Computational Linguistics and Chinese Language Processing
Vol. 18, No. 3, September 2013, pp. 21-44

Understanding Mandarin Prosody: Tonal and Contextual Variations in Spontaneous Conversation

Li-chiung Yang[*], and Richard Esposito[+]

Abstract

Tonal identity and tonal variation in Mandarin have been the focus of intensive research that has long sought to bring out the underlying causes of variations in realized pitch values. Included among the variables studied in tonal variation are syntactic, contextual, emotional, and interactional influences. In the current study, we present results of our comparative research into tonal pitch variation in read speech and spontaneous Mandarin conversations. We acoustically and quantitatively characterize differences in the degree of pitch variability of these two modes of speech, and we present our results on tonal variability, as well as the influence of tone sequencing, syllable amplitude, and contextual factors on realized tonal shape. We show that, although tones are manifested in great diversity of pitch in spontaneous speech, there is a consistency of pitch shape that is dependent on tonal lexical identity.

Keywords: Tone, Prosody, Mandarin, Tonal Variability, Spontaneous Speech

1. Introduction

Previous research on read speech and spontaneous speech in Mandarin has demonstrated the wider variability of pitch movement in the latter, and researchers have attributed the greater variability to a number of different factors. Research on read speech (Xu, 1997; Shih, 1992; Shen, 1990) has predominantly focused on production and perception of tones in read or experimentally elicited speech. In these studies, prominent results have been the elucidation of rules for tone sandhi and the modification of tonal shapes under differing aspects, such as questioning, focus, or emphasis.

[*] English Language Center, College of Arts, Tunghai University, Taichung Taiwan
TEL: 011-886-04-2359-0121x31923 FAX:011-886-04-2359-0232
E-mail: yang_lc@thu.edu.tw
The author for correspondence is Li-chiung Yang.
[+] Spoken Language Research, USA
E-mail: esposito_r@sprynet.com

Recent research has demonstrated the wide divergence between defined tonal pitch values that occur in spontaneous speech (Tseng, 2005) and values in read speech. The interactive and cognitively intense environment of spontaneous speech provides an abundance of differing factors that often lead to tonal sequences with tones that seemingly rarely reach their defined values.

Researchers have suggested several avenues to explain systematic divergences from intrinsic tone shape, including tone sequence patterns, targeting of adjacent tones, and stress and metrical patterns of speech. Prior research on speech prosody in both tone and non-tonal languages (Hirst *et al.*, 1998; Shriberg *et al.*, 2000; Tseng, 2005; 2009; Tseng, 2010) has shown that there are a number of important influences on the prosody of natural speech, including interactive monitoring activity, emotional state, and the level of uncertainty, as well as topic organization and phrasal position.

Our view is that experimental and spontaneous speech corpora are complementary and each has its important role to play in the discovery process. Experimental and read speech data are ideally suited to testing hypotheses on relationships among known variables while controlling for confounding effects of other variables. Study of spontaneous speech, on the other hand, is ideally suited to the discovery of new contributing variables and to the forming of new hypotheses. In addition, the study of spontaneous speech, as in other natural science fields, has the potential to yield valid information on underlying processes that are uncovered through the identification of systematic parallels observed across the available data. This discovery of relevant variables is especially important when a particular phenomenon could arise from more than one underlying cause or from an alternate cause. By combining the results of read or experimental studies with those of spontaneous speech, the relative robustness of experimental results in a spontaneous setting can be used as a marker or clue to the existence of other potentially important variables.

With this view in focus, in the current paper, we study how tonal sequences and amplitude affect the realization of lexical tone shape and compare results obtained from a read speech corpus to results found in spontaneous Mandarin. We further investigate differences between the speech modes, and study the roles of tone sequence, speaker variability, and lexical identity in shaping realized tone values.

Section 2 describes our methodology, including the nature and extent of the data corpora, and the automatic and post-processing steps taken to extract acoustic parameters from the speech signal.

In Section 3, we introduce an innovative technique to facilitate consistent comparison of measures of Chinese tonal f_0 contour in large speech corpora. We describe the importance of spontaneous speech to realized tonal variability and also present a benchmark comparison of spontaneous to read f_0 shape. We explore several important factors in the determination of f_0 shape in spontaneous speech, including tonal identity, word syllable number and position, and tonal sequence patterns, including tone sandhi. We present results of our comparisons grouped by lexical tone, by tonal sequence patterns, and by mono- or di-syllabic words, and we indicate the degree to which our results match prior theories on anticipatory and carryover effects of tone sequences. Section 3.3 introduces the data reduction techniques used to extract comparable measures of f_0 shape and provides graphical representations of the distributions of these measures as important guides to the behavior of syllable f_0 in natural spontaneous conversations. In Section 3.4, we show individual instances of tonal variation in spontaneous speech and suggest mechanisms for the specific variational tendencies revealed by the data.

Section 4 provides a summarization of the key findings of the paper, the importance of tone variability in spontaneous Mandarin, and how future work on tonal variability can be enhanced through the techniques introduced.

2. Data and Methodology

2.1 Data, Participants, and Approach

The data utilized in this study are part of a larger project on Mandarin conversational speech, totaling over 20 hours of speech. For this study, a subset of continuous speech from two conversations, one between two female speakers (Speaker P and Speaker S) and the other between one female (Speaker T) and one male speaker (Speaker B), totaling approximately 40 minutes in duration for spontaneous speech were selected and analyzed. In addition, to provide a baseline comparison between the read and spontaneous speech modes of the same speaker, 10 minutes of read speech, consisting of four read stories, by the same male speaker in the spontaneous conversation, were also analyzed. As previous research on Mandarin has concentrated almost exclusively on read or experimentally controlled speech (with the exception of Tseng, 2004, 2005, 2009), in this study, our goal is to concentrate on exploring tonal and prosodic variations in spontaneous Mandarin Chinese.

The spontaneous conversation data were collected in informal settings, and the read speech corpus was collected in a laboratory. Speech data were recorded using a SONY PCM-M1 DAT recorder with a SONY ECM lavalier microphone at a sampling rate of 22,050 kHz. Data were manually segmented to the phrase, word, and syllable levels using Wavesurfer and ESPS/xwaves for their ease with extended speech data processing capability, and acoustic-prosodic features such as time, amplitude, and pitch (f_0) values were automatically extracted from the speech files using ESPS/xwaves function *get_f0*. F_0 values were further corrected for formant outliers (*e.g.* doubling and halving), and slight errors in segmentation among the different label files were adjusted automatically.

A total of 7,710 lexical syllables from the 40 continuous minutes of segmented spontaneous speech were obtained, after eliminating overlapping syllables for which f_0 values were ambiguous between speakers. For read speech, there was one speaker, and a total of 1,804 syllables of speech. Tables 1-3 show the breakdown by speech mode, corpus, speaker, and tone. Altogether, the corpora investigated contained a total of 9,514 syllables over 50 minutes of speech.

In this study, we investigate the f_0 contours of tones as they are realized in spontaneous speech by utilizing several measures of f_0 contour to facilitate data reduction and comparisons across a large corpus of syllables. Our approach is to examine the data first, without any preconceived assumptions about specific tonal variations, show the patterns that emerge from this large corpus, and then relate our results to previous claims and findings.

The results between spontaneous and read speech are based on spontaneous and read speech corpora from one male speaker. While the sample size of the read speech is small, use of the same speaker highlights differences in syllable f_0 contour while controlling for speaker variability. Focusing on finding metrics to evaluate the variations in f_0 contour in spontaneous speech with respect to defined tonal shape, we present overall results on the consistency of tonal change across speakers in the wider spontaneous corpora.

Table 1. Number of syllables by tone, spontaneous conversation MC1, 2 female speakers, Speaker P and Speaker S.

	Tone 1	Tone 2	Tone 3	Tone 4	Tone 0	Total
P	439	338	455	902	357	2491
S	270	222	297	579	239	1607
Total	709	560	752	1481	596	4098

Table 2. Number of syllables by tone, spontaneous conversation MC2, 1 male speaker, Speaker B and 1 female speaker, Speaker T.

	Tone 1	Tone 2	Tone 3	Tone 4	Tone 0	Total
B	404	353	485	807	340	2389
T	151	182	261	449	180	1223
Total	555	535	746	1256	520	3612

Table 3. Number of syllables by tone, read speech, RS1, male speaker, Speaker B.

	Tone 1	Tone 2	Tone 3	Tone 4	Tone 0	Total
B	356	300	318	502	328	1804

3. Results: Variability of Tonal Shapes

The defined lexical tones in Mandarin are commonly recognized as Tone 1, which has a high pitch level and is flat (55); Tone 2, which starts at a mid to low pitch level, and rises (35); Tone 3, which starts at a mid to high level, falls, then rises (214); Tone 4, which is high and falling (51); and a neutral tone, Tone 0 (Chao, 1968). Tone 3 has two recognized variants in spontaneous speech: a high to low fall with no final rise and a short duration, low pitch value (21). Tone sequence pitch values are also governed by generally accepted tone sandhi rules (Chao, 1968): a Tone 3 that is followed by a Tone 3 takes on a rising pitch (33-->23); Tone 3 when followed by a non-3rd tone takes on a half-Tone 3, without an ending rise; and a Tone 4 that is followed by Tone 4 takes on a less steep fall.

As a simple first measure of relative concordance of syllable shapes with the defined pitch movement for tones, we used the linear slope of f_0 as a simple shape indicator to approximate f_0 values. Syllable f_0 values were extracted automatically and corrected as described in Section 2. The resulting f_0 values of each syllable were fitted using S-plus, and linear slopes and intercepts of each syllable were calculated. A linear slope would be most applicable for Tones 1 and 4, and, to a large degree, rising Tone 2. A quadratic fit would better capture the curvature of all tones, especially Tones 2 and 3, and was also produced for each syllable and used in measuring the effects of amplitude.

3.1 Tonal Values in Read and Spontaneous speech

Read, controlled, or experimental speech data are frequently considered as benchmarks that preserve a number of relatively stable phonological relationships in their realization. Table 4 shows the overall measure of tonal shape for our read speech data, using averages of linear approximations to each syllable's f_0 data, restricted to monosyllables occurring in the pre-pause or at phrase end position, to avoid the influence of adjacent tonal values. This is likely to have some downward bias due to pitch declination at phrase end, and this may cause the slight negative slope of level Tone 1. As expected, Tone 4 has a large falling slope, while Tones 2 and 3 have an overall positive rise. The f_0 minimum values shown in Table 4 indicate the percentage point within the syllable that attains the syllable minimum, and they provide additional information on syllable shape, as they mark the location of syllable f_0 slope direction change. The f_0 minimum point is very informative on the shape as well and is more robust with respect to averaging over different speakers and different speech situations, as this point is defined in percentage terms. In read speech, Tones 1 and 4 reach their minimum pitch point relatively close to the end, while Tones 2 and 3 reach their minimum pitch point very close to the midpoint of the syllable, matching the defined fall-rise shape of Tone 3, and rising Tone 2 with an initial fall.

Table 4. Mean slope in Hz change per second and percentage point of syllable f_0 minimum of each token, pre-pausal monosyllables, read speech, Speaker B.

	Tone 1	Tone 2	Tone 3	Tone 4	Tone 0
Slope	-39.73	36.42	58.98	-212.94	-75.49
f_0 min	0.64	0.48	0.48	0.72	0.55

Table 5 shows the parallel results for this same speaker as he engaged in spontaneous conversation. It is immediately clear that tone values in spontaneous speech diverge widely from their lexically defined values and from their realized values in read speech. In particular, Tone 1 exhibits a greater average fall than expected and Tone 4 displays an average fall that is just barely greater than defined level Tone 1. Notably, lexically rising Tone 2, which on average rises in read speech, also has an overall negative slope in spontaneous speech. Of the four tones, only Tone 3 has an overall rise that is similar to its read speech counterpart. Both Tables 4 and 5 are restricted to monosyllables in pre-pausal position, so the average pitch slopes for Tone 3 are independent of tone sandhi rules.

The f_0 minimum points for Tones 2, 3, and 4 occur earlier than in read speech and occur at nearly the same percentage position for Tone 1. For this speaker, Tones 1 and 2 in spontaneous speech became more negatively sloped, while Tone 3 remained similar to read speech. The most striking change occurs with Tone 4, which falls much less in spontaneous speech than in read speech and has about the same average slope as spontaneous Tone 1. Neutral tone (Tone 0) for this speaker becomes significantly more neutral, that is, *flatter*, in spontaneous speech, with a slope near zero, indicating very little pitch change.

Table 5. Mean slope in Hz change per second and percentage point of syllable f0 minimum of each tone, pre-pausal monosyllables, spontaneous speech, Speaker B.

	Tone 1	Tone 2	Tone 3	Tone 4	Tone 0
Slope	-62.31	-11.86	54.54	-64.97	3.91
f_0 min	0.64	0.41	0.34	0.54	0.38

The overall tone slope results, including mono-, di-, and tri-syllables, for the two participants in the current spontaneous corpus are presented in Table 6.

Table 6 indicates that the slope directions agree for each of the four tones across the two speakers, although the *strength* of directional changes varies, and the tones vary substantially from their defined lexical values. For both speakers, there is a striking similarity in average pitch slope for Tones 1 and 4, but there is also a consistent difference in degree between the speakers. For Speaker B, Tones 1 and 4 fall by about the same amount. For Speaker T, Tone 1 actually falls more than defined falling Tone 4, on average. Relative to each speaker’s pattern for all tones, Table 6 shows that Speaker T’s Tone 2 syllables fall relatively more, highlighting the importance of speaker variation in the realized tone shapes of spontaneous speech.

Table 7 further breaks out the results for the two speakers in spontaneous conversation MC2 by whether the syllable is a monosyllabic word (mono), the 1st syllable in a disyllabic word (D1), or the 2nd syllable in a disyllabic word.

Table 6. Mean slope in Hz change per second over all syllables, spontaneous speech, Speaker B and Speaker T.

	Tone 1	Tone 2	Tone 3	Tone 4	Tone 0
B	-79.84	-15.72	21.47	-77.64	-104.38
T	-188.39	-104.14	7.87	-138.69	1.16

Table 7. Mean slope in Hz change per second over monosyllables, 1st and 2nd syllables of disyllabic words, spontaneous speech, Speaker B and Speaker T.

Speaker	Tone 1	Tone 2	Tone 3	Tone 4	Tone 0
Mono-B	-62.31	-11.86	54.54	-64.97	3.91
Mono-T	-192.31	-124.76	14.43	-150.49	-127.10
D1-B	-90.49	-30.22	-23.42	-105.59	-
D1-T	-189.30	-79.72	-10.39	-227.56	-
D2-B	-52.57	7.24	-16.64	-69.99	-17.70
D2-T	-118.26	-128.95	1.86	-45.33	-72.46

The most striking pattern seen in Table 7 is the evident pervasive strength of a falling pitch, with almost all values showing an average negative slope. Table 7 clearly indicates that this pattern is similar for both speakers. As the two participants differ in overall pitch level, in Table 8 we show the same data normalized to Z-score values with respect to each speaker's overall average pitch mean and standard deviation, calculated across all f_0 values, by speaker.

Table 8. Mean slope in Hz change per second over monosyllables, 1st and 2nd syllables of disyllabic words, normalized to each speaker's average syllable pitch range, spontaneous speech, Speaker B and Speaker T.

Speaker	Tone 1	Tone 2	Tone 3	Tone 4	Neutral
Mono-B	-2.13	-0.41	1.87	-2.22	0.13
Mono-T	-4.36	-2.83	0.33	-3.41	-2.88
D1-B	-3.09	-1.03	-0.80	-3.61	-
D1-T	-4.29	-1.81	-0.24	-5.16	-
D2-B	-1.80	0.25	-0.57	-2.39	-0.61
D2-T	-2.68	-2.92	0.04	-1.03	-2.68

From Table 8, calculated from the spontaneous speech of corpus MC2, we can see that Tones 4 and 1 have the largest negative slope, but the mean values for Tone 4 do not have a consistently greater negative slope than Tone 1 for both speakers. For Speaker B, Tone 4 is marginally more falling than Tone 1, but for Speaker T, Tone 1 falls marginally more than Tone 4 for monosyllables and for the 2nd syllable of a disyllabic word.

The normalized data of Table 8 indicate that Speaker T has a greater propensity for a falling pitch over all tones except Tone 3, and for Tone 4 when it is the 2nd syllable of a disyllabic word; such differences in tonal modification may form a component of a speaker's characteristic or general speech style.

The data from Tables 7 and 8 show that, except for Tone 4, the slope of syllables in spontaneous speech exhibit a divergence from the lexically defined shape, on average. As these results are consistent across syllable types, Table 8 further suggests that the results of divergence from lexically defined tonal values may not be due solely to a pattern that affects only monosyllables or that is effected through syllable position in the word, and may arise from other causes as well.

3.2 Tonal Sequencing and the Influence of Preceding and Following Lexical Tone

Researchers on Mandarin tonal pitch values have proposed that local tone sequences can theoretically affect the realized target in two primary ways: through anticipatory effects of the upcoming syllable or through carryover effects of the preceding syllable. A succeeding tone value with a high f_0 onset should lead to a higher offset for the previous syllable, while a succeeding syllable with a low onset should induce a low offset in the previous syllable, according to the anticipatory theory. Analogously, carryover theory predicts that a high f_0 offset in a preceding syllable should lead to a higher onset for the succeeding syllable, while a preceding syllable with a low offset (half-Tone 3 and Tone 4) should induce a low onset in the succeeding syllable.

For example, Xu (1997) found greater evidence for the strength of carryover effects than for anticipatory effects using balanced sequences in experimental speech for Mandarin, while Chang & Hsieh (2012) found a more balanced effect of carryover and anticipatory effects for the more complex tonal system of Eng Choon Hokkien in their experimental data. In the current study, we were interested in investigating if similar effects on the realized target tones hold in our spontaneous speech data.

In Tables 9 through 12, we compare the average slope of each lexical syllable of two speakers from spontaneous conversation MC1, grouped by immediately preceding and succeeding syllable tone, using the linear regression slopes for the two speakers. The resulting slope coefficients were grouped by the subsequent lexical tone in Tables 9 and 10, and by the preceding tone for Tables 11 and 12.

Table 9. Mean slope in Hz change per second of each tone, by tone of the following syllable, Speaker P.

Slope of	X-Tone 1	X-Tone 2	X-Tone3	X-Tone 4	X-Tone 0
Tone 1	-78.4	-76.2	-125.3	-35.6	-119.5
Tone 2	-66.1	24.7	-115.8	-16.0	-53.2
Tone 3	-185.4	-76.7	-0.4	-149.5	-97.8
Tone 4	-133.6	-254.7	-248.8	-187.1	-202.9

Table 10. Mean slope in Hz change per second of each tone, by tone of the following syllable, Speaker S.

Slope of	X-Tone 1	X-Tone 2	X-Tone3	X-Tone 4	X-Tone 0
Tone 1	71.1	-114.3	-51.3	-53.4	-81.3
Tone 2	-23.7	-48.1	-58.7	52.6	-83.9
Tone 3	-211.1	-145.4	-87.8	-98.0	-205.5
Tone 4	-35.6	-233.8	-177.5	-139.9	-119.9

Table 11. Mean slope in Hz change per second of each tone, by tone of the preceding syllable, Speaker P.

Slope of	Tone 1-X	Tone 2-X	Tone 3-X	Tone 4-X	X-Tone 0
Tone 1	-56.4	-35.2	-67.4	-75.2	-102.6
Tone 2	-160.0	-130.0	30.3	-18.5	17.0
Tone 3	-211.3	-163.6	-118.5	-106.2	-26.7
Tone 4	-225.5	-249.7	-122.9	-201.4	-180.1

Table 12. Mean slope in Hz change per second of each tone, by tone of the preceding syllable, Speaker S.

Slope of	Tone 1-X	Tone 2-X	Tone 3-X	Tone 4-X	X-Tone 0
Tone 1	-116.5	-35.3	57.4	-104.3	-32.1
Tone 2	-124.7	-61.5	34.6	-12.3	-1.7
Tone 3	-332.8	-87.3	-149.9	-163.8	-114.4
Tone 4	-212.2	-127.7	-69.4	-151.9	-27.5

According to the anticipatory theory, high succeeding onset Tones 1 and 4 should show a relatively positive slope on the previous syllable, and conversely for low and low onset tones 2 and 3. In forward assimilation not based on the succeeding onset, a succeeding rising tone 2 can cause a rising effect. We can see from Tables 9-12 that there is considerable variability in the slopes of each tone depending on the tonal sequence; we can also see that this is in part speaker dependent. The data from Tables 9-12 suggest that different tone combinations merge contextually in different ways.

For example, for both speakers, Tables 9-10 show that the anticipatory condition is true for Tone 4 followed by all tones, that is, the following tone influences the pitch slope of the previous syllable in the hypothesized direction, on average. For both speakers, succeeding Tones 3 and 4 induce a deeper fall for Tone 4 than succeeding Tones 1 and 4. A succeeding neutral tone induces a fall for Tone 4 roughly between these two cases. Similarly, high onset succeeding Tone 4 induces a relatively flatter (but still falling) pitch for all Tones except initial Tone 2 for Speaker S in Table 10. Initial Tone 2 achieves an overall average rising pitch only when it is followed by Tone 2 for Speaker P, contrary to the anticipation account, and by Tone 4 for Speaker S, in agreement with the anticipation effect. Initial Tone 2 falls the most for both speakers when it is followed by Tones 2 and 3, again consistent with the anticipation effect. A succeeding Tone 1 appears as the most problematic: for both speakers, an initial Tone 3 achieves its greatest falling pitch when followed by Tone 1. Based on our data, we suggest that the defined level nature of Tone 1 may introduce a discrete jump rather than a transformation to a gradient-sloped succeeding tone.

The tone sequence values seen in these tables also show some evidence of tone sandhi effects on pitch shapes. Tables 9-12 show some evidence for the 3-3 tone sandhi rule, since, for both speakers, the first of two 3rd Tones falls the least of all 3rd Tones, especially for Speaker P. When followed by another tone, Tone 3 exhibits a strong falling contour, as predicted by tone sandhi. Similarly, comparing the 4-4 tonal sequences for both speakers in Tables 9-12 shows that the 1st of two 4th Tones will have a slightly flatter pitch slope than the 2nd. Tone sandhi rules can override assimilation, as seen when Tone 3 is followed by Tone 1, where Tone 3 remains low, rather than assimilating to high Tone 1. However, in our data, succeeding Tone 4 has a relative positive effect on Tone 3 for both speakers.

When looking at the influence of the preceding tone in Tables 11 and 12, we see that a preceding Tone 1 has a negative effect on the subsequent tonal slope. When preceded by Tones 2 or 3, the tonal slope becomes relatively higher for Tone 1 and Tone 2 for both speakers, and these results are in accordance with the carryover predictions. Also, preceding Tone 1 has a negative effect on all tones for Speaker S, and for all tones except Tone 1 for Speaker P. For both speakers, a succeeding Tone 4 has the greatest fall in pitch when it is preceded by high ending Tones 1 and 2, consistent with carryover.

Both anticipatory and lag effects are found in the above tables. Because of the greater consistency of anticipatory results, there may be a marginally greater influence of anticipatory effects than carryover effects, and the overall pattern provides support for Chang and Hsieh’s findings of more balanced effects from both anticipatory and carryover. The differences that are evident between the speakers suggest that the influence of sequencing in speech on tonal targets may be conditioned by a number of speaker factors, such as speech rate and speaking style. The results obtained confirm the results of prior researchers, who have found substantial effects of preceding and subsequent tones on the realized tonal pitch values in read and experimental speech (Xu, 1997; Chang & Hsieh, 2012).

3.3 Variability of Tonal Shape and Amplitude

As seen in the above results, average results to a certain degree can be accounted for by tonal sequencing and by consistency of speaker style in producing the defined lexical tones. Nevertheless, when we look at individual tokens, our data indicate that it is much more difficult to account for the wide range of pitch shapes found in spontaneous speech for a given tone *only through* the factors presented in the tables above. A wide range of phonological and linguistic phenomena has been cited as also affecting the pitch values of syllables. Among the factors affecting pitch are the position in the phrase, phonemic identity, emotional and interactive effects, communicative function, and prosodic environment.

3.3.1 Comparative Measures of Tonal Shape

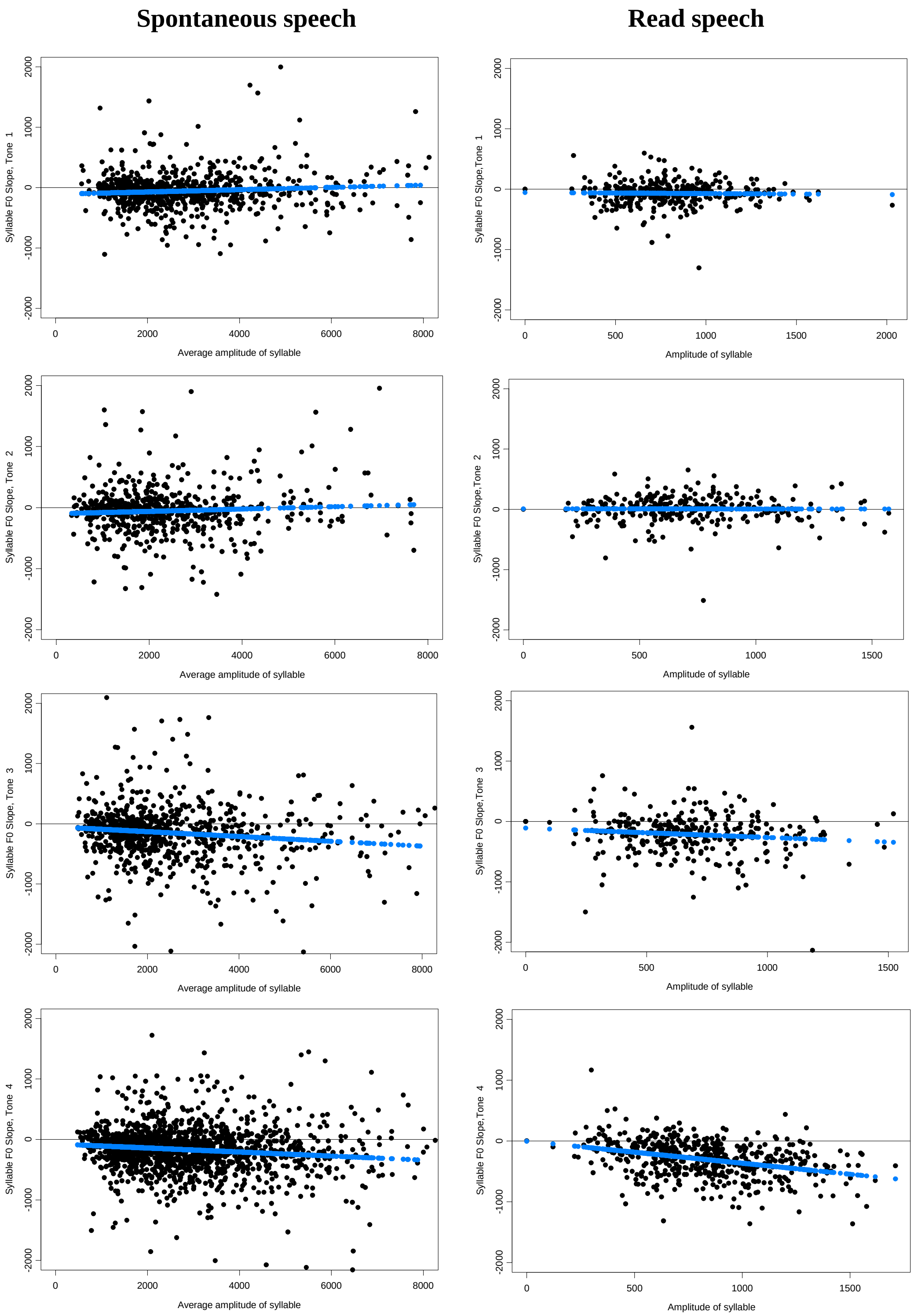
Figures 1 through 8 present a comparison of read speech to spontaneous speech, based on the simple linear slope measure of each syllable in the read speech corpus and the corresponding spontaneous speech counterpart, in which the fitted linear syllable slope is plotted against syllable average amplitude. A further regression line of slope vs. amplitude is superimposed on each figure.

Figures 1-4 show the slopes for spontaneous speech, and Figures 5-8 show the slopes for read speech. Comparison of spontaneous to read speech syllable slopes by tone shows that spontaneous speech is dispersed much more widely around the linear regression line. Furthermore, for each tone, for spontaneous speech the variation of pitch slope extends widely into both negative and positive slope regions, showing that the basic slope direction for spontaneous speech occurs with great frequency as either rising or falling.

By contrast, the slope value points for read speech in Figures 5-8 are clustered more tightly around the regression line, and their slope values are more in accordance with their defined lexical pitch values. In particular, for read speech, the slope values for Tone 1 cluster around zero, consistent with Tone 1’s level pitch definition, and virtually all Tone 4 slopes are falling for read speech. Tone 3 has a preponderance of falling slopes, while Tone 2 has a somewhat greater preponderance of rising pitch values for read speech. The much greater dispersion of spontaneous speech tone pitch values into both negative and positive slope regions indicates the presence and greater influence of contextual variables in determining pitch shapes in spontaneous speech.

These figures indicate that *tonal identity* is still an important factor in the realized pitch shapes of syllables because of the consistencies with respect to tones that exist in the figures for both read and spontaneous speech. For example, comparing across tones for spontaneous speech shows that, like read speech, Tone 1 syllables tend to cluster relatively symmetrically around a zero level pitch slope value, in accordance with the defined flat slope of Tone 1, although, for both read and spontaneous, the overall slope average was slightly below zero. Spontaneous Tone 2 syllables generally do not have the steeper falling slopes associated with Tones 3 and Tones 4. Spontaneous Tone 4 syllables seem to be the most consistent in their adherence to falling pitch, with the great majority of syllables still having a negative slope, although not as consistently as with read speech. The overall tendency for a greater falling pitch for syllables than is predicted by the lexical identity may be the reason that Tone 4 exhibits the greatest resilience and adherence to its defined falling pitch value, as it combines both of these effects in its pitch realization. Pair-wise non-parametric Wilcoxon rank sum tests comparing read and spontaneous speech by tone indicate the existence of systematic contour differences between spontaneous and read speech at high significance levels (See Appendix A).

Results shown in Figures 1-8 indicate that amplitude also correlates with changes in tonal shape. A steeper falling slope for high amplitude syllables is seen in the figures for spontaneous and read speech Tones 3 and 4, and, for Tones 1 and 2, greater amplitude is moderately associated with flatter or rising slope values. Prior research using experimental speech has found that the distribution of energy in syllables is correlated with tone identity in Mandarin tones (Whalen *et al*., 1992). The current finding measures across syllables and finds that in both spontaneous and read speech, *higher average amplitude* of a syllable is associated with a greater degree of adherence to the defined lexical tone shape for the syllable as a whole.

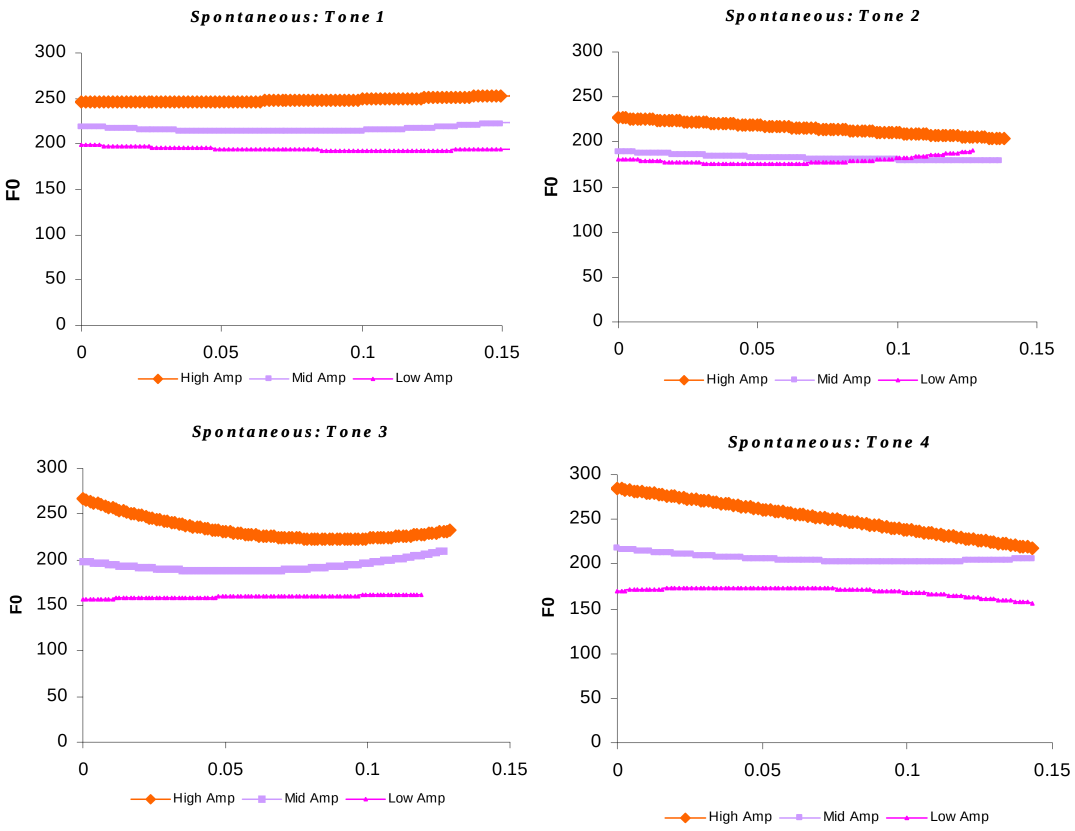

Spontaneous speech

Read speech

Figures 1-8. Slope of f_0 by amplitude for Tones 1-4, spontaneous speech shown in the left column and read speech in the right column. x-axis= syllable amplitude level; y-axis=syllable f_0 slope in Hz change per second

To achieve a more precise representation of the different tonal shapes, we fitted each syllable's normalized f_0 values to a quadratic polynomial. For all syllables in the spontaneous corpora used in this study, all f_0 values output were normalized by calculating Z-score values with respect to each speaker's mean syllable f_0 and standard deviation over all f_0 values for that speaker. After manual segmentation to the syllable level for all speech data, an average amplitude was calculated as the mean of all non-zero amplitude values. For each syllable, a quadratic fit of the f_0 normalized values also was calculated. Within each of the 4 tones' sub-groups, the average syllable amplitudes were categorized as low, medium, or high, such that each amplitude category had equal numbers of syllables within a tone sub-group, that is, each amplitude grouping contained 1/3 of the syllables within that tone. An average duration was also calculated over all syllables within each subgroup, resulting in 12 average syllable duration values. For each quadratic contour, the time dimension then was compressed or extended linearly to the average duration within its tone and amplitude subgroup, in order to avoid averaging across inconsistent segments of each syllable. Within each of the 12 tone and amplitude sub-groups, a model quadratic contour then was calculated as the average of the f_0 values over all quadratic contours within that subgroup, which, because of the normalization to average duration, is accomplished by averaging over the quadratic coefficients. Quadratic contours provide better overall model representations of average syllable shape than simple linear slopes, especially for Tones 2 and 3, because of the varying curvatures of the lexically defined f_0 values of the different tones.

The resulting quadratics are plotted in Figures 9-12. The model pitch slope shapes for spontaneous speech depicted in these figures corroborate the above conclusion that higher amplitude is associated with greater conformity to the tonal target: Tone 1 becomes higher and flatter and Tones 3 and 4 fall more with higher amplitude. Only Tone 2 breaks this pattern: higher amplitude Tone 2 becomes higher in average pitch, but exhibits a relatively greater fall in slope. The average durations are similar across amplitude groups for each tone, except for a slightly shorter average duration for low and mid amplitude syllables for Tone 3 tokens, so that the amplitude effect is not due to greater duration tokens having a fuller manifestation of the lexical shape. The model shapes seen in Figures 9-12 may provide a partial explanation of the similarity of slope between different tones, as seen previously with Tone 1 and Tone 4 in Table 5. The low and mid amplitude model shapes are more similar than dissimilar across tones, and this will cause greater similarity in the overall averages.



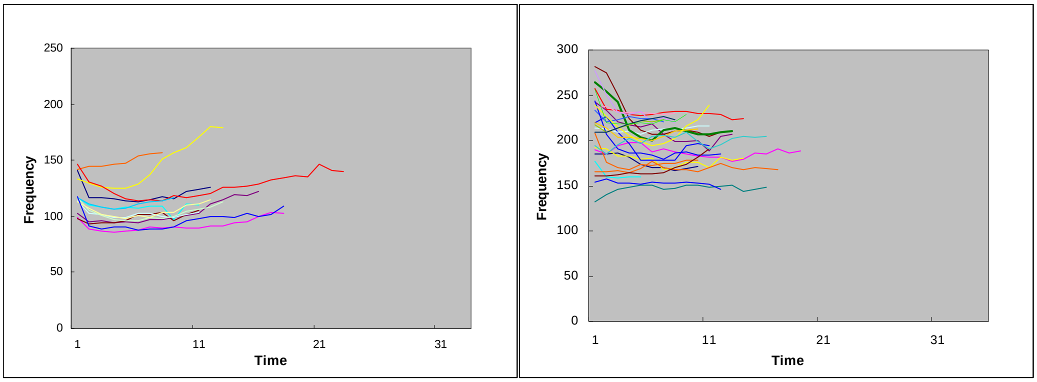
Figures 9-12. Slope of f_0 for Tones 1-4 by amplitude, spontaneous speech, in upper left, right, lower left, right order. x-axis= normalized syllable duration; y-axis= f_0 in Hz

The greater overall conformity to defined shape for read shape may arise because of the relative lack of other contextual influences. However, the greater conformity for greater amplitudes f_0 for both read and spontaneous speech must arise from a different cause, as it also occurs in the highly contextualized environment of interactive spontaneous conversation. Amplitude frequently is associated with *emphasis*, and emphasis arises naturally in both read and spontaneous speech. The above results suggest that one way to give emphasis to a lexical item in a tonal language, such as Mandarin, is to provide a more prominent and distinct lexical tone shape that marks the lexical meaning as more salient.

3.4 Visualizing Tonal Variation in Read and Spontaneous Speech

When viewed without respect to amplitude level, Figures 1-8 also give us a clearer picture of the differences between read and spontaneous speech. A comparison of the read to spontaneous speech in these figures shows that, although the pitch level varies more in spontaneous speech, the level nature of Tone 1 is similar in both read and spontaneous speech.

To analyze the differences between read and spontaneous in greater detail, we compared spontaneous vs. read speech for a number of identical tokens that were among the most frequently used syllables in both the read and the spontaneous corpora. This comparison suggested that a difference in speech mode does not affect all tones equally. The contours from read speech for identical Tone 1 tokens substantiated the result seen in Figures 1 and 5 above that Tone 1 remains essentially flat in both read and spontaneous speech. For Tones 2-4, however, there were more noticeable differences between the speech modes, as seen in the following Tone 2 *hai* 'still' example.



Figures 13-14. f_0 contours of 2^{nd}Tone 'hai' in read (left) and spontaneous speech (right)

3.4.1 Data Example 1: Tone 2 *hai*

The f_0 contours for Tone 2 *hai* 'still' of Figure 13 for read speech show clearly the defined rising nature of Tone 2, while, in the spontaneous speech shown in Figure 14, the pitch contours, for the most part, are much flatter, after an initial drop. Similar results were found for Tone 2 with other tokens. This *flattening* effect for spontaneous Tone 2 can also be seen from Figures 2 and 6 above, with a greater proportion of read speech slope measures in the rising range above zero than spontaneous Tone 2. A similar result occurs for Tones 3 and 4: the read speech slopes depicted in Figures 7 and 8 reflect an overall falling slope for almost all Tone 3 and Tone 4 read syllables, while Figures 3 and 4 show that spontaneous Tones 3 and 4 have greater numbers of syllables that *fall less* and frequently are *rising*.

These results (also see the results in Appendix A) suggest that, while spontaneous speech has more variability in pitch height and pitch contour, it also has a greater *flattening* effect on f_0 adherence to the defined tonal contours. Tone 1 remains overall flatter in spontaneous speech than in read speech. Spontaneous Tone 3 is similar to read speech, while Tone 2 is distinctly flatter in spontaneous speech than in read speech. Spontaneous Tone 3 is less likely to have an ending rise, and Tone 4 falls less than in read speech.

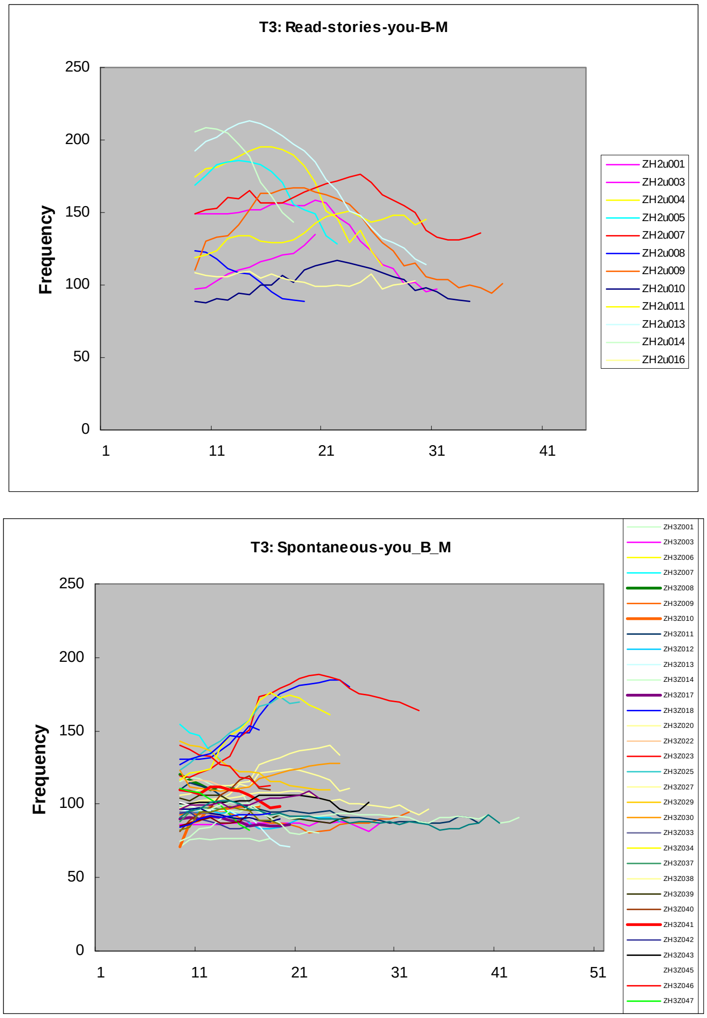
Our analysis suggests that it is the greater multi-functional usage of f_0 variation in spontaneous speech that leads to these results. In conversation, there is a greater use of pitch to indicate topic, provide signals of emotional and cognitive state, and to signal interactive intentions, and that the ‘flattening’ or reduced adherence to the lexically defined shape may be a more efficient use of lexical pitch so that a greater proportion of f_0 variability for the discourse functions is adopted.

3.4.2 Data Example 2: Tone 3 *you*

In Figures 15 and 16, we show individual instances of *you3* ‘have/has’ or ‘is’ for the male speaker as read speech in Figure 15 and in spontaneous conversation in Figure 16. *You* in Mandarin not only has a very high frequency of use, but frequently occurs in environments where cognitive or emotional intensity is high, such as in both questioning and in answers to questions, when uncertainty of information is present. On the other hand, *you* also occurs as a unremarkable syntactic object within a stream of more relevant lexical items or simple statements of facts. From Figures 15 and 16, Tone 3 *you* achieves its lexical full fall-rise shape or half-tone falling shape very rarely for this speaker, even in read speech. Since the read speech is a story in this case, the rise-fall pattern that is most predominant in Figure 15 reflects the communicative functions of emphasis added to the underlying lexical token. In spontaneous speech, *you* is reduced in pitch range in most cases to a flattened pitch contour, with a small number of exceptions that rise and fall as in read speech. In neither the read nor the spontaneous case do the pitch shapes approximate the lexical shape. In spontaneous speech, *you* has a very high frequency of use in interactive exchange, comparable to ‘has/have’ and ‘is’ in English, and its frequency of use in matter-of-fact statements de-emphasizes the need for a prominent signal for lexical comprehension; thus flatter pitch shape would not hinder comprehension and de-emphasis may aid in placing focus on the informative lexical target. Therefore, high frequency tokens, such as *you*, may rarely hit their defined shape: on one hand, de-emphasis flattens the contour, while, when communicating a cognitively or emotionally high intensity content, the commonplace nature of this token may be what allows it to take on primarily communicative prosody.

Spontaneous Tone 3 is similar to read speech Tone 3, while Tone 2 is distinctly flatter in spontaneous speech than in read speech. Spontaneous Tone 3 is less likely to have an ending rise, and Tone 4 falls less than in read speech. From research on experimental data, it has been proposed that flattening effects in speech are correlated with speech rate, and that shorter duration syllables should have greater flattening. The longer duration tokens for *hai* in Figure 13 that approximate continuous ending rise suggest that this may be a partial factor in flattening. However, in the corresponding spontaneous contours of Figure 14, there does not appear to be a correlation between short duration and flatness of slope. The flatter average slopes found for mid and low amplitude tones in the quadratic approximations in Figures 9-12 also do not appear to reflect the influence of duration, as, for all tones except Tone 3, all amplitude classes had approximately the same average duration.

Our preliminary analysis suggests that it may be the greater multi-functional usage of f_0 variation in spontaneous speech instead that leads to these results. In spontaneous speech, there is a greater use of pitch to indicate topic, provide signals of emotional and cognitive state, and to signal interactive intentions than in read speech, as exemplified in the emotional emphasis signaled in the high arching contours in Figure 16. The general 'flattening' or reduced adherence to the lexically defined shape may be a more efficient use of lexical pitch so that a greater proportion of f_0 variability for the discourse functions is adopted.

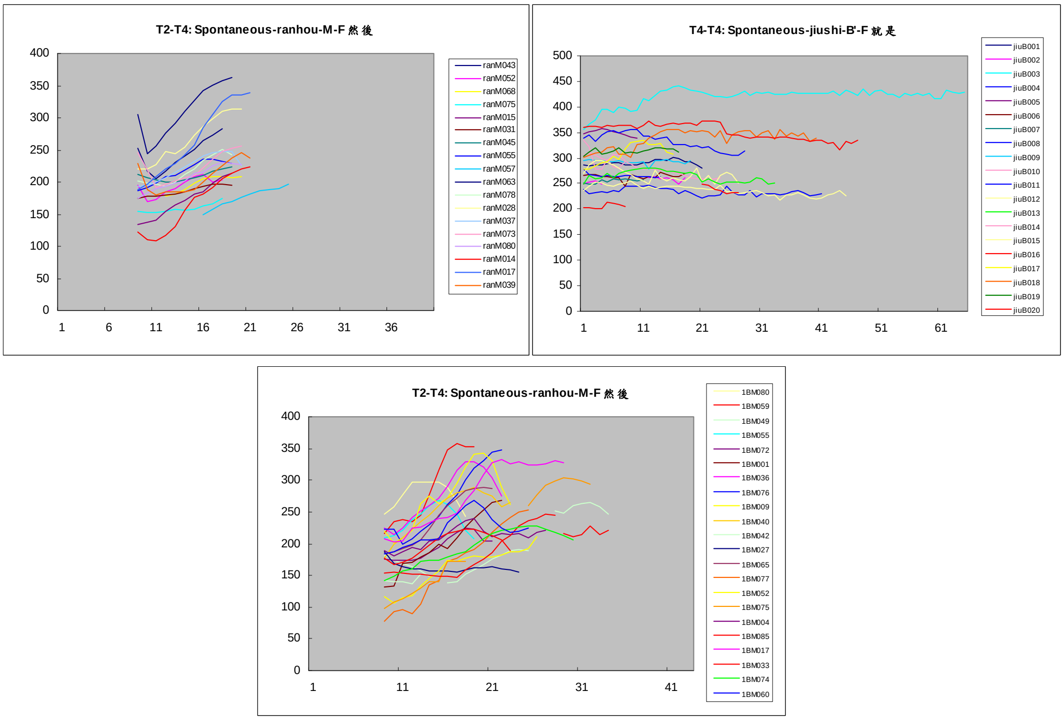


Figures 15-16. f_0 contours of 3rd Tone 'you' of the same speaker in read (top) and spontaneous speech (bottom)

3.5 Lexical Identity and Prosodic Shape

3.5.1 Data Example 3: *ranhou (2-4) & jiushi (4-4)*

It is illuminating to compare individual instances of tone shapes for specific lexical tokens. In Figures 17, 18, and 19, we compare tokens of two frequently used lexical items from our spontaneous corpus: 2-4 tone sequence *ranhou* ‘*then*’, and 4-4 tone sequence *jiushi* ‘*it’s just*’. The tokens illustrate how individual characters or words in Mandarin exhibit differing tendencies or abilities to vary their tonal shape, suggesting that a token’s lexical identity and its propensity to vary its tonal pitch contour are closely related.



Figures 17-19. f_0 contours of 2-4 tone sequence ranhou & 4-4 tone sequence of jiushi in spontaneous speech, in top left, top right and bottom order

In conversation, *ranhou* has a number of frequently used functions: it can act as a simple, direct logical or temporal connector, linking events that are sequential in time or in logical progression, similar to ‘*then*’ in English. As it links lexical and pragmatic meaning, it often takes on a rising form, where the falling Tone 4 ‘*hou*’ has been transformed prosodically. This is especially evident in the very short rising slopes of Figure 17. However, when it functions to emphasize the temporal transition from one previous state to the following one, *ranhou* is often realized as a pitch sequence that mirrors its lexical rise-fall tonal shape, as in Figure 19.

Expressivity through pitch can also depend on speaker or speaker state, as well as the pragmatic communicative functions that a particular lexical token assumes. *Jiushi* is often used as a hesitation marker as a speaker retrieves information for the next step in the communicative sequence, and the long and flattened pitch slopes of Figure 18 frequently occur. When *jiushi* is used to signal varying degrees of affirmation and agreement, the certainty and emphasis is reflected in a greater adherence to falling Tone 4 *shi* as the final syllable.

These examples illustrate a number of important points that provide insight into both the opportunities and the constraints governing tonal shape in spontaneous speech. The very high frequency and familiarity of both *ranhou* and *jiushi* imply that there is low cognitive effort in auditory interpretation, and this feature allows a high reduction in pitch shape under rapid speech or coarticulation. This is most likely to occur when these words are used as simple links between statement sequences that are the semantic focal targets of the communication. Moreover, as links that have specific lexical function, the familiarity and high frequency enable both *ranhou* and *jiushi* to more fully take on prosody that expresses emotional and cognitive qualifications on the nature of the linkage, especially including the marking of hesitation through duration and pitch shape. Conversely, when the meaning in the lexical link itself is the focus, emphasis on that lexical meaning is reinforced through a high, even exaggerated, use of the defined lexical shapes. Thus, the propensity of a given token to conform to defined tonal shapes may be conditioned by its frequency of use and its lexical identity, as well as its intrinsic ability to take on a number of different pragmatic communicative functions.

4. Conclusion

In this study, we have presented a framework for characterizing tonal variability over spontaneous conversations and have compared the degree of adherence to defined tonal values in read and spontaneous Mandarin. The results show that realized tonal pitch shapes in spontaneous speech have greater internal variance as well as greater overall divergence from defined pitch shapes. We have investigated several factors that contribute to this divergence, and we have shown evidence that corroborates, for spontaneous conversations, prior research on the existence and importance of anticipatory and carryover effects. The numerous exceptions to sequencing effects indicate that these effects were not *sufficient* to account for the wide variability of tonal shapes in spontaneous speech. A key finding of the study is that adherence to a defined shape is greater when syllable *amplitude* is relatively high. The study also identified two tendencies for pitch shape in spontaneous speech. First, pitch slope tends to have more negative slopes, compared to the defined shape for Tones 1, 2, and 3. A second related tendency is that slopes in spontaneous speech operate under a flattening tendency, with 4th Tones less steep than in read speech and with level and rising tones more negative.

The current study provides an initial platform from which to extend research to the role of conversational contextual factors, including cognition, emotion, and communicative function, on realized tonal shapes, and we have shown initial evidence on the importance of lexical meaning and speaker state to tonal and prosodic variation. Finally, results of the study indicate that, although there is great variability in individual tone pitch shape, there are also systematic relationships among the tones and in tone sequences that are dependent on tonal identity. Thus, our results suggest that the diverse variations in realized tonal shape are evidence of the great ability of Mandarin to simultaneously express both lexical meaning and speaker state in a unified system of lexical and prosodic form and demonstrate the high potential for expressive prosody in Mandarin.

Acknowledgements

This work was supported in part by the National Science Council of Taiwan under grant number 100-2410-H-029-039 to the first author. We would like to thank the two anonymous reviewers whose comments have substantially improved the paper. We would also like to thank Dr. Shu-Chuan Tseng and Dr. Chiu-yu Tseng of Academia Sinica for very helpful discussions while the first author was a Visiting Scholar at the Institute of Linguistics, Academia Sinica, Taiwan, during the year 2012.

References

Chang, Y.-C, & Hsieh, F.-F. (2012). Tonal coarticulation in Malaysian Hokkien: A typological anomaly? *The Linguistic Review, 29*, 37-73.

Chao, Y. R. (1968). *A grammar of spoken Chinese*, Berkeley, University of California Press.

Gussenhoven, C. (2004). *The phonology of tone and intonation*, Cambridge University Press.

Hirst, D. & Di Cristo, A. eds. (1998). *Intonation systems: a survey of twenty languages*, Cambridge, Cambridge University Press.

Hsieh, F.-F. (2008). Preservation of the marked as slope correspondence in Hangzhou Chinese disyllabic tone sandhi. *Interfaces in Chinese Phonology*, 223-242.

Rose, P. (2012). Two sides of the same coin: between - speaker F0 differences in linguistic-phonetic description and forensic voice comparison. TAL 2012. *Interanational Conference on Tonal Aspects across Tone and Non-tone Languages*, keynote speech, Nanjin, China.

Tseng, C.-Y. (2010). Beyond sentence prosody. Keynote speech, In *Proceedings of Interspeech 2010*, Makuhari, Japan, 20-29.

Tseng, C.-Y. (2010). An f0 analysis of discourse construction and global information in realized narrative prosody. *Language & Linguistics, 11*(2), 183-218.

Tseng, S.-C. (2004). Processing spoken Mandarin corpora. *Traitement automatique des langues. Special Issue: Spoken Corpus Processing*, *45*(2), 89-108.

Tseng, S.-C. (2005a). Syllable contractions in a Mandarin conversational dialogue corpus. *International Journal of Corpus Linguistics*, *10*(1), 63-83.

Tseng, S.-C. (2005b). Mandarin topic-oriented conversations. *International Journal of Computational Linguistics and Chinese Language Processing. Special Issue: Annotated Speech Corpora*, *10*(2), 201-218.

Tseng, S.-C. (2008). Spoken corpora and analysis of natural speech. *Taiwan Journal of Linguistics*, *6*(2), 1-26.

Tseng, S.-C. (2013). Lexical Coverage in Taiwan Mandarin. *International Journal of Computational Linguistics and Chinese Language Processing*, *18*(1), 1-18.

Shen, X. S. (1990). *The Prosody of Mandarin Chinese*, University of California Press.

Shih, C. & Sproat, R. (1992). Variations of the Mandarin rising tone. *IRCS workshop on prosody in natural speech*, 193-200.

Shriberg, E., Stolcke, A., Hakkani-Tur, D. & Tur, G. (2000). Prosody-Based automatic segmentation of speech into sentences and topics. *Speech Communiccation*, *32*(1-2), 127-154 (Special Issue on Accessing Information in Spoken Audio).

Whalen, D. H. & Xu, Y. (1992). Information for Mandarin tones in the amplitude contour and in brief segments. *Phonetica*, *49*, 25-47.

Xu, Y. (1997). Contextual tonal variations in Mandarin. *Journal of Phonetics*, *25*, 61-83.

Xu, Y. (2011). Functions and mechanisms in linguistic research -- Lessons from speech prosody. In *Proceedings of Workshop on Experimental Linguistics*. Paris: 1-10.

Xu, Y., & Wang, Q. (2001). Pitch targets and their realization: evidence from Mandarin Chinese. *Speech Communication*, *33*(4), 319-337.

Yang, L.-C. (1995). *Intonational Structures of Mandarin Discourse*, Ph.D. dissertation, Georgetown University.

Appendix A: Supplementary Statistical Results on Tonal Variation

Table A1 compares the distributional spread of spontaneous speech to read speech of Speaker B. For Tone 1, the standard deviation for spontaneous speech is about twice as large as for read speech, with the kurtosis values for spontaneous tones being greater than for read speech, especially for Tones 1 and 4, indicating a greater occurrence of extreme variations in slope for spontaneous speech.

Table A1. Comparison of syllable of f_0 slope, read speech to spontaneous speech, Speaker B

	Read speech				Spontaneous speech			
	Tone 1	Tone 2	Tone 3	Tone 4	Tone 1	Tone 2	Tone 3	Tone 4
Mean	-29.82	36.95	-22.61	-271.06	-84.33	-14.45	27.20	-95.64
Median	-22.92	51.12	-76.22	-260.38	-62.76	-5.99	16.33	-75.03
SD	155.20	232.12	457.43	319.33	328.65	247.37	334.43	297.83
Kurtosis	9.51	7.34	6.65	1.92	21.11	8.82	8.90	14.49
Skewness	-1.20	-1.47	0.54	0.35	-1.66	-0.62	0.16	0.84
R^2	0.0005	2e-07	0.019	0.1419	0.0419	0.0078	0.0075	0.0358

The skewness and large kurtosis values for both read and spontaneous speech show significant departures from normality, so non-parametric Wilcoxon rank sum tests were computed to test the hypothesis of no shift in average slope between read and spontaneous speech. Table A2 presents the results of the pair-wise Wilcoxon tests between read and spontaneous syllable slope by tone for Speaker B. The very low p-values shown indicate that the alternative hypothesis of a change in syllable slope holds at high significance levels and indicate the existence of systematic contour differences between spontaneous and read speech.

Table A2. Wilcoxon rank sum test comparing read to spontaneous speech by tone, Speaker B

Read vs. Spontaneous	W-value	p-value
Tone 1	73760	6.035e-07
Tone 2	52604	4.882e-06
Tone 3	45319	0.0001075
Tone 4	97593	< 2.2e-16

The Wilcoxon results corroborate the shifts in syllable slope between read and spontaneous speech shown in Table A1. The decrease in median slope of spontaneous speech for Tones 1 and 2 and the increase in median slope for Tones 3 and 4 suggest a tendency towards flatter slopes, on average, in spontaneous speech. The exception to this is Tone 1, which tends to fall more in spontaneous speech than in read speech.