Tone Recognition of Isolated Mandarin Syllables

Zhaoqiang Xie and Zhenjiang Miao

Institute of Information Science, Beijing Jiao Tong University,
Beijing 100044, P.R. China
{08120470,zjmiao}@bjtu.edu.cn

Abstract. Mandarin is tonal language. For Mandarin, tone identification is very important for speech recognition and pronunciation evaluation. Mandarin tone behavior varies greatly from speaker to speaker and it presents the greatest challenge to any speaker-independent tone recognition system. This paper presents a speaker normalizing algorithm which is designed to reduce this influence. Then a basic neural network tone recognizer using recognition features extracted from the processing syllable is introduced. The system employs an improved pitch detector and a powerful tone identification method. Finally, the experimental results show that the tone recognition system classifies four tones very well.

Keywords: Mandarin speech, tone recognition, pitch normalization, pitch contour, BP neural network.

1 Introduction

Mandarin is well known as tonal language with four tones (i.e., 1st tone, high-level; 2nd tone, rising; 3rd tone, falling-rising; and 4th tone, high-falling). There are a total of about 416 phonologically allowed syllables in Mandarin Chinese, if the differences in tones are disregarded. But if the differences in tones are considered, there are about 1345 syllables. Accurate tone recognition is greatly helpful for Chinese recognition systems, because the tones in syllables have lexical meanings in Chinese. So the tone identification is regarded as an important part for the mandarin speech recognition.

It has been shown that the main difference among the tones is in the pitch contours [1]. Hence the performance of pitch extraction can affect the identification of tone very much. Many algorithms for pitch extraction have been studied extensively ever since many years ago, some of which utilized the various characteristics of speech signals in time domain, and some in frequency domain, and some in both [2]. We all know that one of the oldest digital methods for estimating the fundamental frequency (F_0) of voiced speech is auto-correlation analysis. What is proposed in this paper is a simplified analysis technique, based upon an inverse filter formulation, which retains the advantages of both the autocorrelation and cepstral analysis techniques [4].

One of the biggest difficulties of tone recognition is that tone varies seriously, and the reasons that cause tone variations are too many and too complicated to solve, such as speakers and contexts. Even spoken by the same speaker, syllables with the same tone usually don't have similar characteristics. Thus a good method to model tone

A. Elmoataz et al. (Eds.): ICISP 2010, LNCS 6134, pp. 412–418, 2010.

variations is necessary. In this paper, we will introduce our work on reducing the impact of gender.

At the last step of the whole system, the value of tone is identified. There are many tone identification methods based on pitch contours, for example, the method based on fuzzy pattern recognition [5], or the method using HMM and wavelet transform [6], or ANN (Artificial Neural Network) [7]. The ANN has self-organization, fuzziness, redundancy and self-learning characters, and it also has good non-linear distinguishing ability. Therefore its use in tone recognition is effective. In the applications of ANN, the multi-layer perception (MLP) using the back-propagation (BP) algorithm [8] is used very often. In this paper, BP will be adopted as a tone recognizer.

The rest of this paper is organized as follows. Section 2 describes the composition of the tone recognition system. A pitch detection method is given in Section 3. In Section 4, the method of the final tone identification is discussed. Then the experiment and results are illustrated in Section 5. Finally, conclusions are given in Section 6.

2 Tone Recognition System

Tone is mainly described by pitch contour. So an accurate extraction of pitch is necessary in a tone recognition system. In addition, the accuracy of this extraction essentially influences the correct rate of a tone recognition system. For voiced INITIALs or zero INITIALs, the pitch contour begins with the sound of the word, but it begins with the FINAL in unvoiced INITIAL, because the unvoiced INITIAL is much like a noise. So the classification of the voiced INITIAL and the unvoiced INITIAL must be done before the extraction of the pitch contour. There are many methods of pitch extraction, such as classical autocorrelation function method [9], one level cutoff wave method, AMDF and Seneff's method [10], SIFT, etc. There are some multi-time pitch points, divided pitch points and random error points in the extraction pitch sequence. Therefore, a smoothing algorithm is needed to correct such errors. The framework of the tone recognition is show in Fig. 1.

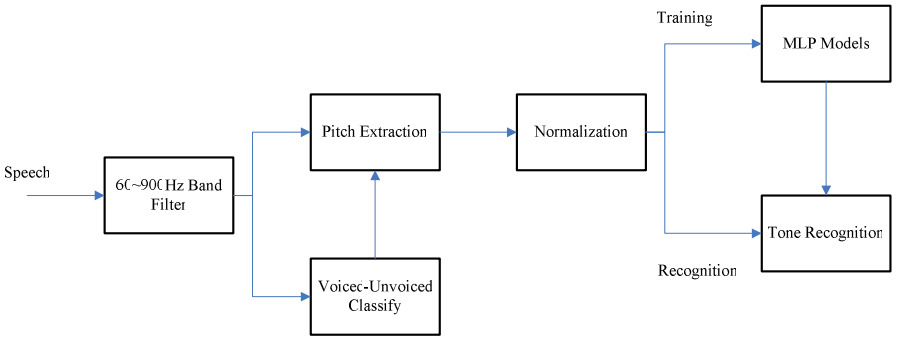


Fig. 1. The Diagram of Tone Recognition System

3 Tone Pitch Detection

3.1 End Point Detection

The unvoiced part of the input speech signal must be well cut off for this part is useless and has serious impact on the effect of recognition. So a special V/UV classification method based on ZCR and STAM is adopted [3]. For isolated Mandarin syllables, it is enough to classify the voiced and unvoiced part.

3.2 Pitch Detection

Mandarin tone is the patterns of pitch variation, so it may be acquired by pitch extraction. Many methods of pitch detection are developed so far; in this paper the pitch detector using SIFT algorithm [4] is adopted. The algorithm is based upon a simplified version of a general technique for fundamental frequency extraction using digital inverse filtering. It is demonstrated that the simplified inverse filter tracking algorithm encompasses the desirable properties of both autocorrelation and cepstral pitch analysis techniques. In addition, the SIFT algorithm is composed of only a relatively small number of elementary arithmetic operations. In machine language, SIFT should run in several times real time while with special-purpose hardware it could easily be realized in real time. That is why we choose it as our pitch detector. Our algorithm procedure is showed by Fig. 2.

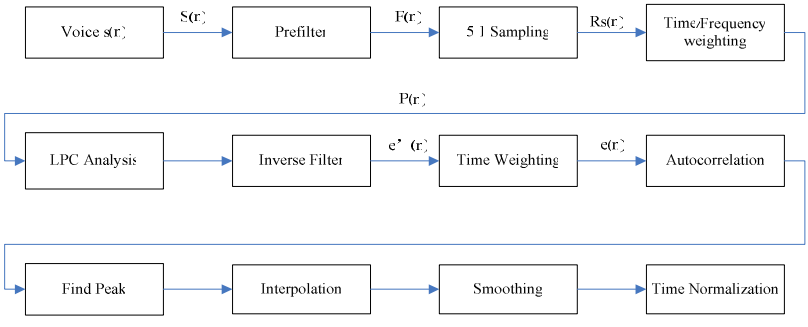


Fig. 2. The SIFT Procedure

The Time/Frequency weighting is to avoid spectral leakage generated by data truncated because of limited frame width. The formula is like this:

$$\begin{cases} rs(n) = 0; n < 1 \\ p(n) = [rs(n) - rs(n-1)] \times [0.54 - 0.46 \times \cos(\dfrac{2\pi(n-1)}{FW / DR})]; n = 1,2,3,..., FW / DR \end{cases} \quad (1)$$

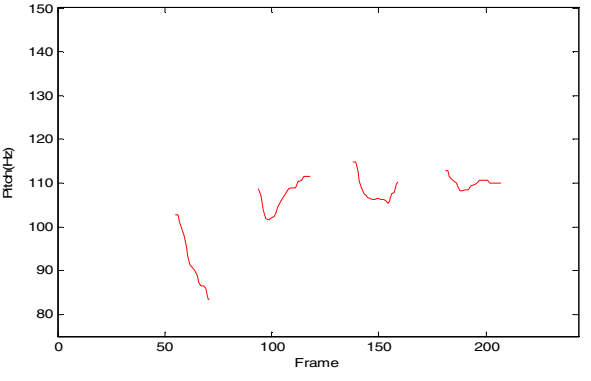
And FW/DR is the frame width after sampling.

The autocorrelation function (Normalization) shows as follows:

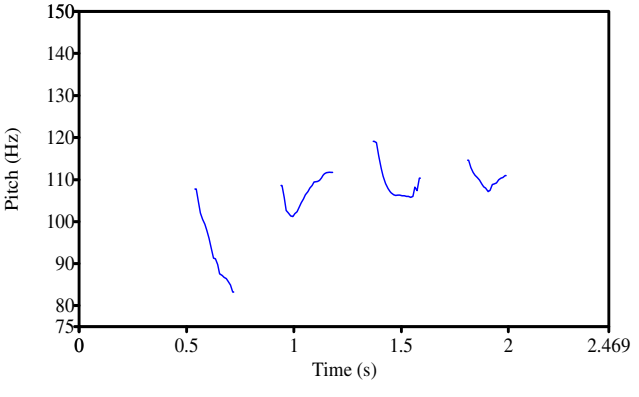
$$\rho_e(i) = \frac{\sum_{n=1}^{N} e(n)w(n)e(n+i)w(n+i)}{\sum_{n=1}^{N} e(n)w(n)e(n)w(n)} \tag{2}$$

And e (n) is an inverse filter output by LPC analysis, w (n) is a Hamming window.

Fig. 3 is an example compared with Praat. (Frame increment is 10ms)



(a)



(b)

Fig. 3. Estimated fundamental frequency curves for utterance "běi jīng jiāo tōng" (a) is gained by the above SIFT method, and (b) is from Praat. We can see the result is satisfactory.

3.3 F0 Smoothing Method

After observing some speech signal waveforms, we found the result of pitch calculation usually has divided frequency, multi frequency and random error points, so it needs smoothing. There are many smoothing methods, such as the linear

smoothing method, the middle point smoothing method and the normal linear interpolation method, etc. These methods cannot treat the continual random error points. From the physical progress of a voice, we can find that the pitch will not change with large magnitude in continual speech, so that we can solve the problem with a statistic method.

Suppose $F_0(1)$, $F_0(2)$, $F_0(3)$……$F_0(N)$ are the N frame frequency points of the FINAL segment, the smoothing algorithm is as below:

1) For all frames, get rid of that frame,
   If $|F_0(i)- F_0(i-3)|> F_0(i)*RUNAWAY$ && $|F_0(i)- F_0(i+3)|> F_0(i)*RUNAWAY$, (RUNAWAY=0.3) then $F_0(i)$ is recognized as a real isolated point, $F_0(i)=0$.
2) For all frames, if $F_0(i) =0$, $F_0(i-1)!=0$, find the first $F_0(k)!=0$ (k=i+1~i+3), then $F_0(i)= (F_0(i-1)+ F_0(k))/2$;

4 Tone Recognition Scheme

4.1 Tone Feature Extraction

The dynamic pitch ranges of different speakers differ greatly; to within 80Hz to 350 Hz. Therefore speaker independent tone recognition must normalize F0. The method here is that the F_0 contour of each sentential utterance is normalized by its own means to reduce inter speaker variability.

$$f_0(i) = F_0(i) / F_{mean}; \quad e(i) = E(i) / E_{mean} \tag{3}$$

$F_0(i)$ is the i^{th} frame frequency point, and E(i) corresponds the i^{th} frame energy.

Based on the previous F0 normalization method, each signal frame was represented by a vector with 5 features [f_0 (t), $\triangle f_0$ (t), $\triangle\triangle f_0$ (t), e (t), $\triangle$e (t)].

$$\Delta f_0(t) = f_0(t+1) - f_0(t) \ ; \Delta\Delta f_0(t) = \Delta f_0(t+1) - \Delta f_0(t)$$
$$\Delta e(t) = e(t+1) - e(t) \tag{4}$$

F0 curve and energy curve are divided into three uniformly sub contours, and we pick up the means of these sub contours as final tone features. So our tone feature vector dimension is fifteen. These features are then fed into a MLP pattern recognizer for tone recognition.

4.2 MLP Model

MLP model consists of three layers: the input layer, the hidden layer and the output layer. The hidden layer may have multi-layers, but in this system, all MLP models have only one hidden layer for simplification purposes [11]. It consists of two output nodes corresponding to tour tones. Each neuron output is the sigmoid function of the weighted summation of inputs. The MLP recognizer is trained by the back propagation (BP) rule, which minimizes the mean squared error between the feed forward outputs and the desired targets.

5 Experiments and Results

Experimental results are carried out on a speaker independent isolative syllable database in Chinese. The database consists of 5 male and 6 female speakers, the sample rate is 16 KHz and the number of bits per sample is 16bit. Every person pronounced 1588 syllables with tone. Each syllable is pronounced four times using 4 tones although some tones are not pronounced in the daily lives. For example, although "tian4" (the number is tone value) is not pronounced in the daily lives it is also recorded. The training set uses only a standard female voice pronunciation. The reason for doing so is to examine whether the normalization method is effective or not.

Table 1 describes the difference of features ([$f_0(t)$, $\triangle f_0(t)$, $\triangle\triangle f_0(t)$ …] by formula (3)(4)) between male and female. Here we take the four tones of syllable "Xin" as an example. As can be seen from the table, after normalized, the difference of the features ([$f_0(t)$, $\triangle f_0(t)$, $\triangle\triangle f_0(t)$ …]) between male and female is so small so that it basically can be ignored. Because of this, even though we use very few training samples and the training set does not contain male samples, the results for both male and female test sets show very good, as can be seen from Table 2.

Table 1. The normalization features of male and female on the same tone

	Male	Female
Xin1	(1.004,0.994,1.005,-0.001,0.024,0.056,-0.017, 0.007,-0.015…)	(1.023,0.978,0.996,-0.276,0.004,-0.033,0.139,0.011,-0.051…)
Xin2	(0.748,0.1.06,1.206,1.62,0.409,0.558, 0.025,0.016,-0.016…)	(0.862,0.966,1.271,1.736,1.996,0.692, 0.005,0.026,0.003 …)
Xin3	(1.008,0.898,1.097,-0.559,0.067,0.714,0.047,0.05,0.003…)	(1.072,0.89,1.024,-0.664,-0.118,0.658,0.093,-0.051,0.022…)
Xin4	(1.216,0.997,0.768,-0.17,-1.506,-0.263,-0.08,-0.02,0.15…)	(1.189,1.024,0.768,-0.268,-1.301,-0.912,-0.109,-0.127,0.195…)

Table 2 shows the confusion matrix of experiment 3. In this table, we can see that tone 3 is prone to be tone 2. In fact, isolative tone 3 is quite similar with tone 2 except that tone 3 is lower than tone 2 and sometimes it becomes creaky voice.

Table 2. The identification accuracy of four tones

Input / Percent / Output	Tone 1	Tone 2	Tone 3	Tone 4
	5832	5892	5384	5230
Tone 1	95.32%	0.45%	0.59%	0.32%
Tone 2	0.49%	92.12%	5.39%	0.08%
Tone 3	2.11%	4.98%	90.93%	1.02%
Tone 4	0.43%	0.12%	1.28%	96.29%

6 Conclusion

Several tone recognition schemes based on artificial neural networks have been discussed in this paper. First, we presents a tone pitch detector using SIFT and a smoothing scheme. In the recognizer, from our analysis we can see, our normalization method can reduce the influence of gender, and experimental results verified that the normalization method works very well. Future work can be focused on how to apply such conclusions on large vocabulary speech recognition system.

Acknowledgment

This work is supported by Innovative Team Program of Ministry of Education of China (IRT0707).

References

[1] Lee, L.S.: Voice dictation of Mandarin Chinese. IEEE Signal Processing Magazine 14(4), 63–101 (1997)
[2] Rabiner, L.R., Cheng, M.L., Rosenberg, A.E., McGonegal, C.A.: A comparative performance study of several pitch detection algorithms. IEEE Trans. Acoust. Speech, Signal Processing, ASSP 24, 399–417 (1976)
[3] Xie, X., Xie, L.: A new approach for tone recognition of isolated Mandarin syllables. Communications and Information Technologies (2006)
[4] Markel, J.D.: The SIFT Algorithm for Fundamental Frequency Estimation. IEEE Trans. AU 20, 361–377 (1972)
[5] Shilin, X.: Fuzzy Tone Recognition for Chinese Speech. Acta Electronica Sinica, China 24(1), 119–121 (1996)
[6] Jun, C., Kechu, Y., Bingbing, L.: A tone recognizer using wavelet transform and Hidden Markov Model. Journal of Electronics and Information Technology, China 19(2), 177–182 (1997)
[7] Fang, S., Guangrui, H.: Chinese Tones Recognition Based on a New Neural Network. Journal of Shanghai JiaoTong University 31(5), 36–38 (1997)
[8] Tang, L., Yin, J.: Mandarin Tone Recognition Based on Pre-Classification. In: Proceedings of the 6th World Congress on Intelligent Control and Automation, June 21-23 (2006)
[9] Rabiner, L.R., Schafer, R.W.: Digital Processing of Speech Signals. Science Press, Beijing (1983)
[10] Seneff, S.: Real-time harmonic pitch detector. IEEE Trans. Acoustics, Speech and Signal Processing 26(4), 358–365 (1978)
[11] Qing, S., Lin, T.: Introduction of Pattern Recognition.National Defense University of Technology Press, Changsha (1991)