Tonal Effects on Intonation

Chilin Shih

Department of East Asian Languages and Cultures
Department of Linguistics
University of Illinois, Urbana-Champaign
cls@uiuc.edu

Abstract

This paper investigates the intricate patterns of tone and intonation interaction by sampling and comparing data across a few interactive terms. We identify situations with different degrees of tonal effects on intonation, and suggest that the prominence level, or prosodic strength, associated with narrow focus is a key factor controlling the magnitude of the effect.

1. Introduction

One of the puzzling dilemma in the literatures of Chinese tone and intonation is the seemingly conflicting claims on the exact relation between the two. On the one hand, numerous authors report that tone and intonation interact. On the other hand, models that assume independence between the two, implicitly or explicitly, are reported as working reasonably well. These models include some of the Chinese intonation models and f_0 generation models used in many speech applications.

There are two obvious possibilities under which the two opposite views may be reconciled. The first is that the magnitude of the interaction effect is small, hence the errors are negligible in a f_0 prediction model without interaction terms. The other is that the interaction effect may be strong in some data space and weak in others, hence researchers make seemingly opposite conclusions based on the data properties they are working with, and the f_0 generation model performs well on data similar to training data. Each report is correct within its own domain, as in the parable *The Blind Men and the Elephant*.

This paper investigates the intricate patterns of tone and intonation interaction by sampling and comparing data across a few interactive terms. The pilot study supports the second interpretation. The seemingly different conclusions from previous studies may stem from sampling different uses of language. Tonal effect on intonation is negligible in plain declarative and interrogative sentences. If the speech recording is limited to non-expressive reading style, and the speech application is for routine reports and routine dialogue acts, a f_0 generation model without explicit modeling of tone and intonation interaction may work quite well. On the other end of the spectrum, tone and intonation interact strongly in sentences with narrow focus. The magnitude of the effect appears to be correlated with the prosodic strength of the focus. When narrow focus lands on different tones, post-focus pitch range variation is big and the effect may extend into following phrases.

2. Background

F_0 variations in a sentence may come from tone, tonal coarticulation, intonation, and the interaction of these factors. The research community has a good handle on the modeling of tone and tonal coarticulation effects and the reported phenomenon is consistent [1, 2, 3]. Although tonal coarticulation effects may happen outside the scope of the tone trigger, the articulatory base of tonal coarticulation is understood, and the effect can be modeled accurately with lexical tone information and their prosodic strength. We have tested the model on all disyllabic tone pairs and read speech with the model Stem-ML [4, 5], and Xu has tested the model PENTA [6].

There are other tone modeling techniques that capture tonal coarticulation effects using rules [7, 8], statistical approaches [9], and neural network [10]. These models capture the surface phenomenon reasonably well because the effects are local, the patterns are limited, and the most important factors, the tonal combination, is given lexically.

F_0 variations from non-lexical factors, such as statement, question, discourse functions, focus and emotional states are all attributed to the *intonation* factor [11].

Some tone and intonation models rest at least implicitly on the assumption that there is no interaction between tone and intonation. More specifically, there is an assumption that intonation comes first, which defines the pitch range within which lexical tones are realized. Gårding [12] is a representative case. Her proposal of drawing sentence-level intonation grid to represent different expressive functions implies that such grids can be isolated at the sentence level, independent of the tonal composition of the sentence. Other similar models include [13], where the goal is to find f_0 contours representing declarative, yes/no question and wh-question, and [14], where word level f_0 contours are taken as the building blocks of sentence intonation. A slightly more complicated model [15] further incorporates focus and syntactic juncture information into the prediction of pitch range.

Nonetheless, experimental studies [16, 17, 6, 1, 18, 19] repeatedly show that there are interaction effects between tone and intonation. One may talk in impressionistic terms that question intonation tends to be higher than declarative intonation, but to estimate how much higher one may need information about the tonal composition of the sentence.

This finding presents potential problem to f_0 generation models, which need to predict the tone effects, intonation effects, and interaction effects. Precise measurement and parameter estimation is required to generate believable intonation for unrestricted input text. Interaction effect increases the complexity of the problem. If tone and intonation *do not* interact, we could get a fairly realistic survey of Chinese tone and intonation by studying factors controlling tonal variations and factors controlling intonation variations and derive the combination of the two by a simple function. The research project in this case would have been quite manageable. If tone and intonation in-

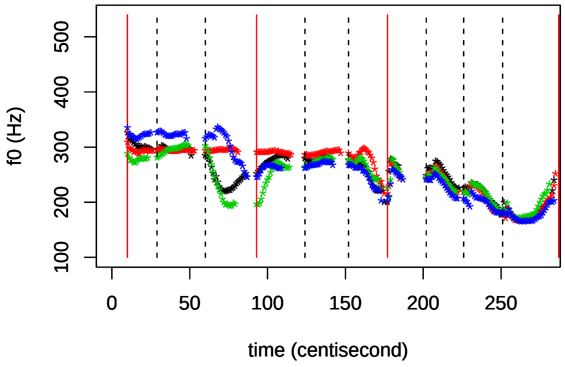

Figure 1: *Plain declarative intonation without narrow focus.*

teract, in principle we would need to laboriously plow through all the possible combinations, crossing tone factors with intonation factors, in order to get an equally comprehensive survey. The explosion of factor combination is one of the main reason why f_0 generation models are not yet successful in generating expressive intonation [7, 20, 8, 21, 10], for any moderate improvement requires large database and prohibiting resources.

With this background in mind, we design experimental data to investigate some of the hypothesis space of tone and intonation interaction. The eventual goal is to understand the nature of the interaction, and to build f_0 generation model that can handle the interaction effect.

3. Experimental Design

Much of the observation in this paper is drawn from a pilot study designed to investigate the nature of tone and intonation interaction. The following factors are considered:

- Lexical tones: tone1, tone1, tone3, tone4.
- Intonation types: declarative intonation and question intonation.
- Prosodic strength: with or without narrow focus on one digit.

The stimuli are 10 digit strings simulating U.S. telephone number in the following phrasing:

ddd-ddd-dddd.

We used digit strings to eliminate as much as possible any uncontrolled complication coming from syntax, semantics and discourse functions. Every digit in the telephone number is equally important. We expect that the speakers have full awareness that any error in any digit will render the whole string useless.

The tone of the third position rotates through all four lexical tones. This is where we plant the trigger of tonal variations. This is a position that is sufficiently far away from the reported loci of strong intonation effect, which is strongest at the sentence final position, and possibly with weak effect in the sentence initial position.

A few tone1's are used in the second, fourth, and fifth position, around the site with the planted tonal trigger. The digits in all other positions are randomly chosen with replacement.

The following digit strings are the primary source of discussion in this paper.

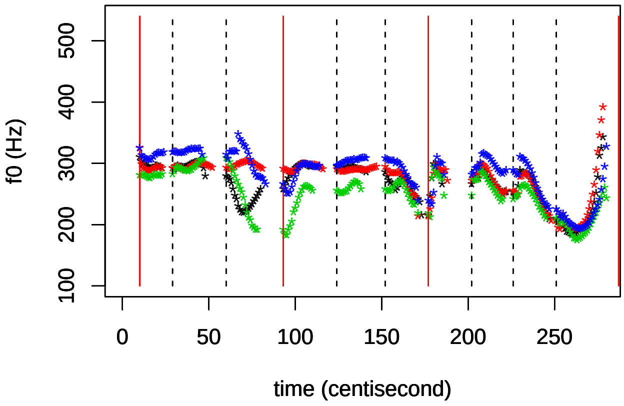

Figure 2: *Plain question intonation without narrow focus.*

Digit string	Tone sequence
811-112-7660	111-114-1442
810-112-7660	112-114-1442
815-112-7660	113-114-1442
816-112-7660	114-114-1442

Each digit string is conveyed in six different ways, yielding 24 stimuli.

- Plain (broad focus) declarative intonation, in answering to the question *What is your telephone number?*
- Plain (broad focus) question intonation, asking whether the telephone number is correct.
- Narrow focus contrastive intonation, asserting the third digit (with varying tones)
- Narrow focus contrastive intonation, asserting the fourth digit (with tone1).
- Narrow focus confirmation intonation, asking whether the third digit is correct (with varying tones).
- Narrow focus confirmation intonation, asking whether the fourth digit is correct (with tone1).

The stimuli were mixed with equal number of randomly generated digit strings as fillers, randomized, and recorded by 4 speakers, two repetitions each. The data reported in this paper comes from the second repetition of the first speaker, where the recording session was completed without errors.

The recording was made with Computerized Speech Lab (CSL) Model 4300B by Kay Elemetrics, and stored in 16 bit, 44100 Hz sampling rate .wav format.

The speech recording were cut into sentences using Praat. Syllable boundaries were labeled in Praat using auditory feedback, waveform, spectrogram, pitch and intensity displays.

4. Comparing tonal effects on intonation

We inspect the tone and intonation interaction effect by juxtaposing sentences with different lexical tones but with the same intonation intent. These sentences have identical tone sequence for all positions but the third syllable, hence we expect the intonation contour to be similar away from the third syllable if tone and intonation do not interact.

F_0 displays in this section are raw f_0 tracks. Syllable boundaries are drawn with black vertical dashed lines. Phrasing

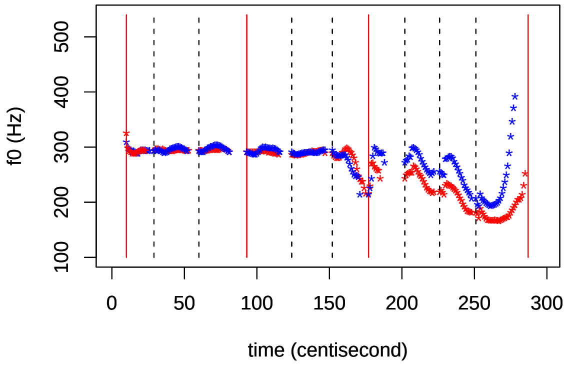

Figure 3: *Comparing f_0 contours of matched declarative sentence and question. The third position has tone1.*

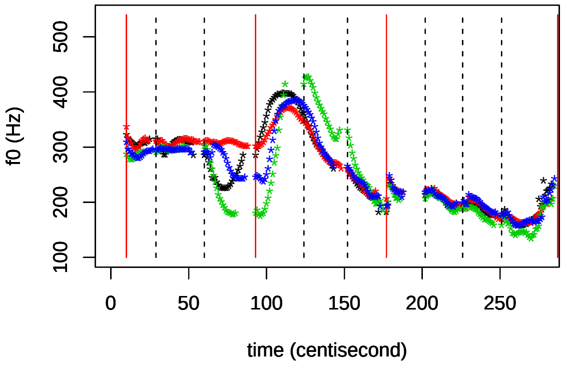

Figure 4: *Declarative intonation with focus on the fourth position, which has tone 1.*

boundaries are drawn with red vertical solid lines. The syllable duration used in the plots are the maximum syllable duration of that position in the database. F_0 display of each syllable starts with the syllable boundary line. This display method allow us to align raw f_0 contours of different sentences by syllable without altering the original duration.

4.1. Plain sentences

In plain sentences, tonal effects on intonation appears to be weak and local.

Figure 1 shows plain declarative sentences. These sentences were used to answer the question *What is your telephone number?*. Figure 2 shows plain questions These questions were asked to confirm a telephone number, which have the connotation *Do I get your telephone number right?*. In both cases there is no narrow focus.

In these two plots, f_0 values outside the third and fourth positions are comparable. The only notable difference is in Figure 2, the plain question, where f_0 contour in the phrase after tone 3 (green) is lower.

We expect the f_0 pattern on the third syllable to be different, since f_0 is the primary acoustic correlates of lexical tones. The noticeable f_0 difference of the fourth position is consistent with our understanding of the tonal co-articulation effect. Their patterns are predictable from local tonal information such as the current tone and the preceding tone.

It seems that the difference between question and statement can be straightforwardly derived by taking the difference between paired statement/question with identical lexical composition. Figure 3 shows one of these examples where the third position has tone1. The f_0 curve of the question is shown in blue, and the f_0 curve of the statement is shown in red. Question intonation floats slightly above the declarative sentence in the last phrase. In this plot, the final rising tone has a steeper rising slope in question, at the same time it rises to a higher pitch level. Note that final intonation contour doesn't necessarily rise in a Chinese question if the the lexical tone of the last syllable is different.

If we take the difference of f_0 contours from matched question/declarative sentence pairs, the results are similar despite the tonal difference on the third syllable. This level of consistency is the basis of many previous intonation models that represent tone and intonation separately.

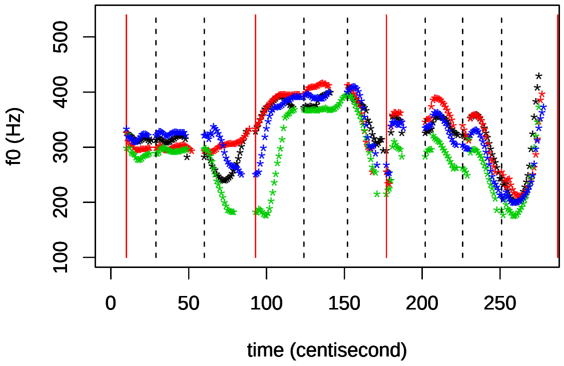

Figure 5: *Question with focus on the fourth position, which has tone 1.*

4.2. Narrow focus on like tones

Now we move to sentences with narrow focus. A typical usage of declarative sentence with narrow focus is to correct one of the digits, as in speaker B's reply:

> *A: Is your number 812-612-7660?*
>
> *B: No. It is 812-***1***12-7660..*

Conversely, a typical usage of question with narrow focus is to try to confirm one of the digit.

> *Is your number 812-***1***12-7660?*

In this case, the speaker indicated that he/she is uncertain about the fourth digit and requested confirmation only of that position.

Figures 4 and 5 show f_0 contours of declarative sentences and questions with narrow focus, respectively. Each plot includes four sentences where the lexical tones of the third position differ, and a narrow focus was placed on the fourth digit, which has a high level tone (tone1).

The last phrase *7662* has very similar f_0 values in declarative sentences, as in Figure 4. Even though there are different tones early on, the tonal effects is blocked by the narrow focus. We see more variations in this region in questions, as in Figure 5. There are some variations in the pitch height of the tone1 under narrow focus, and this difference is maintained in the second phrase. A general observation is that the narrow focus alters the intonation contour depending on its own strength.

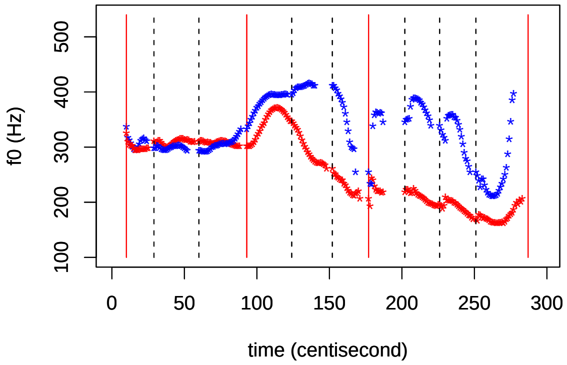

Figure 6: *The contrast between statement and question where there are narrow focus on the fourth position. The third position has tone1.*

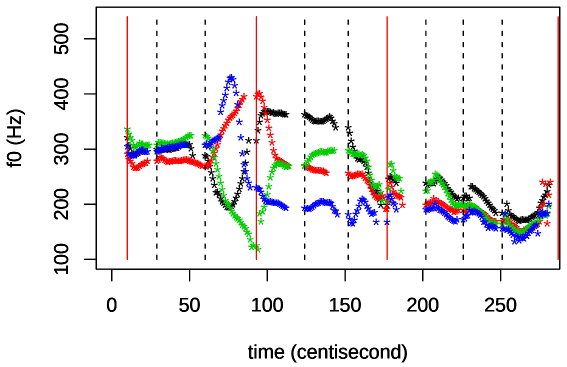

Figure 7: *Declarative intonation with focus on the third position, where lexical tones vary.*

The tones before the narrow focus is relatively weak and their effects is not strong enough to have an impact beyond the narrow focus.

Another observation is that the f_0 differences in matched declarative and question sentence pairs are not the same as those in the plain sentences. Figure 6 shows one of the matched pair examples where the third position has tone1. This means that the conversion function between question and statement is more complicated than the picture presented in the plain sentences. It appears that we would need multiple transfer functions to convert between statement and question, one for plain sentences, one for sentences with narrow focus.

4.3. Narrow focus on different tones

Figures 7 and 8 show yet another case of tone and intonation interaction. The narrow focus lands on the third digit where the lexical tones vary. Under these circumstances, tonal effect on intonation is strong and global. The tonal influence persists throughout the entire sentence.

These two figures show strong interaction between tone and intonation in the following sense:

- One need to know the tonal category of the third syllable in order to predict f_0 values later on, even in the last phrase.

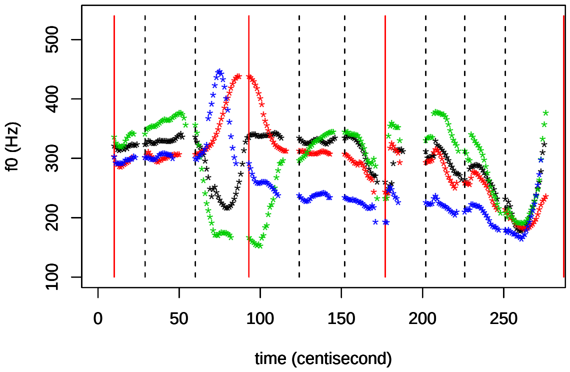

Figure 8: *Question with focus on the third position, where lexical tones vary.*

- F_0 differences in paired statement/question are no longer consistent across sentences with different tones. One needs the tonal information to estimate the magnitude and the direction of the difference.
- F_0 differences in paired statement/question with narrow focus is not the same as the F_0 differences calculated from plain sentence pairs. Therefore, we can no longer define a consistent transform function that converts a statement to a question and vice versa.

5. Discussion

Looking at these plots jointly, we can piece together a coherent pattern of Chinese tonal effect on intonation.

Narrow focus, and its associated prosodic strength and articulatory force, is a key factor controlling the magnitude of the effect. In plain sentences, there is little evidence of long term tonal effect. When narrow focus lands on different lexical tones, we see divergent intonation patterns after the focus that cannot be predicted by a pre-determined intonation pattern defined in abstraction of tonal information. Tonal variation immediately before the focus affects the realization of the focus tone due to tonal coarticulation, but has little influence on materials after the focus. This reduces a lot of the complexity in the system.

In our previous prosody learning and generation works based on the prosodic model Stem-ML [4, 5], we model f_0 production as a compromise between articulatory effort and communication needs. Furthermore, we introduce a term representing prosodic strength, which controls the interaction of physiological effort and communication needs. In the model, the shifting prosodic strength from unit to unit is what controls surface variations in f_0 generation. The model successfully accounts for tonal co-articulation effect in controlled experiments and in natural speech.

The finding of this paper suggests that the more global effect of Chinese intonation and its interaction with lexical tone is also mediated through prosodic strength. We will test this idea with Stem-ML modeling on the experimental data in the near future.

6. Conclusion

This paper investigates some of the hypothesis space of tone and intonation interaction.

Even though we find circumstances in which tone and intonation interaction effect is strong, tonal effect on intonation is largely predictable and manageable. Our data show that one of the major factor controlling tonal effect on intonation is the strength of the tone. Tones in plain sentences do not have a global effect on intonation while strong tones have strong effects. The same lexical tones with comparable prosodic strength exert similar influence onto the rest of the sentence. This situation is much more manageable than a unconstrained model.

This paper only investigated limited hypothesis space. There are inherent danger in generalizing this conclusion to unseen data.

7. Acknowledgement

I thank Jiahong Yuan and Greg Kochanski for comments and discussion.

8. References

[1] Chilin Shih, "Tone and intonation in Mandarin," in *Working Papers of the Cornell Phonetics Laboratory, Number 3: Stress, Tone and Intonation*, 83–109. Cornell University, 1988.

[2] Xiao-Nan Susan Shen, "Tonal coarticulation in Mandarin," *Journal of Phonetics*, vol. 18, 281–295, 1990.

[3] Yi Xu, "Contextual tonal variations in Mandarin," *Journal of Phonetics*, vol. 25, 61–83, 1997.

[4] Greg Kochanski and Chilin Shih, "Prosody modeling with soft templates," *Speech Communication*, vol. 39, no. 3-4, 311–352, 2003.

[5] Greg Kochanski and Chilin Shih, "Quantitative measurement of prosodic strength in Mandarin," *Speech Communication*, vol. 41, no. 4, 625–645, 2003.

[6] Yi Xu, "Transmitting tone and intonation simultaneously – the parallel encoding and target approximation (PENTA) model," in *International Symposium on Tonal Aspects of Languages – with Emphasis on Tone Languages*, 2004.

[7] Chilin Shih and Richard Sproat, "Issues in text-to-speech conversion for Mandarin," *Computational Linguistics and Chinese Language Processing*, vol. 1, no. 1, 37–86, 1996.

[8] Lin-Shan Lee, Chiu-Yu Tseng, and Ching-Jiang Hsieh, "Improved tone concatenation rules in a formant-based Chinese text-to-speech system," *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 3, 287–294, 1993.

[9] Sin-Horng Chen, Saga Chang, and Su-Min. Lee, "A statistical model based fundamental frequency synthesizer for Mandarin speech," *Journal of Acoustical Society of America*, vol. 92, no. 1, 114–120, 1992.

[10] Jianhua Tao, Xin Ni, and Lianhong Cai, "Clustering and feature learning based f_0 prediction for Chinese speech synthesis," in *Proceedings of the 7th International Conference on Spoken Language Processing*, Denver, 2002.

[11] Yuen Ren Chao, "Tone and intonation in Chinese," *Bulletin of the Institute of History and Philology*, vol. 4, 2121–2134, 1933.

[12] Eva Gårding, "Speech act and tonal pattern in standard Chinese: Constancy and variation," *Phonetica*, vol. 44, 13–29, 1987.

[13] Xiao-Nan Susan Shen, *The Prosody of Mandarin Chinese*, University of California Press, 1990.

[14] Zong-ji Wu, "A new method of intonation analysis for Standard Chinese: frequency transposition processing of phrasal contours in a sentence," in *Analysis, Perception and Processing of Spoken Language*, K. Hirose, Ed., 255–268. Elsevier Science and Technology Books, 1996.

[15] Jiong Shen, "Beijinghua shengdiao de yinyu he yudiao (Pitch range and intonation of the tones of Beijing Mandarin)," in *Beijing Yuyin Shiyan Lu (Acoustic Studies of Beijing Mandarin)*, 73–130. Beijing University Press, 1985.

[16] Jiahong Yuan, *Analysis and Modeling of Intonation in Mandarin Chinese*, Ph.D. thesis, Cornell University, 2004.

[17] Yi Xu, "Effects of tone and focus on the formation and alignment of f0 contours," *Journal of Phonetics*, vol. 27, 55–105, 1999.

[18] Chilin Shih, "A declination model of Mandarin Chinese," in *Intonation: Analysis, Modeling and Technology*, Botinis A., Ed., 243–268. Kluwer Academic Publishers, 2000.

[19] Yiya Chen, *The Phonetics and Phonology of Contrastive Focus in Standard Chinese*, Ph.D. thesis, State University of New York at Stony Brook, 2003.

[20] Sin-Horng Chen, Shaw Hwa Hwang, and Chun-Yu Tsai, "A first study of neural net based generation of prosodic and spectral information for Mandarin text-to-speech," in *Proceedings of IEEE ICASSP*, 1992, vol. 2, 45–48.

[21] M. Chu, H. Peng, and E. Chang, "A concatenative Mandarin TTS system without prosody model and prosody modification," in *Proceedings of the 4th ISCA Workshop on Speech Synthesis*, Scotland, 2001.