Guiyang Mandarin Tone Sandhi Report

Introduction

Standard Mandarin is a prototypical tone language, where different tones rely on pitch contours to distinguish word meanings. In order to describe the tone pattern on each syllable in Mandarin more consistently and accurately, linguists often use the “five-degree tone value system” (Chao 1930). The core idea of this system is not to compare an absolute frequency “how many Hertz” directly, but rather to divide each speaker’s individual pitch range from lowest to highest into five equal levels, represented by 1–5, so that pitch can be standardized across different speakers. For example, a high rising tone can be written as “35”, a falling tone as “31”, and an even level tone as “44”.

Guiyang Mandarin belongs to Southwestern Mandarin. Its phonological system has been shaped by multiple waves of migration and contact with many minority languages, and it has very distinctive tonal properties. However, past work on Guiyang has mainly focused on segmental aspects such as vowels, consonants, and rimes. Research on tones and tone sandhi has been very limited. In particular, work before the 1980s mostly relied on auditory “listen-and-describe” methods (Wang Ping). After 2000, a small number of acoustic studies have revised earlier descriptions, but overall the literature is still lagging behind. There is still no unified conclusion about the citation tone values of Guiyang.

Tone sandhi is a very typical phonological phenomenon in tone languages. It means that when two syllables are pronounced together, the tones do not keep their original shapes, but instead change in systematic ways. Tone sandhi in Guiyang is quite complex, and at present there is a lack of systematic statistical analysis. The goals of this project are:

1） To determine the tone values of Guiyang Mandarin based on collected data;
2） To describe the tone sandhi patterns in Guiyang AA reduplicated forms (currently limited to kinship terms);
3） To build a probabilistic tone sandhi model based on real speech data;
4） To use Monte Carlo simulation to test whether the model can reproduce real-world behavior.

Data Description

The data for this project come from a 26-year-old female speaker from Guiyang. The target items are common kinship AA reduplications such as baba ‘dad’, mama ‘mom’, jiejie ‘older sister’, gege ‘older brother’, etc.

The stimulus materials have two main parts:

1）Monosyllabic part (for determining tone values):

First, I designed a set of ma-series monosyllables with the same syllable but different tones as a reference. The following 16 characters were then divided into four groups according to tone, and each tonal group contained several characters with different segments but the same tone (for example, a high-level group, rising group, falling group, etc.). The purpose of this part is to determine the standard citation tone values of the four tones for this speaker using acoustic methods.

2）AA reduplication part (for sandhi analysis):

After confirming the tone values, I recorded 17 kinship AA reduplicated forms (such as baba, nainai, jiujiu, etc.) to investigate tone sandhi in connected speech.

Each token contains the following:

(1) citation tone (the tone when pronounced in isolation);
(2) index (A1 or A2);
(3) surface tone (the tone in connected speech);
(4) acoustic information (F0).

The data cleaning procedures include:

(1) manually segmenting syllables in Praat (TextGrid);
(2) extracting F0 using Python (parselmouth);
(3) smoothing F0 and correcting obvious outliers;
(4) converting pitch contours into tone values using the five-degree tone system;
(5) semi-manual cleaning and correction of automatically generated tone values;
(6) removing non-reduplicated or non-kinship items and tokens with clearly abnormal pronunciation.

Methods

1. F0 Extraction

F0 is the fundamental frequency of the voice, it is the number of times the vocal folds vibrate per second. The higher the perceived pitch, the larger the F0. Tone research must rely on F0, because a tone is essentially a curve of pitch over time.

Technically, I used the parselmouth interface to Praat to automatically extract F0 from each syllable. The specific steps are:

(1) Read in the audio file and its corresponding TextGrid.
(2) For the time interval of each syllable, compute the F0 contour over that interval.
(3) Set appropriate pitch floor and pitch ceiling values (based on the female speaker's pitch range) to avoid mistaking harmonics or noise for the fundamental.
(4) Smooth the resulting F0 contour, removing sudden jumps and zero values.
(5) Within each syllable, select several representative sampling points (for example, onset, midpoint, offset), and compute the average or overall trend of these points to convert them to five-degree tone values.
(6) Put simply, the computer first draws a "pitch polyline" for each syllable, from which I derive tone values from these polylines.

2. Tone Labeling

Because automatic algorithms are relatively weak at recognizing Guiyang tonal shapes, this study adopts a "manual + semi-automatic" approach:

(1) Citation tone: all disyllabic characters are labeled manually based on the speaker's intuition and the literature, to ensure that the tone of each character when read in isolation matches the expected citation tone at the cognitive level.
(2) Surface tone: F0 is first converted into 1–5 tone heights using the T-value method and the five-degree system; then the tonal pattern is automatically classified based on the contour shape, and clearly unexpected tokens are manually corrected.

In determining the final standard citation tone values of the monosyllables, the following process was employed:

(1) First, I grouped the characters into four tone categories based on the speaker's metalinguistic awareness.
(2) For each group, I examined the distribution of their five-degree tone values.
(3) If the tone values of all characters in a group were very similar, I directly took the most common pattern in that group as the citation tone value.
(4) If there were different patterns within the same group (for example, some characters showed 35 and others were closer to 34), I first chose the most frequent pattern (mode) as the representative tone value. If there were multiple modes (for example, 35 and 34 occurring with equal frequency), I preferentially chose the pattern whose contour shape was more typical and whose change was more salient, so that the representative value would better match perception.

Using this procedure, I finally determined the four citation tones as:

- Tone 1: 35
- Tone 2: 31
- Tone 3: 44
- Tone 4: 14

3. Tone Sandhi Dataset Construction

After confirming tone values and completing tone labeling, I filtered the original table to obtain the subset truly used for sandhi analysis, with the following criteria:

- keep only kinship terms;
- keep only AA reduplication structures;
- keep only syllables with index = 1 (A1) and 2 (A2);
- remove tokens with poor recording quality or unclear boundaries.

In the resulting dataset, each syllable is annotated with:

- the character (e.g., “爸”, “妈”);
- the word (e.g., “爸爸”);
- index (1 or 2);
- citation tone (1–4);
- surface tone (1–4).

4. Exploratory Analysis

Three types of figures are plotted to observe sandhi patterns.

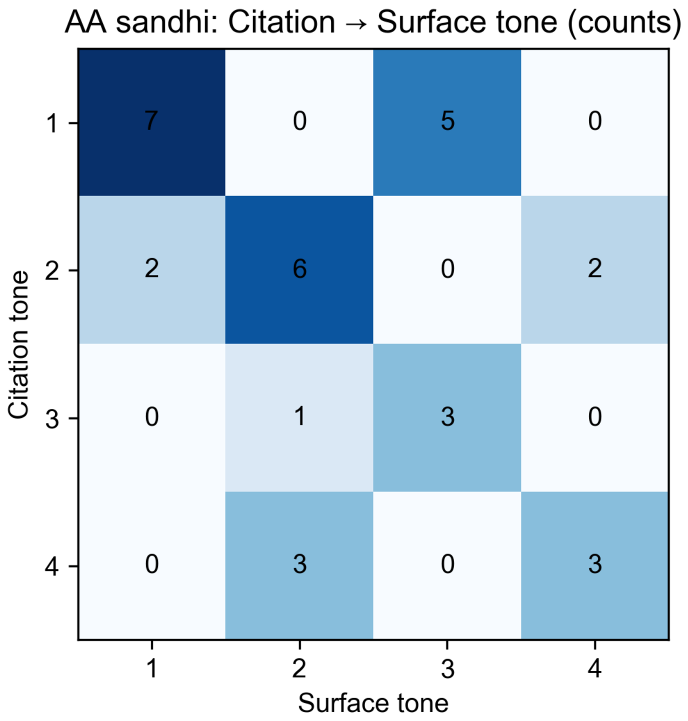


Figure 1: Citation → Surface transition matrix (heatmap)

From the heatmap we can see that:

citation 1 (35) mostly stays as 1 or changes to 3 (higher tones);

citation 2 (31) mostly changes to surface tone 2 (mid tone);

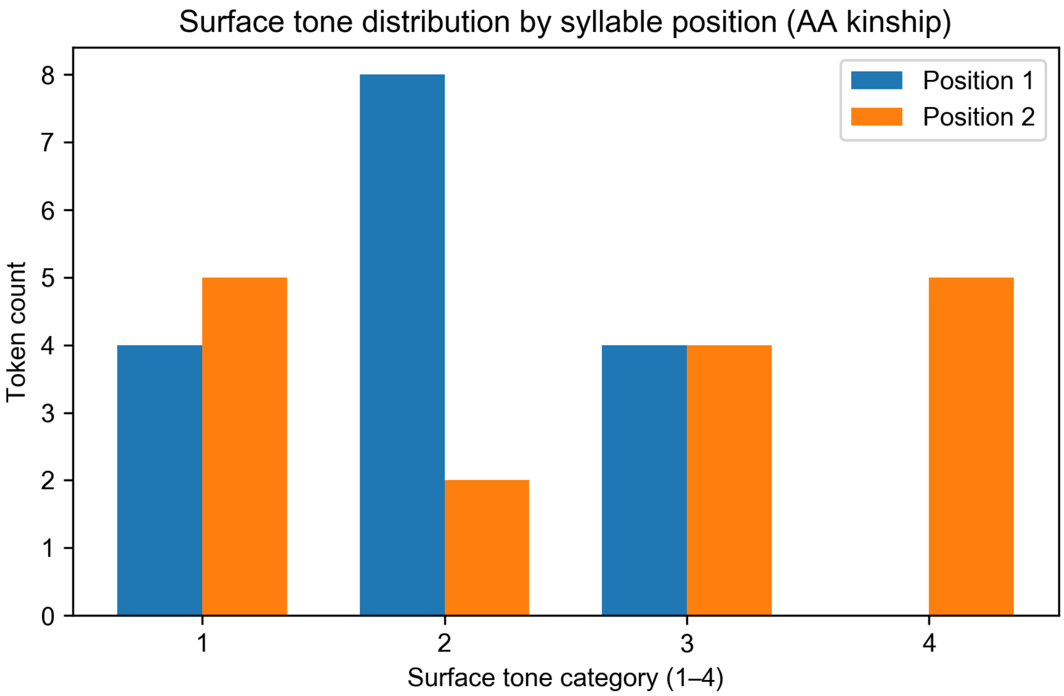


Figure 2: Surface tone distribution for A1 vs A2

This figure shows that:

- A1 tones almost always keep their citation tones;
- the overall distribution of A2 tones is clearly lower, with more tokens concentrating in Tone 1 and Tone 2;
- this indicates a tendency for tone lowering and neutralization in A2 position in Guiyang AA reduplications, suggesting that A2 is a relatively weak position.

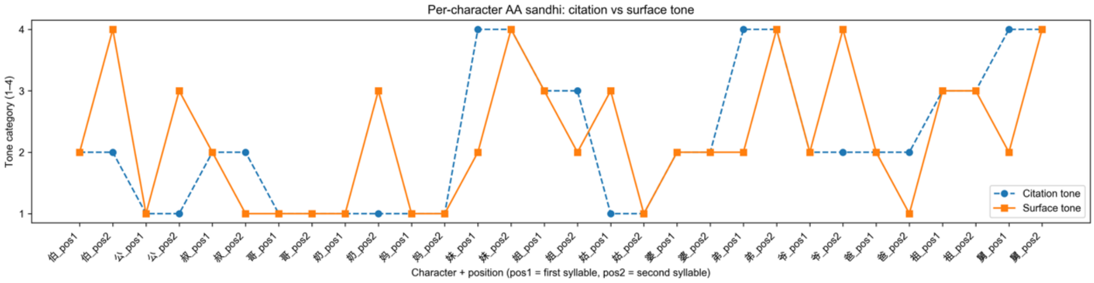


Figure 3: Per-kinship word A1/A2 tone comparison

This per-character plot shows that:

- for almost all words, A1 is very stable as its citation tones veer close to the citation tone of the character;
- for almost all words, A2 shows clear tone changes with Tone 2 and Tone 3 in particular being more likely to be lowered;
- for a small number of characters (such as “弟” and “婆”), the surface tone differs across meanings (e.g., 弟弟 vs 弟 3, 婆婆 vs 婆 3/婆 4), suggesting possible semantic-conditioned tone sandhi.
- These exploratory findings provide important clues for the subsequent probabilistic modeling.

5. Probabilistic Tone Sandhi Model

The model used in this study is:

$$p(surface\_tone \mid citation\_tone, position)$$

For each (citation tone, position) combination, I count how many tokens correspond to each of the four surface tones in the real data, and normalize these counts to probabilities. For example: among all tokens with citation tone 2 and position A2, what proportions become surface tones 1, 2, 3, and 4?

This yields a simple but interpretable tone sandhi probability table.

There are three reasons for using such a probabilistic model:

- Guiyang sandhi is not a hard rule like “2 always becomes 1”; the same citation tone can behave differently in different words or contexts;
- A probabilistic model can distinguish between “frequent” and “occasional” tone changes, reflecting variation in real speech;
- A probabilistic model can serve as a generative model, allowing us to “let the computer speak” in later Monte Carlo simulations and generate additional possible data.

In other words, this model does not write a discrete rule of the form “if … then always …”; instead it answers: “Under these conditions, what is the most common outcome, and what are the probabilities of the other outcomes?”

6. Monte Carlo Simulation

Using the probability model above, I randomly generated 5,000 “hypothetical” AA reduplicated surface tones. The specific steps are:

(1) Randomly draw a citation tone (1–4) and a position (A1 or A2);
(2) According to the probability distribution in the corresponding row, randomly draw one of the four possible surface tones;
(3) Repeat this process 5,000 times to obtain a “simulated corpus.”

The simulated data are therefore not a simple copy of the original data, but new samples freely generated based on the model.

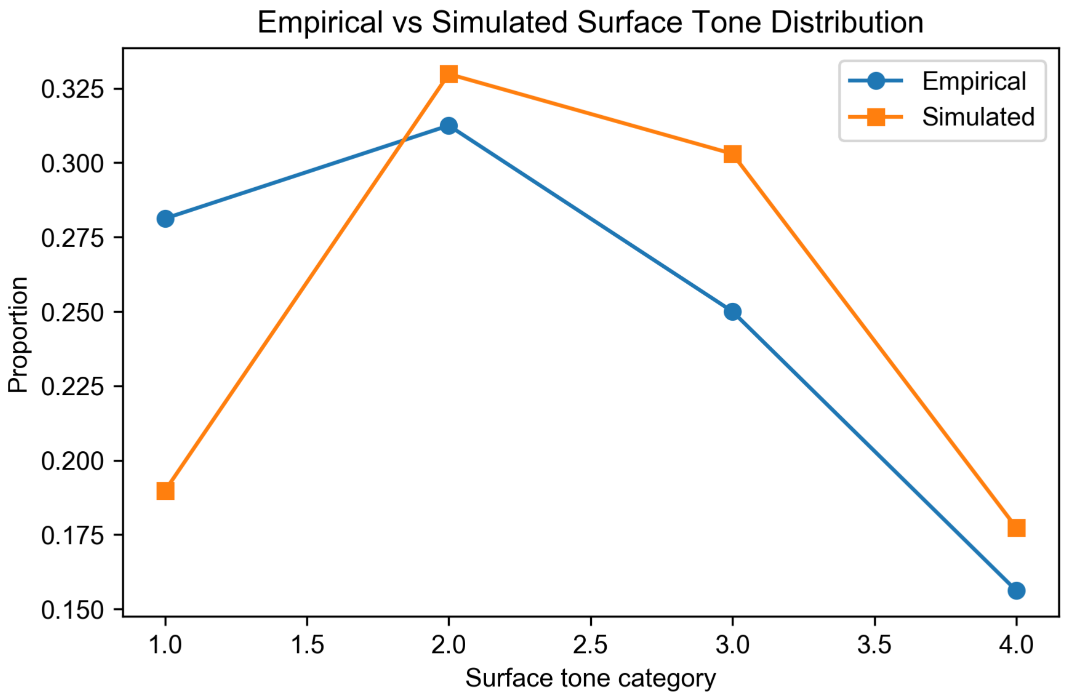
The Monte Carlo simulation has two main functions:

Model checking: If the overall distribution of simulated tones is very different from the real data, the model has not captured the key properties; if they are similar, the model’s characterization of the system is reasonable.

Smoothing small-sample noise: The original dataset is small, and some extreme cases might be accidental. Large-scale simulation lets us see the stable tendencies that would appear under “infinitely many repetitions of the experiment.”

7. Model Evaluation

I plotted the surface tone distributions of the real data and the simulated data on the same graph:



- The two curves have similar overall shapes: the proportion of Tone 2 is highest, followed by Tones 1 and 3, with Tone 4 being the least frequent;
- The simulated results are slightly smoother in the mid tones (Tones 2 and 3), which is the natural effect of converting sparse counts to probabilities and then drawing many samples;
- This shows that the simple probabilistic model has successfully captured the main tendencies of tone sandhi in Guiyang AA reduplications.

The differences mainly stem from:

- the small dataset size, with some (citation, position) combinations having very few tokens;
- particular words (such as “婆”, “弟”) that may undergo semantic-conditioned sandhi, whose behavior has a larger impact on the overall distribution when the sample is small;
- the model includes only coarse features (citation tone and position) and thus can only fit macro-level patterns, not more detailed effects such as lexical identity or speech rate.

Overall, under the current data size, this model is reasonable, interpretable, and has passed the simulation-based check.

Discussion

This study reveals several interesting patterns in Guiyang kinship AA reduplications:

(1) Stable A1, lowered A2: A1 largely preserves the citation tone, while A2 is more often lowered or neutralized;

(2) Neutralization of Tones 2 and 3 in A2: these mid or higher tones are easily pulled down to mid or even lower levels in the weak position;

(3) Different tone values for some characters across meanings (e.g., "婆"): this suggests that tone sandhi may be related to lexical meaning or grammatical role, not just phonological environment;

(4) Clear probabilistic nature of sandhi: syllables in the same category do not always produce a single deterministic outcome. Rather, this variability is quantified through the probabilistic model.

However, the limitations of the model include:

- small dataset with only one speaker, which cannot represent the whole Guiyang area;
- automatic F0 extraction is still affected by noise, recording conditions, and boundary labeling;
- the five-degree tone algorithm can be biased when boundaries are unclear or there is strong creaky voice, requiring further manual checking;
- the model uses only citation tone and position as features and does not incorporate other potentially important variables such as lexical identity, speech rate, or context.

Future Work

(1) Expand the corpus: create a more diverse dataset through the inclusion of more speakers and more lexical types to ultimately test whether the patterns found here hold at the community level;

(2) Improve F0 extraction and tone-value algorithms: strengthen smoothing and outlier detection to reduce potential human-error with manual correction;

(3) Explore semantic-conditioned sandhi: systematically compare forms like 婆 3 vs 婆 4, 弟 3 vs 弟 4 to determine whether semantic roles systematically influence sandhi;

(4) Develop more complex models: on top of the current probabilistic model, try

HMMs, neural networks, and other methods to explore whether finer-grained patterns can be captured;

(5) Compare with other Southwestern Mandarin dialects: such as Chengdu, Chongqing, Zunyi, etc., to situate Guiyang tone sandhi within the broader Southwestern Mandarin landscape.

Conclusion

Through acoustic analysis, tone-value conversion, probabilistic modeling, and Monte Carlo simulation, this study provides an initial description of tone sandhi in Guiyang kinship AA reduplications. The results show that A2 has a clear tendency toward tone lowering and neutralization, and that tone sandhi is better described using probabilistic methods than traditional “hard rules.”

Although the dataset is small, this project demonstrates the feasibility of applying statistical methods to dialect tone research and offers a reusable technical pipeline for future, more systematic studies of Guiyang tone.

References

Xu, X. (2011). An Introduction to Phonetics and Phonology.

Duanmu, S. The Phonology of Standard Chinese.

Tonal Sandhi Patterns Across Chinese Dialects.

Li, R. & Wang, P. (1994). Guiyang Dialect Dictionary. Jiangsu Education Press.

Bei, X. (2012). “Tone patterns and vowel patterns in Mandarin.” Wuling Journal, 131–136.

Chen, D. (2013). “Phonological variation in Guiyang Mandarin.” Journal of Guizhou Normal College, 92–99.

Luo, R. (2018). “Acoustic study of tone values and tone length in Guiyang Mandarin.” Journal of Guizhou Institute of Engineering, 63–67.

Shi, F. (2002). “The vowel pattern of Beijing Mandarin.” Nankai Linguistics, 30–36.

Shi, F. (2010). “On phonological patterns.” Nankai Linguistics., 1–14.

Tu, G. (1982). “Comments on ‘The Phonetic System of Guiyang Dialect’.” Dialect, 229–233.

Tu, G. (1987). “Noun reduplication in Guiyang.” Dialect, 202–204.

Wang, P. (1981). “The phonetic system of Guiyang dialect.” Dialect, 122–130.