



# Perceptual Relevance of Pitch Contours of Mandarin Tones and its Efficacy in Prosody Generation of Speech Synthesis

*Shi-Han Chen and Chih-Chung Kuo*

Advanced Technology Center, ICL, ITRI, Hsinchu, Taiwan

korochen@itri.org.tw

## Abstract

Modeling Mandarin tones is one of the most important issues in speech synthesis. However, established knowledge is mainly focused on the “production” aspect. In this paper, we first characterized relative pitch levels of tones. Next, two perceptual experiments were designed to investigate “perceptual” relevance of pitch levels and shapes in Mandarin. Results showed that relative pitch levels of tones were perceptually more important than exactness of pitch shapes, and humans could not perceptually distinguish tonal variations in synthesized Chinese names.

**Index Terms:** speech synthesis, perception, prosody, tone

## 1. Introduction

Prosody generation is important for speech synthesis because it largely determines naturalness of synthesized speech. Besides, prosody also affects intelligibility of tonal languages like Mandarin. Modeling pitch contours of Mandarin tones has always been a major issue since there are clear tonal variations in continuous speech.

There are basically four different tones in Mandarin [1]. When syllables are pronounced in isolation, pitch contours are rather stable and have well defined canonical forms [2]. However, when produced continuously, these contours show clear variations depending on tonal contexts [3][4][5]. Besides, the variations happen both in voiced syllables and syllables with voiceless consonants [6]. Furthermore, results in [6] indicate that voiceless consonants lead to additional local raisings of pitch contours at voice onset.

Many researchers try deal with tonal variations by adopting rules or statistical models in speech synthesis systems [4][5][7]. However, related knowledge is mainly established in the “production” domain, which tries to formalize mechanism of prosody production. “Perceptual” importance of tonal variations, however, is still unclear when we are dealing with speech synthesis. As concluded by Xu [6], for example, coarticulation affects mainly the beginning parts of tone contours, and various pitch contours of a particular tone still converge to an identical pattern. While humans can produce tone contours in a variety of ways, do we really have ability to perceptually distinguish such differences? Is there any governing factor that affects naturalness and intelligibility of speech in the perceptual aspect?

In this initial study, we tried to investigate perceptual relevance of pitch contours of tones by analyzing prosody of trisyllabic Chinese names. Chinese names were used because problems could be limited in word level, and we could still preserve rich tonal variations because degrees of coarticulation of syllables are all different in a trisyllabic word. In addition, name synthesis is also a strongly demanded application in practice. We first analyzed characteristics of relative pitch levels of tones, and results showed that average

pitch level ratio of each tone combination was consistent across speakers. Next, two perceptual experiments were designed to clarify the perceptual relevance of pitch levels and shapes. The first one was a tone recognition task for humans, and we found that relative pitch levels of tones were more important than exactness of variable pitch shapes due to tonal variations. We further verified the result by synthesizing two groups of speech. The first group was synthesized by using prosody extracted directly from true speech, and another was produced by replacing pitch contours of true prosody with canonical pitch shapes and the average pitch level ratios derived from database statistics. Perceptual quality of the two groups was very similar. It indicated that humans might not be able to perceptually distinguish tonal variations. Since contextual tonal variations complicate the task of modeling Mandarin tones, our findings might be helpful in designing prosody models of speech synthesis.

## 2. Analysis of Relative Pitch Levels

Mandarin tones are defined both by pitch shapes and levels. Chao [2] uses five pitch levels to describe tones and the representation is widely used in this field. For example, tone 1 is characterized by its steady pitch trajectory and relatively high pitch level in the representation. Therefore, there might be some stable relationships between pitch levels of different tones. In this section we tried to characterize relative pitch levels of tones in a restricted prosodic environment, Chinese names, and two different databases were analyzed.

### 2.1. Databases description

#### Database 1 (Female A):

Pitch contours of Mandarin tones depend on tonal contexts, syllable types, and positions in prosodic phrases. In order to collect as many variations as possible, our trisyllabic Chinese name database included all 64 different tone combinations. Each tonal combination further consisted of 20 names covering different types of consonants, which depended on voicing types of the second and the third syllables. There were totally 1280 names in our database, and these names were randomly chosen from 469,558 Chinese students’ names. The database was a 16 KHz female voice. Speech segmentation was first automatically done by tools developed previously [8]. Next, we manually checked every syllable’s boundary and corrected it if necessary. Pitch tracking was done by Praat [9], and we corrected about 320 names having irregular pitch contours.

#### Database 2 (Female B):

In order to verify analysis results of database 1, we collected this database using another female’s voice. This database consisted of 297 names that were randomly chosen from database 1, and it was processed by the same procedure mentioned above.

## 2.2. Relative pitch levels of tones

In order to investigate relationships of pitch levels of tones, *mean values* of log-pitch contours of syllables in database 1 were calculated to represent pitch levels. Next, we categorized the database into 16 disyllabic tone combinations, and *Pearson's Correlation (R)* of each tone combination was calculated. In Table 1 we gave distribution of correlation values. All correlation values were higher than 0.3, and 40.0% of the values were higher than 0.75. In Figure 1 we gave some examples to illustrate correlations of pitch levels between consecutive syllables of different tone combinations. Relationships between the first and the second syllables (blue x-marks) were similar with that between the second and the third syllables (red circles). And also, people tended to lower their pitch levels in the end of words.

After investigation of the results, we found that tone combinations that included tone-3 contributed the lowest correlations ( $0.3 < R < 0.5$ ). It might be due to the relatively stronger variations in pitch contours of tone-3. Standard deviation of log-pitch levels of tone-3 was about 1.26, which was nearly double of other tones ( $\sim 0.69$ ) in our database. Correlations between 0.5 and 0.75 were contributed by those having tone-4 at the end of names. As shown by the red circles in the lower-right part of Figure 1, some instances had pitch levels unusually lower than average. Because tone-4 is a sharp downward accent from high to low level, it makes mean values of pitch vary easily from syllable durations. This is especially true for the last syllable of a word, which usually has larger variation in duration.

While degrees of correlation of pitch levels had some inconsistent across different tone combinations, we calculated *average pitch level ratio* of consecutive syllables of each tone combination, and results were compared with another speaker (database 2) in Figure 2. It clearly showed that there was a strong correlation and highly consistent relationship between the two speakers. In fact, Pearson's Correlation of the two sets of average pitch level ratios is higher than 98%.

<i>R</i>	>0.9	>0.75	>0.5	>0.3
% in range	6.7	40.0	60.0	100.0

Table 1: Distribution of Correlations of pitch levels

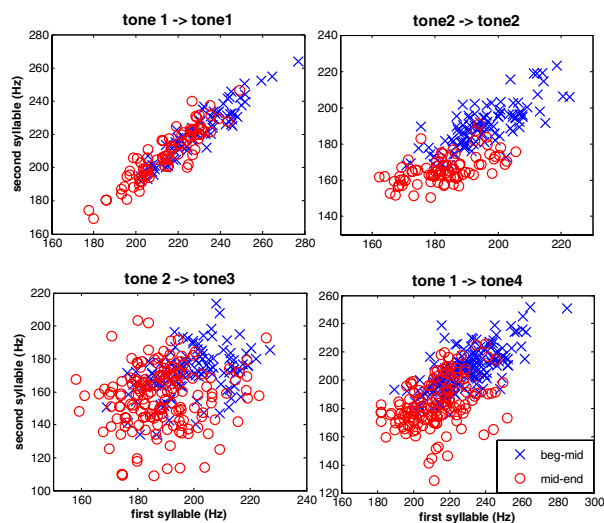


Figure 1: Correlations of pitch levels of consecutive syllables of different tone combinations

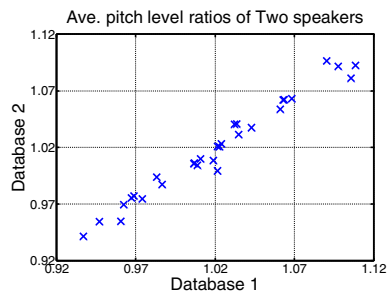


Figure 2: Correlation of average pitch level ratios of tone combinations of two speakers

The above results showed that although “exact” relationships of pitch levels of tones might depend on more sophisticated factors than merely on types of tone combinations, the average pitch level ratios, which represented “average” relationships of pitch levels of tones, were relatively consistent across speakers. In Section 4 we will determine whether humans could distinguish such variations in relative pitch levels by speech synthesis.

## 3. Perceptual relevance of pitch levels and shapes in a tone recognition task

In order to investigate the perceptual relevance of pitch levels and shapes in term of “intelligibility”, a tone recognition task was performed in this section. And in the next section, we will examine the perceptual relevance of tone contours in term of “naturalness” by our speech synthesis system.

### 3.1. Material

In Figure 3 we listed 15 disyllabic Mandarin words used in the experiment, which covered all possible tone combinations (without 3-3 because of sandhi rules [1]). Among them, only shen-3 shen-4, “thoughtful”, is a word in Mandarin, and others are nonsense words and unnatural combinations. These words were pronounced by a male speaker, and they served as *original words* ( $O(i)$ ,  $1 \leq i \leq 15$ ). Next, we produced *pseudo words* ( $P(i, j)$ ,  $1 \leq i, j \leq 15$ ,  $i \neq j$ ) whose pitch levels were same as those of  $O(i)$ , but having pitch shapes of  $O(j)$ . This was performed by modifying pitch level of each syllable in  $O(j)$  by PSOLA [10], and preserving pitch shapes of  $O(j)$ . Therefore,  $P(i, j)$  and  $O(i)$  both shared the same pitch level ratio  $R(i)$ , which differed from  $R(j)$  of  $O(j)$ . There were totally 15 original words and 210 pseudo words.

/shen/ was used because its unvoiced consonant prevents discontinuity in pitch between syllables after pitch modification. Since in [6] it shows that voiceless consonants do not change original contextual tonal variations, we believed this set of words was suitable for the experiment.

### 3.2. Listening procedure

20 Taiwanese people who speak Mandarin were asked to perform the tone recognition task. In the task, listeners were asked to select a word from Figure 3 after a stimulus was played by a computer program. They listened to 15 original words first, and the computer program would stop the task if more than four syllables were given wrong answers. In this way we could ensure that each participant has adequate ability to perform the task, and only one failed to pass the test. 210 pseudo words were separated into three groups, and each participant listened to one of the groups in a test. The whole task took about seven minutes, and orders of stimuli were random.

shen-1 shen-1	shen-1 shen-2	shen-1 shen-3	shen-1 shen-4
深深	深神	深审	深慎
shen-2 shen-1	shen-2 shen-2	shen-2 shen-3	shen-2 shen-4
神深	神神	神审	神慎
shen-3 shen-1	shen-3 shen-2	shen-3 shen-4	
审深	审神	审慎	
shen-4 shen-1	shen-4 shen-2	shen-4 shen-3	shen-4 shen-4
慎深	慎神	慎审	慎慎

Figure 3: Disyllabic words used in tone recognition

### 3.3. Results

Results were shown in Table 2 and Table 3, and data of the one who failed the initial test was excluded. Here we defined recognition rate as percentage of syllables of  $P(i, j)$  that were given same tone responses as those of  $O(j)$ . In other words, it represented percentage of syllables that were not influenced by changes in pitch level ratio. As indicated by Table 2, recognition rate was almost perfect when there was no change in pitch level ratio (listening to true words). However, when listening to pseudo words, recognition rates decreased when changes in pitch level ratio increased.

In Table 3 we separated responses of  $P(i, j)$  by tone types of  $O(j)$ . When pitch level ratios were altered, tone-3 syllables were easily perceived as tone-4, and tone-2 syllables were easily identified as tone-1. Although tone-1 and tone-4 syllables were relatively robust to changes in pitch level ratios, there were still about 30% of syllables being perceived as tone-2 and tone-3, respectively. On the other hand, listeners hardly recognized tone-1 and tone-2 syllables as tone-3 and/or tone-4 syllables. Tone-3 and tone-4 syllables were seldom perceived as tone-1 and/or tone-2 syllables, either.

In order to clarify the perceptual relevance of relative pitch levels in recognizing tones, we analyzed pitch level ratio  $R(k)$  of response  $O(k)$  after listening to stimulus  $P(i, j)$ , whose pitch level ratio was  $R(i)$ . Distribution of differences between  $R(k)$  and  $R(i)$ , and distribution of all possible differences of pitch level ratios (between  $O(i)$  and  $O(j)$  when producing  $P(i, j)$ ) were listed in Table 4. While only about 50% of all possible differences of pitch level ratios were below 0.2, more than 90% of differences between  $R(k)$  and  $R(i)$  were smaller than 0.2. It showed that listeners might tend to choose a tone combination having pitch level ratio closer to that of the given stimulus.

Finally, while the above results demonstrated that pitch shape of a tone might be able to “morph” another tone if relative pitch levels were correct, we would like to know whether pitch shape of a tone in a particular tonal context could be perceived as the same tone in different tonal context. As shown by Xu [3], for example, pitch shapes of tone-1 are very different when preceded by tone-1 and tone-3. In Table 5 we listed proportions of listeners who considered tones of second syllables were unchanged when second syllables of  $P(i, j)$  and  $O(j)$  had the largest tonal variations [3]. We did not observe clear perceptual influence caused by tonal variations. It is worth noting that in the case  $P(i, j) = \{\text{tone-3, tone-2}\}$  and  $O(j) = \{\text{tone-2, tone-2}\}$ , six listeners considering the second syllables were changed all gave responses of  $\{\text{tone-2, tone-1}\}$ . The reason might be due to the fact that pitch shapes of tone-2 are the flattest when being preceded by tone-2, and pitch level ratios of  $\{\text{tone-3, tone-2}\}$  resembled  $\{\text{tone-2, tone-1}\}$  a lot in our previous analysis of databases.

Word Type	$O(i)$	$P(i, j)$				
Abs. $R(i)-R(j)$	0	< 0.1	< 0.3	< 0.5	< 0.7	
Rec. rate (%)	98.8	89.3	73.8	64.1	61.6	

Table 2: Tone recognition rates in different degrees of change in pitch level ratio

Ori. Tones	Tone responses (%)			
	Tone1	Tone2	Tone3	Tone4
1	60.6	30.2	8.9	0.3
2	47.4	49.6	2.8	0.2
3	5.3	0	31.4	63.3
4	0.1	0.6	28.3	71.0

Table 3: Tone responses of pseudo words

Pitch level ratio diff.	< 0.1	< 0.2	< 0.3	< 0.7
Abs. $R(k)-R(i)$ (%)	61.7	91.4	98.9	100
All (%)	22.9	49.5	71.4	100

Table 4: Distribution of differences between pitch level ratios of recognized words and pseudo words

Tone	$P(i, j)$	$O(j)$	(%)	$P(i, j)$	$O(j)$	(%)
1	1,1	3,1	6/7	3,1	1,1	8/8
2	2,2	3,2	7/8	3,2	2,2	2/8
3	2,3	4,3	8/8	4,3	2,3	8/8
4	1,4	3,4	8/8	3,4	1,4	8/8

Table 5: Proportions of listeners giving unchanged tone responses when there were tonal variations

### 3.4. Discussion of the results

We listed some of our views of the results here for discussion.

1. Pitch level ratio has clear influence on tone recognition. It might be a perceptually important factor used to distinguish tone-1 from tone-2, as well as tone-3 from tone-4.
2. Pitch shape might be mainly used to perceptually distinguish the group of tone-1 and tone-2 from the group of tone-3 and tone-4. Listeners seldom misjudged the two groups.
3. Listeners might identify tones of a word by using pitch shapes to distinguish the two groups mentioned above, and choosing a tone combination having similar pitch level ratio with that of the given stimulus.
4. Although tonal variations are important for applications like automatic speech and tone recognition, they might be perceptually less important for a tone recognition task done by humans. If this is true, then we might be able to take advantages of the perceptually indistinguishable tonal variations in speech synthesis.

## 4. Perceptual relevance in speech synthesis

In previous section we demonstrated that tonal variations in pitch shapes were perceptually less important, and in Section 2 we also showed that average pitch level ratio of each tone combination was consistent across speakers. In this section, we performed Mean Opinion Score (MOS) test [11] to assess naturalness of synthesized speech produced by results from previous sections.

### 4.1. Synthesis material

100 trisyllabic Chinese names were randomly selected from database 2, and we generated units by our male-version

corpus-based text-to-speech (TTS) system, using true prosody extracted from true speech. Next, we produced two groups of prosody-modified speech using the same units mentioned above. Prosody of the first group was modified to fit the true prosody using PSOLA. These names served as a reference group. Next, we generated simplified prosody for the second group by replacing pitch levels and shapes in the true prosody:

1. Pitch level of the first syllable was fixed to average pitch of our TTS corpus. Pitch levels of the next two syllables were calculated using average pitch level ratios derived from statistics of database 1. Pitch level ratios depended on types of tone combinations and positions in words, which were discussed in Section 2.
2. Pitch shape of each syllable was chosen from canonical shapes, which were generated for each tone by averaging all pitch shapes in database 1 without considering tonal variations. Time-normalized vectors described in [12] were used to represent pitch contours, Pitch shapes depended on positions in words and tone types.

Similarly, PSOLA was used to modify prosody for the second group. Since PSOLA occasionally produces noticeable distortion after prosody modification, speech having severe distortion was automatically excluded by method developed previously [13]. Then we manually selected speech with unnoticeable distortion from remaining speech. Finally there were 63 stimuli for each group.

#### 4.2. Listening procedure

20 people were asked to perform the test. Among them nine are speech experts and others are volunteers. We used a five-scale scoring criteria in the test. Listeners were first given a short demonstration to show range of speech quality. After that listeners began to judge stimuli played in a random order. In each test, there were totally 63 stimuli consisted of equal amount of synthesized speech from the two groups and true speech. There were two sessions in each test. Listeners gave scores without seeing characters of stimuli in the first session. Next, they were asked to indicate in which parts the stimuli were unnatural and/or incorrect by choosing from five categories, which were pitch, speed/tempo, loudness, continuity, and others, after seeing characters of the stimuli. In this way we could ascertain efficacy of the simplified pitch contours of tones without harming fairness of MOS test. The whole task took about sixteen minutes.

#### 4.3. Results

MOS results were shown in Table 6. Synthesized speech with simplified prosody had similar quality with that using true prosody. Quality of true prosody was even slightly worse, and it might be due to the fact that some syllable boundaries of true speech included unnecessary silences, which led to flatter pitch shapes after pitch vectors were decoded using longer syllable durations. As for correctness of the simplified pitch contours, only four out of 63 stimuli were considered unnatural in pitch by the speech experts. While MOS results included responses of all listeners, judgments of correctness of prosody from volunteers were excluded here because we found they easily confused pitch, speed/tempo, and continuity with each other.

By maintaining only the average pitch level ratios of tone combinations, and using only canonical pitch shapes without considering tonal variations, we produced synthesized names having similar quality with that using true prosody. It indicated that humans might not be able to perceptually distinguish variations in pitch levels and shapes. Note that

true prosody was extracted from female B, and simplified prosody was generated from female A. In other words, what we observed in pitch contours of tones might be consistent across different speakers.

Group	True speech	True prosody	Simplified prosody
MOS	4.8	3.6	3.8

Table 6: MOS of true speech and synthesized speech using different pitch contours of tones

## 5. Conclusions

Tonal variations affect both pitch levels and shapes. However, we showed that relative pitch levels of tones might be perceptually more important than variations of pitch shapes. Average pitch level ratios of different tone combinations were shown to be consistent across speakers. Using only canonical shapes and the average pitch level ratios might be sufficient to synthesize speech in word level. Although the results were limited in name synthesis, we will try to apply them in general speech synthesis in the near future.

## 6. Acknowledgements

This paper is a partial result of Project 6301XS2610 conducted by ITRI, under sponsorship of the Ministry of Economic Affairs, Taiwan, R.O.C.

## 7. References

- [1] C. H. Lee, H. Li, L. S. Lee, R. H. Wang, Q. Huo, *Advances in Chinese Spoken Language Processing*, World Scientific Publishing Company, 2006
- [2] Chao, Y.R., "A Grammar of Spoken Chinese", *University of California Press, (Berkeley)*, 1968
- [3] Yi Xu, "Contextual tonal variations in Mandarin," *Journal of Phonetics*, 25:61-83, 1997
- [4] C.-L. Shih and G. Kochanski, "Chinese Tone Modeling with Stem-ML." *ICSLP 2000*, vol.2, 67-70, Oct. 2000
- [5] L. S. Lee, C. Y. Tseng, and M. O.-Young, "The synthesis rules in a Chinese text-to-speech system," *IEEE Trans. on Acoust. and Speech Signal Proc.* pp. 1309- 1320, 1989.
- [6] Ching X. Xu and Yi Xu, "F<sub>0</sub> perturbations by consonants and their implications on tone recognition." *ICASSP 2003*
- [7] S. H. Chen, Wen-hsing Lai, Yih-Ru Wang, "A statistics based pitch contour model for Mandarin speech", *J. Acoust. Soc. Am.*, Vol. 117, No. 2, pp. 908-925, 2005
- [8] Chih-Chung Kuo and Chi-Shiang Kuo. "Automatic Speech Segmentation and Verification for Concatenative Synthesis," *Eurospeech 2003*, Geneva, Sep. 1-4, 2003
- [9] Boersma, Paul & Weenink, David Praat: doing phonetics by computer, <http://www.praat.org/>
- [10] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text to speech synthesis using diphones," *Speech Communications*, Vol. 9, pp. 453-476, December 1990.
- [11] ITU-T Recommendation P.800, "Methods for subjective determination of transmission quality", ITU, 1996
- [12] S. H. Chen and Y. R. Wang, "Vector quantization of pitch information in Mandarin speech," *IEEE Trans. Commun.*, vol. 38, pp. 1317-1320, Sept. 1990.
- [13] S.-H. Chen, S.-J. Chen, C.-C. Kuo, "Perceptual Distortion Analysis and Quality Estimation Of Prosody-Modified Speech For TD-PSOLA," *ICASSP 2006*, Toulouse, May 14-19, 2006