Designing Algorithms to Recognize Tones in Mandarin

Min-Wen Du
Dept. of Information Science
National Chiao-Tung Univ.
Hsinchu, Taiwan, ROC
Mwdu@cis.nctu.edu.tw

Wen-Horng Yang
Dept. of Information Science
National Chiao-Tung Univ.
Hsinchu, Taiwan, ROC

Abstract -- *The Chinese language is a tonal language. The accuracy rate of a Chinese speech recognition system will be affected directly by the accuracy rate of its tone recognition algorithm. The recognition problem of tones deserves a closer look if we want to build a highly reliable Chinese speech recognition system.*

In this paper, we studied the tone recognition problem in Mandarin. We have designed four features to distinguish the tones. We have also developed three major ways t o classify tones. We have performed experiments to test the classification methods on large test samples (1390 syllables for each speaker with 19 speakers in total). The best classification method is the mixture one where we use the Euclidean distance on a multi-level decision tree. Its recognition rate is 89.35% for males and 80.90% for females of the mixture method. Individually, the recognition rate of the mixture method can reach higher than 90%. For one male the rates are 94.7%, 92.9 %, 89.7% 92.4% for tone 1 to tone 4.

Index terms – Mandarin, tone, speech recognition

1. Introduction

1.1 Motivation

In this paper, we will study the problem of tone recognition in Mandarin, and develop algorithms to recognize tones.

The Chinese language is a tonal language. It is also monosyllabic. Every Chinese character is pronounced with a tone, possibly with a consonant, as well as a vowel or a diphthong. In Mandarin, which is the main dialect of Chinese language, there are five tones. Without tone components, many Chinese phrases become indistinguishable phonetically. For example, the phrases 一生 (一丶ㄕㄥ), 醫生(一 ㄕㄥ), 益生(一丶 ㄕㄥ), 一省(一丶 ㄕㄥ∨), will be the same in Mandarin, if the tones in the phrases are disregarded. Therefore, the accurac rate of a Chinese speech recognition system will be affected directly by the accurac rate of its tone recognition algorithm. The recognition problem of Chinese tones deserves a closer look if we want to build a highly reliable Chinese speech recognition system.

1.1 Tones in Mandarin

In this paper we will focus our study on the tone recognition problem of Mandarin. There are five tones in Mandarin, from tone 1 to tone 4 and a neutral tone (referred as tone 0). It has been shown [Howie, 1976] that the primary differences among the four lexical tones lie in their pitch contours. Tone 0 is light and short and is harder to judge. Since tone 0 is used less frequently (less than 2% [3]), we will study only tone 1 to 4 in this paper.

Traditionally, most tone recognition studies are based on the fundamental frequency sequences of the speech signals. In our study, we will use the pitch period sequences as the base instead. The pitch period sequence of a speech wave corresponding to a syllable is the sequence of period lengths in the wave of the syllable. It can be represented as the sequence of the number of samp les between the consecutive pitches plus one in the sound wave. In our study we took samples from the input wave at the rate of 16000 samples per second. Therefore, the relation between the pitch period and the fundamental frequency can be given below as

$$pitch\_period = \frac{16000}{fundametal\_frequenc}$$

Figure 1 shows the typical pitch period sequence patterns of tone 1 to tone 4 in Mandarin. Tone 1 is a horizontal line. Tone 2 rises at the beginning and falls gradually to the end. Tone 3 has relatively long pitch periods (higher position in the di agram) at the beginning and rises gradually to the end. Tone 4 has relatively short pitch periods and is horizontal at the beginning for quite a while then rises to the end. An algorithm for tone recognition ordinarily extracts features from various portions of the pitch period sequence (or fundamental frequency sequence) to classify which tone that piece of the sound wave belongs to.

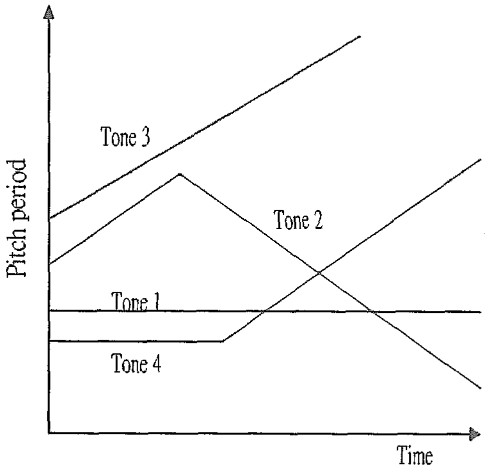

Fig. 1.Pitch period sequence patterns of tone 1 to tone 4.

1.2 Previous Research

Several studies on Mandarin tone recognit ion have been conducted in the past few years. Since each tone has its own special pitch contour pattern, all the algorithms developed extract features from the pitch contours. Chang [5] applied statistical methods to this problem by using a number of parameters (pitch mean, duration, pitch slopes) extracted from eight sectioned segments of pitch contours.

You [4] matches the pitch contours by using orthogonal polynomial representations. Chen [3] and Yang [6] applied the Hidden Markov Model (HMM) method to solve tone recognition problem. In the Golden Mandarin (I) system developed by Lee and his group [1], DHMM method is used to do the tone recognition work.

1.3 Our Approach

We choose to apply the statistical method to recognize tones. Since there are only five tones in Mandarin, reliable statistical data can be gathered easily and quickly in practical applications. The major works done in statistical approach lie in discovering useful features that can represent the special properties of the subjects, and developing good classification methods that make better use of the feature values to distinguish the subjects. In this paper, we have designed four tone features based on the patterns of the pitch period sequences of tones. W have also developed three differ ent classification methods to perform tone classification task.

2. Feature Design

We have designed many features for tone recognition in Mandarin. Among them, we have found that four are most useful for the distinguishing purpose. We describe them in the following sub-sections.

In the discussions of this Section, we assume that the pitch period sequence from which we want to extract features is P[i], for 0 <= i < n.

2.1 Feature 1 (Angle after turning point)

Intuitively, the slope of the line connecting the starting and ending points of a pitch period sequence seems to be a good feature to distinguish the four tones (see Fig. 1). We measure the slope by the angle between that line and the horizontal line. That angle of tone 1 tends to be zero. For tone 2 it is a negative value. For tone 3 it is positive. For tone 4 it is also positive but smaller than that of tone 3.

This observation can be carried one step further. When we examine Fig.1 again, we can discover that tone 2 and tone 4 each has a easily visible turning point, while tone 1 and tone 3 do not have. If we calculate the angle of each tone after the turning point, the resulting angle will be enlarged and it will be more effective to use to distinguish the tones. So we do the following. We use a line that is in parallel with the line connecting the starting and ending points. We move it up or down to find the turning point on the curve (Fig. 2.). We use the turning point as the starting point together with the same ending point to calculate the angle as the feature value.

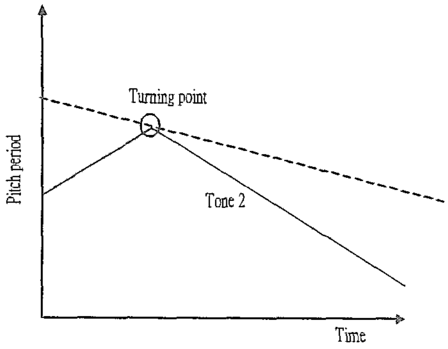

Fig. 2. Turning point calculation.

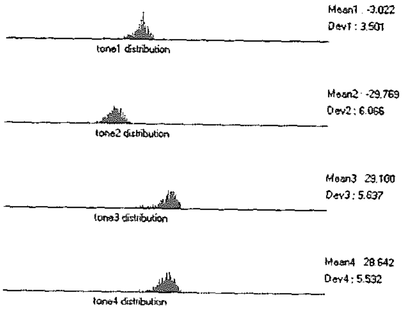

Fig. 3. Feature 1 distributions.

From the distribution, we see that Feature 1 separates tone 1 and tone 2 from tone 3 and tone 4 quite well. It is not good to be used to distinguish tone 1 from tone 2 or tone 3 from tone 4, however.

2.2 Feature 2 (angle after turning point multiplied by deviation of points before turning point)

Here we try to improve Feature 1. From the diagram in Fig. 3, we see that the ranges of the distributions of tone 1 and tone 2 overlap each other, and that of tone 3 and tone 4 overlap each other even much more. Can we move the overlap distributions apart? Actually we can, by noticing that tone 1 and tone 4 both are close to a horizontal line before (to the left) the turning point, while tone 2 and tone 3 are not. So we can enhance the distinguishing capability of Feature 1 by multiplying it with the deviation of the segment before the turning point. We call the resulting value Feature 2.

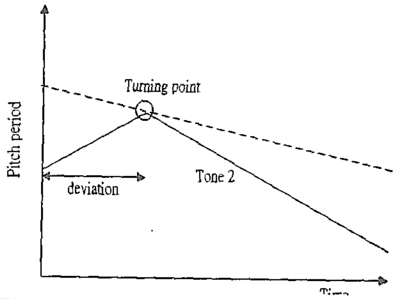

Fig. 4. The feature 2 description figure.

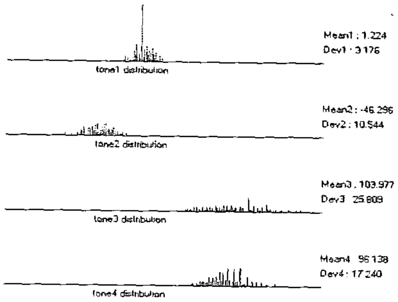

Fig. 5. Feature 2 distributions.

Fig. 5 shows the distributions of the Feature 2 values for the testing data of a person.

2.3 Feature 3 (Up or down percentage after turning point)

We know that tone 1 is nearly a st raight horizontal line. Also, after the turning point from left to right, tone 2 is falling continuously. Tone 3 is rising continuously. And tone 4 is also rising continuously. We may assign an up-down value to every consecutive two points P[i] and P[i+1] after turning point in the following way.

Define the up-down value of the of P[i] and P[i+1] =

$$\begin{array}{rl} 1 & \text{if } P[i+1] > P[i]; \\ 0 & \text{if } P[i+1] = P[i]; \\ -1 & \text{if } P[i+1] < P[i]. \end{array}$$

The sum of these up-down values in a tone 1 pitch period sequence will tend to be zero. It will tend to be a negative value for a tone 2 sequence. It will tend to be a positive value for a tone 3 sequence. It will also tend to be a positive value for a tone 4 sequence, with a somewhat smaller value than that of a tone 3 sequence.

We define s_count to be the sum of positive counts plus the zero counts minus the negative counts of the up-down values. We define the Feature 3 value feature_ = s_count *100/total_count, where total_count is the total number of points after the turning point.

Fig. 6 shows the distributions of Feature 3 values for the testing data of a person.

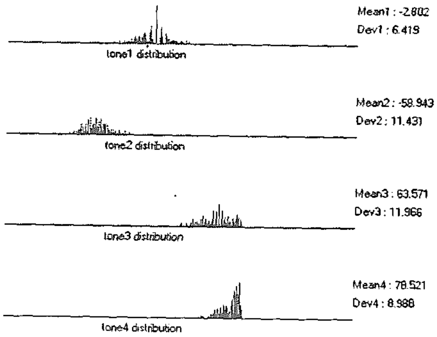

Fig. 6. Feature 3 distributions.

We can see that this feature has the ability to distinguish tone 1 and tone 2. It can also distinguish tone 1 and tone 2 from tone 3 and tone 4.

2.4 Feature 4 (period value at the maximum density point)

If we project the period values of the pitch period sequences of four tones onto the vertical axis, we can see that both tone 1 and tone 4 will have a small concentration point, while tone 2 and tone 3 do not have (Fig. 12). This property leads us to another feature: the center of the concentration region of the vertical axis projection of t he pitch period sequence.

This center will be somewhat random for tone 2 and tone 3, however. This values will be very useful to distinguish tone 3 and tone 4 because the center of the concentration point of tone 4 will be smaller than any period value in a tone 3 sequence, as can be seen from Fig. 7.

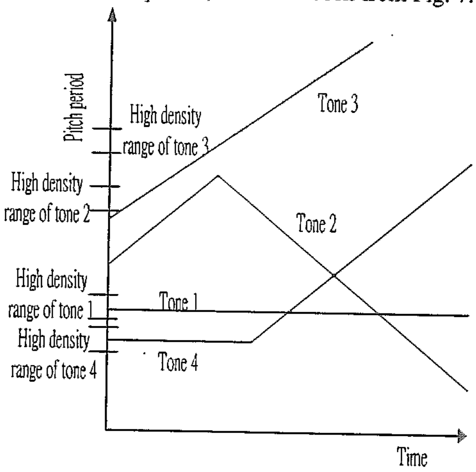
Pitch period
High density range of tone 3
Tone 3
High density range of tone 2
Tone 2
High density range of tone 1
Tone 1
High density range of tone 4
Tone 4
Time

Fig. 7. Feature 4 description figure.

Fig. 8 shows the distributions of Feature 4 values for the testing data of a person.
It can·be seen that Feature 4 is especially capable of distinguishing tone 3 from tone 4.

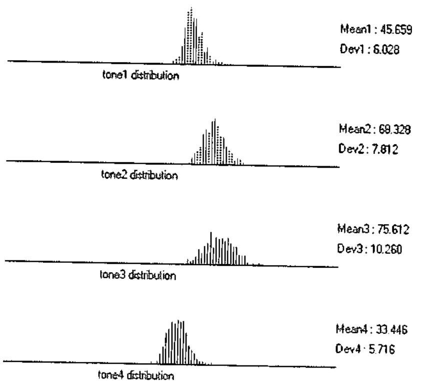

Fig.8. Feature 4 distributions.

3. Classification Method Design

In this Section we develop classification methods for tone 1 to tone 4 based on the features designed in Section 2.

3.1 Euclidean distance classification method

It can be easily observed that a single feature cannot distinguish all the four tones well. This is because that each feature is designed according to a specific property of one or two tones. For example, Feature 1 focuses on the slope of the sequence curve. It is good at di stinguishing tone 1 and 2 from tone 3 and 4. But since the curves of tone 3 and tone 4 have similar positive slops, tone 3 and tone 4 cannot be distinguished by Feature 1 effectively. To distinguish all the four tones, we need to use several features together.

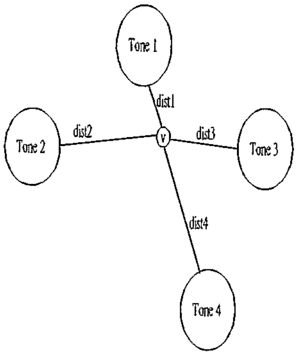

Fig. 9. The Euclidean distance method.

Method description

Given the feature values v1, v2, v3 and v4 calculated from a pitch period sequence P. We calculate the Euclidean distance of P to the centers of the four tones by the following formula :

$$dis\tan ce = \sqrt{\left(\frac{v_1 - mean_1}{deviation_1}\right)^2 + \ldots.. + \left(\frac{v_4 - mean_4}{deviation_4}\right)^2}$$

Where $mean_i$ is the average of Feature i values, and dev_i is the deviation of Feature i values, for $1 <= i <= 4$.

The tone with the minimum distance to P is the tone we select for P, see Fig.9.

3.2 Multi- layer classification method

From the feature value distribution diagrams we can observe that a feature may be good t distinguish certain tones but not good to distinguish others. A natural thought is to apply the features in a sequence to perform the classification work. For example, Feature 1 can separate tone 1 and tone 2 from tone 3 and tone 4. If we apply feature 1 to a sample sequence and find that it must be tone 3 or tone 4, since we know that Feature 4 is good at distinguishing tone 3 from tone 4, we may use Feature 4 to make the second-layer (final) decision.

Method description

The method described above results in a decision tree. There are many ways to assign features on different layers on the decision tree to make the decision work.

By observing the distribution diagrams of the four features for different test data sets, we have found that Feature 2 can classify the tones into subgroups {1}, {2} and {3, 4}. Tone 3 and tone 4 can be distinguished by using Feature 4. Therefore, we assigned Feature 2 to the first layer, and Feature 4 to the second layer to make the decision tree, as shown in Fig. 10.

Note that we may also apply Feature 1 or Feature 3 at the first layer with Feature 4 at the second layer.

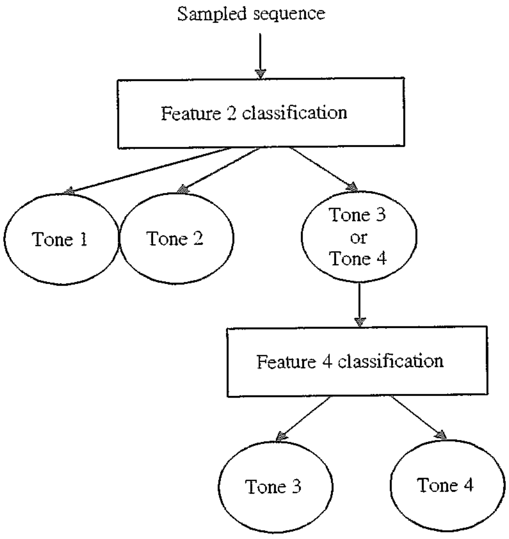

Fig. 10. Multi-layer decision tree method.

3.3 Mixed Euclidean Distance and Multi-layer method

Primitive experiments have shown that the Euclidean distance classification method is better than the classification method by using a single feature. Can we replace Feature 2 assigned on layer 1 in the multi-layer decision tree method to achieve better classification results? This idea leads us to a mixed Euclidean distance and multi-layer decision tree method as shown in Fig. 11.

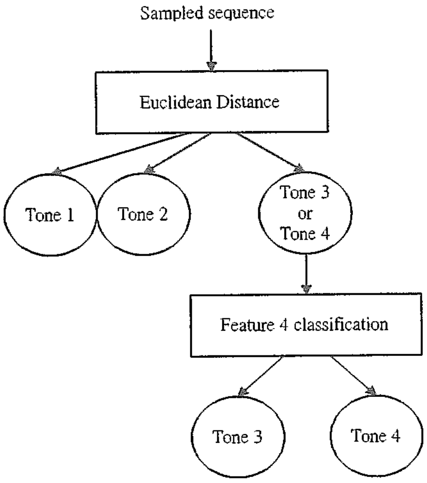

Fig. 11. Mixed Euclidean distance and multi-layer decision tree classification method.

4. Experimental results

We have applied the classification methods developed in Section 3 to a large collection of testing data for various speakers. We present these experimental results in the following sub-sections.

4.1 Test set collections

The voice data we used in the experiments were provided by 19 speakers, including 10 males and 9 females. We will denote the 10 males as M1, M2, ..., M10, and the 9 females as F1, F2, ... , F9. Each speaker was asked to pronounce the 1412 distinct syllable three times, in an isolated fashion. Because we discarded all the 22 tone 0 syllables, we have three sets of voice data each containing 1390 syllables available to use for each person in the experiments. We used two sets of these voice data of 1390 syllables for training and the remaining set for testing for each person in each experiment. Table 1 shows the distribution of the four tones in the training and the testing data for each speaker in the experiments.

	Tone 1	Tone 2	Tone 3	Tone 4	Total
Training Data	680	712	666	722	2780
Testing Data	340	356	333	361	1390

Table 1. The distribution of the four tones in the voice data in each experiment.

4.2 Euclidean distance classification method result

In this experiment we use the method described in Section 3.1, the Euclidean distance classification method, to perform the test. Fig. 12 gives the results.

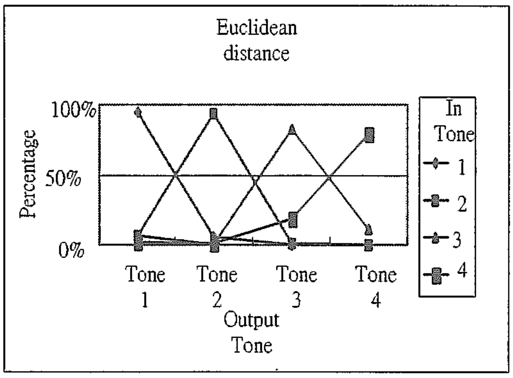

Fig. 12. Euclidean distance classification distribution.

4.3. Multi- layer classification method

In this section we test the multi -layer classification method. In experiment A, we used Feature 1 on the first layer of the decision tree. Fig. 13 gives the results. In experiment B, we used Feature 2 on the first layer of the decision tree. Fig. 14 gives the results. In experiment C, we used Feature 3 on the first layer of the decision tree. Fig. 15 gives the results.

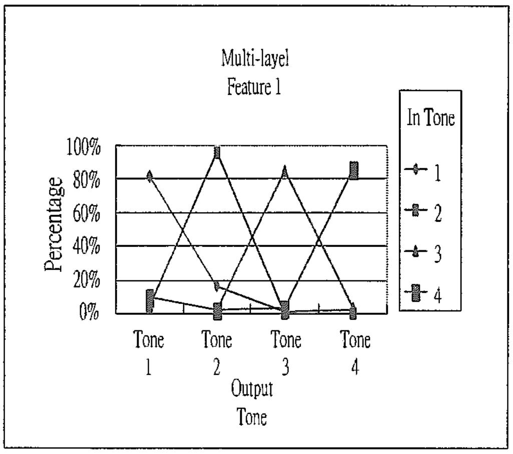

Fig. 13. Multi-layer feature 1 at layer 1 distribution.

B. Feature 2 with Multi -layer method

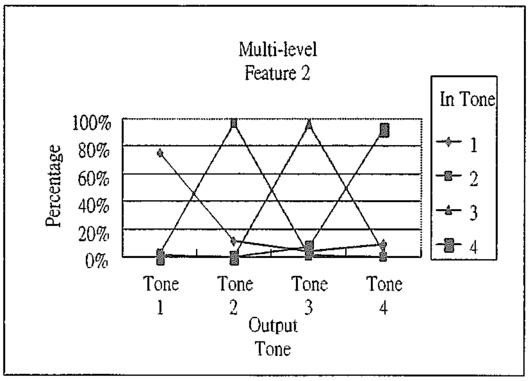

Fig. 14. Multi-layer feature 2 at layer 1 distribution.

C. Feature 3 with Multi -layer method

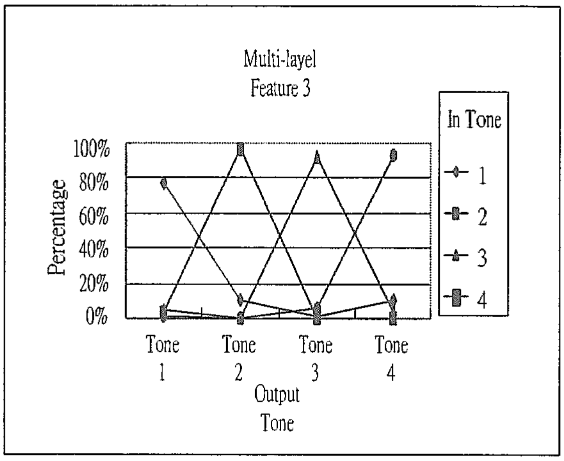

Fig. 15. Multi-layer feature 3 at layer 1 distribution.

4.3. Mixed Euclidean distance and Multi -layer method

In this section we test the Euclidean distance and multi-layer classification method. Fig. 15 gives the results.

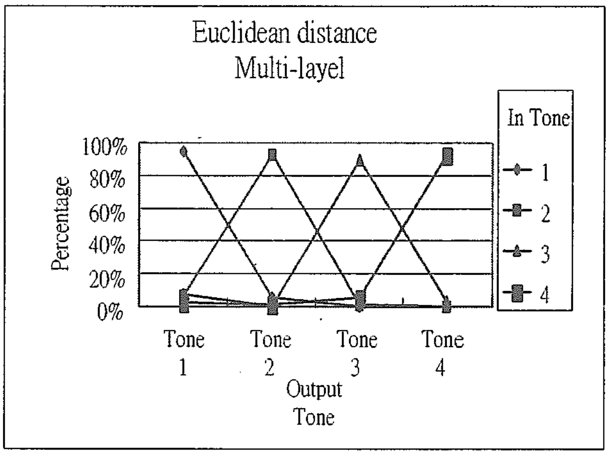

Fig. 16. Mixed Euclidean distance and Multi-layer distribution.

4.4. Results summary

Fig. 17. Summarizes all the experimental results in this Section. Here we classify the speakers into female and male groups. We take the average of the recognition rates over all the four tones and all the persons in a group for each of the classification method presented in Section 3. The symbols used for the different classification methods are listed in Table 2.

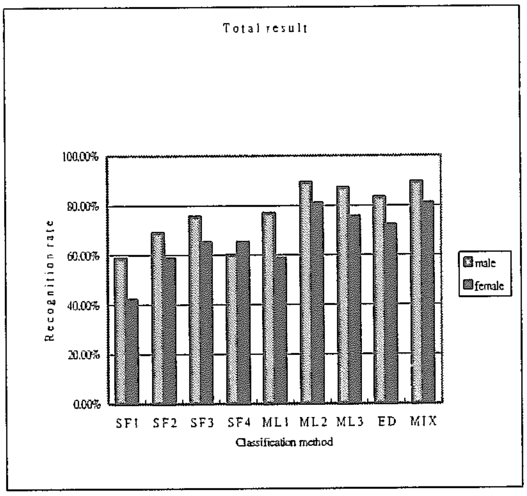

Fig. 17. Summarization of all the experimental results.

Symbol	Meanin
SF1	Single Feature 1
SF2	Single Feature 2
SF3	Single Feature 3
SF4	Single Feature 4
ML1	Multi-layer feature 1
ML2	Multi-layer feature 2
ML3	Multi-layer feature 3
ED	Euclidean distance
MIX	Euclidean distance Multi -layer

Table 2. Symbol of Fig. 17.

5. Conclusions and Future Work

In this paper, we have developed methods for tone recognition in Mandarin. We designed features based on observing the special properties of the pitch period sequences of tone 1 to tone 4 to·distinguish them. Different features are good at distinguishin g different pairs of the tones. This can be summarized in Table 3. For example, Features 1, 2 and 3 are good at distinguishing Tone 1 and Tone 3, while only Feature 4 can distinguish Tone 3 from Tone 4 well.

	Tone1	Tone 2	Tone 3
Tone 2	F2		
Tone 3	F1,F2,F3	F1,F2,F3	
Tone 4	F1,F2,F3	F1,F2,F3	F4

Table 3. Features that are good for distinguishing different pairs of the tones.

We have proposed three major classification methods—Euclidean distance calculation, multi-layer decision tree, and the mixture of the two. We have performed extensive experiments to test the classification methods proposed on large collections of test sets. Not one of the classification methods is superior to other methods in every case. However, the method in which we used the Euclidean distance calculation on the first-layer of the decision tree (the mixture method) performed better than other methods in general.

The mixture method has a recognition rate higher than 90% for many individuals. For example, its recognition rates for Male 1 are 94.7, 92.9, 89.7, and 92.4 percents for tone 1 to tone 4, respectively The total average recognition rates of the mixture method is 89.4% for males, and 80.9% for females.

The recognition rates for females are lower than that for males for all the proposed methods. The differences are quite significant in all the experiments. Typically a 10% difference can be observed. Closer examinations need be done to find out the real reasons.

We believe that a much better algorithm for tone recognition in Mandarin can be achieved through the following directions.

1. Refining the pitch period sequence calculation routine.
2. Design better features to distinguish the tones. We designed only four features. There must be other good ones to be discovered. We need also develop techniques to enhance the distinguishing capabilities of the features.
3. Develop better classification method. There are various ways to define distance measures and assigning features to the different layers on a decision tree. More experiments need be done to test different ways to combine the classification methods to achieve an even better one.

References:

1. L. S. Lee, C.Y. Tseng, H.Y. Gu, F. H. Liu, C. H. Chang, Y. H. Lin, Y. M. Lee, S. L. Tu, S. H. Hsieh, and C. H. Chen. "Golden Mandarin (I)-A Real-Time Mandarin Speech Dictation machine for Chinese Language with Very Large Vocabulary," IEEE Trans. On Speech and Audio Processing, vol. 1 no 2, p158-179, APRIL 1993.
2. C. H. Lin, C. H. Wu, P. Y. Ting, H. M. Wang. "Frameworks for recognition of Mandarin syllables with tones using sub-syllabic units," Speech Communication, vol.18, no.2, p175-190, April 1996.
3. Y.R. Wang, J. M. Shieh, and S. H. Chen. "Tone recognition of continuous Mandarin speech based on hidden Markov model," Int. J. Pattern Recog. Artific. Intell., 8:p233-246, 1994.
4. R. S. You, " Tone recognition of mandarin speech," Masters Thesis, National Chiao Tung University,Taiwan, 1988.
5. P.C.Chang,"Tone Recognition and Initial Analysis of Isolated Mandarin Word," Master thesis, National Chiao Tung University,1986.
6. W. J. Yang, J. C. Lee, Y. C. Chang, and H. C. Wang, "Hidden Markov model for Mandarin lexical," IEEE Trans. Acoust. Speech, Signal Process. 36, 7: p988-992, 1988.
7. Y.R. Wang, and S. H. Chen. "Tone recognition of continuous Mandarin speech based on neural network," IEEE Trans. On Speech and Audio Processing,3(2): p146-150,Mar. 1995.
8. M. CB, J. A. "Speaker normalization in the perception of Mandarin Chinese tones," Journal of the Acoustical Society of America, vol. 102, no. 3, Sept. 1997.
9. Q. J. Fu, F. G. Zeng, S. RV, S. SD. "Importance of tonal envelope cues in Chinese speech recognition," Journal of the Acoustical Society of America, vol. 104, no. 1, p505 -510, July 1998.
10. T. Lee, P. C. Ching, L. W. Chan, Y. H. Cheng, and B. Mak. "Tone Recognition of Isolated Cantonese Syllables," IEEE Trans. On Speech and Audio Processing.Vol. 3, no. 3, p204 -209, May 1995.
11. K. F. Lee and H. W. Hon, "Speaker-independent phone recognition using hidden Markov models," IEEE Trans. Acoust. Speec Signal Process. 37, 11: p1641 -1648, 1989.
12. J.D. markel,"The SIFT Algorithm for Fundamental Frequency estimation," IEEE Trans. On Audio and Electroacoustics, vol. AU-20, no.5, p367-377, 1972.
13. [13]. J. Neter, W. Wasserman, G. A. Whitmore, Applied Statistics, Allyn & Bacon, Reading, Boston, 1988.
14. M.J. Ross, H.L. Schaffer, Andrew Cohen, "Average magnitude Difference Function Pitch Extractor," IEEE Trans. On Acoustics, Speech, Signal Processing, Vol. ASSP-22, No.5, Oct. 1974.
15. A. Michael Noll, " Cepstrum Pitch Detection," J. Acoust. Soc. Am., Vol. 41, pp.293-309, Feb. 1967.
16. John S. Devitt, Calculus with Maple V, Reading, Pacific Grove, Calif, Brooks/Cole Pub. Co., 1993.