Top-down and Bottom-up Processing of Familiar and Unfamiliar Mandarin Dialect Tone Systems

Liang Zhao[1]*, Shayne Sloggett*[1]*, and Eleanor Chodroff*[1]

[1]University of York, Department of Language and Linguistic Science

liang.zhao@york.ac.uk, shayne.sloggett@york.ac.uk, eleanor.chodroff@york.ac.uk

Abstract

Speech processing involves active integration of bottom-up and top-down information types. In the present study, we investigated the relative weighting of top-down expectedness and bottom-up lexical tone in the perception of familiar and unfamiliar lexical tone systems. Standard Mandarin and Chengdu Mandarin are mutually intelligible language varieties with comparable segmental and highly distinct tonal realizations. In a spoken semantic-plausibility judgment task, we manipulated whether a word was high-surprisal or low-surprisal given the preceding context and dialect-specific tone. All participants were native Standard Mandarin speakers with minimal Chengdu Mandarin experience. Lower judgment accuracy was observed when the stimulus was Chengdu Mandarin, and suggested that expectedness (i.e., top-down) information overrides tonal (i.e., bottom-up) information in sentence plausibility judgments. However, judgment response times to sentence surprisal were uniform across stimuli from both dialects, suggesting that speakers are aware of the surprisal conveyed by a non-standard tone, even if not used in their final decision. These findings reveal listener sensitivity to both top-down expectedness and bottom-up tone regardless of the initial tone reliability. For unfamiliar tone systems, top-down influence overrides bottom-up processing to access utterance meaning, but bottom-up processing is indeed present and may reflect rapid learning of the unfamiliar tone system.

Index Terms: lexical tone, tone perception, lexical access, speech perception, Mandarin dialects

1. Introduction

In addressing how the speech signal is processed to decode the intended utterance, several prominent theoretical frameworks have identified two high-level mechanisms: top-down processing and bottom-up processing. Early models of speech perception, such as The Cohort Model [1, 2], Direct Perception [3] and Direct Realism [4] have often assumed a privileged role of bottom-up information in the perceptual system. However, subsequent extensions of these theories have taken the influence of top-down information into consideration (e.g., TRACE [5], Acoustic Landmarks and Distinctive Features [6]). Others have eschewed any role for top-down processing (e.g., Shortlist [7]; Merge [8]). Current models tend to incorporate both top-down and bottom-up processes in speech perception, but the relative weighting and integration of these sources of information remains unclear.

Moreover, previous work on speech perception has been historically segment-oriented and concentrated on non-tonal languages, leaving the mechanisms for tonal speech perception relatively under-investigated. For tonal languages such as Mandarin dialects, lexical tones are crucial to differentiate the meanings of lexical items. The fact that tonal information is heard and perceived alongside, and not independent from, segmental information has given rise to several studies on the relative weighting of lexical tone and segmental information for lexical access. Some have argued that segmental information is more salient than tonal information in sub-lexical processing as tonal information is accessed later or with lower accuracy than segments [9–12]. Specifically, Taft and Chen [9] reported substantial difficulty in the judgment of homophones between written Chinese characters when the characters were paired with identical phonemes and different tones. Longer latencies and lower accuracies were also found for tonal contrasts relative to segmental contrasts in word/nonword decision, same-different character judgment [10], word monitoring [11] and word reconstruction [12].

Others dispute this seemingly inferior status of tones in lexical access, and contend that lexical tones could have an equal or even greater contribution to lexical access relative to segments given appropriate top-down feedback [13, 14]. Lexical judgments on disyllabic words and idioms were equally accurate for segmental and tonal manipulations in Liu and Samuel's study [13]. Results from an eye-tracking study, in which participants matched a spoken word to an array of pictures, also indicated a comparable contribution of segmental and tonal information in lexical access [14]. More recently, the Reverse Accessing Model [15] reported a distinct advantage for segments over lexical tones, suggesting that tone information is processed only *if necessary*.

While researchers generally agree that both top-down and bottom-up information are used in tone processing, there is little consensus on the relative weighting of these sources of information and how they interact. To what extent do speaker expectations guide (or indeed, override) attendance to bottom-up segmental and pitch information? To test this, we manipulated the reliability of tonal information in high and low surprisal (lexical expectedness) sentences using natural regional variation among Mandarin dialects. Mandarin dialects provide a natural testbed for research on tonal speech perception due to their comparable segmental inventories, but distinct tone systems. We focus on Standard Mandarin and Chengdu Mandarin. Both dialects have a four-tone system and the same mapping between phonological tone category and lexical category. They differ, however, in the phonetic implementation of each phonological tone category: Standard Mandarin has tone 1 (Chao tone numerals: 55), tone 2 (35), tone

3 (215), and tone 4 (51), whereas Chengdu Mandarin has tone 1 (55)[1], tone 2 (21), tone 3 (53) and tone 4 (213) [16].

For the familiar tone system (Standard Mandarin), we expect speakers to use both top-down and bottom-up information, as the sentential context and tone representations are both reliable cues for listeners. For the unfamiliar tone system (Chengdu Mandarin), we expect the dominance of top-down information from sentential context, and little or no use of lexical tone due to the unfamiliarity of the tone system. Our findings suggest that speakers seem to attend to tone even if they do not always use it in determining word identity. When tone information was reliable, speakers correctly detected semantic implausibility, suggesting that they attend to tone even when the context introduces a strong bias against a particular lexical item. However, when tone information was unreliable, they relied on the sentential context in making their decisions. Interestingly, in both cases, response times were longer for sentences containing tones which would increase sentence surprisal, indicating that speakers are sensitive to tone even if they do not use it in determining lexical identity.

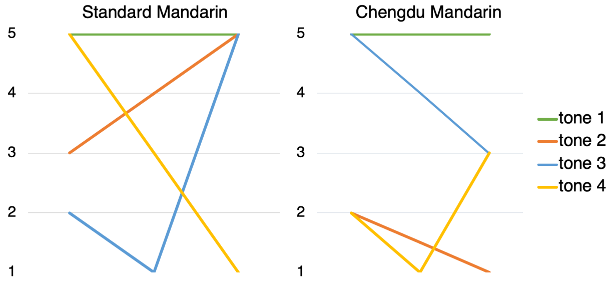

Figure 1: *Schematic tone contours of Standard Mandarin and Chengdu Mandarin.*

2. Methods

A 2×2 factorial design was used to assess the effects of sentence semantic plausibility (high surprisal vs. low surprisal) and dialect familiarity (native Standard Mandarin vs. non-native Chengdu Mandarin) on the accuracy of semantic plausibility judgments and response times.

2.1. Participants

Twenty-one native speakers of Standard Mandarin who reported little or no knowledge of Chengdu Mandarin participated in the experiment. No participant reported hearing or reading impairments.

2.2. Materials

24 sentences were created manipulating Mandarin Dialect in a between-item design (12 Standard Mandarin and 12 Chengdu Mandarin sentences). Within these sentences, the lexical tone of a critical word was manipulated resulting in either a semantically plausible (low surprisal) or a semantically implausible (high surprisal) sentence. Participants heard different sets of sentence items in Standard Mandarin (native dialect) and Chengdu Mandarin (non-native dialect) trials. Half the critical words were sentence-medial and half were sentence-final. Tone combinations were counterbalanced across items.

Table 1 gives an example pair of high and low surprisal sentences, presented in *Pinyin* symbols with its tone category—tone 1, tone 2, tone 3, and tone 4. The phonetic tone realizations of these words in Chengdu Mandarin are considered unknown or unfamiliar to speakers of Standard Mandarin (see Figure 1). Surprisal was manipulated by altering the tone of a critical word in which the segments were rendered intact, but were paired with different tones. In the example here, /fei1/ (plausible: *"There is an eagle in the sky flying"*) contrasted with /fei2/ (implausible: *"There is an eagle in the sky gaining weight"*). Participants heard both renditions of each sentence for a total of 48 trials.

The Standard Mandarin stimuli were produced by a female native speaker of Standard Mandarin (aged 26) and Chengdu Mandarin stimuli were produced by a male native speaker of Chengdu Mandarin (aged 29). A 10ms-silence was inserted at the beginning of each sentence, and the audio file was scaled to an intensity of 70dB.

Table 1: *An example sentence item across surprisal conditions.*

low-surprisal sentence	a) 有 一只 鹰 在 天上 飞 *You3 yi4 zhi1 ying1 zai4 tian1 shang4 fei1* *There is an eagle in the sky flying* "There is an eagle flying in the sky"
high-surprisal sentence	b)* 有 一只 鹰 在 天上 肥* *You3 yi4 zhi1 ying1 zai4 tian1 shang4 fei2** *There is an eagle in the sky gaining weight** "There is an eagle gaining weight in the sky"

2.3. Procedure

The experiment was run online using Gorilla Experiment Builder [17]. Participants were asked to complete the sentential semantic plausibility judgment task on a device with internet access in a quiet environment, and with headphones if possible. They were first briefed on the purpose and content of the experiment. The participants were made aware that they would be listening to sentences spoken in either Standard Mandarin or another Mandarin dialect. Then they were presented with a test audio and adjusted the volume of the sound output to a comfortable level.

In the practice phase, participants listened to two example pairs of high-surprisal and low-surprisal sentences in Standard Mandarin, answered the question "*Does this sentence make sense*", and then received feedback regarding their answer in the form of a written version of the sentence. Specifically, participants were instructed to click the "play" button to start the audio and then click on the "yes" or "no" button on the screen to answer the question. The correct answer and the sentence in standard simplified Chinese characters were then presented. The "yes" and "no" buttons were presented closely adjacent to each other at the center of the screen with the "yes" button on the left side and the "no" button on the right side.

In the test phase, the presentation of trials was fully randomized. The procedure was identical to the familiarization stage except that no feedback was provided.

[1] F0 measurement and perception of Chengdu dialect recordings suggest that Chengdu tone 1 is more likely a high-rising tone (45) than a high-flat tone (55), as in Standard Mandarin [18].

2.4. Data analysis

Accuracy and response time from the test phase were analyzed as dependent variables across manipulations of dialect (Standard Mandarin vs. Chengdu Mandarin) and surprisal (high-surprisal vs. low-surprisal).

“Yes” responses to low-surprisal (i.e., plausible) sentences and “No” responses to high-surprisal (i.e., implausible) sentences were coded as “correct” responses. Response time was calculated as the interval between the end of the audio file and the click registering a judgment. Five trials were excluded from the analysis due to missing or negative response times, likely due to internet connectivity issues. One sentence pair in each dialect was also omitted due to experiment error. This left a total of 919 trials (range of 43–44 trials per participant) for analysis.

Accuracy was modeled with a Bayesian logistic mixed-effects regression, and response time with a Bayesian log-normal mixed-effects regression, both with weakly informative priors [19]. Each model included fixed effects of surprisal, dialect, trial number, and the full set of interactions. The random effect structure for participant included an intercept and slopes for surprisal, dialect, trial number and the interaction between surprisal and dialect, and for sentence frame, an intercept and random slope for dialect. Priors for main effects and interactions were Normal distributions centered on 0 with a standard deviation of 20 for the accuracy model and Normal distributions centered on 0 with a standard deviation of 1 for the response time model. The prior for the intercept was $\mathcal{N}(0, 20)$ for accuracy and $\mathcal{N}(7, 1)$ for response time. The model was run for 2000 iterations with a burn-in period of 1000 iterations. Surprisal and dialect were sum-coded (*surprisal:* high-surprisal = 1, low-surprisal = − 1; *dialect:* Chengdu = 1, Standard Mandarin = −1), and trial number was centered on the mean. If the 95% credible interval for an estimated effect excluded 0 (i.e., no effect), then it was deemed to be *credible* in its direction of influence on the respective dependent variable.

3. Results

3.1. Accuracy in semantic plausibility judgment

As shown in Figure 2, accuracy was near ceiling for both surprisal conditions in Standard Mandarin (high: 98%, low: 92%), but differed considerably by surprisal in Chengdu Mandarin (high: 20%, low: 94%). For the familiar speech (Standard Mandarin), the overall high accuracy suggests that participants understood the task in general, validating the plausibility of the surprisal manipulation in the experiment.

Correspondingly, the model revealed credible main effects of surprisal, dialect and the interaction between surprisal and dialect on accuracy. Specifically, accuracy was higher in the low-surprisal condition than in the high-surprisal condition (*surprisal:* β = −2.21, 95% CI = [−3.43, −1.34]). In addition, accuracy was higher for sentences spoken in Standard Mandarin than in the Chengdu dialect (*dialect:* β = −1.96, 95% CI = [−2.86, −1.22]). A credible interaction was also observed between surprisal and dialect, indicating an even lower accuracy for sentences spoken in the Chengdu dialect in the high-surprisal condition (*surprisal x dialect:* β = −1.00, 95% CI = [−1.88, −0.09]). Trial number and its interactions with surprisal and dialect were not reliable in the direction of their effects, indicating that accuracy did not reliably improve in any condition across the course of the experiment (*trial:* β = 0.03, 95% CI = [−0.01, 0.07], *trial x surprisal:* β = 0.02, 95% CI = [−0.01, 0.05], *trial x dialect:* β = −0.01, 95% CI = [−0.04, 0.02], *trial x surprisal x dialect:* β = −0.01, 95% CI = [−0.04, 0.02]).

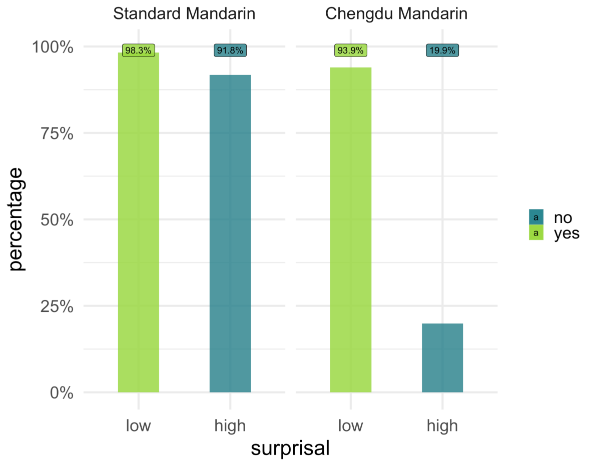

Figure 2: *Percentage of “correct” responses across dialect and surprisal conditions. “Yes” (plausible; lime green) is treated as correct for low-surprisal conditions, and “no” (not plausible; blue green) for high-surprisal conditions.*

3.2. Response time

The distributions of participant-specific response times for each condition are presented in Figure 3. Reliable main effects were observed for surprisal, dialect, and the interaction between surprisal and dialect. Response times were reliably slower for high-surprisal than low-surprisal sentences (*surprisal:* β = 0.21, 95% CI = [0.13, 0.30]); they were also slower for sentences spoken in Chengdu Mandarin than in Standard Mandarin (*dialect:* β = 0.13, 95% CI = [0.02, 0.25]). The interaction between surprisal and dialect also reliably modulated the contrast in response times between high and low surprisal conditions within each dialect: this difference was enhanced for Standard Mandarin, and slightly diminished for Chengdu Mandarin (*surprisal x dialect:* β = −0.13, 95% CI = [−0.21, −0.05]). Notably, the magnitude of the main surprisal effect exceeded its interaction with dialect, indicating listener sensitivity to the high-low surprisal contrast even in the unfamiliar Chengdu Mandarin. Based on the transformed marginal means, the estimated mean difference between high and low conditions for Chengdu Mandarin was approximately 297 ms, whereas for Standard Mandarin it was about 875 ms. While the surprisal effect was substantially larger for Standard Mandarin than Chengdu Mandarin, high surprisal nevertheless led to reliably longer response times in both dialects.

The remaining effects of trial and its interactions with surprisal and dialect were not reliable in their direction of influence (*trial:* β = 0.0032, 95% CI = [−0.0005, 0.0067]; *trial x surprisal:* β = −0.0006, 95% CI = [−0.0025, 0.0014]; *trial x dialect:* β = 0.0007, 95% CI = [−0.0013, 0.0027]), except for the interaction between trial, surprisal and dialect (*trial x surprisal x dialect:* β = 0.0021, 95% CI = [0.0001, 0.0041]). Though response times decreased in the Standard high-surprisal condition, particularly in the initial trials, the marginal means indicate that the interaction is driven by a reliable slowdown in the Chengdu high-surprisal condition over the course of the experiment.

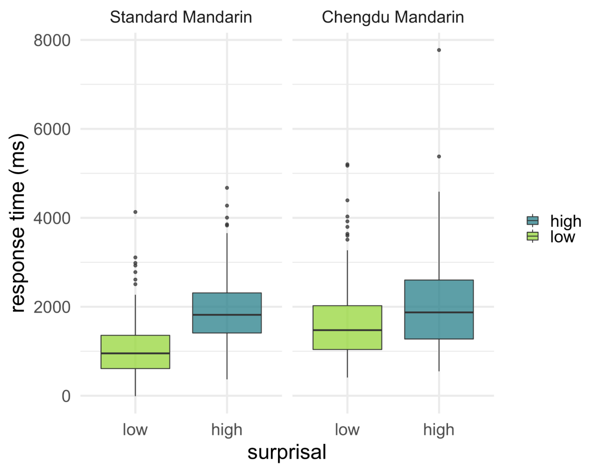

Figure 3: *Response times across dialect and surprisal conditions.*

4. Discussion

The present study investigated the relative weighting of top-down and bottom-up information in processing familiar and unfamiliar tone systems. For the familiar tone system, accuracy results suggested that listeners have strong representations of segments and tones and thus were able to use this bottom-up information, together with sentential context, to achieve high accuracy in the semantic plausibility judgment task. Response time results also suggested listeners' sensitivity to the surprisal manipulation using both bottom-up and top-down information.

For the unfamiliar Chengdu speech, accuracy results suggested an overriding effect of top-down information in determining sentence meaning as the listeners' judgments were overwhelmingly biased towards semantically plausible sentences based on the sentential context alone, despite any mismatch in tone. Low accuracy for the high-surprisal Chengdu sentences indicates a major bottom-up failure in identifying unexpected tones of the unfamiliar tone system. However, the overall lower accuracy for Chengdu speech compared with Standard Mandarin does not denote difficulty in understanding Chengdu Mandarin in general. Listeners consistently understood Chengdu speech well enough to correctly judge plausible sentences; they simply under-valued tone information in high-surprisal environments. For any discrepancies between observed and expected tones in the unfamiliar tone system, sentential context (i.e., top-down information) overwhelmingly guided lexical access towards a plausible judgment.

With respect to response time, a slowdown in response times in the high-surprisal condition was present to a reliable degree in both Standard Mandarin and Chengdu Mandarin. Though the magnitude of the surprisal effect was indeed greater for Standard than Chengdu Mandarin, the presence of the effect revealed listeners' sensitivity to the implausibility indicated by a high-surprisal tone in Chengdu speech. This suggests that listeners indeed attend to the bottom-up tone information, even in unfamiliar systems, despite their ultimate bias towards a response of semantic plausibility in the Chengdu Mandarin condition.

Though top-down information dominates lexical access in unfamiliar speech, differences in response times across surprisal conditions indicates an unexpected integration of bottom-up information. This contrasts with Gao et al. [15]'s proposal that tone information is processed only if necessary. In our study, the listeners were exposed to a cumulative one minute of unfamiliar Chengdu speech, but they seemed capable of retuning the tone category–contour mapping to the extent that the measured response times were reliably different between surprisal conditions across all trials.

One plausible interpretation is that listeners extract lexical tone information even in unfamiliar speech given exposure to the full utterances. This information could then be used to update the mapping from phonological tone category to its corresponding phonetic realization. It is unclear whether listeners build long-lasting representations of the dialect- or talker-specific tone category–contour mapping, or only temporary, task-specific representations. Whether this online adaptation persists over time and might constitute learning remains to be seen, but the awareness of the surprisal contrast, as suggested in the response time data, indicates some degree of online adaptation.

With regard to semantic plausibility judgment, listeners may have been less confident about the tone acoustics of the unfamiliar speech, and thus top-down information overrode the output of tone-level processing to access the sentence meaning. Although listeners failed to report the tone mismatch for the high-surprisal sentences in the unfamiliar speech, they somehow constructed tone representations using bottom-up information and responded differently in terms of response time.

Moreover, for potential learning of the unfamiliar tone system, the reliable slowdown over the course of the experiment in the high-surprisal Chengdu condition suggests gradually raised attention and awareness of bottom-up information. However, the lack of credibility of trial and its interactions with dialect or surprisal suggests a consistent slow response time in the high-surprisal conditions for both familiar (Standard) and unfamiliar (Chengdu) speech. This indicates that listeners may be very rapidly learning or adapting to a novel tone category–contour mapping, possibly as soon as the experiment commenced.

5. Conclusion

The present study tested the relative role of tonal information for sentence interpretation in two Mandarin dialect varieties. We found that contrary to previous suggestions, phonetic tonal information seems to be processed, even when its mapping to the phonological tone categories is unfamiliar. Further research is needed to address dialect- and tone-specific perception with carefully balanced tone contrast for a broader range of Mandarin group dialects other than Chengdu Mandarin. The current experimental design with the surprisal and dialect manipulations could also be expanded by introducing an exposure phase to the unfamiliar dialect to explore the potential perceptual adaptation to or learning of unfamiliar tone systems.

6. Acknowledgements

We would like to thank Paul Foulkes and Phil Harrison for helpful discussion, as well as the audiences at PARLAY 2021 and the University of York Psycholinguistics Lab for their questions and feedback.

7. References

[1] Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. Cognitive psychology, 10(1), 29-63.

[2] Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. Cognition, 25(1-2), 71-102.

[3] Gibson, J. J. (1954). A theory of pictorial perception. Audiovisual communication review, 2(1), 3-23.

[4] Fowler, C. A. (1986). An event approach to the study of speech perception from a direct–realist perspective. Journal of phonetics, 14(1), 3-28.

[5] McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. Cognitive psychology, 18(1), 1-86.

[6] Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. The Journal of the Acoustical Society of America, 111(4), 1872-1891.

[7] Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. Cognition, 52(3), 189-234.

[8] Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. Behavioral and Brain Sciences, 23(3), 299-325.

[9] Taft, M., & Chen, H. C. (1992). Judging homophony in Chinese: The influence of tones. In Advances in psychology (Vol. 90, pp. 151-172). North-Holland.

[10] Cutler, A., & Chen, H. C. (1997). Lexical tone in Cantonese spoken-word processing. Perception & Psychophysics, 59(2), 165-179.

[11] Ye, Y., & Connine, C. M. (1999). Processing spoken Chinese: The role of tone information. Language and cognitive processes, 14(5-6), 609-630.

[12] Wiener, S., & Turnbull, R. (2016). Constraints of tones, vowels and consonants on lexical selection in Mandarin Chinese. Language and speech, 59(1), 59-82.

[13] Liu, S., and Samuel, A. G. (2007). The role of Mandarin lexical tones in lexical access under different contextual conditions. Lang. Cogn. Process. 22, 566–594.

[14] Malins, J. G., & Joanisse, M. F. (2010). The roles of tonal and segmental information in Mandarin spoken word recognition: An eyetracking study. Journal of Memory and Language, 62(4), 407-420.

[15] Gao, X., Yan, T. T., Tang, D. L., Huang, T., Shu, H., Nan, Y., & Zhang, Y. X. (2019). What makes lexical tone special: A Reverse Accessing Model for tonal speech perception. Frontiers in psychology, 10, 2830.

[16] 李荣 (Li, Rong.). (2002). 现代汉语方言大词典 (The Modern Dictionary of Chinese Dialects). 江苏教育出版社 (Jiangsu Education Press).

[17] Anwyl-Irvine, A.L., Massonié J., Flitton, A., Kirkham, N.Z., Evershed, J.K. (2020). Gorilla in our midst: an online behavioural experiment builder.

[18] Zhao, L., & Chodroff, E. (2022). ManDi: Mandarin Chinese Dialect Corpus. Retrieved from https://osf.io/69fx5/.

[19] Bürkner, P. C. (2018). "Advanced Bayesian Multilevel Modeling with the R Package brms." The R Journal, 10(1), 395–411. doi:10.32614/RJ-2018-017.