Process vs. Result: Evidence for Coarticulatory Process Difficulties in French-speaking Learners of Mandarin

Yingyu Yan [1], *Jorina Brysbaert* [2], *Anne-Catherine Simon* [3]

[123] University of Louvain

yingyu.yan@uclouvain.be, jorina.brysbaert@uclouvain.be, anne-catherine.simon@uclouvain.be

Abstract

Mastering tonal coarticulation is a key challenge for L2 learners of Mandarin. This study investigates the demanding T4-T1 tonal sequence to tease apart difficulties in motor planning (process) from those in target attainment (result). We compare 139 sequences produced by eight native French-speaking learners of Mandarin and eight native speakers of Mandarin, elicited through a controlled question-response task. French L1 learners were selected as a representative non-tonal L1 group. Whether the same patterns hold for other non-tonal L1 backgrounds remains to be tested. Using linear mixed-effects models, we analyze (i) the timing of cross-syllable F0 transitions, (ii) the within-syllable F0 dynamics, and (iii) the resulting T1 target realization, with all measures automatically extracted via a Praat script.

Results reveal a consistent process-level disruption among learners of Mandarin. First, learners produce longer T4–T1 transition times, indicating slower cross-syllable coordination. Second, although the magnitude of the T4 fall is native-like, learners reach the T4 valley significantly earlier within the syllable. This premature valley alignment appears to trigger a marked mid-syllabic F0 "sag" (i.e. dip) in the first T1 syllable. This process-level error, in turn, is associated with a deeper mid-syllable sag in L2 T1a although the group difference in overall T1 height was not significantly significant. Overall, the findings indicate that L2 tone production difficulties stem more from imprecise motor planning and coarticulatory timing across syllables than from limits in reaching the tonal target.

Index Terms: tonal coarticulation, L2 prosody, Mandarin Chinese, carry-over effect

1. Introduction

In Mandarin, each syllable (character) carries a specific lexical tone [1]. For L2 learners whose native languages are non-tonal, such as French, mastering these tones poses a significant challenge. However, an even greater difficulty may lie in producing them accurately in continuous speech, which requires complex tonal coarticulation.

Mandarin comprises four lexical tones, each with a distinct pitch contour. According to Chao's 5-level system [1], Tone 1 (T1) is a static High-Level target, while Tone 4 (T4) is dynamic High-Falling. In natural speech, tones are rarely produced in isolation. Instead, they interact with adjacent lexical tones through a process known as cross-syllabic tonal coarticulation [2], primarily manifested as carry-over effect (the mechanical inertia of a preceding tone influencing the onset of the following tone) and anticipatory effects (the regression of a following tone influencing the offset of a preceding tone). This study focuses on the T4-T1 sequence, which represents a particularly demanding coarticulatory challenge: the speaker must execute a rapid pitch fall (T4) and immediately engage the vocal folds to maintain a high stable pitch (T1).

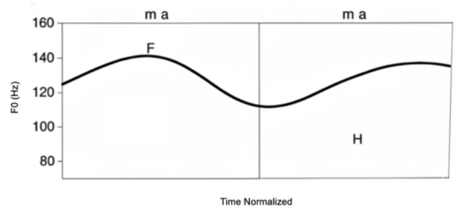

Figure 1: Schematic representation of the F0 contour for the Tone 4 (High-Falling) to Tone 1 (High-Level) sequence. The labels "F" and "H" denote the Falling and High tonal targets, respectively [2].

L2 tonal acquisition has been widely studied, but most of these studies have focused on measurements of tones produced in isolation [3, 4, 5, 6, 7]. Such designs minimize articulatory constraints, failing to capture the complexity of motor planning in connected speech. Consequently, it remains unclear whether L2 learners' difficulties stem from an inability to form the correct tonal targets (the result) or from the execution of coarticulation within the allotted time (the process).

To tease apart these mechanisms, we adopt the theoretical framework of "Pitch target and their realization" proposed by Xu [2]. Under this view, a tone is defined as an underlying pitch target, which represents the intended result. In contrast, the surface F0 contour is the tonal implementation of approaching this target, representing the process. Following this logic, we operationalize "inter-syllable transition timing" and "dynamic contour shape" as measures of the process (motor planning and execution), while "target attainment" (such as the T4 fall magnitude and T1 plateau height) represents the result of that execution.

To address this question, we compared productions from L1 and L2 speakers of Mandarin to test the following Research Questions (RQs) concerning the process and the result:

1. Does the realized T4 fall magnitude and T1 F0 plateau height differ between groups?

Hypothesis (H1): L2 speakers will show a flatter T4 fall and a similar (or only slightly lower) T1a plateau height.

2. Does the F0 dynamic contour shape (the process) in the T4-T1 sequence differ between L1 and L2 speakers?

Hypothesis (H2): L2 speakers will show a later valley alignment and a non-native "F0 sag" within the T1 interval.

2. Methods

2.1. Participants

Two groups participated in the study. The learner group comprised eight native speakers of French (age range: 20–23) learning Mandarin as a foreign language. They were recruited from university language programs in Paris (France), Louvain-la-Neuve, and Mons (Belgium). All held HSK Level 4 and reported no extensive residence (defined as >12 months) in a Chinese-speaking environment. The control group consisted of eight native speakers of Mandarin (age range: 21–24) who study in Shanghai (China).

2.2. Materials and task

The data analyzed in this study come from a larger production experiment on prosodic focus in French and Mandarin. The full corpus consists of question-answer dialogues elicited under several focus conditions. For the present study, we used the neutral-focus condition only, which is also called no-focus condition, to avoid focus-related confounds.

The experimental materials comprised 27 scripted question-answer dialogues which fit into a general context of a party. Each of these 27 items contained a context question and a target response. For example, in the following dialogue, the participant will hear a recorded question (ex. Q)? and answer with the provided response (ex. A):

- Q: 发生什么事了？ What happened?
- A: 芬在窗边吃瓜。 Fen is eating melon by the window.

Each target response was designed to embed a trisyllabic T4﹣T1a﹣T1b sequence in a simple syntactic environment. For example, within the response in A, 在﹣窗﹣边 (zaì chuāng biān, "next to the window") instantiates the T4-T1a-T1b tonal sequence (syllables labeled T4, T1a, T1b). All the analyzed tonal sequences are in the middle of the utterance and preceded by a syllable which carries the first tone T1 to avoid the influence of the preceding tones. Crucially, we utilized a trisyllabic sequence ending in a High tone (T1b) rather than a disyllabic pair. The presence of the final T1b serves two purposes: first, it prevents the target T1a from undergoing phrase-final declination or creakiness; second, it ensures that any F0 sag observed in T1a is attributable to the carry-over effect of the preceding T4, rather than an anticipatory dip induced by a following low tone.

Participants were tested individually in quiet rooms using a Roland R-05 recorder. On each trial, participants first heard and saw a pre-recorded context question (spoken by a female native Mandarin speaker) and then read the on-screen answer aloud. They were asked to repeat the answer if they hesitated. Items were pseudo-randomized to avoid potential session effects, such as participant speaking style, or slight day-to-day changes in the recording environment.

Across speakers, this elicitation task produced a dataset of 139 usable T4-T1a-T1b realizations in total: 81 for the L1 group and 58 for the L2 group. For quality control, tokens exhibiting production errors (mispronunciation or disfluencies) or technical issues (F0 tracking failure) were excluded.

2.3. Acoustic Analysis

The target responses were automatically transcribed using Whisper [8] and segmented into syllables using SPPAS [9] for forced alignment. The boundaries of the target T4, T1a, and T1b syllables were manually corrected in Praat [10] through auditory and spectrographic inspection. Acoustic analysis was performed using a custom Praat script. For each participant, all F0 values (Hz) were converted to semitones (st) relative to that speaker's median F0 to ensure cross-speaker comparability.

The script was specifically built to automatically extract measures aligned with our research questions. RQ1 is addressed by t4_drop_range_st, quantifying T4 fall magnitude and by F0mean_T1a_60_90_st, addressing the height of the T1 plateau. RQ2 is assessed by T4 valley timing and T1a F0 sag magnitude (calculated as the difference between the onset of T1a and its minimum F0). For each syllable, F0 maximum, F0 minimum, and duration were also extracted.

Table 1. Summary of acoustic measures for the T4-T1a-T1b sequences.

RQ	Measure	Explanation
1.Target realization	t4_drop_range_st	F0 (st) difference between the 20% and 80% time points of the T4 syllable (avoid the voiceless onset and creaky offset)
	F0mean_T1a_60_90_st	Mean F0 (st) over the 60%–90% window of the T1a interval (avoid perturbation of consonant and exclude protentional final unstable F0 lowering)
2.F0 dynamics	t4_min_time_pct	Normalized temporal position (0.0–1.0) of the F0 valley within the T4 syllable
	t1a_max_min_st	F0 (st) difference between the F0 maximum in the first half and the F0 minimum in the second half of the T1a syllable (=magnitude of F0 excursion during the plateau phase)

2.4. Statistical Analysis

All statistical analyses were performed in R [11] using the lme4 package [12] to fit linear mixed-effects models (LMMs). We constructed separate models to address the two dimensions of our research questions: Target Realization (Result) and F0 Dynamics (Process). To address RQ1 (Target Realization), we examined whether learners successfully attained the tonal targets. The dependent variables were the T4 fall magnitude (t4_drop_range_st) and the T1 plateau height (F0mean_T1a_60_90_st). To address RQ2 (F0 Dynamics), we investigated the articulatory process and contour shape. The dependent variables were the temporal alignment of the T4 valley (t4_min_time_pct) and the magnitude of the F0 sag within the T1a syllable (t1a_max_min_st).

For all models, Group (L1 vs. L2) was entered as the fixed effect of interest. To control for variations in speaking rate, the z-scored duration of the corresponding syllable was included as a covariate. We included random intercepts for Speaker and Item to account for by-participant and by-item variability. Visual inspection of residual plots was performed to verify assumptions of homoscedasticity and normality. Significance of the fixed effects was determined using the lmerTest package [13], which provides p-values via Satterthwaite's degrees of freedom method.

3. Results

Before turning to the quantitative analyses for each research question (Sections 3.1-3.2), we first present the group-averaged, time-normalized F0 contours for the T4-T1a-T1b sequence (Figure 2). This visualization provides a global impression of how L1 and L2 speakers differ in the shaping and timing of the tonal sequence. Overall, both L1 and L2 speakers show similar inter-syllabic transition durations from the T4 valley to the T1a peak. However, two qualitative differences emerge: L2 productions display (i) differences in the timing to reach the valley of T4 and (ii) a more pronounced mid-syllable F0 sag ("dip") within T1a.

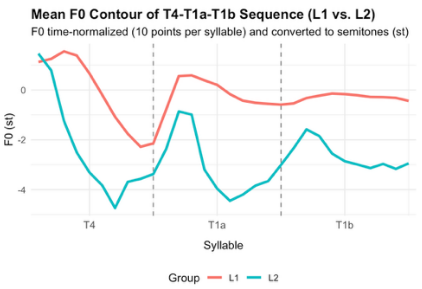

Figure 2. Group-averaged F0 contours for the T4-T1a-T1b sequence. Each panel shows one syllable (T4, T1a, T1b). The x-axis represents normalized time (log10 scale); the y-axis shows F0 in semitones (st) relative to each speaker's median F0. Curves represent smoothed mean F0 trajectories for L1 speakers (red) and L2 learners (teal).

3.1. "Result": Target Realization

To address RQ1 (the "Result"), we analyzed the magnitude of the fall in T4 and the realized F0 height of the T1a plateau (measured as the mean F0 in the 60%–90% window of the T1a syllable). Figure 3 illustrates the T4 fall magnitude (H1a) and the T1 plateau height (H1b).

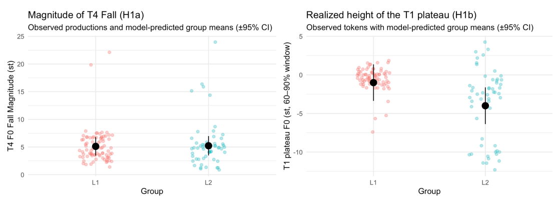

Figure 3. Acoustic measures of T4-T1a target realization for L1 (native) and L2 (learner) groups. (A) T4 fall magnitude (in st); (B) Realized height of the T1a plateau. Small light-colored dots represent individual tokens black dots show model-predicted group means with 95% confidence intervals.

First, regarding the magnitude of the fall in T4 (t4_drop_range_st), the linear mixed-effects model revealed no significant main effect of Group ($\beta = 0.11$, SE = 1.23, $p = 0.928$). The average magnitude of the T4 fall was 5.12 semitones (st) for the L1 group and 5.23 st for the L2 group (Figure 4A). This indicates that L2 learners produced a T4 fall comparable in size to native speakers.

Second, regarding the late T1a window measure (F0mean_T1a_60_90_st), the linear mixed-effects model revealed a marginal main effect of Group ($\beta = -2.996$, SE = 1.69, $p = 0.080$). Since this measure averages 60-90% of the syllable, it largely captured a mid-syllable sag in L2 productions rather than a stable plateau. The negative estimate indicates that the L2 group produced a mid-syllable sag that was, on average, nearly 3 semitones (st) lower than that of the L1 group. Although this difference did not reach the conventional threshold for statistical significance ($p < 0.05$), it constitutes a strong trend in the hypothesized direction (supporting H1).

The model's R2 values help contextualize this trend. The marginal R^2 was 0.150, indicating that Group alone accounts for 15% of the total variance in plateau height, which is substantial for an F0 measure. By contrast, the high conditional R^2 (0.935) and ICC (0.924) show that a very large proportion of the variance is attributable to between-speaker differences, especially within the L2 group. This large speaker-level variability likely contributes to the borderline p-value. As shown in Figure 3B, individual L2 productions are more dispersed and generally lower than L1.

3.2. "Process": F0 Dynamics

To investigate RQ2, we analyzed two acoustic measures capturing dynamic aspects of the T4-T1a F0 contour (see Figure 4): the temporal alignment of the T4 valley (H2a), and the degree of mid-syllable sag in T1a (H2b).

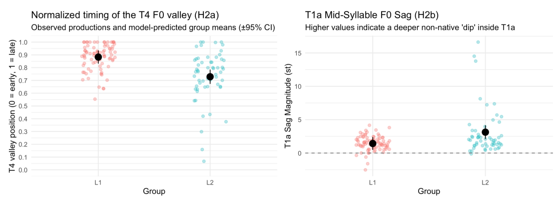

Figure 4. Acoustic measures of T4-T1a F0 contour dynamics for L1 (native) and L2 (learner) groups. (A) T4 valley timing (normalized time position 0–1); (B) T1a mid-syllable sag (st). Light dots represent individual tokens; black dots show model-predicted group means with 95% confidence intervals.

First, for the temporal alignment of the F0 valley in T4 (t4_min_time_pct), the model revealed a significant main effect of Group ($\beta = -0.15$, SE = 0.04, $p < 0.001$). Importantly, the direction of the effect was opposite to our prediction in H2b, which expected L2 learners to reach the valley later than L1 speakers. Instead, as shown in Figure 4B, native speakers aligned the F0 valley near the syllable offset (mean = 88.3% normalized time), whereas L2 learners reached the F0 valley significantly earlier (mean = 72.9%). This suggests that L2 learners terminate the T4 falling movement prematurely, which in turn indicates difficulty with T4 valley timing.

Second, regarding the magnitude of the mid-syllable sag in T1a (t1a_max_min_st), the model revealed a significant main effect of Group ($\beta = 1.701$, SE = 0.68, $p = 0.014$) (Figure 4B). The model's intercept, representing the L1 baseline, was 1.428 st, indicating that native speakers exhibit a small, natural F0 declination of approximately 1.4 st within T1a. The positive coefficient for the L2 group ($\beta = 1.701$) shows that learners exhibited a sag that was 1.7 st larger than the L1 baseline, yielding an estimated sag of approximately 3.13 st for L2 productions (1.428 + 1.701). This pattern provides support for a clear, non-native F0 contour dynamic.

4. Discussion

This study investigated whether the difficulty observed in French-speaking learners of Mandarin with the T4-T1 sequence reflects a problem with target attainment ("result") or with the execution of the tonal transition ("process"). The evidence consistently points to a process-based deficit: L2 learners relatively successfully form the tonal targets but fail to coordinate them temporally across the syllable boundary.

A key finding is that the "Result" of T4 is largely attained. The magnitude of the T4 fall did not differ between groups, indicating that learners do not suffer from a deficit in reaching tonal targets; they are physiologically capable of executing the rapid pitch excursion required for Tone 4. This rules out explanations in terms of inaccurate end points of T4 and shifts attention to how the trajectory folds. Regarding the T1a plateau, although learners produced a numerically lower mean F0 in the late portion of T1a than native speakers (showing a marginal trend, $p = 0.08$), this difference did not reach statistical significance. This implies that learners generally aim for the correct "High" target and do not suffer from a fundamental inability to access the upper register of their voice. However, the fact that the T1 plateau was consistently lower suggests that, while the target is correct, its implementation is compromised by other factors.

The "Process" dimension reveals why the T1 realization is compromised. Learners reached the T4 valley significantly earlier within the syllable than native speakers (72.9% vs. 88.3%). This premature termination of the T4 fall suggests that learners are rushing the pitch movement, effectively desynchronizing the tonal gesture from the syllable boundary. This premature alignment appears to trigger a mechanical chain reaction, explaining the "F0 sag" observed in the following T1a syllable. According to the Target Approximation Model [2], native speakers utilize the syllable boundary to facilitate the sharp T4 drop, maintaining tension into the onset of T1. In contrast, by finishing the T4 fall too early, learners likely allow the vocal folds to enter a brief state of relaxation before re-engaging the high tension required for T1. This results in the observed "sag"—a dip in pitch before the rise to the T1 plateau.

The presence of the T1b (High tone) in our experimental design rules out alternative explanations. Since the sequence ends with a high target, the observed sag in T1a cannot be attributed to anticipatory lowering induced by a following low tone. It is, therefore, a pure carry-over effect stemming from the timing mismanagement of the preceding T4.

5. Conclusions

The present study provides acoustic evidence that L2 tone production difficulties in the T4-T1 sequence stem more from imprecise temporal coordination (the process) than from an inability to reach pitch targets (the result). L2 speakers of Mandarin appear to aim for the correct underlying targets but approach them inefficiently. Specifically, they execute the T4 fall too early, creating a temporal gap that destabilizes the onset of the following High tone.

Pedagogically, these findings suggest that instruction should move beyond static pitch height drills. Training should emphasize syllable synchronization — teaching learners to delay the T4 fall onset so that the pitch momentum carries directly into the next syllable without an intervening "sag".

Limitations of this study include the modest sample size, which may have limited the statistical power for the T1 plateau measure. Future work should expand the sample and integrate articulatory measures (such as EGG) to directly test whether the earlier valley alignment reflects a premature cessation of laryngeal activity.

6. Acknowledgements

This work is supported by China Scholarship Council (CSC), Wallonie-Bruxelles International (WBI), and Fédération Wallonie-Bruxelles (FWB).

7. References

[1] Y.R. Chao, *Language and symbolic systems*, Cambridge,U.K.: Cambridge Univ. Press, 1968.

[2] Y. Xu, and Q.E. Wang, "Pitch targets and their realization: Evidence from Mandarin Chinese," *Speech communication*, vol. 33, no.4, pp. 319-337, 2001.

[3] C. Yang, *The acquisition of L2 Mandarin prosody: From experimental studies to pedagogical practice*. Amsterdam, The Netherlands: John Benjamins, 2016.

[4] D. Francolino, and W. Cao, "Prosodic challenges among Italian speakers in L2 Mandarin: Preliminary evidence from falling tone production," in *Proc. 5th International Symposium on Applied Phonetics*, Tartu, Estonia, 2024.

[5] T. Bent, "Perception and production of non-native prosodic categories," Ph.D. dissertation, Northwestern Univ, Evanston, IL, USA, 2005.

[6] Y.C. Hao, "Second language acquisition of Mandarin Chinese tones by tonal and non-tonal language speakers," *Journal of Phonetics*, vol. 40, no. 2, pp.269–279, 2012.

[7] T. Zou, J. Zhang, and W. Cao, "A comparison study on f0 distribution of tone 2 and tone 3 in Mandarin disyllables by native speakers and Japanese learners," in *Proc. Speech Prosody*, pp. 226–229, 2012.

[8] A. Radford, et al., *"Robust Speech Recognition via Large-Scale Weak Supervision,"* arXiv:2212.04356, 2022.

[9] B. Brigitte, "SPPAS-multi-lingual approaches to the automatic annotation of speech," *The Phonetician. Journal of the International Society of Phonetic Sciences*, vol. 111, ISSN: 0741-6164, p. 54-69, 2015.

[10] P. Boersma and D. Weenink, *Praat: Doing Phonetics by Computer* [Computer program], version 6.4.23, 2024.

[11] R Core Team, *R: A language and environment for statistical computing* [Computer software]. Version 3.6.1.Vienna, Austria: R Foundation for Statistical Computing, 2019.

[12] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4", *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015, doi: 10.18637/jss.v067.i01.

[13] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest Package: Tests in Linear Mixed Effects Models", *Journal of Statistical Software*, vol. 82, no. 13, pp. 1–26, 2017, doi: 10.18637/jss.v082.i13.