f_i - f_{i+1}| < TH1$

then $f_i = f_i$

else if $|f_i * 2 - f_{i-1}| < TH2$

then $f_i = f_i * 2$

else if $|f_i/2 - f_{i-1}| < TH2$

then $f_i = f_i/2$

else if $|f_i/3 - f_{i-1}| < TH2$

then $f_i = f_i/3$

else if $|f_i - f_{i-1}| > TH1$ & $|f_i - f_{i+1}| > TH1$

then $f_i = (f_{i-1} + f_{i+1})/2$

else $f_i = f_{i-1} + f_{i+1} - f_i$

With smoothing algorithm method, the error type is checked to see if it is half, double, triples point or bad point with wrong value. TH1 and TH2 in judge condition are thresholds determined by experiments. With different error point type, the modification method is also different: For half point, the fundamental frequency should be double of f_i; for double point, the fundamental frequency should be half of f_i; for triples point, the fundamental frequency should be one thirds of f_i. If the fundamental frequency point is with wrong value, then check $|f_i - f_{i-1}|$ and $|f_i - f_{i+1}|$. If both of them are larger than threshold TH1, $f_i = (f_{i-1} + f_{i+1})/2$; otherwise $f_i = f_{i-1} + f_{i+1} - f_i$.

The goal of this smoothing approach is to replace the error fundamental frequency points with the corresponding correct or approximate value. By using this smoothing method, all the jump points in the fundamental frequencies can disappear in theory.

4. Tone Recognition

In the last decades, many approaches have been applied to tone recognition for tone languages, such as Hidden Markov Models [8, 9], Neural Networks [10], Decision-tree Classification[11], Support Vector Machine [12]. Among these approaches, Support Vector Machine is an excellent static pattern classifier, which is first introduced by Vapnik. It is a binary-class classifier, which has unique advantages in solving classification problems of small samples [13]. It can solve three cases classification: linear separable, linear non-separable and non-linear separable. It can also be extended to multi-class classifier.

Assume a set of training data $\{(x_i, y_i)\}$ belongs to two classes. x_i represents the input training vector, the value of y_i denotes the class of x_i. y_i equals to 1 or -1. Figure 4 shows the linear separable case. The solid and hollow dots represent two different class of samples. A two class classifier based on Support Vector Machine is used to identify the two lines that can separate these samples, and at the same time the margin between the two lines is at its minimum.

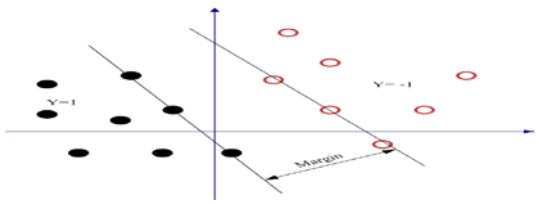

Figure 4. Linear Separable Case

For the non-linear separable case, mapping the input vector into a high dimensional feature space can turn it into a linear separable case. A kernel function can map the non-linear case into a linear feature space. The basic form of non-linear separable Support Vector Machine classifier [14] which classifies an input vector $x \in R^n$ by the formula:

$$f(x) = \mathrm{sgn}(\sum_{SVs} \alpha_i y_i K(x_i, x) + b) \tag{5}$$

Where x_i is the i-th training sample; y_i is the i-th class label ($y_i \in \{1,-1\}$); α_i is the Lagrange multiplier, b is a bias, and K is a kernel function; $f(x)$ is the predicted value. There are three kinds of commonly used kernel functions: Polynomial Function, Radial Basis Function, and Sigmoid Nuclear Function. In this paper, RBF is used as the inner production function for tone recognition.

In order to classify four tones in the tone recognition program, SVM should be extended to multi-class classifier. Normally there are two ways: one is to build a "one vs. all" classifier for each class; the other is to build a "one vs. one" classifier for each pair of class. For tone recognition, four "one vs. all" classifiers were constructed: tone 1 vs. tone 2, tone 3 and tone 4; tone 2 vs. tone 1, tone 3, and tone 4; tone 3 vs. tone 1, tone 2, and tone 4; tone 4 vs. tone 1, tone 2, and tone 3.

SVM is a static classifier based on statistics learning theory. The dimension of model in SVM should be consistent. However the dimension of each syllable feature vectors is variable. Therefore, the feature vectors should be converted to the new one with the same dimension. The feature vectors are normalized by curve fitting method, and a same-dimension feature vector is generated by composing the fitting coefficients. The least square method based on the Legendre Polynomials basis functions is used for curve fitting. The top 6 orders Legendre Polynomials are as follows:

$$P_0(x) = 1 \tag{6}$$

$$P_1(x) = x \tag{7}$$

$$P_2(x) = \frac{1}{2}(3x^2 - 1) \tag{8}$$

$$P_3(x) = \frac{1}{2}(5x^3 - 3x) \tag{9}$$

$$P_4(x) = \frac{1}{8}(35x^4 - 30x^2 + 3) \tag{10}$$

$$P_5(x) = \frac{1}{8}(63x^5 - 70x^3 + 15x) \tag{11}$$

The top 4 orders Legendre Polynomials are used only, which means that the dimension of new feature vector should be a multiple of four. The feature vectors commonly used for tone recognition are fundamental frequency, first order difference of fundamental frequency, logarithm energy, and first order difference of logarithm energy. This sixteen feature vector can be the x_i or x in formula 5, its corresponding class label should append according to the classifier it was attached to. For example, the label of syllable "mai" with tone two should be 1 for the classifier of tone 2 vs. tone 1, tone 3, and tone4; it should be -1 for the classifier of tone 1 vs. tone 2, tone 3, and tone4.

5. Experimental Results

To confirm the smoothing algorithm, 10,080 isolated Mandarin speeches are tested. The experimental training and testing data are recorded by 10 male and 4 female whose ages range from 20 to 35. The speech signals are sampled at a rate of 16 kHz and saved with 16 bit. The isolated words include 168 syllables with four tones, which are recorded twice per person for training and testing.

The speech files are saved in wav file format. The wav file can be converted to fundamental frequencies by autocorrelation method, at the same time the logarithm energies are computed. A curve fitting algorithm can be used to change the fundamental frequencies and logarithm energies of variable length to a sixteen dimension vector. This sixteen dimension vector is x_i or x in formula 2. The label y_i for tone classifier also needs to append. For example, in tone 1 vs. tone 234 mode, y_i of tone 1 equals 1, y_i of tone 2, 3, and 4 equals -1. The Lagrange multiplier α_i and bias b can be obtained after training period. In the testing period, the predicted result $f(x)$ will get through formula 2. For classifier of tone 1 vs. tone 2, 3, and 4, $f(x)$ equals 1 means the input sixteen feature vector represents tone 1; $f(x)$ equals -1 means the input sixteen feature vector represents tone 2, tone3, or tone 4.

In order to see how many error fundamental frequency points have in the entire fundamental frequency points, a grogram is written for counting the number of error points. The counting grogram runs on fundamental frequencies both before and after smooth. The half, double and triple fundamental frequency point are named as “odd point”; the error fundamental frequency points are named as “bad point”, Table 1 shows that there are 9,275 error points in all the fundamental frequency points from both training and testing speech file. All of these two kinds of points disappeared after smooth.

Table 1. Number of the Odd Point and Error Point

	odd points	bad points
Before smooth	2484	6,791
After smooth	0	0

Table 2. Isolate Word Tone Recognition Error Rate

		Tone1	Tone2	Tone3	Tone4
Train set	without smooth	0.26%	0.10%	0.16%	0.16%
	with smooth	0.16%	0.05%	0.05%	0.04%
Test set	without smooth	10.5%	11.75%	10.75%	9.75%
	with smooth	8.75%	8.65%	8.31%	7.52%

In order to see whether the smooth algorithm decrease the error rate of tone recognition, the tone classifier based on SVM theory is performed on the fundamental frequencies both before and after smooth. The four tone classifier should be trained and tested separately. Before the fundamental frequency points are smoothed, the four tone classifiers are trained and tested on the sixteen dimension vector from the fundamental frequencies by autocorrelation method. After the fundamental frequency points are smoothed, the four tone classifiers are trained and tested again on the new sixteen dimension vector from the smoothed fundamental frequency points. The error rate of tone recognition can be computed as below:

$$\text{error rate of tone recognition} = \frac{\text{the number of error classified tones}}{\text{the total number of tones in testing}} \quad (12)$$

Table 2 show the tone recognition result of isolated words. No matter for training set or for testing set, the recognition error rates decreased after smooth. For the training set, error rate of tone 1 decreased by 0.10%; error rate of tone 2 decreased by 0.05%; error rate of tone 3 decreased by 0.11%; error rate of tone 4 decreased by 0.12%. The recognition error rate is reduced by 0.095% on average. For the test set, error rate of tone 1 decreased by 1.75%; error rate of tone 2 decreased by 3.1%; error rate of tone 3 decreased by 2.44%; error rate of tone 4 decreased by 2.23%. The recognition error rate is reduced by 2.38% on average.

6. Conclusions

The four lexical tones of Mandarin Chinese play an important role in Mandarin speech recognition. The commonly used feature of tone recognition is fundamental frequencies. Fundamental frequency contour carries tone information, but the fundamental frequencies always have some error points when calculated by autocorrelation method. The new pitch smoothing method proposed can deal with different kinds of error point appropriately. The pitch contour smoothed by this method can be more accurate and smoother. The experiment result shows that all the error points disappeared after smooth. It is also confirmed that the smoothed pitch contour for tone recognition can reduce the error rate of tone recognition.

References

[1] M. J. Ross, H. L. Shaffer, A. Cohen, R. Freudberg and H. J. Manley, “Average Magnitude Difference Function Pitch Extractor”, IEEE Transactions of Acoustic, Speech, Signal Processing, vol. ASSP-24, **(1976)**, pp. 2-8.
[2] J. D. Markel, “The SIFT Algorithm for Fundamental Frequency Estimation”, IEEE Transactions Audio Electroacoust, vol. AU-20, **(1972)**, pp. 367-377.
[3] L. R. Rabiner, “On the use of Autocorrelation Analysis for Pitch Detection”, IEEE Trans. on Acoust. Speech, and Signal Processing, vol. ASSP-25, no. 1, **(1977)**, pp. 24-33.
[4] R. W. Schafer and L. R. Rabiner, “System for Automatic Formant Analysis of Voiced Speech”, Journal of Acoustic Society Amer, vol. 47, **(1970)**, pp. 634-648.
[5] L. Jun, X. Zhu and Y. Luo, “An Approach to Smooth Fundamental Frequencies in Tone Recognition”, ICCT, **(1998)**, pp. 101-105
[6] X. Zhao, D. O’Shaughnessy and M. Nguyen, “A Processing Method for Pitch Smoothing Based on Autocorrelation and Cepstral F0 Detection Approaches, ISSSE, **(2007)**, pp. 59-62
[7] R. L. Rabiner and J. M. Cheng, “A Comparative Performance Study of Several Pitch Detection Algorithms”, IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-24, no. 5, **(1976)**, pp. 399-418.
[8] X. Chen, C. Cai, P. Guo and Y. Sun, “A Hidden Mardov model applied to Chinese four-tone recognition”, Proc. International Conference on Acoustics, Speech, and Signal Processing, **(1987)**, pp. 797-800.
[9] W. Yang, “Hidden Markov Model for Mandarin lexical tone recognition”, IEEE Trans. Acoust, Speech Signal Process, vol. 36, **(1988)**, pp. 988-992.

[10] M. Emonts and D. Lonsdale, "A memory-based approach to Cantonese tone recognition", Proc. 8th European Conference on Speech Communication and Technology, **(2003)**, pp. 2305-2308.
[11] Y. Cao, "Tone Recognition in Mandarin using Focus", INTERSPEECH, **(2005)**, pp. 3301-3304.
[12] P. Gang, S. William and Y. Wang, "Tone recognition of continuous Cantonese speech based on support vector machines", Speech Communication, vol. 45, **(2005)**, pp. 49-62.
[13] S. Wang, Z. Tang, Y. Zhao and S. Ji, "Tone Recognition of Continuous Mandarin Speech Based on Binary-Class SVMs", ICISE, **(2009)**, pp. 710-713
[14] J. Chen and L. Jiao, "Classification Mechanism of Support Vector Machines", Proceedings of ICSP, **(2000)**, pp. 1556-1559.