There materials include:

1) *Additional details on participants*
2) *Additional discussion regarding the frequency of unsystematic tone errors in L2 speech*
3) *Additional details regarding stimuli*
4) *Additional details about procedures*
5) *Additional details about statistical models*
6) *Exploratory analyses of adaptation over the course of the experiment*
7) *Additional results of post-experiment questions*
8) *Note about Chinese language history questionnaire*
9) *Stimuli for critical trials*

1. Additional details on participants

Excluded participants: Two participants who completed the task were replaced due to scoring <80% accuracy on critical unrelated trials, a third was replaced for failing to cooperate with instructions.

*Contact with L2 speaker*s: A post-experiment survey indicated most participants considered themselves to have little experience speaking to non-native Mandarin speakers, with responses as follows: 50 people indicated "very rarely", 12 "relatively rarely", 11 "occasionally", 6 "relatively often" and 1 "very often".

Mandarin language: Though all listeners identified Modern Standard Mandarin (*Pǔtōnghuà* 普通话) as their native language (*mǔyǔ* 母语), over half (45 out of 80) also indicated that they often spoke one or more regional dialects. We chose not to be strict in this regard, as we wanted to generalize beyond purely monolingual Mandarin speakers. When accounting for regional dialects of Mandarin—common across northern and southwest China (cf. Ramsey, 1987)—the subset of strictly 'monodialectical' Mandarin speakers is small and not representative of most Chinese people with whom typical L2 speakers interact.

2. Additional discussion regarding the frequency of unsystematic tone errors in L2 speech

Here we address the nature and frequency of L2 tone errors in more detail. As noted in the main text, numerous studies have provided evidence of the frequency of tone errors in L2 speech within carefully controlled experiments (e.g., reading words or sentences from prompts). There are several factors that likely contribute to the frequency of tone errors. They include difficulty with coarticulation of tones in disyllabic words (Hao, 2018), inaccurate pedagogical descriptions of tones (He et al., 2016; H. Zhang, 2014), interference from L1 prosody (Yang, 2016; Yang & Chan, 2010), and gaps in L2 speakers' memory of tones (Pelzl, 2018). Because of the controlled elicitation methods used in most previous studies, they seem likely to underestimate the frequency of tone errors, as one of the major sources of errors (gaps in memory) are not relevant. However, the one study we are aware of that analyzed tone errors in relatively spontaneous L2 speech (Winke, 2007, p. 34), reports numbers that are surprisingly low (roughly 12% tone errors overall) given that participants were novice learners. This seems to be at odds with the higher error rates found with more controlled elicitation methods (e.g., Chen et al., 2016), as well as the anecdotal experience of teachers and students themselves. In short, more research is needed to better understand how prevalent tone errors are in L2 speech at various proficiency levels.

While we do not have precise estimates of the prevalence of unsystematic tone errors, Pelzl's (2018) results suggest even advanced learners have incomplete or incorrect tone knowledge for as much as 20% of the vocabulary they know. For less proficient learners, this percentage could be even higher. These words will, by definition, be produced in an unsystematic fashion, as each individual L2 speaker will vary in the errors they make and the consistency of those errors (e.g., if a person does not know a word's tones, they might randomly vary in producing it each time the word comes up). It is conceivable learners also resort to some sort of 'default' tone for unknown items, but to our knowledge no research indicates this to be the case. It would add yet another layer of complexity for listeners trying to find patterns in L2 tone errors.

In summary, while there is plenty of reason to believe unsystematic errors are common in L2 tone production, an empirical study of their frequency has yet to be conducted. We acknowledge that, if unsystematic errors are very infrequent, this would reduce the ecological validity of the current study. Given our results, a lower frequency in the occurrence of such errors would make an (indirect) effect even less likely.

3. Additional details regarding stimuli

Primes: Both sets of critical primes had three words for each of the possible two-syllable tone combinations (Tone 1+Tone 1, Tone 1+Tone 2, etc.).

No initial syllables were repeated between contextualizing primes and critical primes, but we did not control repetition between the contextualizing primes themselves. Because of the large number of nouns needed, and natural asymmetries in the distribution of tone frequencies in the Mandarin lexicon (see Duanmu, 2007, p. 253), it was also not possible to have equal distribution of each of the four tones across the contextualizing primes, but we did achieve a rough balance in the occurrence of each tone in the two sets of contextualizing stimuli (Set 1: 19% T1, 28% T2, 9% T3, 45% T4; Set 2: 18% T1, 27% T2, 10% T3, 46% T4).

Real word targets: Critical visual targets for unrelated trials utilized 48 high frequency Chinese words that share no characters with any other stimuli in their set (and none in the contextualizing stimuli). They were balanced for frequency and paired with primes so that there was never a syllable in the prime that was also in the target.

Nonword targets: We verified that none of the nonwords occurred in the SUBTLEX-CH corpus. They were also inspected by several highly educated native Chinese speakers, and any item they thought could plausibly be a word was replaced. Finally, all contextualizing targets were checked against the critical stimuli to avoid any repetition of characters between them, though repetition between targets within the contextualizing stimuli was not avoided.

We did not attempt any strict control of character stroke counts or phonological or orthographic neighborhood density. Because critical comparisons were between conditions and all items were rotated across speakers and conditions, any item-level differences should be consistent across speakers and conditions. That is, if a word with many neighbors or complex characters would be recognized more slowly in the systematic condition, it would also be recognized more slowly in the unsystematic condition.

Creation of auditory stimuli: The L2 speakers were chosen according to two criteria. First, they had noticeably different voice quality, so that listeners could easily differentiate them from one another. Second, they had sufficient control of tones to be able to produce the stimuli accurately given our elicitation procedures.

Spoken stimuli were recorded using a Fostex DC-R302 in a sound-attenuated room using the following procedures. Each spoken item was produced by a model speaker—a proficient L2 Mandarin speaker and former Mandarin teacher—and then imitated by the experimental speaker. If the model speaker judged a production to be problematic, for example due to inaccurate tones, clear segmental errors (e.g., a /b/ produced as a /p/), or otherwise distorted (e.g., by lip-smacks or other noise), the model speaker prompted the experimental speaker to produce the item again. In this way the categorical accuracy or inaccuracy of tones was carefully controlled, but accent-shifted features of L2 pronunciation were not controlled. This approach resulted in more natural productions than if stimuli had been read from prompts, and also encouraged more similarity in speech rate between the two experimental speakers (*critical prime duration in ms:* Speaker 1 m= 844, sd=72; Speaker 2 m= 812, sd=92). Both (female) experimental L2 speakers produced all stimuli in both conditions. A third (male) L2 speaker was recorded for use in practice trials.

After recording, all items were cut from the original audio files, and intensity was normalized to 70dB using *Praat* (Boersma & Weenink, 2018). After inspection of the audio files by the first author (a former teacher of Mandarin), it was judged that the tones of some items were not accurate, or contained the incorrect type of tone error, so a second recording session (following the same procedures as the original) was held with each of the L2 speakers to elicit acceptable tokens. The final result of these procedures was a total of 480 unique audio files produced by each of the L2 speakers (i.e., a total of 960 files).

4. Additional details about procedures

E-Prime 2.0 (Psychology Software Tools, Inc.) was run on a PC running Windows XP. Audio was played through over-ear headphones (Edifier H840). All instructions were presented in spoken Mandarin or written in Chinese characters. Participants were allowed to take a self-paced break between blocks and sub-blocks.

5. Additional details about statistical models

Modeling details

Data were processed and analyzed using *R* (3.6.1) (R Core Team, 2018) and the *lme4* (1.1-21) package (Bates et al., 2015). Accuracy and response time (RT) data from 80 participants were submitted to (generalized) linear mixed effects models, using the *glmer* and *lmer* functions respectively. For accuracy, the dependent variable was accuracy (1,0), with fixed effects for condition (Error Free, Tone Error) and trial type (identical, unrelated) and their interaction. For RT models, the dependent variable was RT (continuous), with fixed effects for (Error Free, Tone Error) and trial type (identical, unrelated) and their interaction.

All models were selected starting with the most complex random effects structure, and simplifying to select the best fitting and most parsimonious model using the *step()* function of *lmerTest* (Kuznetsova et al., 2017), but retaining all fixed effects as they were of theoretical interest.

Accuracy results

A generalized linear mixed effect model provided no evidence of differences in the accuracy of decisions due to the contextualizing Error Free/Tone Error conditions, though there was a small effect of trial type, suggesting some listeners were occasionally lured into accepting target nonwords as real words.

Note: In all results below "unsys" is short for 'unsystematic' and indicates the Tone Error condition.

##

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerM
od']
 Family: binomial  ( logit )
Formula: score ~ cond * trialType + (1 | subj) + (1 | item)
   Data: criticalTrialsACC
Control: glmerControl(optimizer = "bobyqa")

     AIC      BIC   logLik deviance df.resid
  1950.2   1991.9   -969.1   1938.2     7674

Scaled residuals:
     Min       1Q   Median       3Q      Max
-10.9501   0.0663   0.1093   0.1768   1.0863

Random effects:
 Groups Name        Variance Std.Dev.
 item   (Intercept) 0.9907   0.9954
 subj   (Intercept) 0.4069   0.6379
Number of obs: 7680, groups:  item, 96; subj, 80

Fixed effects:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                    5.1800     0.3059  16.935  < 2e-16 ***
condunsys                      0.4358     0.3608   1.208    0.227
trialTypeunrelated            -1.9211     0.3362  -5.714 1.11e-08 ***
condunsys:trialTypeunrelated  -0.2088     0.3883  -0.538    0.591
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr) cndnsy trlTyp
condunsys   -0.467
trlTypnrltd -0.801  0.424
cndnsys:trT  0.436 -0.928 -0.467
```

##

Additional details of RT analyses for the indirect effect of Tone Error

Below we report full model output for main analysis of RTs (Error Free vs. Tone Error). This model aligns with that reported in Table 5 and Figure 5 in the main text. Further below we also report model results with transformed (inverse) RTs and after outliers were removed. None of these procedures had substantive effects on outcomes.

##

raw RTs

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: RT ~ cond * trialType + (cond + trialType | subj) + (1 | item)
   Data: criticalTrials
Control: lmerControl(optimizer = "bobyqa")

REML criterion at convergence: 91701.5

Scaled residuals:
    Min      1Q  Median      3Q     Max 
-3.5526 -0.5834 -0.1448  0.3653 11.1584 

Random effects:
 Groups   Name               Variance Std.Dev. Corr       
 item     (Intercept)         1304.4   36.12              
 subj     (Intercept)         6044.1   77.74              
          condunsys           1219.9   34.93   -0.50      
          trialTypeunrelated   699.9   26.46   -0.22  0.19
 Residual                    12691.2  112.66              
Number of obs: 7413, groups:  item, 96; subj, 80

Fixed effects:
                              Estimate Std. Error        df t value Pr(>|t|)    
(Intercept)                   550.1897    10.4598  129.3775  52.600   <2e-16 ***
condunsys                      -0.5374     5.3467  131.5060  -0.101     0.92    
trialTypeunrelated             99.4757     8.7699  133.6853  11.343   <2e-16 ***
condunsys:trialTypeunrelated   -1.0570     5.2394 7082.5662  -0.202     0.84    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) cndnsy trlTyp
condunsys   -0.421              
trlTypnrltd -0.431  0.189       
cndnsys:trT  0.122 -0.476 -0.300
```

##

inverse RTs

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: invRT ~ cond * trialType + (1 + cond * trialType | subj) + (1 |item)
   Data: criticalTrials
Control: lmerControl(optimizer = "bobyqa")

REML criterion at convergence: 2762.9

Scaled residuals:
    Min      1Q  Median      3Q     Max
-9.7247 -0.6036 -0.0242  0.5699  4.6105

Random effects:
 Groups   Name                         Variance Std.Dev. Corr
 item     (Intercept)                  0.007566 0.08698
 subj     (Intercept)                  0.061899 0.24880
          condunsys                    0.017295 0.13151  -0.47
          trialTypeunrelated           0.012456 0.11161  -0.84  0.54
          condunsys:trialTypeunrelated 0.005895 0.07678   0.41 -0.98 -0.39
 Residual                              0.077122 0.27771
Number of obs: 7413, groups:  item, 96; subj, 80

Fixed effects:
                               Estimate Std. Error        df t value Pr(>|t|)
(Intercept)                  -1.910e+00  3.118e-02  1.081e+02 -61.253   <2e-16 ***
condunsys                    -3.148e-04  1.724e-02  7.866e+01  -0.018    0.985
trialTypeunrelated            3.054e-01  2.355e-02  1.483e+02  12.966   <2e-16 ***
condunsys:trialTypeunrelated  1.667e-03  1.551e-02  1.236e+02   0.108    0.915
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr) cndnsy trlTyp
condunsys   -0.430
trlTypnrltd -0.668  0.343
cndnsys:trT  0.285 -0.766 -0.344
```

##

These models were re-run after removing outliers. Outliers were calculated for each participant separately as any trials that were greater than +/- 2.5 std. dev. outside that participant's average RT.

##

raw RTs with outliers removed

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: RT ~ cond * trialType + (1 | item) + (cond + trialType | subj)
   Data: criticalTrimmed
Control: lmerControl(optimizer = "bobyqa")
```

```
REML criterion at convergence: 87947.2

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.9495 -0.6303 -0.1210  0.4681  7.1882

Random effects:
 Groups   Name               Variance Std.Dev. Corr
 item     (Intercept)         852.3   29.19
 subj     (Intercept)        5941.3   77.08
          condunsys          1023.2   31.99    -0.48
          trialTypeunrelated  677.8   26.03    -0.35  0.40
 Residual                    9001.1   94.87
Number of obs: 7309, groups:  item, 96; subj, 80

Fixed effects:
                              Estimate Std. Error        df t value Pr(>|t|)
(Intercept)                   545.2721     9.8391  115.2207  55.419   <2e-16 ***
condunsys                       0.3773     4.7267  124.4213   0.080    0.937
trialTypeunrelated             93.3335     7.3452  142.4231  12.707   <2e-16 ***
condunsys:trialTypeunrelated    1.4394     4.4456 6978.8189   0.324    0.746
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr) cndnsy trlTyp
condunsys   -0.421
trlTypnrltd -0.432  0.257
cndnsys:trT  0.109 -0.455 -0.304
```

##

inverse RTs with outliers removed

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: invRT ~ cond * trialType + (1 + cond * trialType | subj) + (1 | item)
   Data: criticalTrimmed
Control: lmerControl(optimizer = "bobyqa")

REML criterion at convergence: 1919.7

Scaled residuals:
    Min      1Q  Median      3Q     Max
-5.7239 -0.6135 -0.0064  0.6028  4.4869

Random effects:
 Groups   Name                         Variance Std.Dev. Corr
 item     (Intercept)                  0.006343 0.07964
 subj     (Intercept)                  0.061910 0.24882
          condunsys                    0.015083 0.12281  -0.47
          trialTypeunrelated           0.013075 0.11434  -0.83  0.59
          condunsys:trialTypeunrelated 0.004952 0.07037   0.41 -0.98 -0.46
 Residual                              0.068962 0.26261
Number of obs: 7309, groups:  item, 96; subj, 80
```

```
Fixed effects:
                              Estimate Std. Error         df t value Pr(>|t|)    
(Intercept)                  -1.917887   0.030703 103.773726 -62.466   <2e-16 ***
condunsys                     0.003021   0.016178  78.406445   0.187    0.852    
trialTypeunrelated            0.299250   0.022451 150.079418  13.329   <2e-16 ***
condunsys:trialTypeunrelated  0.002339   0.014604 127.196419   0.160    0.873    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr) cndnsy trlTyp
condunsys   -0.434              
trlTypnrltd -0.675  0.387       
cndnsys:trT  0.283 -0.759 -0.373
```

##

Exploratory analyses of the direct effect of tone error

Below we report the full output from the exploratory analysis of the direct effect of tone errors. This model aligns with that reported in Table 6 and Figure 6 in the main text. The model included the dependent variable RT (continuous), with fixed effects for prime type (stimType: no tone errors, tone errors) and trial type (tialType: identical, unrelated) and their interaction. We also tested a model with inverse RTs.

Note: In the output the label "filler" corresponds to "tone errors".

##

Direct tone errors: raw RTs

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: RT ~ stimType * trialType + (stimType + trialType | subj) + (1 |      item)
   Data: unsysTrials
Control: lmerControl(optimizer = "bobyqa")

REML criterion at convergence: 63163.4

Scaled residuals: 
    Min      1Q  Median      3Q     Max 
-3.4283 -0.5850 -0.1395  0.3698 10.8312 

Random effects:
 Groups   Name               Variance Std.Dev. Corr
```

```
 item     (Intercept)         1401.1   37.43
 subj     (Intercept)         4757.2   68.97
          stimTypefiller       211.2   14.53    0.86
          trialTypeunrelated   739.1   27.19   -0.24  0.14
 Residual                    13052.5  114.25
Number of obs: 5089, groups:  item, 131; subj, 80

Fixed effects:
                                  Estimate Std. Error      df t value Pr(>|t|)
(Intercept)                        549.678      9.772 140.065  56.248  < 2e-16 ***
stimTypefiller                      52.588     11.611 122.148   4.529 1.39e-05 ***
trialTypeunrelated                  98.476      9.040 138.667  10.894  < 2e-16 ***
stimTypefiller:trialTypeunrelated  -57.418     16.463 121.233  -3.488  0.00068 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr) stmTyp trlTyp
stimTypfllr -0.223
trlTypnrltd -0.471  0.350
stmTypfll:T  0.224 -0.691 -0.487
```

##

Direct tone errors: inverse RTs

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: invRT ~ stimType * trialType + (trialType | subj) + (1 | item)
   Data: unsysTrials
Control: lmerControl(optimizer = "bobyqa")

REML criterion at convergence: 1979

Scaled residuals:
    Min      1Q  Median      3Q     Max
-9.6904 -0.5922 -0.0190  0.5592  4.5368

Random effects:
 Groups   Name               Variance Std.Dev. Corr
 item     (Intercept)        0.007947 0.08914
 subj     (Intercept)        0.046689 0.21608
          trialTypeunrelated 0.008727 0.09342  -0.73
 Residual                    0.077428 0.27826
Number of obs: 5089, groups:  item, 131; subj, 80

Fixed effects:
                                   Estimate Std. Error        df t value Pr(>|t|)
(Intercept)                        -1.90988    0.02810 122.14634 -67.960  < 2e-16 ***
stimTypefiller                      0.15528    0.02750 122.83220   5.647 1.07e-07 ***
trialTypeunrelated                  0.30696    0.02289 156.32506  13.412  < 2e-16 ***
stimTypefiller:trialTypeunrelated  -0.16499    0.03938 123.78773  -4.190 5.27e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr) stmTyp trlTyp
stimTypfllr -0.267
```

```
trlTypnrltd -0.607  0.328
stmTypfll:T  0.186 -0.698 -0.460
```

##

6. Exploratory analyses of adaptation over the course of the experiment

As previous studies revealed adaptive effects by examination of change over the experiment (e.g., from first to second half in Witteman, Weber, & McQueen, 2014), we also conducted an exploratory analysis of adaptation over trials. Compared to our primary analysis, these models are underpowered, and should be interpreted with caution. Whereas our main analysis had approximately 1920 observations per cell (24 trials * 80 participants for each condition and each trial type before removal of incorrect trials), these analyses have half (for the by-half models) or even fewer (an average of 13 observations per trial in the by-trial model). Nevertheless, as we expect some readers will be curious about this aspect of the data, we have included these analyses here.

By-half analyses

Models included fixed effects of condition (Error Free, Tone Error), trial type (identical, unrelated), and half (A = first, B = second). As above, lmerTest was used to select the best fitting model. Below we report the model for the untransformed raw data We also tested models for inverse RTs, and then the same models again after removal of outliers. Results were not substantively different, so we are not including them here.

##

By-half adaptation: raw RTs

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: RT ~ cond + trialType + half + (cond + trialType + half + cond:half | subj) +
```

```
         (1 | item) + cond:trialType + cond:half + trialType:half + cond:trialType:half
   Data: criticalTrials
Control: lmerControl(optimizer = "bobyqa")

REML criterion at convergence: 91611.7

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.4794 -0.5746 -0.1417  0.3622 11.2830

Random effects:
 Groups   Name               Variance Std.Dev. Corr
 item     (Intercept)         1309     36.18
 subj     (Intercept)         7616     87.27
          condunsys           2404     49.03   -0.55
          trialTypeunrelated   713     26.70   -0.24  0.29
          halfB               1012     31.81   -0.66  0.59  0.25
          condunsys:halfB     1911     43.72    0.41 -0.77 -0.34 -0.73
 Residual                    12429    111.48
Number of obs: 7413, groups:  item, 96; subj, 80

Fixed effects:
                                    Estimate Std. Error        df t value Pr(>|t|)
(Intercept)                         552.7329    11.6437  125.8469  47.470   <2e-16 ***
condunsys                            -0.3486     7.4965  132.1935  -0.047    0.963
trialTypeunrelated                  105.5373     9.5102  184.0219  11.097   <2e-16 ***
halfB                                -5.1422     6.2315  173.4208  -0.825    0.410
condunsys:trialTypeunrelated         -7.3241     7.3366 6926.5888  -0.998    0.318
condunsys:halfB                      -0.2841     8.7262  175.8919  -0.033    0.974
trialTypeunrelated:halfB            -12.0245     7.3491 6933.1204  -1.636    0.102
condunsys:trialTypeunrelated:halfB   12.5389    10.3698 6928.4276   1.209    0.227
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr) cndnsy trlTyp halfB  cndn:T cndn:B trlT:B
condunsys   -0.487
trlTypnrltd -0.428  0.249
halfB       -0.494  0.529  0.265
cndnsys:trT  0.153 -0.475 -0.387 -0.287
cndnsys:hlB  0.323 -0.714 -0.217 -0.716  0.408
trlTypnrl:B  0.153 -0.238 -0.386 -0.572  0.500  0.408
cndnsys:T:B -0.109  0.336  0.273  0.405 -0.707 -0.577 -0.709
```

##

Figure S2 depicts the change over halves for raw RTs.

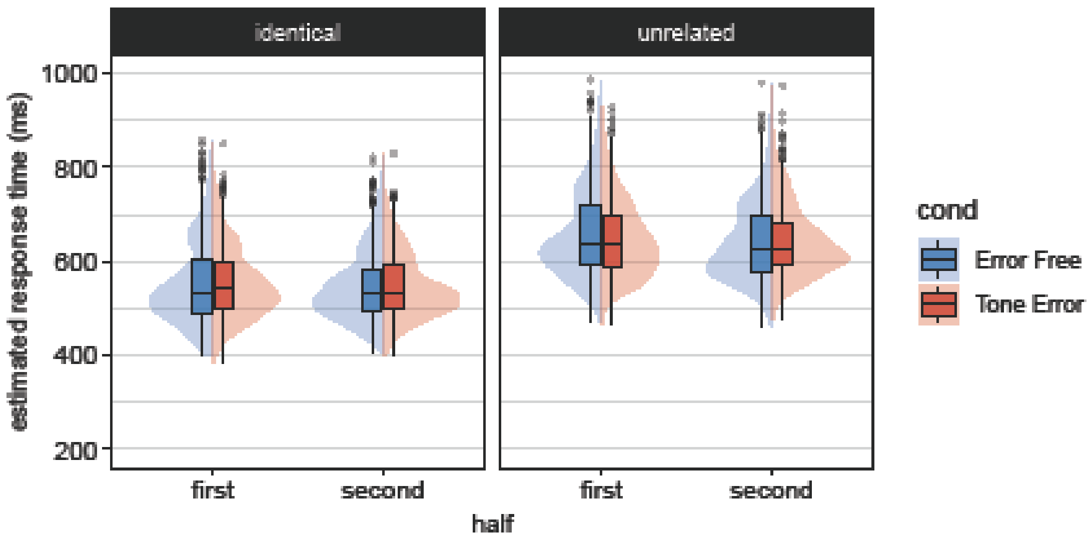

Figure S2. *Boxplots of model estimates for change over experiment halves for the indirect effect of tone errors. Shaded areas behind boxplots indicate the estimated distribution of responses.*

By-trial analyses

Models included fixed effects of condition (Error Free, Tone Error), trial type (identical, unrelated), and trial (1-144). Trial was not included in random effects due to convergence issues. As above, lmerTest was used to select the best fitting model. There appear to be small but substantive differences in models for raw RTs, inverse RTs, and when outliers are removed.

##

By-trial adaptation: raw RTs

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: RT ~ cond * trialType * trial + (cond + trialType | subj) + (1 | item)
   Data: criticalTrials
Control: lmerControl(optimizer = "bobyqa")

REML criterion at convergence: 91698.8

Scaled residuals:
    Min      1Q  Median      3Q     Max
```

```
-3.5131 -0.5754 -0.1493  0.3612 11.1852

Random effects:
 Groups   Name              Variance Std.Dev. Corr
 item     (Intercept)        1301.6   36.08
 subj     (Intercept)        6046.1   77.76
          condunsys          1226.9   35.03   -0.50
          trialTypeunrelated  700.5   26.47   -0.22  0.19
 Residual                   12667.1  112.55
Number of obs: 7413, groups:  item, 96; subj, 80

Fixed effects:
                                    Estimate Std. Error         df t value Pr(>|t|)
(Intercept)                        5.501e+02  1.137e+01  1.808e+02  48.367  < 2e-16 ***
condunsys                          4.360e+00  8.318e+00  7.280e+02   0.524  0.60030
trialTypeunrelated                 1.180e+02  1.091e+01  3.201e+02  10.814  < 2e-16 ***
trial                              5.759e-04  6.241e-02  7.094e+03   0.009  0.99264
condunsys:trialTypeunrelated      -2.001e+01  1.055e+01  7.097e+03  -1.896  0.05798 .
condunsys:trial                   -6.828e-02  8.879e-02  7.100e+03  -0.769  0.44194
trialTypeunrelated:trial          -2.573e-01  9.049e-02  7.100e+03  -2.843  0.00448 **
condunsys:trialTypeunrelated:trial  2.636e-01  1.273e-01  7.103e+03   2.071  0.03835 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) cndnsy trlTyp trial  cndn:T cndns: trlTy:
condunsys   -0.461
trlTypnrltd -0.479  0.318
trial       -0.393  0.538  0.410
cndnsys:trT  0.222 -0.614 -0.487 -0.424
cndnsys:trl  0.277 -0.765 -0.288 -0.703  0.604
trlTypnrlt:  0.271 -0.371 -0.596 -0.690  0.617  0.485
cndnsys:tT: -0.193  0.534  0.424  0.491 -0.868 -0.698 -0.711
```

##

Figure S3 depicts the linear change over trials for raw RTs.

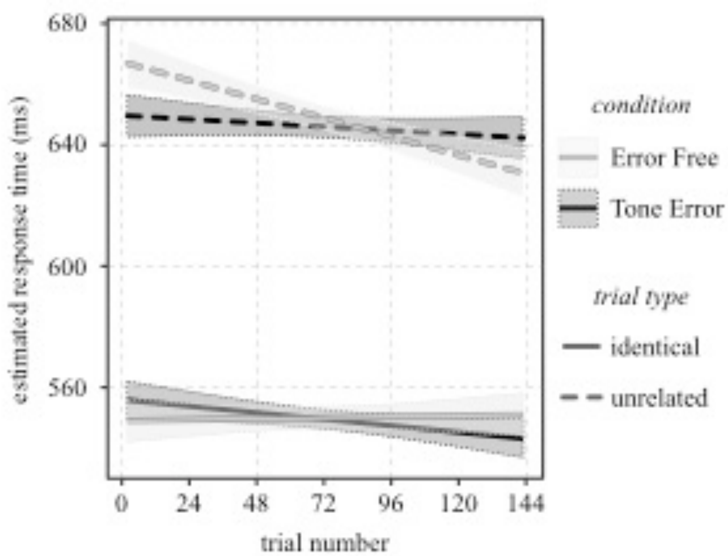

Figure S3. *Model estimates of linear change in response time across trials (raw RTs, no removal of outliers).*

##

By-trial adaptation: inverse RTs

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: invRT ~ cond * trialType * trial + (cond + trialType | subj) + (1 | item)
   Data: criticalTrials
Control: lmerControl(optimizer = "bobyqa")

REML criterion at convergence: 2840

Scaled residuals: 
    Min      1Q  Median      3Q     Max 
-9.8559 -0.6000 -0.0174  0.5607  4.5335 

Random effects:
 Groups   Name               Variance Std.Dev. Corr       
 item     (Intercept)        0.007565 0.08698             
 subj     (Intercept)        0.058443 0.24175             
          condunsys          0.008997 0.09485  -0.42      
          trialTypeunrelated 0.010569 0.10281  -0.78  0.28
 Residual                    0.077430 0.27826             
Number of obs: 7413, groups:  item, 96; subj, 80

Fixed effects:
                                     Estimate Std. Error         df t value Pr(>|t|)    
(Intercept)                        -1.918e+00  3.242e-02  1.438e+02 -59.154  < 2e-16 ***
condunsys                           2.828e-02  2.102e-02  6.176e+02   1.345  0.17896    
trialTypeunrelated                  3.471e-01  2.811e-02  3.503e+02  12.347  < 2e-16 ***
trial                               1.171e-04  1.543e-04  7.089e+03   0.759  0.44800    
condunsys:trialTypeunrelated       -5.751e-02  2.610e-02  7.095e+03  -2.204  0.02758 *  
condunsys:trial                    -3.991e-04  2.196e-04  7.094e+03  -1.818  0.06915 .  
```

```
trialTypeunrelated:trial          -5.834e-04  2.237e-04  7.100e+03  -2.607  0.00914 **
condunsys:trialTypeunrelated:trial  8.298e-04  3.147e-04  7.101e+03   2.637  0.00837 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr) cndnsy trlTyp trial  cndn:T cndns: trlTy:
condunsys   -0.418
trlTypnrltd -0.617  0.334
trial       -0.341  0.527  0.394
cndnsys:trT  0.193 -0.601 -0.468 -0.424
cndnsys:trl  0.240 -0.749 -0.277 -0.703  0.604
trlTypnrlt:  0.235 -0.363 -0.572 -0.690  0.617  0.485
cndnsys:tT: -0.168  0.523  0.407  0.491 -0.868 -0.698 -0.711
```

##

By-trial adaptation: raw RTs with outliers removed

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: RT ~ cond * trialType * trial + (cond + trialType | subj) + (1 |      item)
   Data: criticalTrimmed
Control: lmerControl(optimizer = "bobyqa")

REML criterion at convergence: 87950.2

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.0387 -0.6265 -0.1252  0.4617  7.2137

Random effects:
 Groups   Name               Variance Std.Dev. Corr
 item     (Intercept)         852.3   29.19
 subj     (Intercept)        5942.0   77.08
          condunsys          1027.1   32.05    -0.48
          trialTypeunrelated  679.3   26.06    -0.34  0.40
 Residual                    8989.1   94.81
Number of obs: 7309, groups:  item, 96; subj, 80

Fixed effects:
                                     Estimate Std. Error         df t value Pr(>|t|)
(Intercept)                         545.95172   10.54091  151.71770  51.794   <2e-16 ***
condunsys                             3.45045    7.17080  629.93644   0.481   0.6306
trialTypeunrelated                  105.79370    9.20400  348.75603  11.494   <2e-16 ***
trial                                -0.00950    0.05283 6985.78113  -0.180   0.8573
condunsys:trialTypeunrelated         -9.87348    8.97276 6993.47286  -1.100   0.2712
condunsys:trial                      -0.04282    0.07520 6994.99487  -0.569   0.5691
trialTypeunrelated:trial             -0.17209    0.07701 6998.11747  -2.234   0.0255 *
condunsys:trialTypeunrelated:trial    0.15689    0.10814 7000.03620   1.451   0.1469
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr) cndnsy trlTyp trial  cndn:T cndns: trlTy:
condunsys   -0.449
trlTypnrltd -0.469  0.352
trial       -0.359  0.528  0.411
```

```
cndnsys:trT  0.202 -0.600 -0.493 -0.422
cndnsys:trl  0.252 -0.752 -0.289 -0.703  0.601
trlTypnrlt:  0.246 -0.362 -0.603 -0.686  0.618  0.482
cndnsys:tT: -0.175  0.523  0.429  0.489 -0.869 -0.696 -0.712
```

##

By-trial adaptation: inverse RTs with outliers removed

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: invRT ~ cond * trialType * trial + (cond * trialType | subj) +      (1 | item)
   Data: criticalTrimmed
Control: lmerControl(optimizer = "bobyqa")

REML criterion at convergence: 1974.3

Scaled residuals:
    Min      1Q  Median      3Q     Max
-5.7081 -0.6150 -0.0048  0.5948  4.4745

Random effects:
 Groups   Name                         Variance Std.Dev. Corr
 item     (Intercept)                  0.006338 0.07961
 subj     (Intercept)                  0.061905 0.24881
          condunsys                    0.015097 0.12287  -0.47
          trialTypeunrelated           0.013039 0.11419  -0.83  0.59
          condunsys:trialTypeunrelated 0.004925 0.07018   0.41 -0.98 -0.46
 Residual                              0.068918 0.26252
Number of obs: 7309, groups:  item, 96; subj, 80

Fixed effects:
                                     Estimate Std. Error         df t value Pr(>|t|)
(Intercept)                        -1.923e+00  3.244e-02  1.293e+02 -59.276   <2e-16 ***
condunsys                           2.377e-02  2.202e-02  2.660e+02   1.080   0.2813
trialTypeunrelated                  3.316e-01  2.718e-02  3.217e+02  12.198   <2e-16 ***
trial                               6.930e-05  1.464e-04  6.987e+03   0.474   0.6359
condunsys:trialTypeunrelated       -4.275e-02  2.605e-02  1.144e+03  -1.641   0.1011
condunsys:trial                    -2.895e-04  2.084e-04  6.996e+03  -1.389   0.1649
trialTypeunrelated:trial           -4.479e-04  2.132e-04  6.991e+03  -2.101   0.0357 *
condunsys:trialTypeunrelated:trial  6.258e-04  2.994e-04  6.997e+03   2.090   0.0366 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr) cndnsy trlTyp trial  cndn:T cndns: trlTy:
condunsys   -0.456
trlTypnrltd -0.652  0.418
trial       -0.323  0.476  0.386
cndnsys:trT  0.280 -0.701 -0.505 -0.403
cndnsys:trl  0.227 -0.678 -0.271 -0.703  0.573
trlTypnrlt:  0.222 -0.327 -0.565 -0.687  0.589  0.483
cndnsys:tT: -0.158  0.472  0.402  0.489 -0.828 -0.696 -0.712
```
##

Summary: adaptation over the course of the experiment

The by-half analysis revealed no evidence of differences between halves of the experiment. The pattern of results across models for the by-trial analysis is unstable. Models with outliers included suggest some adaptation for unrelated trials in the Error Free condition, such that responses grew faster across the experiment, but this effect grows weaker or becomes insignificant when the outliers are removed. Given the small number of observations per trial, we do not place much trust in this particular trend. To reliably test for adaptation across trials, a much larger sample of participants would be required.

7. Additional results of post-experiment questions

Due to space limitations, we did not report all of the post-experiment questions in the main text. Here we report the remaining two. The effect for ratings of intelligibility is largely similar to what was observed for accentedness, with lesser intelligibility being attributed when the speaker made tone errors (Figure S4). The effect of tone errors on ratings of pleasantness is less pronounced (Figure S5).

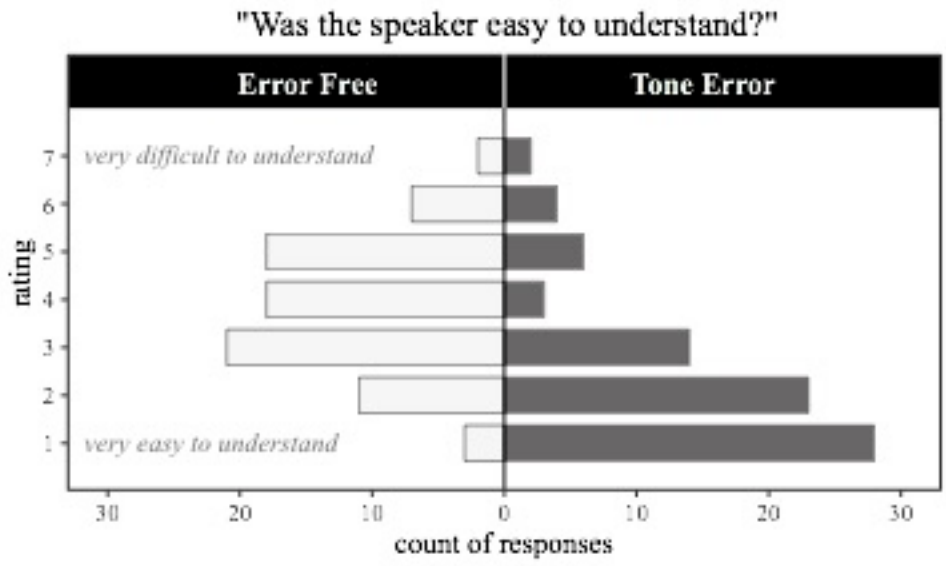

Figure S4. *Intelligibility ratings for the speakers without tone errors (left) and with tone errors (right).*

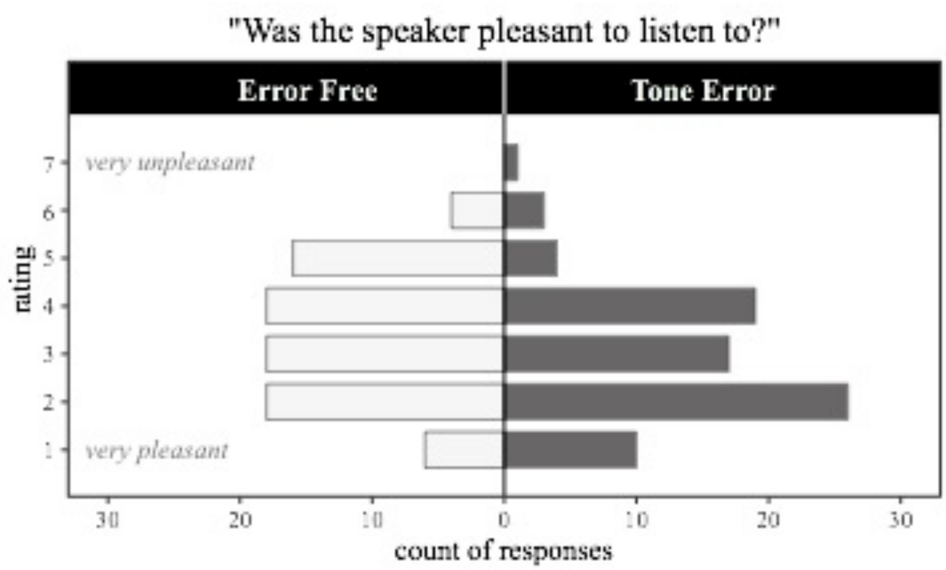

Figure S5. *Pleasantness ratings for the speakers without tone errors (left) and with tone errors (right).*

8. Note about Chinese language history questionnaire

The Chinese questionnaire used to explore participants' language history was adapted from materials graciously shared by colleagues at University of Kansas. A unique focus of this questionnaire was participants' previous Chinese dialect usage and their experience with foreign-accented Mandarin. For additional details, please contact the corresponding author.

9. Stimuli for critical trials

PinyinTone	English gloss	Prime Freq	Target	Target Freq	Trial Type
<u>*Critical Set A*</u>					
xīnwén	*news*	3.2095	新闻		identical
hénjì	*trace*	2.8727	痕迹		identical
liúmáng	*hoodlum*	2.5599	流氓		identical
línghún	*spirit*	3.0542	灵魂		identical
lèqù	*delight*	2.7177	乐趣		identical
zhuānyè	*profession*	3.0508	专业		identical
jiāngjūn	*general*	2.699	将军		identical
quánlì	*power*	3.0913	权利		identical
nǎodài	*brain*	3.1399	脑袋		identical
nányǒu	*boyfriend*	2.8639	男友		identical
biǎoqíng	*expression*	3.0035	表情		identical
qiánbāo	*wallet*	2.8089	钱包		identical
chǎnpǐn	*product*	2.6776	产品		identical
huàxué	*chemistry*	2.6031	化学		identical
chǒngwù	*pet*	2.6294	宠物		identical
cèsuǒ	*toilet*	3.0199	厕所		identical
zūnyán	*honor*	2.5024	尊严		identical
jiàzhí	*value*	3.0799	价值		identical
gēshǒu	*singer*	2.8062	歌手		identical
bèndàn	*idiot*	3.1028	笨蛋		identical

chènshān	*shirt*	2.7474	衬衫		identical
huǒchē	*train*	2.8041	火车		identical
bēijù	*tragedy*	2.7143	悲剧		identical
nǚshén	*goddess*	2.415	女神		identical
zhèngfǔ	*government*	3.1617	穿着	2.8028	unrelated
bùmén	*department*	2.9786	奶酪	2.6702	unrelated
xiāngcūn	*countryside*	2.574	嘴巴	2.7275	unrelated
shèqū	*community*	2.7101	生日	3.1136	unrelated
jīnglǐ	*manager*	2.8657	灯光	2.5966	unrelated
míngxīng	*celebrity*	3.0512	顾客	2.7657	unrelated
lǎohǔ	*tiger*	2.316	白痴	3.2482	unrelated
niánjí	*age*	2.8837	线索	3.1433	unrelated
duìxiàng	*target*	2.9106	广告	2.9832	unrelated
zhǔtí	*subject*	2.7716	团队	2.8274	unrelated
zāinàn	*disaster*	2.7796	森林	2.6385	unrelated
wūdǐng	*roof*	2.6721	马桶	2.4265	unrelated
zhànzhēng	*war*	3.0584	基础	2.6532	unrelated
huànzhě	*patient*	2.5145	羞耻	2.5198	unrelated
hūnyīn	*marriage*	3.0208	类型	2.8055	unrelated
lǚguǎn	*motel*	2.9253	语言	2.8722	unrelated
mǎijiā	*buyer*	2.316	糖果	2.5302	unrelated
jiǔdiàn	*hotel*	2.9504	阶段	2.752	unrelated
máojīn	*towel*	2.5051	咖啡	3.2851	unrelated
tóngshì	*coworker*	3.0048	良心	2.574	unrelated
méitǐ	*media*	2.8727	种族	2.601	unrelated
shǎguā	*fool*	3.0973	秘书	2.5416	unrelated
píngwěi	*evaluator*	2.5092	母亲	3.3736	unrelated
tiāntáng	*paradise*	2.9355	儿童	2.8797	unrelated
	mean (sd)	2.82 (0.23)		2.81 (0.26)	

<u>*Critical Set B*</u>

yīngxióng	*hero*	3.1065	英雄		identical
móguǐ	*devil*	2.7889	魔鬼		identical
xiǎochǒu	*clown*	2.6884	小丑		identical
dírén	*enemy*	3.0116	敌人		identical
tiáojiàn	*conditions*	3.0374	条件		identical
shǒuxí	*seat of honor*	2.4757	首席		identical
fūfù	*husband & wife*	2.7235	夫妇		identical
táicí	*lines*	2.5623	台词		identical
yǎnyuán	*actor*	3.0588	演员		identical
bàngqiú	*baseball*	2.7084	棒球		identical
pífū	*skin*	2.8848	皮肤		identical
guòchéng	*process*	3.0885	过程		identical
hǎitān	*beach*	2.8041	海滩		identical
fǎlǜ	*law*	3.1477	法律		identical
diàntī	*elevator*	2.721	电梯		identical
wǎngzhàn	*website*	2.6532	网站		identical
èmèng	*nightmare*	2.7451	噩梦		identical
kōngqì	*air conditioner*	2.9731	空气		identical
āyí	*aunt*	2.5933	阿姨		identical
bàozhǐ	*newspaper*	2.9917	报纸		identical
zhōngyāng	*center*	2.6998	中央		identical
lánsè	*color*	2.9133	蓝色		identical
shùzì	*numeral*	2.9096	数字		identical
guāndiǎn	*viewpoint*	2.847	观点		identical
zǒuláng	*hallway*	2.7686	财产	2.7952	unrelated
zhuàngtài	*status*	3.1119	礼拜	2.8136	unrelated
jiǎodù	*viewpoint*	2.9595	提要	3.0334	unrelated
zázhì	*magazine*	3.0199	目标	3.2639	unrelated

nèiróng	*topic*	2.9675	粉丝	2.6693	unrelated
chuánzhǎng	*captain*	2.4914	珠宝	2.4713	unrelated
jiǎndāo	*scissors*	2.2227	玉米	2.5809	unrelated
cuòshī	*measure*	2.6839	范围	3.0191	unrelated
huángjīn	*gold*	2.4786	优势	2.6425	unrelated
dàjiē	*street*	2.945	冰箱	2.7412	unrelated
zhīpiào	*check*	2.8488	原则	2.6665	unrelated
shāngkǒu	*wound*	2.8739	味道	3.2047	unrelated
wǎncān	*dinner*	3.1242	身材	2.7118	unrelated
dǔchǎng	*casino*	2.2625	警察	3.4447	unrelated
gōngchǎng	*factory*	2.6693	耳朵	2.9004	unrelated
yínháng	*bank*	3.0082	领导	2.786	unrelated
fēnggé	*style*	2.9518	厨房	3.0228	unrelated
bànlǚ	*companion*	2.4928	牛奶	2.7243	unrelated
xīzhuāng	*suit*	2.5658	费用	2.658	unrelated
yáchǐ	*tooth*	2.7275	联邦	2.9513	unrelated
línjū	*neighbor*	3.0422	姓名	2.5832	unrelated
hàomǎ	*number*	3.185	士兵	2.7853	unrelated
zǒngtǒng	*president*	2.9703	技巧	2.7604	unrelated
sījī	*driver*	2.9079	癌症	2.6749	unrelated
	mean (sd)	2.82 (0.23)		2.83 (0.24)	

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Boersma, P., & Weenink, D. (2018). *Praat: Doing phonetics by computer* (Version 6.0.42) [Computer software]. www.praat.org

Chen, N. F., Wee, D., Tong, R., Ma, B., & Li, H. (2016). Large-scale characterization of non-native Mandarin Chinese spoken by speakers of European origin: Analysis on iCALL. *Speech Communication*, *84*, 46–56. https://doi.org/10.1016/j.specom.2016.07.005

Hao, Y.-C. (2018). Contextual effect in second language perception and production of Mandarin tones. *Speech Communication*, *97*, 32–42. https://doi.org/10.1016/j.specom.2017.12.015

He, Y., Wang, Q., & Wayland, R. (2016). Effects of different teaching methods on the production of Mandarin tone 3 by English speaking learners. *Chinese as a Second Language*, *51*(3), 252–265.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13

R Core Team. (2019). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. http://www.R-project.org/

Ramsey, S. R. (1987). *The Languages of China*. Princeton University Press.

Winke, P. M. (2007). Tuning into Tones: The Effect of L1 Background on L2 Chinese Learners' Tonal Production. *Journal of the Chinese Language Teachers Association*, *42*(3), 21–55.

Witteman, M. J., Weber, A., & McQueen, J. M. (2014). Tolerance for inconsistency in foreign-accented speech. *Psychonomic Bulletin & Review*, *21*(2), 512–519. https://doi.org/10.3758/s13423-013-0519-8

Yang, C. (2016). *The Acquisition of L2 Mandarin Prosody: From experimental studies to pedagogical practice*. John Benjamins Publishing Co.

Yang, C., & Chan, M. K. M. (2010). The Perception of Mandarin Chinese Tones and Intonation. *Journal of the Chinese Language Teachers Association*, *45*(1), 7–36.

Zhang, H. (2014). The Third Tone: Allophones, Sandhi Rules and Pedagogy. *Journal of the Chinese Language Teachers Association*, *49*(1), 117–145.