Cross-linguistic perception of clearly spoken tense and lax vowels based on auditory, visual, and auditory-visual information

Keith Leung and Yue Wang
Department of Linguistics, Simon Fraser University

Charles Redmon, Allard Jongman, and Joan Sereno
Department of Linguistics, University of Kansas

The 5th Joint Meeting of the Acoustical Society of America and the Acoustical Society of Japan: Honolulu, 2016

Background

- Plain versus clear speech modifications involve articulatory and acoustic modifications[1–4], which can enhance auditory and visual speech intelligibility in native (L1) and non-native (L2) languages[5–7]. However, excessively exaggerated speech resulting in overlap of phonemic categories, or attention to incorrect auditory or visual cues, may inhibit intelligibility [10].
- English tense-lax vowel contrasts and plain-to-clear vowel modifications share similar articulatory and acoustic features[3,4,8,9]. These overlapping cues may affect intelligibility of tensity contrasts in clear speech, especially for L2 perceivers whose L1 does not have vowel tensity contrasts (e.g., Mandarin Chinese[11]).

Present Study

- Examines auditory-visual perception of clearly versus plainly produced English tense and lax vowels by native English and Mandarin perceivers.
- Our articulatory study showed greater articulatory movements in clear than plain productions for both tense and lax vowels[4]. Our acoustic study further revealed greater temporal modifications for tense vowels and greater spectral modifications for lax vowels in clear speech[3].
- Based on these findings, we ask:
  1. How does vowel tensity interact with clear speech in auditory and visual perception?
  2. How do clear speech modifications influence L1 versus L2 vowel tensity perception?

Methods

Participants:
- 23 native English perceivers
- 30 native Mandarin perceivers, late intermediate-level learners of English

Stimuli:
- Vowels [i, ɪ, ɑ, ʌ, u, ʊ] in [k_d] minimal pairs: 'keyed', 'kid', 'cod', 'cud', 'cooed', 'could'
- Produced by 6 native English speakers (3F) in plain and clear speech[4]
- Elicitation of stimuli involves a simulated computer speech recognition program[Fig 1]
- Presented in audio (A), audio-visual (AV), and visual (V) modes[Fig 2]
- Audio was embedded in cafeteria noise (−15 dB SNR)

Task:
- Six-alternative forced choice (6AFC)

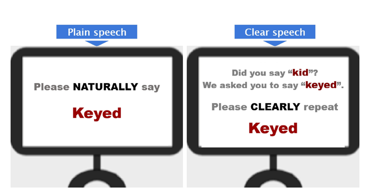

Figure 1: Elicitation of plain and clear speech

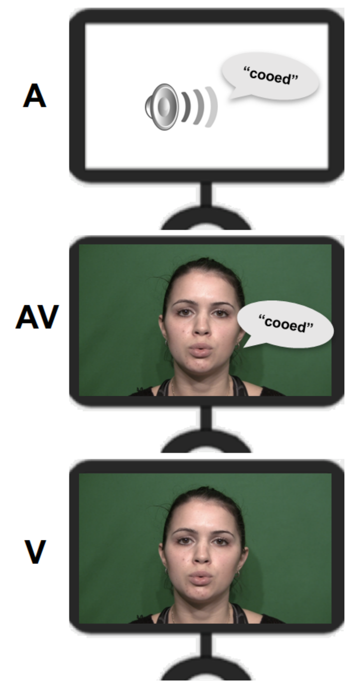

Figure 2: Presentation of stimuli

Results: Accuracy

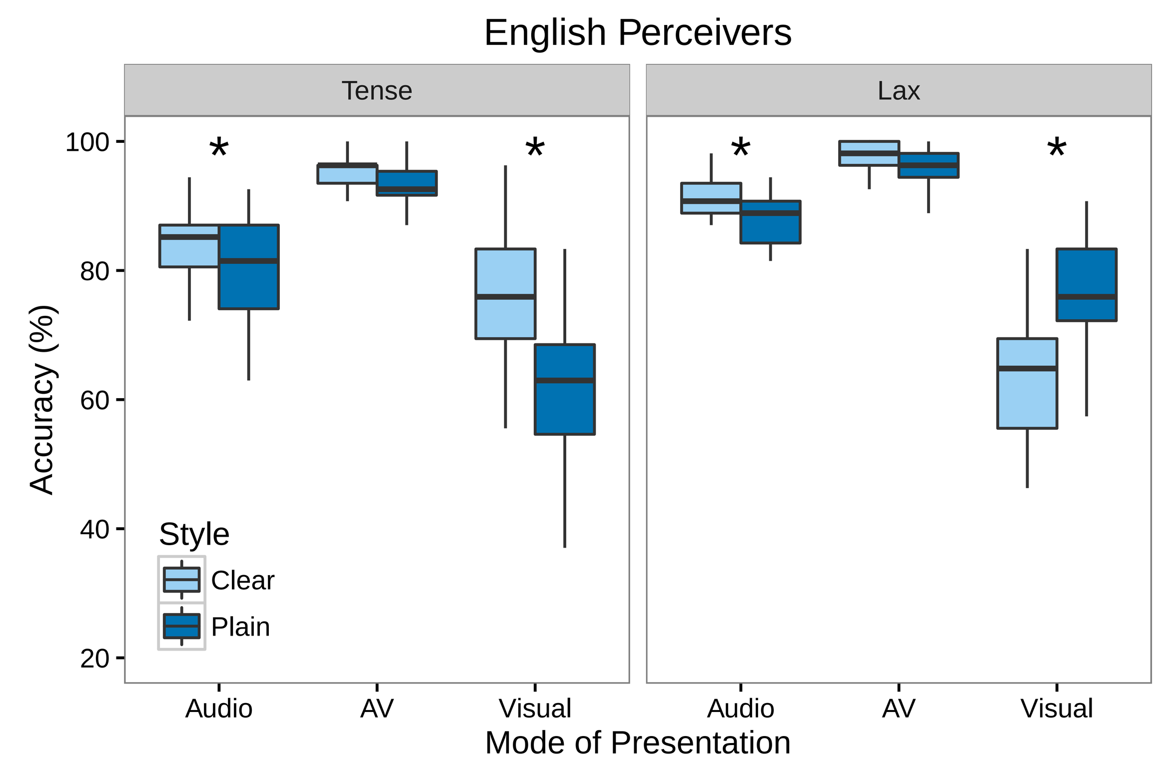

Figure 3: Overall accuracy of English perceivers by Stimulus Mode (A, AV, V), Speech Style (clear, plain), and Vowel Tensity (tense, lax).

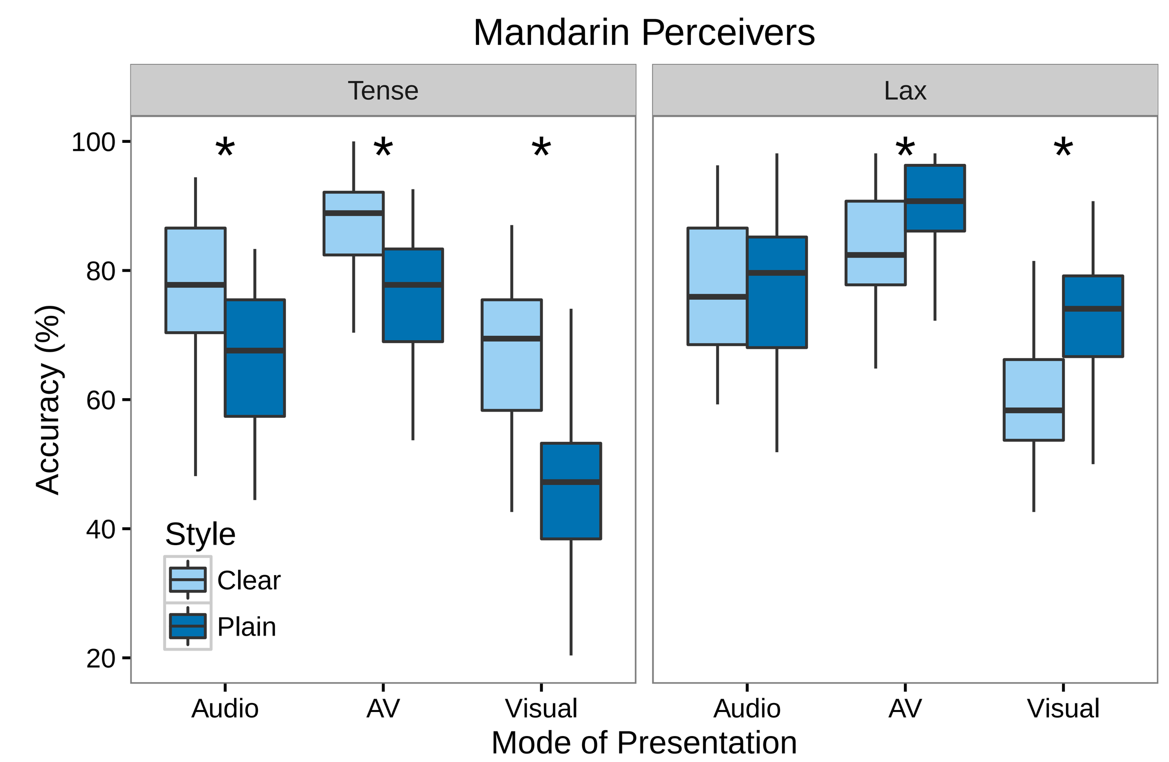

Figure 4: Overall accuracy of Mandarin perceivers by Stimulus Mode (A, AV, V), Speech Style (clear, plain), and Vowel Tensity (tense, lax).

- Mixed-effects logistic regression run on Accuracy with Subject as a random effect in English and Mandarin perceiver groups

	Clear > Plain	Plain > Clear
English perceivers	in A and for tense vowels in V	in V for lax vowels
Mandarin perceivers	for tense vowels	in V and AV for lax vowels

Results: Errors

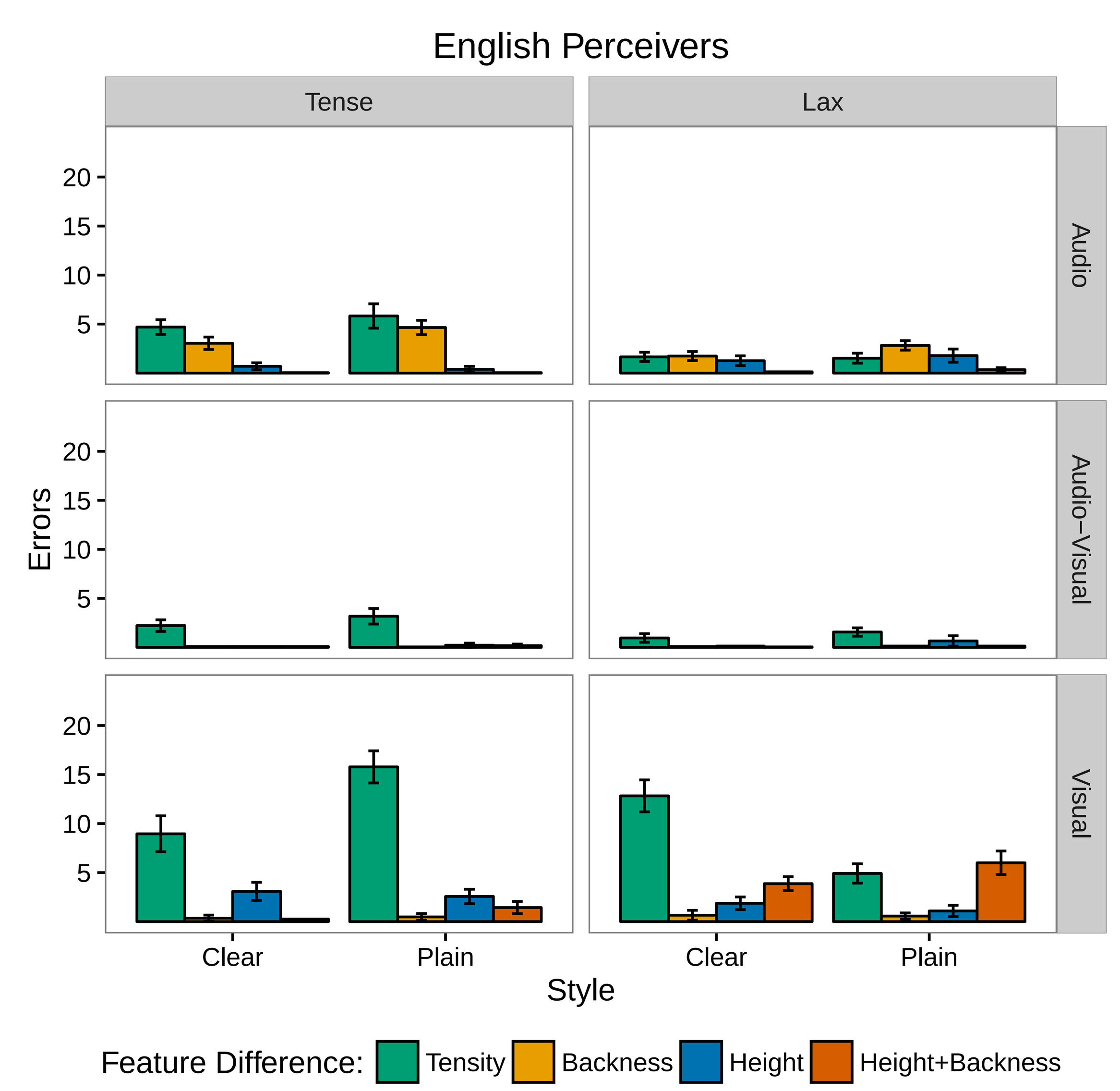

Figure 5: Distribution of errors among English perceivers by Stimulus Mode, Speech Style, Vowel Tensity, and Featural Difference characterizing the error (tensity, backness, heigh, height+backness).

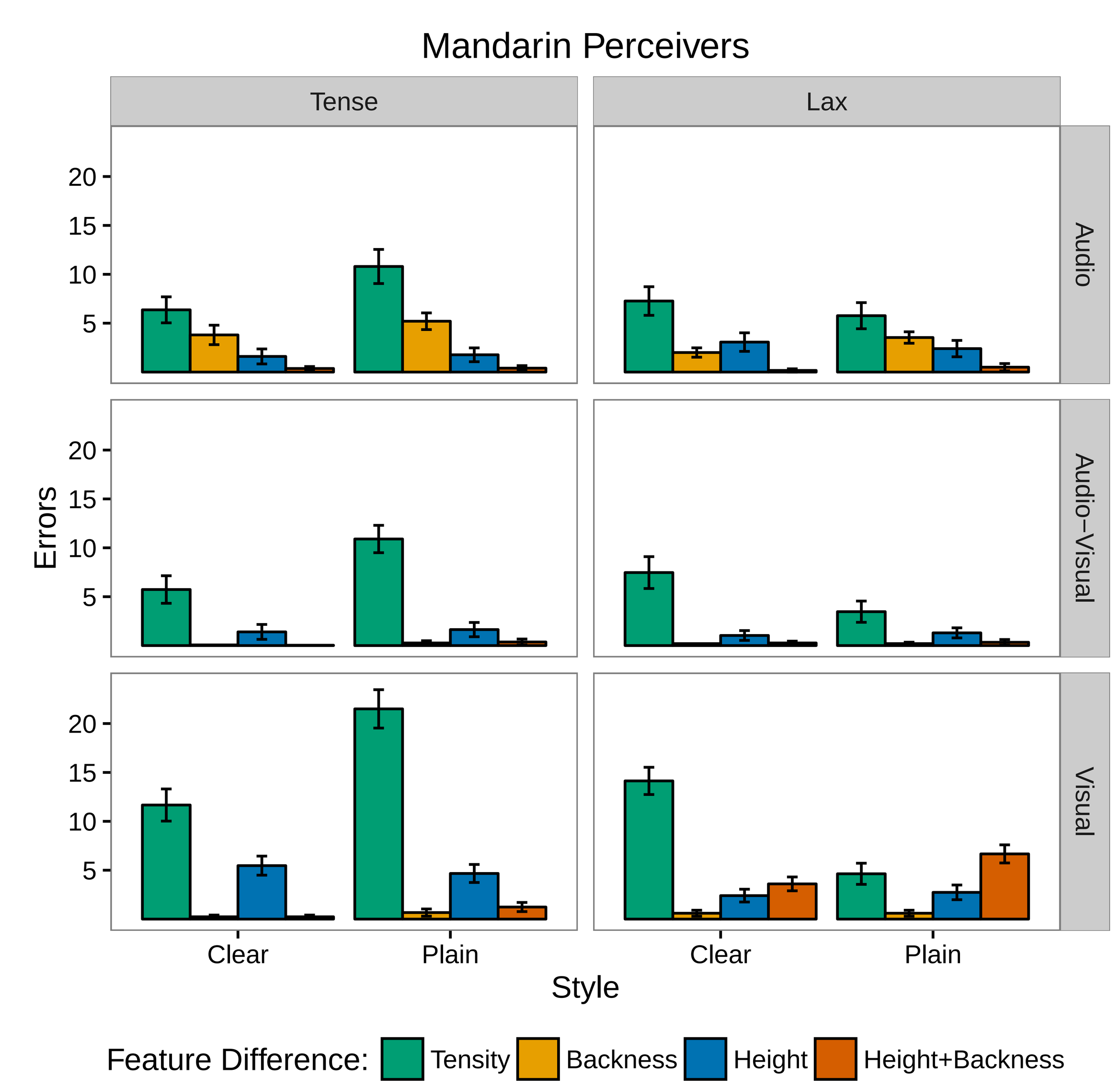

Figure 6: Distribution of errors among Mandarin perceivers by Stimulus Mode, Speech Style, Vowel Tensity, and Featural Difference characterizing the error (tensity, backness, heigh, height+backness).

- Separate mixed-effects negative binomial regression models run on Error Counts within Mode and Language group
- Tensity errors predominate; consistent with the accuracy results of the clear < plain pattern in visual lax stimuli

Relative Cue Contributions

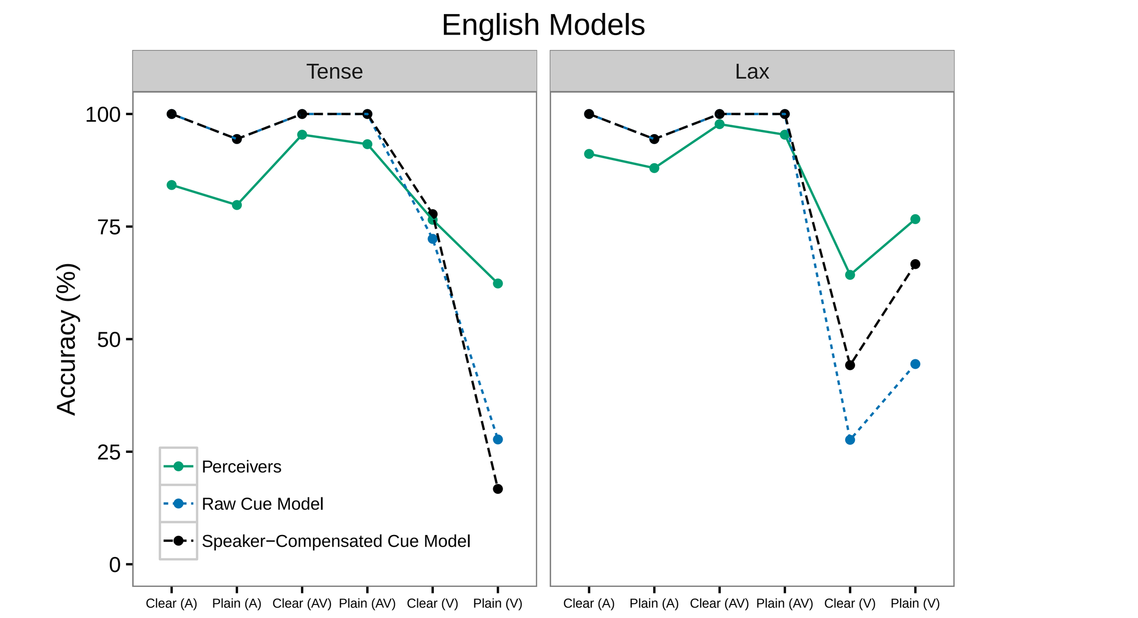

Figure 7: Predicting English perception from acoustic[3] and articulatory[4] cues.

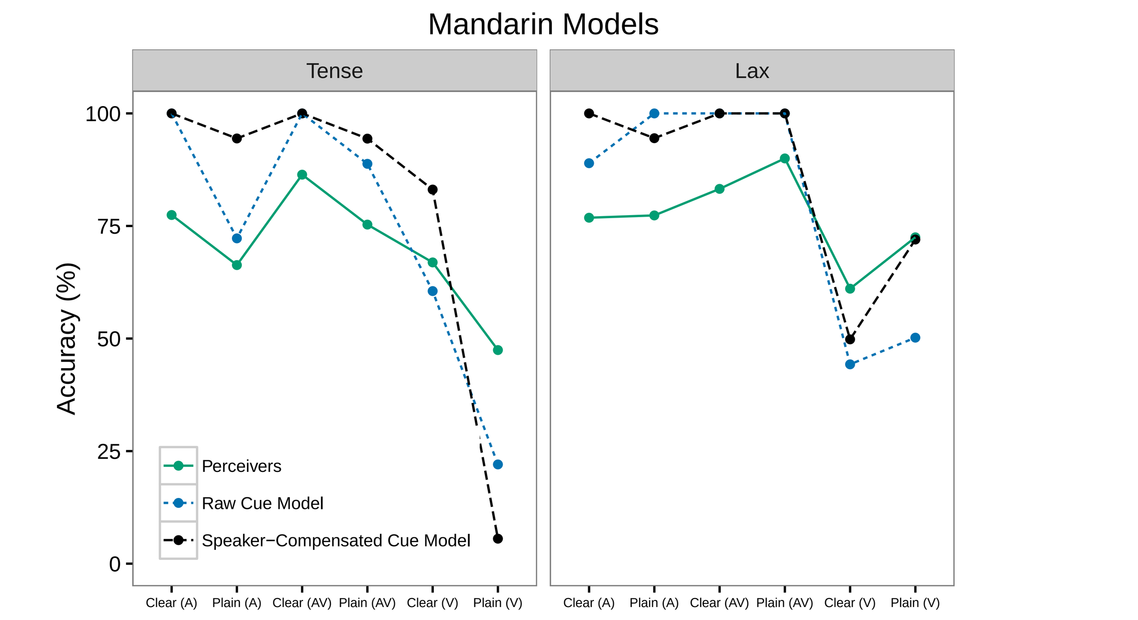

Figure 8: Predicting Mandarin perception from acoustic and articulatory cues.

- F1 was the most important† parameter in Audio models for both English and Mandarin perceivers
- No visual parameters were predictive of English Audio-Visual responses, whereas all four visual parameters contributed to the Mandarin model
- Lip rounding was more predictive in the English Visual model, while Jaw Displacement appeared to be a more prominent cue for Mandarin perceivers

Conclusions

- Clear speech benefits perception when its cues are compatible with those that characterize phonemes (e.g., tense vowels), but inhibits perception when its hyper-articulated features are in conflict with phoneme-intrinsic characteristics (e.g., lax vowels).
- Nonnatives benefit from clear speech for familiar (tense) vowels in their L1, but perception of non-familiar L2 (lax) vowels can be more impaired by incompatible clear-speech and phonemic features.

Acknowledgements

We thank Dr. Bob McMurray for his input on the C-Cure analysis. We also thank Quentin Qin, Lisa Tang and members of the Language and Brain Lab for their assistance in stimulus development and data collection. This project was supported by a research grant from the Social Sciences and Humanities Research Council of Canada (SSHRC Insight Grant 435-2012-1641).

References

1 Ferguson & Kewley-Port 2007. *JSLHR, 50*, 1241-1255.
2 Kim & Davis 2014. *Comput Speech Lang, 28*, 598-606.
3 Leung et al. 2016. *J Acoust Soc Am, 140*, 45-58.
4 Tang et al. 2015. *Speech Commun, 75*, 1-13.
5 Ferguson & Kewley-Port 2002. *J Acoust Soc Am, 112*, 259-271.
6 Gagné & Charest 2002. *Speech Commun, 37*, 213-230.
7 Smiljanić & Bradlow 2009. *Lang Linguist Compass, 3*, 236-264.
8 Hillenbrand et al. 1995. *J Acoust Soc Am, 97*, 3099-3111.
9 Smiljanić & Bradlow 2008. *J Phon, 36*, 91-113.
10 Kirchhoff & Schimmel 2005. *J Acoust Soc Am, 117*, 2238-2246.
11 Lin 2001. *A grammar of Mandarin Chinese.* Muenchen: Lincom Europa.

*Holm-corrected Wilcoxon signed rank tests were used for all pairwise comparisons.
†Parameter importance was determined from relative AIC reduction in the full model.