ISCA Archive
http://www.isca-speech.org/archive

International Symposium on Chinese Spoken
Language Processing (ISCSLP 2004)
Hong Kong
December 15-18, 2004

# TONE RECOGNITION FOR CHINESE SPEECH: A COMPARATIVE STUDY OF MANDARIN AND CANTONESE

*Gang PENG*[*], *Hongying ZHENG*[*] *and William S-Y. WANG*[†]

[*]Department of Electronic Engineering,
City University of Hong Kong, Kowloon, Hong Kong
[†]Department of Electronic Engineering,
The Chinese University of Hong Kong, Shatin, Hong Kong
gpeng@ee.cityu.edu.hk wsywang@ee.cityu.edu.hk

## ABSTRACT

This paper presents a comparative study on automatic continuous tone recognition for Mandarin and Cantonese. Compared with Mandarin, Cantonese has a much more complex tone system. The effects of $F_0$ normalization on tone recognition of Mandarin and Cantonese will be studied. Furthermore, the two tone systems will be compared from an engineering point of view. Tone recognition accuracies of 71.50% and 83.06% have been obtained for Cantonese and Mandarin respectively. These results compare favorably with results reported for other tone recognition experiments on the same (for Cantonese) and similar databases (for Mandarin).

## 1. INTRODUCTION

Tone is an essential component for word formation in all tone languages, and is used to build words much as consonants and vowels do. So speech recognition of tone languages depends not only on the articulatory composition but also on tone patterns.

During the last two decades, many approaches have been proposed for tone recognition. Hidden Markov Models, Neural Network, and Fujisaki's model have been applied to recognize tones in tone languages, such as Mandarin, Cantonese and Thai. For isolated tone recognition, very high recognition accuracy has been obtained. However for tone recognition in continuous speech, although relatively high tone recognition accuracy has been achieved in [1] and [2] for Mandarin and Thai respectively, manual segmentation has to be done before training the tone models, which is not suitable for automatic speech recognition. For automatic tone recognition, recognition score of 77.27% has been reported earlier for Cantonese with simple phonological constraints [3]. However, without phonological constraints, lower recognition scores of 70.01% [4] and 71.5% [3] have been reported for Mandarin and Cantonese respectively.

Following the methods proposed in [3] for continuous Cantonese tone recognition, Mandarin tones will be studied. Then these two different tone systems will be compared and analysed from an engineering point of view.

The Mandarin and Cantonese continuous speech databases being used in this study are described in the next section. Then the tone feature selection and normalization will be briefly introduced in Section 3. Experimental results are presented in Section 4. Finally, conclusions are given in Section 5.

## 2. SPEECH DATABASE

The Cantonese database we used is CUSENT database [5]. In this database, 5,100 training and another exclusive 600 test sentences were separately selected from five local newspapers of Hong Kong. The training sentences were evenly divided into 17 groups, each containing 300 unique sentences. Each group of sentences was read by four speakers (2F, 2M). Thus, a total of 20, 400 (300×4×17) training utterances were obtained from 68 speakers. The 600 test sentences were divided into 6 groups. Each group was read by one male and one female speaker (not drawn from the population of the training speakers). The total number of test utterances is 1,200.

For Mandarin database, the experiments are performed on the later developed part of the database from Chinese Project 863. An earlier portion of this database was reported in [6]. 1,560 sentences were selected from "The People's Daily". They are divided into three groups: group A with 521 sentences, group B with 519 sentences, and group C with 520 sentences. Group A were read by 27 (13F, 14M) speakers; group B by 28(14F, 14M) speakers; and group C by 27 (14F, 13M) speakers. The speakers for the above

This work described in this paper was supported by a CERG Grant (No. 9040676) awarded to Guanrong CHEN.

**Table 1**. Speech databases for Mandarin and Cantonese, where '#' stands for 'Number'.

| | Cantonese | | Mandarin | |
|---|---|---|---|---|
| Properties | Training Data | Test Data | Training Data | Test Data |
| # of Speakers | 68(34F, 34M) | 12(6F, 6M) | 76(38F, 38M) | 6(3F, 3M) |
| # of Syllables | 215,604 | 11,677 | 510,791 | 40,334 |
| # of Sentences | 20,378 | 1,198 | 39,519 | 3,120 |

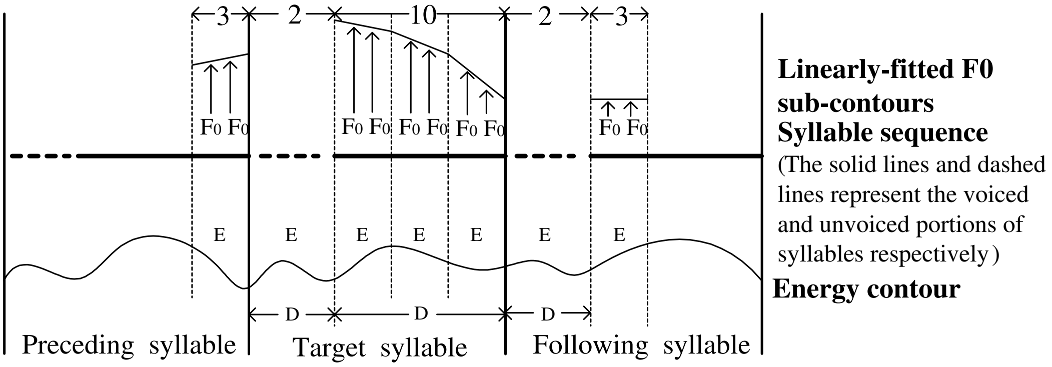
3
2
10
2
3
F0 F0
F0 F0 F0 F0 F0 F0
F0 F0
Linearly-fitted F0 sub-contours
Syllable sequence
(The solid lines and dashed lines represent the voiced and unvoiced portions of syllables respectively)
E
E
E
E
E
E
E
Energy contour
D
D
D
Preceding syllable
Target syllable
Following syllable

**Fig. 1**. Schematic diagram the 20 tone features. $E$ and $D$ represent energy and duration respectively. The numbers in the upper part of this figure indicate the numbers of features extracted from the corresponding speech segments.

three groups are exclusive. The utterances from randomly selected 6 (3F, 3M) speakers are used for testing. Table 1 summarizes the Cantonese and Mandarin database. Please note some corrupted utterances are excluded from this table by the database developer.

## 3. TONE FEATURE EXTRACTION AND NORMALIZATION

The tone of a syllable is mainly determined by its $F_0$ contour. The duration and energy are also related to the tones. Since the details of tone feature extraction and normalization have been presented in [3], we only enumerate the tone feature as follows:

(1) Duration of the $F_0$ contour of the target syllable; $F_0$ values at both the 1/3 and 2/3 time points of each of the three uniformly divided linearly-fitted $F_0$ sub-contours; the means of the three corresponding log-energy sub-contours.

(2) The same three features (i.e., two $F_0$ values, mean of the log-energy) of the last sub-segment of the preceding $F_0$ contour and the corresponding log-energy sub-contour, and the first sub-segment of the following $F_0$ contour and the corresponding log-energy sub-contour.

(3) Log-energy and duration of unvoiced/silent segments both before and after the $F_0$ contour of the target syllable.

As illustrated in Fig. 1, the ten features in (1) are all extracted from the target syllable; the six features in (2) are used to consider the tonal coarticulation effect from the neighboring tones, while the 4 features in (3) are used to implicitly represent the degree of mutual influence between the target tone and its neighboring tones.

Please note a log-scale 5-level transformation is first applied to raw $F_0$ values according to:

$$F_0'(i) = \frac{log_{10}(F_0(i)/Min)}{log_{10}(Max/Min)} \times 4 + 1. \qquad (1)$$

Then three schemes have been reported in [3] to determine the $Max$ and $Min$ in Formula (1). In the Scheme 1, the $Max$ and $Min$ represent the minimum and maximum $F_0$ values with the normalization window (Extending to the past 0.5 second and the future 1 second of the target syllable.) respectively. The Scheme 2 was used to capture the dynamic $F_0$ range of individual speaker, which is a self-adaptive normalization method. While the Scheme 3 was proposed to further consider the declination effect on tone recognition. The log-energy here is depicted as

$$E = 10log_{10}[R(0)], \qquad (2)$$

where $R(0)$ is the zero$^{th}$-order autocorrelation coefficient of the discrete time signal of speech. Then the log-scale energy is further re-scaled by the average log energy within the above normalization window. The energy normalization strategy is the same for three schemes.

**Table 2**. Training and test tone tokens for Mandarin and Cantonese

| | # of Training Tokens | # of Test Tokens |
|---|---|---|
| Mandarin | 76,718 | 38,334 |
| Cantonese | 59,903 | 11,141 |

**Table 3**. Recognition results with *level* feature set

| | Tone Recognition Accuracy | |
|---|---|---|
| | Mandarin | Cantonese |
| Scheme 1 | 80.81% | 64.28% |
| Scheme 3 | 83.06% | 71.50% |
| No Normalization | 77.02% | 60.82% |

## 4. TONE RECOGNITION RESULTS AND ANALYSIS

To obtain the training and test tone tokens for tone recognition, forced alignment by the HMMs was applied to obtain Initial-Final segmentation for all training and test utterances.

Cantonese tone tokens, i.e., the 20-dimension tone-feature vectors, extracted from 5,992 training utterances, from 20 (10 M, 10 F) randomly selected speakers, are used to train the Cantonese tone classifiers. Then tokens extracted from all utterances in the test set of CUSENT [5] are used to evaluate the performance of the tone classifiers. Mandarin tone tokens extracted from 6,239 training utterances, from 12 (6 M, 6 F) randomly selected speakers, are used to train the Mandarin tone classifiers. Then tokens extracted from 3,120 utterances, from 6 (3M, 3F) randomly selected speakers, re used to evaluate the performance of the tone classifiers. Table 2 summarizes the tone tokens used in this study. Please note the population for test is absolutely exclusive from the population for training.

The SVM-based (Support Vector Machine) tone recognizer proposed in [3] is reused here for Mandarin tones, and the recognition results for Cantonese tones presented in [3] will be adopted here for comparison.

For comparison, the tone-feature set introduced in Section 3 is so-called "*level* feature set". Herein, we introduce another feature set by only changing the two $F_0$ values of each $F_0$ sub-contour to the average $F_0$ value and the *slope* of the corresponding $F_0$ sub-contour. This feature set is so-called "*slope* feature set", because it includes *slope* features of $F_0$ sub-contours.

Table 3 summarizes the tone recognition results for both Mandarin and Cantonese. For Mandarin, acceptable accuracy (77.02%) has been obtained even without any normalization. For all three methods, there are considerable performance gaps between Mandarin and Cantonese. As shown in Figures 2 and 3, Cantonese tones are much more crowded in their tone spaces, especially at the lower part of Cantonese tone space. Please note the Mandarin neutral tone may make Mandarin tones a little bit more crowded than as is shown in Fig. 2. But due to the scarce distribution of neutral tone in Mandarin, the effect of neutral tone on the overall accuracy is not large. From an engineering point of view, the recognition of Mandarin tones is easier than that of Cantonese tones. And this statement is confirmed by the results listed in Table 3. Liu [8] has systematically studied the variation of Mandarin and Cantonese tones. She concluded that Cantonese tones vary in a smaller space area on average than do Mandarin tones. In other words, Cantonese tones are less tolerant in variation. Therefore, the tone variation makes the boundaries among Cantonese tones more uncertain.

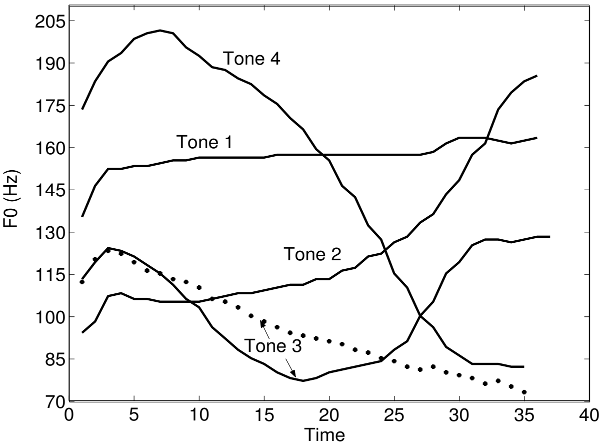

**Fig. 2**. $F_0$ contours of lexical tones of Mandarin uttered by a male speaker. The Tone 3 has three surface forms in continuous speech: Tone 2 according to tone-sandhi rule [7]; half Tone 3 (the dotted line) when at non-final position and no tone-sandhi effect; and Tone 3, when at final position of a phrase or sentence. Please note: here we only consider the surface forms, so the Tone 2 and the Tone 2 changed from Tone 3 by sandhi rule are the same tones in our experiment. The neutral tone is highly context-dependent tone, whose contour in completely determined by the preceding tone. Therefore, it is not presented here.

**Table 4**. Recognition results with *slope* feature set

| | Tone Recognition Accuracy | |
|---|---|---|
| | Mandarin | Cantonese |
| Scheme 1 | 80.34% | 60.18% |
| Scheme 3 | 82.56% | 68.62% |
| No Normalization | 75.39% | 57.80% |

As shown in Table 4, the recognition accuracies for Man-

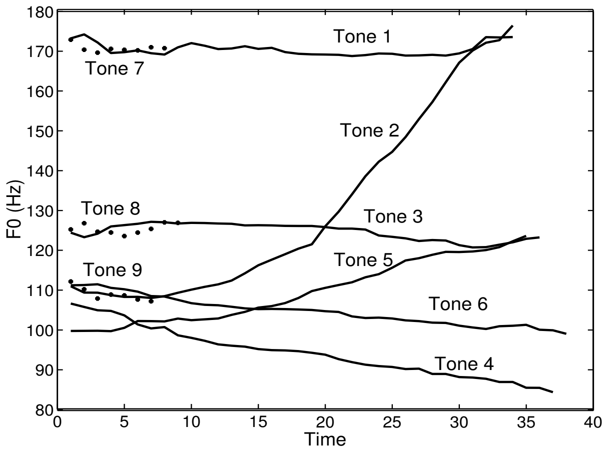

**Fig. 3**. $F_0$ contours of lexical tones of Cantonese uttered by a male speaker. The solid lines are for long tones on unchecked syllables, so-called non-entering tones, while the dotted lines are for short tones on checked syllables, so-called entering tones. In our experiments, the short tones 7, 8 and 9 are labeled according to their long counterparts 1, 3 and 6 respectively, because each pair locates at the same $F_0$ levels.

darin tones drops less than do Cantonese tones when the slope feature set is used. Mandarin is more like a contour system which uses distinctive tone shapes to contract with each other [9]. So the slope feature set preserves good distinction among Mandarin tones. (In $level$ feature set, two $F_0$ point have been adopted. They also implicitly include the $slope$ characteristics of Mandarin tones.) While Cantonese is closer to a register system, which uses distinctive pitch levels to distinguish tones [9]. So the $level$ feature set does considerably better than does $slope$ feature set, in which case the five useless $slope$ elements in the $slope$ feature set may reduce the distinction among the tone-feature vectors from the several level Cantonese tones.

## 5. CONCLUSIONS

A comparative study on tone recognition between Mandarin and Cantonese has been presented. Both Cantonese and Mandarin are tone languages. However, there exists significant difference between their tone systems. Different features may make an obvious differences in tone recognition accuracies. For instance, the $level$ feature set does much better than $slope$ feature set for Cantonese tone recognition. In this case the selection of tone features depends on the structure of the tone system. Moreover, due to the larger tone inventory size and much more complex tone structure, especially several level tones coexist, it makes continuous tone recognition for Cantonese more challenging.

## 6. REFERENCES

[1] S. H. Chen and Y. R. Wang, "Tone recognition of continuous Mandarin speech based on neural networks," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 2, pp. 146–150, 1995.

[2] S. Potisuk, M. P. Harper and J. Gandour, "Classification of Thai tone sequences in syllable-segmentated speech using the analysis-by-synthesis method," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 1, pp. 95–102, 1999.

[3] G. Peng and W. S.-Y. Wang, "Tone recognition of continuous Cantonese speech based on Support Vector Machines," *Speech Communication*, forthcoming.

[4] Y. Cao, Y. G. Deng, H. Zhang, T. Y. Huang and B. Xu, "Decision tree based Mandarin tone model and its application to speech recognition," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3, pp. 1759–1762 2000.

[5] T. Lee, W. K. Lo, P. C. Ching and H. Meng, "Spoken language resources for Cantonese speech processing," *Speech Communication*, vol. 36, no. 3-4, pp. 327–342, 2002.

[6] R. H. Wang, "National performance assessment of speech recognition system for Chinese," in *Proceedings of the Oriental COCOSDA Workshop'99*, pp. 41–44, 1999.

[7] W. S.-Y. Wang and K.-P. Li, "Tone 3 in Pekinese", *Journal of Speech and Hearing Research*, vol. 10, no. 3, pp. 629-636, 1963.

[8] J. Liu, *Tonal Behavior in Some Tone Languages*, Ph.D. Dissertation, City University of Hong Kong, 2001.

[9] C. John and C. Yallop, *An Introduction to Phonetic and Phonology*, Cambridge, MA: Basil Blackwell, Inc., 1990.