A Declination Model of Mandarin Chinese

Chilin Shih

1 Introduction

Declination has been studied intensively from production data and from perceptual experiments. There is a clear trend in many languages that F_0 values tend to drift down in an utterance, particularly in declarative sentences (t' Hart and Cohen 1973; Maeda 1976; Thorsen 1980a; Cohen et al. 1982). Perceptually, listeners compensate for such a downtrend: given two peaks of equal F_0 value, the later one is interpreted as having higher prominence, or for two peaks to be perceived as equal, the second peak should have lower F_0 value than the first (Pierrehumbert 1979; Gussenhoven and Rietveld 1988; Terken 1991; Ladd 1993; Terken 1993). There is also an extensive literature offering physiological explanations of declination (Lieberman 1967; Titze 1989; Strik and Boves 1992). Nonetheless, there are some obvious difficulties in interpreting declination data. Most of the languages being investigated have constraints on accent combination so that the declination slope has to be estimated from sparsely located data points such as F_0 peaks or valleys, while the phonological status of the observed peaks and valleys may be unclear. In addition, an observable F_0 contour of a sentence is the result of the combination of many factors. There is no unique solution in decomposing an observed complex F_0 pattern into individual effects. To study one effect one often needs to make strong assumptions about others, which turns out to be a major cause of disagreements in the intonation literature.

This paper proposes an experimental design which uses Mandarin high level tones to study the declination effect. Mandarin offers an easy way out of some of the problems cited above: it is possible to string sequences of high level tones together and thus allows for a syllable by syllable profile of the intonation topline. Some of the factors affecting intonation, to be discussed momentarily, are naturally absent in these high tone sentences, thus the experiment provides a unique opportunity to observe the declination effect more directly.

Previous studies of Mandarin intonation reported an overall downtrend in declarative sentences with mixed tones (Shen 1985; Gårding 1987) and in natural conversation (Tseng 1981; Yang 1995). There were also a few small scale studies of Mandarin high tone sequences. Shih (1988) and Liao (1994) found declination effect in sequences of

Mandarin high level tones, Shen (1985) emphasized the use of pitch raising to signal the beginning of syntactic boundaries, however, his data showed a downtrend within each syntactic phrase. Xu and Wang (1997) suggested that the observed downtrend comes from discourse effects.

We will start with a brief summary of the major factors affecting surface intonation contours, and comment on how these effects may impact the current study and how at least some of the problems can be addressed in the experiment:

- Declination: A global downtrend referring to the tendency of F_0 to decline over the course of an utterance (t' Hart and Cohen 1973; Maeda 1976; Pierrehumbert 1979; Cohen et al. 1982; Fujisaki 1983; Ladd 1984; Strik and Boves 1995).

- Downstep: A lowering effect that is triggered by a low (L) accent or a L tone, resulting in a descending step-like function in F_0 contour (Liberman and Pierrehumbert 1984; Pierrehumbert and Beckman 1988; Shih 1988; Prieto et al. 1996).

- Final Lowering: An additional lowering effect near the end of a sentence (Maeda 1976; Liberman and Pierrehumbert 1984; Pierrehumbert and Beckman 1988; Herman 1996).

- Accents and tones: Accents and tones create local F_0 excursions. Each accent and tone could be emphasized or de-accented, reflecting the meaning and the structure of the sentence, and the speaker's rendition of the sentence (Liberman and Pierrehumbert 1984; Eady and Cooper 1986; Jin 1996).

- Segmental effects: A lot of the observed F_0 movements are caused by segmental effects. Voiceless fricative and aspirated stops raise F_0, while sonorants typically lower F_0. Low vowels have intrinsically lower F_0 than high vowels (Peterson and Barney 1952; Lea 1973; Silverman 1987).

- Intonation type: Sentence intonation such as declarative, exclamation or question intonation may interact with any of the aforementioned effects. Some intonation models treat declination as the property of declarative intonation (Thorsen 1980a; Gårding 1987).

- Discourse structure: Pitch is typically raised in the discourse initial position and lowered in the discourse final position, and topic initialization is typically associated with high pitch (Hirschberg and Pierrehumbert 1986; Sluijter and Terken 1993; Nakajima and Allen 1993; Yang 1995).

There are many usages of the term “declination”, associated with different schools of thought. Ladd (Ladd 1993) classifies them as the *overt decline approach* and the *implicit decline approach*. The IPO model (t' Hart and Cohen 1973) is the strongest proponent of the overt declination school, where the declination slope is directly observable from the surface intonation contour. A line fitted through the peaks (if prominence variation can be controlled) reflects the topline declination, while a line fitted through the valleys reflects the baseline declination. Other studies that can be subsumed under the overt approach include Maeda (1976) and Umeda (1982).

The implicit declination school includes Fujisaki's (1983) mathematical intonation models where the surface intonation contour is the sum of the phrase component and the accent component. Declination is not directly observable from the intonation contour, but is accounted for by the phrase command, which is a damped function responding to an impulse command, calculated logarithmically. Pierrehumbert's (1980) model offers another implicit decline approach, where declination refers to a global, gradual effect which is modeled as a declining straight line (the baseline) which creates a downward tilt of an intonation contour, and accounts for the perceptual equivalence of early and late peaks with different F_0. The baseline is not a line fitted through the valleys in the intonation contour, nor does it parallel the topline since the topline is subject to other lowering effects and prominence variation. In later models, declination is calculated as the residual downtrend when the downstep effect is factored out (Liberman and Pierrehumbert 1984; Pierrehumbert and Beckman 1988). Liberman and Pierrehumbert (1984) made the strong claim that after the downstep effect was factored out, there was no evidence of declination in English. This position was softened in (Pierrehumbert and Beckman 1988) where both declination and downstep were incorporated in the modeling of Japanese intonation.

Mandarin offers a case where the implicit declination effect as defined in Liberman and Pierrehumbert (1984) can be observed overtly: First, in a sequence of high (H) level tones (tone 1), low (L) tonal targets are absent both phonologically and phonetically, so by definition there will be no trigger for any downstep effect. Secondly, the tone shape is level, which means that the contribution from the accent and the tone to the observed F_0 movement is minimal. There are still a few factors at work, of which final lowering and segmental effects are relatively easy to control, and the local prominence effect can be averaged out or smoothed out to some extent, when the sample size is large and the locations of uncontrolled local prominence distribute randomly in different sentences. Strictly speaking, isolated sentences read under experimental conditions are under the influence of the single-sentence discourse structure. However, the magnitude of discourse-raising or lowering in an isolate sentence, if they can be considered as such, is small comparing to what happens in the reading of long text with several paragraphs and in natural conversation. A natural extension of the current study will be to study paragraphs in high level tones to find out the difference between simple and complex discourse structure, and to model pitch range variations in complex discourse structure.

Under the current experimental conditions, the surface F_0 contour of the experiment sentences is a close approximation of the topline of the Liberman and Pierrehumbert model, and the observed downtrend on the surface is very close to the residue declination effect without downstep.

Note that the proposed experimental materials do not offer a direct observation of the phrase command of the Fujisaki model. The observed F_0 values of a given syllable are the combination (in the log domain) of the phrase command and the accent command corresponding to the tone, plus any previous accent commands still in effect (Fujisaki et al. 1990; Wang et al. 1990). If the tonal contribution (accent command) from each syllable is the same, the observed $F0$ will be parallel to the phrase command after the initial effect subsides.

Plain, final	Lao3-Wang2 zheng1 gua1
Plain, non-final	Lao3-Wang2 zheng1 gua1 le0
Focus, final	(Shei2 zheng1 gua1?) Lao3-Wang2 zheng1 gua1
Focus, non-final	(Shei2 zheng1 gua1?) Lao3-Wang2 zheng1 gua1 le0

Table 1: Experiment sentences in four test conditions. In addition, there are ten length variations.

2 Experiment design

This experiment investigates the pattern of declination in Mandarin and the possible interaction of declination with sentence length, final lowering and prominence. The database consists of 640 sentences: 10 test sentences with 10 length variations, 2 focus conditions, 2 final conditions, 4 repetitions, and 4 speakers (two females and two males, two from northern China and two from Taiwan). Each of the 10 test sentences starts with a two-syllable sentence frame in a low tone (tone 3) and a rising tone (tone 2) *Lao3 Wang2* “Old Wang”. Tone 3 has a low target (L) and tone 2 rises from low to high (LH), so the frame provides a reference to the speaker's pitch range. Sequences of high-level tones (tone 1, H) ranging from 2 to 11 syllables long follow the frame. The sentences share the theme of a person *Lao3 Wang2* cooking a winter melon in various ways. Sonorant and unaspirated consonant are used as much as possible to minimize consonantal effects on F_0. Many of the syllables have mid or low vowels, and end in a nasal coda [n] or [ng].

The final condition refers to a set of the test sentences which have tone 1 in the utterance final position. A non-final condition is created by adding a sentence final perfective particle *le* to the test sentences so that the last tone 1 syllable is no longer utterance final. If there is a final lowering effect affecting the utterance-final unit, whether the unit being syllable, word, or a fixed time interval, the last tone 1 syllable in the sentences without *le* should be lower than the ones in the sentences with *le*.

These 20 sentences were presented in two ways to the speakers: unadorned, plain sentences intended to elicit unmarked reading style, and sentences with a leading question such as *Shei2 zheng1 gua1?* “Who steams the melon?”, intended to elicit a narrow focus reading with the prominence landing on the frame *Lao3 Wang2*. Note that in the focus sentences all test syllables (tone 1 syllables) are actually not in focus. The four test conditions are exemplified in Table 1 with the sentence with two tone 1 syllables *Lao3-Wang2 zheng1 gua1* “Old Wang steams the melon”.

Figure 1 shows the pitch tracks of a set of test sentences. Each sentence begins with the tone 3–tone 2 (L–LH) frame, labeled as “frame”. There are ten test syllables in these sentences, which are labeled as “test syllables”. All test syllables have tone 1 (H). Sentences in the *final condition* end in a tone 1, while sentences in the *non-final condition* end in a neutral tone syllable *le0* labeled as “le”. A neutral tone syllable is unstressed and is realized with low pitch.

Ta1	shuo1	KEYWORD	san1	bian4
He	said	KEYWORD	three	times

Table 2: Sentences for the supporting experiment on segmental effects.

Although the end of the frame is tonally equivalent to the first test syllable, both have the tonal specification H, it is common in Mandarin that the end of a tone 2 doesn't reach the high level if the following tone is also high (Shih 1988; Jin 1996; Xu 1997).

Three of Speaker C's sentences were read incorrectly and were discarded, resulting in a total of 637 sentences and 4139 tone 1 syllables in the database. One F_0/time measurement is taken from each syllable: the lowest point of tone 3, the highest point of tone 2, and near the center of the rhyme of tone 1, at least 50 msec after the onset of the rhyme.

3 Supporting experiment: segmental effects

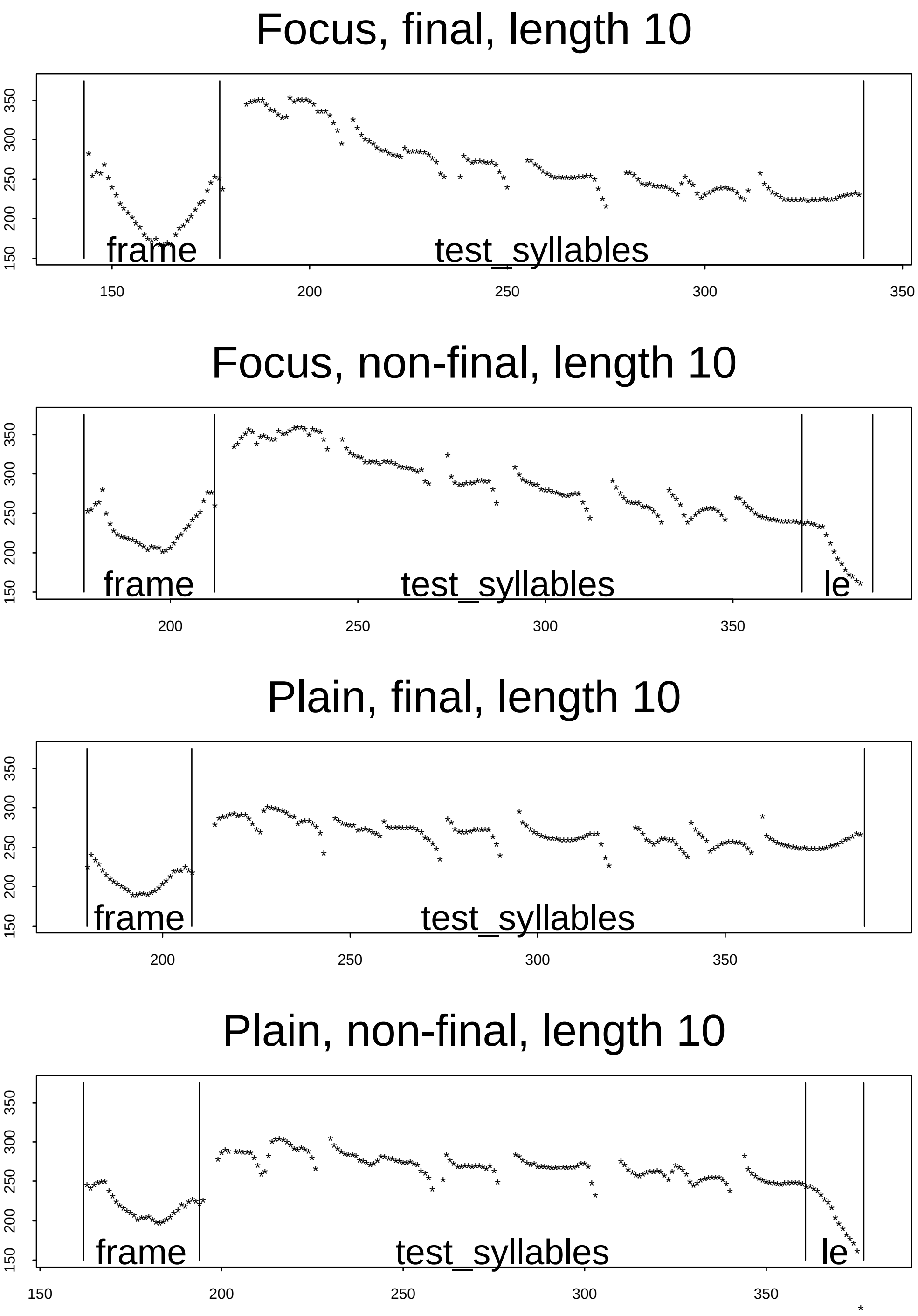
One disadvantage of using real, meaningful sentences is the complication of segmental effects, since different syllables must be used and measurements taken from various syllables may not be comparable. A supporting experiment was done to examine the effect of segmental perturbation to evaluate how comparable these syllables are to each other, and which regions are under the most heavy influence of segmental perturbation. If necessary, data from this study can be used to normalize segmental effects.

One of the speakers (Speaker B) read the sentences described in Table 2 in addition to the experiment sentences described earlier. The keywords were the test syllables used in the experiment *zheng1, dong1, gua1, gang1, zhong1, yi1, jin1, tian1, bang1* plus *ma*, the favorite syllable used in reiterant speech. Each sentence was repeated six times.

The F_0 contour of each keyword were sampled in 30 points. Figure 2 plots average trajectories of the rhyme regions of some of the keywords with the F_0 value in Hz plotted as a function of sample number. Each line in the plot represent the average of six tokens.

The two top panels compare syllables with similar rhymes. As the initial consonant effect subsides, the remaining pitch contours are very similar. Syllables with a low vowel [a] (top left panel) have lower F_0 values in the vowel region than the syllables with mid vowel [e] or [o] (top right panel). The similarity of the three trajectories in the top right panel suggests that there is no difference between front and back mid vowel [e] and [o]. The lower left panel compares *gua* with *gang*, The [u] region of *gua* is high due to the high vowel intrinsic effect, while the [a] regions of these two syllables have comparable F_0 values even though one occurs early in the rhyme and one late in the rhyme. The lower right panel compares the syllable *jin* and the syllable *tian*. The first 30% of the rhyme region of *tian* has higher F_0 than *jin*, due to the raising effect of [t]. The main vowel of *tian*, a fronted low vowel, has lower intrinsic F_0 than the main vowel [i] in *jin*. The intrinsic F_0 differences of *jin* and *tian*, which form the word “today”, become an issue that will be addressed in Section 4.2.

Unaspirated voiceless stops, represented as [b], [d], [g], have rather short effects. Aspirated stop [t], retroflex affricate [zh], and fricative [f] raise the pitch considerably and



Focus, final, length 10

Focus, non-final, length 10

Plain, final, length 10

Plain, non-final, length 10

Figure 1: Pitch tracks of one set of experiment sentences in the focus/plain, final/non-final conditions. The sentences shown have ten test syllables.

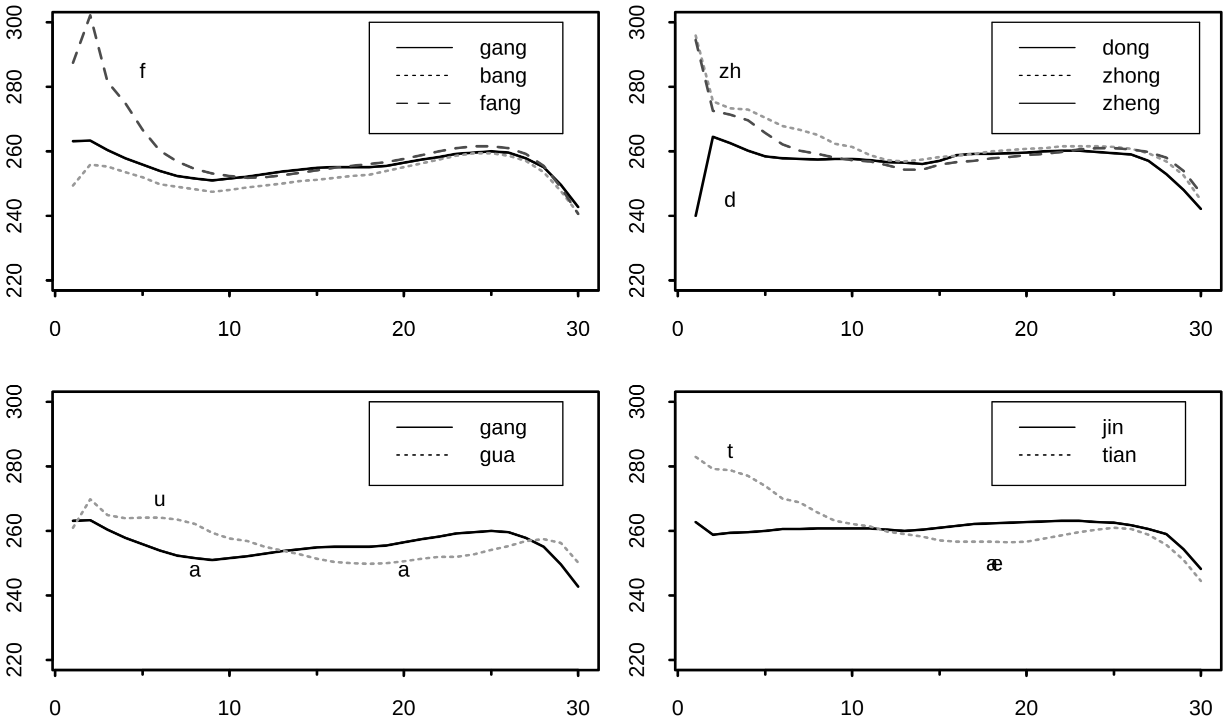


Figure 2: Segmental effects on high tone syllables.

the effects extend into the 10th sample, or the first 30% of the rhyme region. The intrinsic vowel effects were largely consistent with the previous finding, and were predictable from vowel height, with the low vowel [a] having the strongest lowering effect.

On the whole the result is encouraging in that at least the averaged consonantal effects appear to be manageable. The effects are predictable by phone classes and the strongest effects can be excluded from the measurement if the edges of the rhyme are avoided. The effects are surprisingly consistent, therefore they can be normalized if need be. For the purpose of this paper, we will avoid normalization and use instead the raw unnormalized data.

4 Results and discussions

ANOVA analysis was performed to test what factors have an effect on the scaling of F_0 values. Preliminary inspection of the data suggests that the location of word boundary and the location of the verb have some effects, so these two factors were coded in addition to the three factors in the experimental design. The five factors were used in the ANOVA analyses to model the difference in F_0 value between adjacent tone 1 syllables: sentence length (10 levels, from 2 to 11 tone 1 syllables), focus (2 levels, with or without narrow focus in the frame), final (2 levels, with or without final *le*), word (2 levels, whether the two syllables in question straddle a word boundary), verb (3 levels, the second syllable is a verb, the first syllable is a verb, or neither is a verb). The results are given in Table 3.

Speaker A, Female from Beijing

	Df	Sum of Sq	Mean Sq	F	Pr(F)
length	9	750.89	83.43	1.79	0.07
final	1	17.48	17.48	0.38	0.54
focus	1	10.62	10.62	0.23	0.63
word	1	113.96	113.96	2.45	0.12
verb	2	3633.28	1816.64	38.98	0.00
Resid.	865	40315.07	46.61		

Speaker B, Female from Taiwan

	Df	Sum of Sq	Mean Sq	F	Pr(F)
length	9	4408.15	489.79	4.83	0.00
final	1	489.67	489.67	4.83	0.03
focus	1	32564.18	32564.18	321.09	0.00
word	1	939.20	939.20	9.26	0.002
verb	2	4559.81	2279.91	22.48	0.00
Resid.	865	87727.46	101.42		

Speaker C, Male from Taiwan

	Df	Sum of Sq	Mean Sq	F	Pr(F)
length	9	908.18	100.91	1.96	0.04
final	1	55.25	55.25	1.07	0.30
focus	1	75.26	75.26	1.46	0.23
word	1	312.69	312.69	6.08	0.014
verb	2	9672.35	4836.18	94.06	0.00
Resid.	847	43551.47	51.42		

Speaker D, Male from Tianjin

	Df	Sum of Sq	Mean Sq	F	Pr(F)
leng	9	5654.57	628.29	9.35	0.00
final	1	62.57	62.57	0.93	0.34
focus	1	3299.44	3299.44	49.12	0.00
word	1	6932.94	6932.94	103.22	0.00
verb	2	4838.29	2419.15	36.02	0.00
Resid.	865	58098.01	67.17		

Table 3: Results of ANOVA analysis.

The ANOVA results show that sentence length, verb, and word have significant effect in at least three of the four speakers. Final condition has no effect, and focus has an effect on two of the speakers. These conditions are discussed in more detail below.

4.1 Final lowering

Figure 3 plots the averaged F_0 trajectory of the four speakers, including the the two frame syllables at position 1 (tone 3, L) and position 2 (tone 2, LH). The F_0 profiles in Figure 3 are obtained by averaging F_0 values of each speaker by position. The early positions represent more observations than later ones. A point at positions one to four represents 80 samples, while a point at position 13 represents only 8 samples. The zigzag pattern from positions 11 to 13 is caused by both the smaller sample size and the similar syntactic construction used in those three positions.

Tone 1 syllables (H) occur from position 3 and on. The final and non-final conditions are plotted separately: the final condition without *le* in solid lines and the non-final condition with *le* in dotted lines. The two populations match closely within each speaker. Not surprisingly, the difference between the final and non-final conditions are not significant in the ANOVA analysis. Since all data samples were included in the ANOVA analysis, and it is reasonable to assume that final lowering only affects the final section of an utterance, the lack of significance for final lowering in Table 3 could be a result of including sample points from other positions. However, when just the last tone 1 syllables from the final vs. non-final conditions are compared, the differences are still not significant for all four speakers. The results here show a lack of final lowering effect on Mandarin high tones under the assumption that final lowering has the strongest effect on the last syllable of an utterance, a reasonable assumption that is backed by plausible physiological causes (Maeda 1976; Strik and Boves 1992), and also in light of data from Kipare (Herman 1996) where final lowering was found as a separate effect from declination which affects up to three syllables from the end, and the effect increases in magnitude as the sentence approaches the end.

In the berry-list experiment of Liberman and Pierrehumbert (1984), the final lowering effect was estimated from the final stressed syllable of an utterance. Although they assumed that final lowering affects the final stretch of an utterance, the result was consistent with a hypothesis that final lowering affects the last *stressed* syllable in an utterance. This latter interpretation could in principle accounts for the lack of effect in the two conditions here: The added syllable *le* in the non-final condition is an unstressed particle, so the syllables being tested in the final and non-final conditions may be subject to the same amount of final lowering under this alternative hypothesis. The relevant test would then involve sentence pairs as in Table 4, where sentence (b) is formed by adding one stressed syllable to the end of sentence (a). Under this alternative hypothesis, the prediction is that the underlined syllable *gua* should have lower F_0 value in (a) than in (b). There are three sentence pairs like this in the experiments, but again no significant effect is found from any of the four speakers.

The conclusion here is that there is no final lowering effect on Mandarin high tones under the described experimental conditions. The final and non-final conditions are thus

a. Lao3-Wang2 zheng1 dong1 gua1
b. Lao3-Wang2 zheng1 dong1 gua1 zhong1

Table 4: Test sentences for the alternative hypothesis of final Lowering.

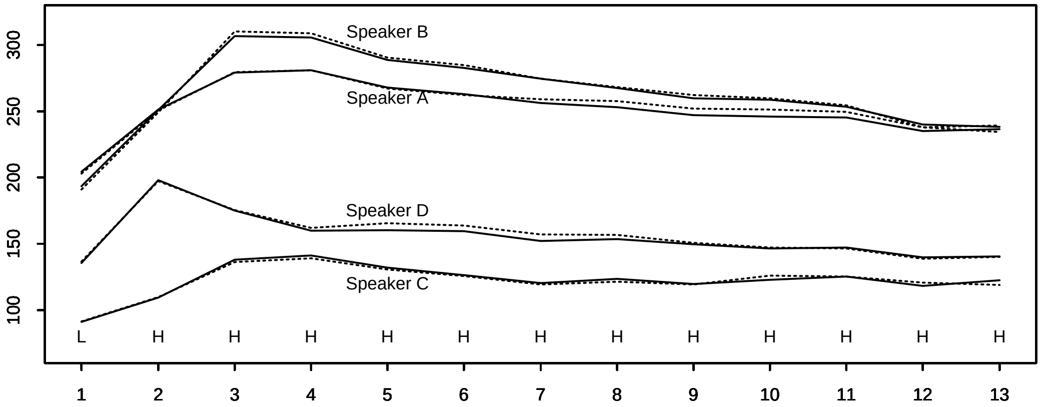


Figure 3: Averaged F_0 trajectories of sentences in the final (without *le*) and non-final (with *le*) conditions. The solid lines represent sentences in the final condition, the dotted line represent sentences in the non-final condition.

collapsed for subsequent analysis of the declination effect.

4.2 Declination

We now turn to the plain sentences to investigate the declination effect. In this section, we will show that there is a declination effect in Mandarin sentences, and propose a model to account for the observed data. The model has some useful properties that makes it attractive to a text-to-speech system: First, it is pitch range independent. The model gives reasonable prediction under changes of pitch ranges. Second, it handles arbitrarily long text gracefully, which is an important property for a text-to-speech system that needs to process unrestricted text input. Finally, as we will discuss in Section 4.3 and Section 4.4, the model successfully captures the variations in declination slopes resulting from the use of focus and changes in sentence length. We believe that variations resulting from discourse structure can be accommodated as well.

Figure 4 show the average trajectories of the plain sentences for four speakers. All four speakers show a decline in tone 1 values starting from the first or the second tone 1 syllable, or the third or fourth position in a sentence. The last values in position 13, although phonologically equivalent to those in position 3, lie around the mid point between the initial L and the highest H. The declination rate is faster in the beginning and slows down as the sentence progresses. This pattern of declination is reminiscent of the downstep equation proposed in Liberman and Pierrehumbert (1984), where each successive peak is modeled as a fraction of the distance of the previous peak and the reference line

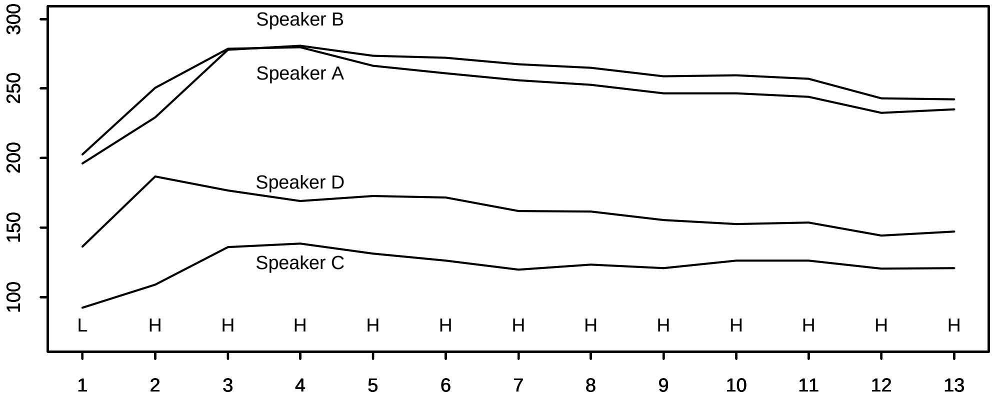


Figure 4: Averaged F_0 trajectories of sentences in the plain condition

Speaker	α	μ	Correlation (R^2)
A	0.90	21.77	0.83
B	0.91	20.82	0.73
C	0.75	29.29	0.51
D	0.81	27.63	0.62

Table 5: Fitted coefficient values obtained from the Plain sentences of the four speakers by Equation (1).

plus a constant.

An exponentially decaying declination model can handle long sentences more gracefully than a time constant model which deducts a fixed amount of F_0 value over the same amount of time. Such a model allows pitch to quickly drop below a level that is possible for the human speech apparatus. In particular, the magnitude of pitch decline we observe here is much larger than the 10Hz/second drop previously proposed (Pierrehumbert and Beckman 1988; Gooskens and van Heuven 1995). The average declination rate calculated from the plain sentences are 34 Hz/sec, 25 Hz/sec, 18 Hz/sec, and 17 Hz/sec for speakers A, B, C, D respectively.

The equation below is used to estimate the F_0 value of a given syllable P_i from the F_0 value of the preceding syllable P_{i-1}. The coefficients α and μ fitted for each speaker are given in Table 5, together with the correlation of the observed values and the predicted values.

$$P_i = \alpha P_{i-1} + \mu \tag{1}$$

Figure 5 plots the predicted values by the four models for sentences 50 syllables long, using the averaged values of the first tone 1 syllables of the plain sentences of each speaker

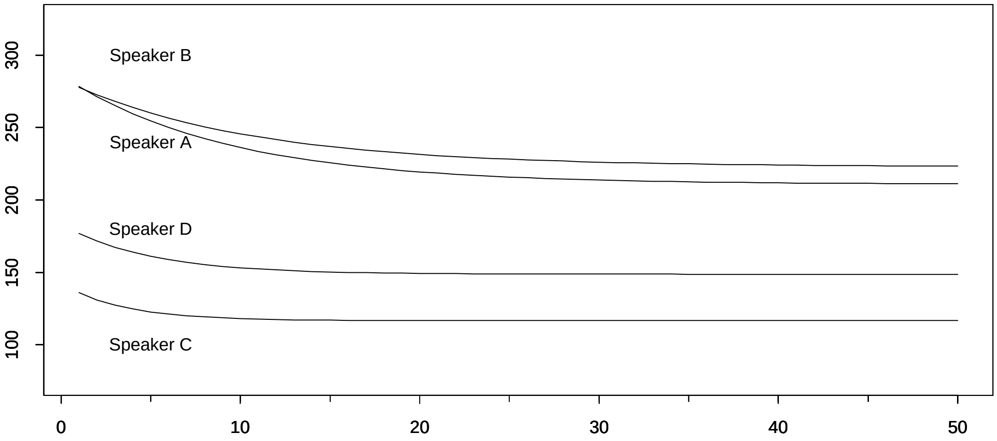


Figure 5: Declination slopes of the four speakers predicted by the declination models.

as the initial values P_1. F_0 values decline faster in the early section of an utterance, then asymptote to a value which is appropriate for the speaker' s pitch range.

The coefficient α in Table 5 controls how fast F_0 value approaches the asymptote, the smaller the value of α, the steeper the decline. The coefficient μ controls the asymptote value: when μ equals 0 the F_0 values asymptote to 0; when μ equals the difference of the initial value P_1 and $\alpha(P_1)$, the asymptote value is P_1, with zero declination effect. The coefficients α and μ are fitted optimally with the observed value of P_1. The model is not very forgiving when the value of P_1 changes. For example, if P_1 is lowered so that μ is bigger then $P_1 - \alpha(P_1)$, F_0 values will rise instead of decline. This is not a desirable property for a text-to-speech system where the value of P_1 needs to be independently manipulated to model discourse effect (Hirschberg and Pierrehumbert 1986), among other things.

A revised model in Equation (3) addresses this problem by explicitly incorporating P_1 in the equation, where μ is expressed as a fraction of $P_1 - \alpha(P_1)$. Given the P_1 value that Equation (1) is fitted for, the value μ in Table 5 can be straightforwardly transformed to the value β in Table 6 by Equation (2). The means of the initial tone 1 for each speaker are the P_1 values, which are included in Table 6. For the given P_1, the two models make the same prediction.

$$\beta = \mu/(P_1 - \alpha(P_1)) \tag{2}$$

$$P_i = \alpha P_{i-1} + \beta(P_1 - \alpha(P_1)) \tag{3}$$

Equation (3) gives more reasonable prediction than Equation (1) when the P_1 value changes. Minimally there will be no F_0 upstep when P_1 value decreases. Furthermore,

Speaker	α	β	P_1	Correlation (R^2)
A	0.90	0.76	278.46	0.83
B	0.91	0.80	277.86	0.73
C	0.75	0.86	135.88	0.51
D	0.81	0.84	176.78	0.62

Table 6: Fitted coefficient values for the four speakers by Equation (3), the revised declination model.

P_1, α and β are intuitively linked to the initial value, the rate of decline, and the asymptote value, three parameters that can be fitted or hand-tuned to generate variations in the declination slope that correspond to naturally occurring intonation variations, including the lack of declination, where β equals 1.

Everything being equal, increasing P_1 gives higher asymptote value, which equals βP_1, consistent with the impression that when speakers raise their pitch register, the entire pitch contour floats on top of a comparable sentence spoken in a lower pitch register. A lower α value gives a steeper decline where the asymptote value is reached earlier. A lower β value gives a lower asymptote value. Being able to control these parameters is very convenient for a Text-to-Speech system. In later sections it will be shown that the variations in sentence length can by modeled simply by varying P_1. Focus, to be discussed in Section 4.3, is a complex phenomenon requiring the combination of raised/expanded pitch range for the word in focus, as well as steeper decline and lower asymptote value in the post-focus material, which can be modeled by increasing P_1 value while lowering α and β values.

Before we turn to the subject of focus and sentence length, we will address one unexpected peculiarty that can be seen in Figure 4. Most declination theories predict that the effect starts from the beginning of the sentence. It is not entirely clear why the declination effect seems to start one syllable late for speakers A, B, and C. On average the second tone 1 syllable appears to be at least maintaining the height of the first tone 1 syllable. The proposed model predicts that the biggest F_0 decline in terms of Hz value should be from the first to the second tone 1 syllable.

There are several possible explanations for this phenomenon. First, it could be an artifact: The syllable *tian1* was used as the second tone 1 syllable in sentences with five or more tone 1 syllables (sentence length seven syllables and up). It was the only aspirated consonant used in the experiment sentences, which was shown in the supporting experiment to have a strong consonantal effect raising F_0 in the first 30% of the rhyme, see Figure 2. It is possible that there are some remaining effect at the location of F_0 measurement. This consonantal effect could be at least partially responsible for the lack of observable declination effect on the second tone 1 syllable, since the problem was most obvious in sentences with five or more tone 1 syllables, exactly where the syllable *tian1* was used. See Figure 7 in Section 4.4 where the sentence trajectories are plotted by length.

Second, what we observe here may be a general property of like tone sequence (especially high tones), that the intended pitch height is reached later than what is expected from the lexical specification. This phenomenon can be observed in Mandarin in tone 2–tone 1 (LH–H) combinations (Shih 1988; Xu 1997). Phonologically the end of tone 2 is the same as tone 1, but the end of tone 2 rarely reaches the high level when the following tone starts high. In the experiment, the second syllable (the frame) is a tone 2 followed by tone 1, and in three of the four speakers, the end of tone 2 is considerably lower than the following tone 1, see Figure 3. A more dramatic version of this effect is clearly documented in Yoruba high tone sequences (Laniran 1992) where the transition from a low tone to a string of high tones spans several syllables. In alternating high and low tone sequences, in Mandarin as well as Yoruba, no such delay is observed. It could be that the second tone 1 in this experiment represents the actual target of the high tone level while the first tone 1 is still in transition, and that the effect shows up only in long sentences because the slow transition is made possible only when the high tone sequence reaches a critical length.

Alternatively, there is a straight forward re-interpretation of this point from the Fujisaki model (Fujisaki et al. 1990; Wang et al. 1990). If the tone 1 syllables are sentence initial, the higher second syllable reflects the shape of the phrase command, which rises initially from the rest position after the impulse command becomes effective. If the tone 1 sequence follow a low or rising tone, as in our experiment sentences, the rising F_0 from the first to the second tone 1 is the result of the the rising accent command still effective after the low or rising tone.

Third, two of the three speakers with non-declining second tone 1 also show less F_0 drop within a word. Since in most of the test sentences the first and second tone 1 syllables form a word *jin1-tian1* “today”, it is possible that the effect is related to the within-word status.

A follow up experiment was done comparing 18 sentences: Two frames: one with two tone 1 syllables and one with the same frame as the main experiment (low and rising tones). Three lengths: 6, 8, and 10 syllables in the sentences. Three different syllables as the second tone 1, including one with *tian1*. Three different syntactic structures so that the first and second tone 1 form a word in two of the three sentences. One of the word conditions used *jin1 tian1* “today”, the same as most of the long test sentences in the main experiment. The 18 sentences were randomized and repeated five times each by one speaker. The results show no effect from the frame and the length variations. The syllable *tian1* were higher than the other two test syllables in the same sentence position, and the F_0 decline is less in the second tone 1 if it is in the same word as the previous syllable. The follow up experiment suggests that the first and third explanations above were supported, while the slow transition hypothesis is not supported.

4.3 Focus

Figure 6 plots the averaged F_0 trajectory of the plain sentences (solid lines) and the sentences with focus (dotted lines) for three of the speakers. Speaker A's trajectories in this case are similar to the ones shown in Figure 3, so her data are omitted from the plot. Again, the two frame syllables are included in the plot.

Speakers A and C did not use F_0 to differentiate focus vs. plain reading, and the two

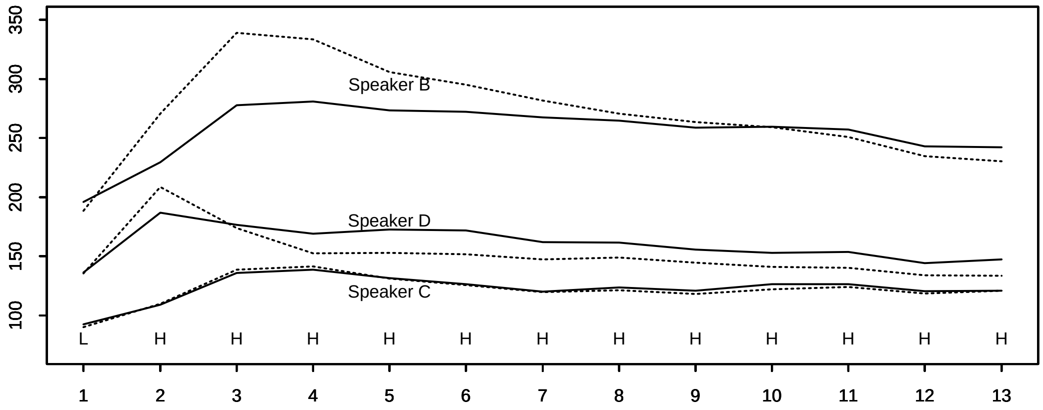


Figure 6: F_0 trajectories of sentences in plain (solid lines) and focus (dotted lines) conditions.

focus conditions of these two speakers have nearly identical F_0 trajectories. Speakers B and D both raised F_0 to signal prominence. Speaker D raised F_0 on the word with narrow focus, and the highest F_0 value was reached by the end of the rising tone *Wang2*, after that F_0 started to decline and in two syllables it reached a level lower than the plain sentence and remained low throughout the sentence. Speaker B showed a similar pattern except that the highest point was on the first tone 1 syllable, or the third syllable position of the sentence, which is one syllable *after* the narrow focus. It was not until the 10th position, eight syllables after the narrow focus, that the F_0 trajectory crossed over the plain sentence trajectory. After the crossover the pitch value remained below the plain sentence until the end of the sentence. This observation is compatible with the claim that narrow focus is realized with expanded pitch range, and after the focus the pitch range is compressed (Gårding 1987). However, one difference is in how soon the compression of pitch range occurs. Garding's claim is that the compression happens immediately after the narrow focus, which is true if there is a low tone following the focus (Shih 1988), or if there is a phrase break after the focus. But in the case of post-focus high tone sequences in the same phrase, the compression effect is gradual.

All four speakers lengthened the frame syllables *Lao3 Wang2* in the focus condition (Jin 1996). The durational difference between the focus and the plain populations is statistically significant in three speakers except Speaker A, suggesting that speakers B, C, and D successfully followed the cues and put emphasis on *Lao3 Wang2*. Speaker A lengthened *Lao3 Wang2* in some of the focus sentences, but didn't do so consistently. However, only two speakers, B and D, raised the pitch range in addition to lengthening in their rendition of focus. This observation is consistent with the report that F_0 cues does not have to be present for the perception of emphasis in Mandarin (Shen 1993).

The top panel in Table 7 compares the mean duration in msec for the syllable *Wang2* in the focus and the plain conditions. The duration of *Wang2* is longer in the focus condition than in the plain condition for all speakers. Also note that the duration of *Wang2* from speaker D is much longer than the other three speakers. This is because speaker D placed a short phrase break after the subject *Lao3 Wang2*, which led to considerable phrase-final lengthening.

Mean Duration of the syllable Wang				
Speaker	Focus	Plain	df	p value
A	178	176	878	p=0.0985
B	170	145	878	p¡0.05
C	167	146	860	p¡0.05
D	262	238	878	p¡0.05

Mean Duration of all test syllables				
Speaker	Focus	Plain	df	p value
A	174	173	878	p=0.7595
B	161	166	878	p=0.0084
C	166	168	860	p=0.618
D	166	170	878	p=0.0923

Table 7: Top panel:Mean duration (in msec) of the frame syllable Wang. Lower panel: Mean duration of all test syllables.

Speaker	Focus Condition	α	β	P_1	Correlation (R^2)
B	Plain	0.91	0.80	277.86	0.73
B	Focus	0.89	0.42	339.16	0.87
D	Plain	0.81	0.84	186.97	0.62
D	Focus	0.59	0.78	208	0.47

Table 8: Declination coefficients for the focus and plain Conditions

The average duration data from all test syllables is included in the lower panel to show that Speaker D's speaking rate is otherwise comparable to Speaker B and C. The phrase break is the most plausible cause of the lengthening of the syllable *Wang*, and of the difference in the post-focus declination pattern of Speakers B and D.

There is also a trend to shorten the post-focus materials (Jun and Lee 1998), shown by the shorter duration of the test syllables (lower panel) in the focus condition, in contrast to the plain condition. This effect is more robust in speakers B and D, the two speakers who happen to raise pitch range to signal focus. For Speaker B, who also keeps the post focus materials in the same phrase as the frame word in focus, the durational difference in focus and plain conditions is statistically significant.

For Speakers B and D, who consistently use expanded pitch range to convey focus, the post-focus declination slopes are different from the plain sentences. Models were fitted separately for the focus and plain conditions for these two speakers, and the results are summarized in Table 8 in the same format as Table 6.

For both speakers, the α and β values fitted for the focus condition are lower than those fitted for the plain condition, suggesting a steeper decline and a lower asymptote for the post-focus declination slope. The post focus $F0$ patterns reported here is consistent with other studies of Mandarin (Shih 1988; Liao 1994; Jin 1996; Xu and Wang 1997). Speaker D reaches asymptote much faster in the focus condition, and this is captured by his very small α value.

4.4 Sentence length

Sentence length has a clear effect on the scaling of Tone 1's. The question is what to change when sentence length changes. Figure 7 plots the sentence trajectories from all speakers. Sentences with different length are plotted separately in each plot. Plain sentences are plotted in the left column and focus sentences in the right column. Each line represents eight repetitions of the same sentence (final and non-final conditions combined).

In this plot, individual characteristics of some sentences show up clearly. We will discuss some effects such as word boundary effect and part of speech effect in the next section.

The initial tone 1's are higher in long sentences than in short sentences, as speakers took a deeper breath to prepare for the delivery of longer materials. Although there are ten variations in sentence length in the experiment, the speakers seem to aim at roughly three pitch range settings. This is reasonable, since we can hardly expect speakers to count syllables before they start reading. The four syllable sentences with two tone 1 syllables, the shortest in the experimental sentence set, were a notch lower than the rest. The medium length sentences started with medium pitch range, while long sentences had the highest pitch range in the initial portion of the sentences. Ths same phenomena was reported in (Thorsen 1980b). The low tones of the sentence initial tone 3's are slightly lower in short sentences, but the fluctuation is not as dramatic as in the high tone range.

Comparing long and short plain sentences, the short sentences start low and end high. To model this phenomenon the starting point of each sentence (P_1 of Equation (3)) should be assigned with reference to sentence length. The higher ending of the short sentences and the lower ending of the long sentences fall out from the model, where longer sentences accumulate more declination effect.

Given the same sentence delivered many times in a similar style, the most common variations we see in the data is a difference in the overall pitch range. When speakers start higher, they also end higher. This aspect of the natural variation in speech can be modeled by Equation (3) with different P_1 values. One example is given in Figure 8, which shows eight repetitions from Speaker A for the sentence *Lao3 Wang2 zheng1 dong1 gua1 zhong1*, “Lao3 Wang steams the winter melon pot”.

One interesting question regarding sentence length is whether longer sentences have more gradual declination slopes than short sentences as expressed by increasing values of α. We know that the slopes would be more gradual in longer sentence in our data if we follow Thorsen's model of Danish (1980b) by fitting a least squares regression line through all data points representing tone 1, since the data seems to suggest, and our model also predicts that most of the decline concentrates in the early section of an utterance, a well-known phenomenon that have already been accounted for in many different ways (Maeda 1976; Cooper and Sorensen 1981). There are two alternative ways to test this inquiry: fit the value of α by sentence length, and compare linear declination slopes by

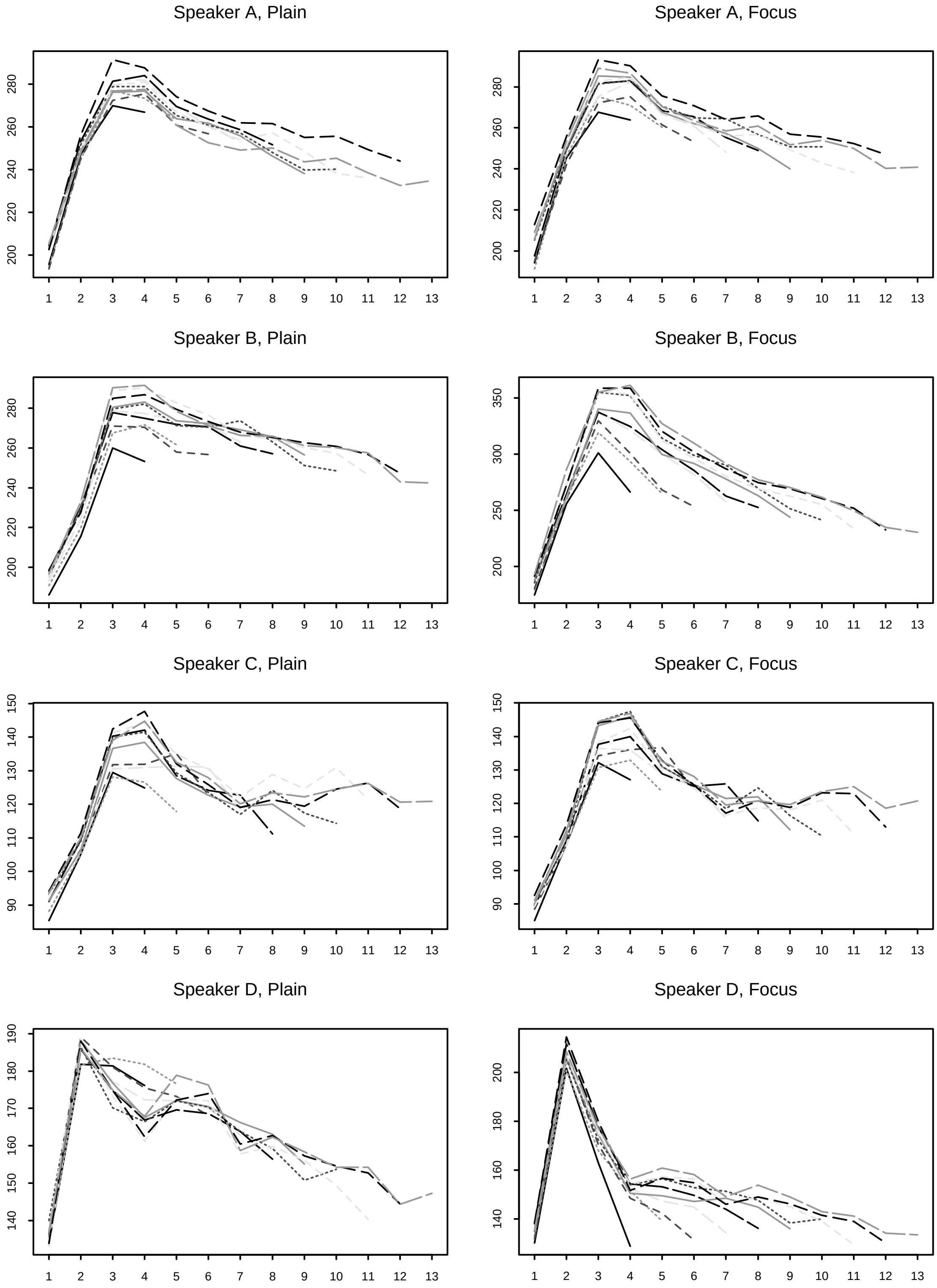


Figure 7: The effect of sentence length: F_0 trajectories plotted by focus condition and by sentence length.

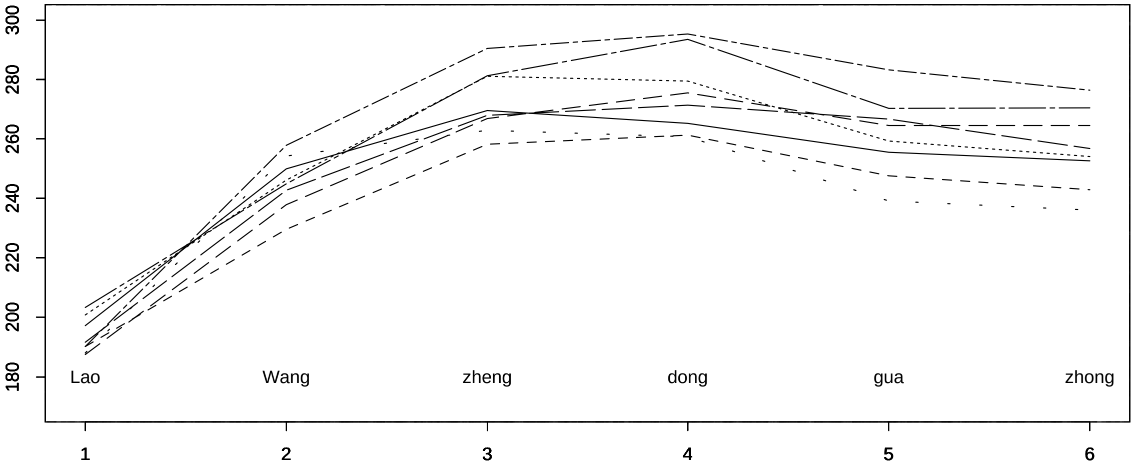


Figure 8: Pitch range variations in eight repetitions of a sentence

sentence length from the same sentence positions (say from the 4th to the 7th syllable). Both tests fail to yield a consistent picture that declination slope become more gradual as sentence length increases. It appears that P_1 value is the only parameter that varies with sentence length consistently, although as noted earlier, speakers vary initial pitch value according to a rough classification of sentence length.

Although there is a significant effect of sentence length on the F_0 value of the last syllable in plain sentences, as shown in the left column plots of Figure 7, the effect is weak in the focus condition for Speaker B and D, the second and the last plots in the right column of Figure 7. The ending F_0 values in these plots are more similar to each other than the corresponding plain sentence plots to the left. A plausible explanation is that in the focus condition the utterance reaches asymptote sooner than in plain sentences, as a result of a lower α value in the model, therefore the ending F_0 values of long sentences are similar.

4.5 Lexical and part of speech categories

Word boundary has a weak effect on F_0 scaling but the effects are not consistent across speakers: two speakers show more F_0 drop across word boundaries and two speakers show more F_0 drop within words. Also in the follow-up experiment (one speaker) where the effect of word boundary was specifically tested, it appears that the speaker attempted to maintain F_0 height within the word, and there is less F_0 drop within words than across words.

The effect of verbs is consistent across all four speakers: verbs have lower F_0 values than surrounding polysyllabic nouns, creating shallow F_0 valleys in those cases. Many of the zigzag patterns in Figure 7 correspond to pitch lowering on verbs. The most plausible explanation of this phenomenon is that verbs are metrically weaker than object nouns in Mandarin. All verbs used in this experiment are monosyllables, while most of the selected nouns are disyllabic. This asymmetric distribution mirrors the natural distribution in the language: most of the colloquial verbs are indeed monosyllabic while nouns in Mandarin have been going through a disyllabification process since Middle Chinese. It is tempting to link both the monosyllabicity of verbs and their lower F_0 values to a weaker prosodic weight than nouns.

No F_0 valleys are observed in cases of monosyllabic verbs followed by monosyllabic nouns. In those cases the verb and noun construction may have been incorporated into one single prosodic word and is for all practical purposes being treated as a single word.

The verb effect can be captured in F_0 modeling by assigning a low default prominence setting to a monosyllabic verb which is followed by a polysyllabic object noun.

Furthermore, we note that the interesting humps in some sentences at positions 4 and 5 of Speaker D are caused by his occasionally emphasis placement on the adverb *gang1-gang1*, “just now”. Speaker C occasionally destressed the final syllable *gua* “melon”, changing its tonal category to neutral tone, which caused some of the final syllable to drop in pitch. This is a lexical issue, unrelated to final lowering.

5 Conclusion

In this paper we have shown that Mandarin has a clear declination effect. The F_0 decline is more pronounced near the beginning of the utterance, and the effect can be modeled as an exponential decay. The model has three parameters: P_1, the initial value; α, the parameter controlling the rate of declination; and β, which controls the asymptote value. The variations in long and short sentences are primarily controlled by varying the P_1 value, where longer sentences have higher initial pitch and shorter sentences have lower initial pitch. Declination also interacts with focus. In the data of one speaker, post-focus pitch range compression of high tone sequence appears to be gradual. Declination slope is steeper in post-focus materials than in plain sentences, which is reflected in a lower α value in the post-focus declination slope. The F_0 values eventually asymptote to a lower value than in plain sentences, which is expressed by the lower value of β.

6 Acknowledgments

I would like to thank Yi Xu and Sun-Ah Jun for their insightful comments on this paper.

REFERENCES

Cohen, A., R. Collier, and J. 't Hart. 1982. Declination: Construct or Intrinsic Feature of Speech Pitch? *Phonetica*, Vol. 39, pp. 254–273.

Cooper, W. E. and J. M. Sorensen. 1981. *Fundamental Frequency in Sentence Production*. Springer, Berlin.

Eady, S. J. and W. E. Cooper. 1986. Speech Intonation and Focus Location in Matched Statements and Questions. *Journal of the Acoustical Society of America*, Vol. 80, pp. 402–415.

Fujisaki, H., K. Hirose, P. Halle, and H. Lei. 1990. Analysis and Modeling of Tonal Features in Polysyllabic Words and Sentences of the Standard Chinese. In *ICSLP*, (Kobe, Japan), pp. 841–844.

Fujisaki, H. 1983. Dynamic Characteristics of Voice Fundamental Frequency in Speech and Singing. In *The Production of Speech* (MacNeilage, P. F., editor), pp. 39–55, Springer-Verlag.

Gårding, E. 1987. Speech Act and Tonal Pattern in Standard Chinese: Constancy and Variation. *Phonetica*, Vol. 44, pp. 13–29.

Gooskens, C. and V. J. van Heuven. 1995. Declination in Dutch and Danish: Global Versus Local Pitch Movements in the Perceptual Characterisation of Sentence Types. In *Proceedings of ICPhS 95*, Vol. 2, (Stockholm), pp. 374–377.

Gussenhoven, C. and A. C. M. Rietveld. 1988. Fundamental Frequency Declination in Dutch: Testing Three Hypotheses. *Journal of Phonetics*, Vol. 16, pp. 355–369.

Herman, R. 1996. Final Lowering in Kipare. *Phonology*, Vol. 13, No. 2, pp. 171–196.

Hirschberg, J. and J. Pierrehumbert. 1986. The Intonational Structuring of Discourse. In *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics*, Vol. 24, pp. 136–144.

Jin, S. 1996. *An Acoustic Study of Sentence Stress in Mandarin Chinese*. PhD thesis, Ohio State University.

Jun, S.-A. and H.-J. Lee. 1998. Phonetic and Phonological Markers of Contrastive Focus in Korean. In *Proceedings of the International Conference on Spoken Language Processing*, (Sydney, Australia).

Ladd, D. R. 1984. Declination: A Review and some Hypotheses. *Phonology Yearbook*, Vol. 1, pp. 53–74.

Ladd, D. R. 1993. On the Theoretical Status of ”The Baseline” in Modelling Intonation. *Language and Speech*, Vol. 36, No. 4, pp. 435–451.

Laniran, Y. O. 1992. *Intonation in Tone Languages: The Phonetic Implementation of Tones in Yorubá*. PhD thesis, Cornell University.

Lea, W. 1973. Segmental and Suprasegmental Influences on Fundamental Frequency Contours. In *Consonant Types and Tones* (Hyman, L., editor), pp. 15–70, Los Angeles: University of Southern California.

Liao, R. 1994. *Pitch Contour Formation in Mandarin Chinese*. PhD thesis, Ohio State University.

Liberman, M. Y. and J. B. Pierrehumbert. 1984. Intonational Invariance under Changes in Pitch Range and Length. In *Language Sound Structure* (Aronoff, M. and R. Oehrle, editors), pp. 157–233, Cambridge, Massachusetts: M.I.T. Press.

Lieberman, P. 1967. *Intonation, Perception, and Language*. MIT Press, Cambridge, Massachusetts.

Maeda, S. 1976. *A Characterization of American English Intonation*. PhD thesis, MIT.

Nakajima, S. and J. F. Allen. 1993. A Study on Prosody and Discourse Structure in Cooperative Dialogues. *Phonetica*, Vol. 50, pp. 197–120.

Peterson, G. E. and H. L. Barney. 1952. Control Methods Used in a Study of the Vowels. *Journal of the Acoustical Society of America*, Vol. 24, No. 2, pp. 175–184.

Pierrehumbert, J. and M. Beckman. 1988. *Japanese Tone Structure*. The MIT Press, Cambridge, Massachussetts.

Pierrehumbert, J. 1979. The Perception of Fundamental Frequency Declination. *Journal of the Acoustical Society of America*, Vol. 66, No. 2, pp. 363–369.

Pierrehumbert, J. 1980. *The Phonology and Phonetics of English Intonation*. PhD thesis, MIT.

Prieto, P., C. Shih, and H. Nibert. 1996. Pitch Downtrend in Spanish. *Journal of Phonetics,* Vol. 24, No. 4, pp. 445–473.

Shen, J. 1985. 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)*, pp. 73–130, Beijing University Press.

Shen, X. S. 1993. Relative Duration as a Perceptual Cue to Stress in Mandarin. *Language and Speech*, Vol. 36, No. 4, pp. 415–433.

Shih, C. 1988. Tone and Intonation in Mandarin. In *Working Papers of the Cornell Phonetics Laboratory, Number 3: Stress, Tone and Intonation*, pp. 83–109, Cornell University.

Silverman, K. E. 1987. *The Structure and Processing of Fundamental Frequency Contours*. PhD thesis, University of Cambridge.

Sluijter, A. M. C. and J. M. B. Terken. 1993. Beyond Sentence Prosody: Paragraph Intonation in Dutch. *Phonetica*, Vol. 50, pp. 180–188.

Strik, H. and L. Boves. 1992. Control of Fundamental Frequency, Intensity and Voice Quality in Speech. *Journal of Phonetics*, Vol. 20, No. 1, pp. 15–25.

Strik, H. and L. Boves. 1995. Downtrend in F0 and Psb. *Journal of Phonetics*, Vol. 23, pp. 203–220.

Terken, J. 1991. Fundamental Frequency and Perceived Prominence of Accented Syllables. *JASA*, Vol. 89, No. 4, pp. 1768–1776.

Terken, J. 1993. Baselines Revisited: Reply to Ladd. *Language and Speech*, Vol. 36, No. 4, pp. 453–459.

t' Hart, J. and A. Cohen. 1973. Intonation by Rule: A Perceptual Quest. *Journal of Phonetics*, Vol. 1, pp. 309–327.

Thorsen, N. 1980. A Study of the Perception of Sentence Intonation – Evidence from Danish. *Journal of the Acoustical Society of America*, Vol. 67, pp. 1014–1030.

Thorsen, N. 1980. Intonation Contours and Stress Group Patterns in Declarative Sentences of Varying Length in ASC Danish. In *Annual Report of the Institute of Phonetics, University of Copenhagen*, Vol. 14, pp. 1–29.

Titze, I. R. 1989. On the Relation between Subglottal Pressure and Fundamental Frequency in Phonation. *JASA*, Vol. 85, No. 2, pp. 901–906.

Tseng, C.-Y. 1981. *An Acoustic Phonetic Study on Tones in Mandarin Chinese*. PhD thesis, Brown University.

Umeda, N. 1982. F_0 Declination is Situation Dependent. *Journal of Phonetics*, Vol. 10, pp. 198–210.

Wang, C., H. Fujisaki, and K. Hirose. 1990. Analysis and Modeling of Tonal Features in Polysyllabic Words and Sentences of the Standard Chinese. In *ICSLP*, (Kobe, Japan), pp. 221–224.

Xu, Y. and Q. E. Wang. 1997. What Can Tone Studies Tell Us about Intonation? In *Intonation: Theory, Models and Applications* (Botinis, A., G. Kouroupetroglou, and C. G., editors), (Athens, Greece), pp. 337–340, ESCA.

Xu, Y. 1997. Contextual Tonal Variations in Mandarin. *Journal of Phonetics*, Vol. 25, pp. 61–83.

Yang, L.-C. 1995. *Intonational Structures of Mandarin Discourse*. PhD thesis, Georgetown University.