Wednesday Bushong and T. Florian Jaeger: JASA Express Letters https://doi.org/10.1121/1.5119271 Published Online 8 August 2019

Dynamic re-weighting of acoustic and contextual cues in spoken word recognition

Wednesday Bushong[a] **and T. Florian Jaeger**
Department of Brain and Cognitive Sciences, University of Rochester, Rochester, New York 14627, USA
wbushong@ur.rochester.edu, fjaeger@ur.rochester.edu

Abstract: Listeners integrate acoustic and contextual cues during word recognition. However, experiments investigating this integration disrupt natural cue correlations. It was investigated whether changes in correlational structure affect listeners' relative cue weightings. Two groups of participants engaged in a word recognition task. In one group, acoustic (voice onset time) and contextual (lexical bias) cues followed natural correlations; in the other, cues were uncorrelated. When cues were correlated, cue weights were stable throughout the experiment; when cues were uncorrelated, contextual cues were down-weighted. Listeners thus can re-weight cues based on their statistical structure. Studies failing to account for re-weighting risk over/under-estimating cue importance.

[Q-JF]
Date Received: April 25, 2019 **Date Accepted:** July 6, 2019

1. Introduction

During spoken word recognition, listeners integrate acoustic and contextual cues in order to infer the intended message. For example, multiple acoustic properties affect the recognition of voicing in English word-initial stops (e.g., "ban" vs "pan"), including the voice onset time (VOT) and fundamental frequency of the stop, as well as the duration of the following vowel (Lisker and Abramson, 1967). Recognition is also affected by lexical context (e.g., listeners would be more likely to hear pan after "frying," cf. Ganong, 1980; Kalikow *et al.*, 1977). The integration of acoustic and contextual cues is a key feature of models of speech perception (e.g., McClelland and Elman, 1986; Norris and McQueen, 2008; Oden and Massaro, 1978). How cue integration proceeds—including questions about what determines the weights of different cues, and whether certain types of cues are considered at all—has been an important theme in this literature.

Typical paradigms in this line of research present participants with speech stimuli that are manipulated to fall along a multi-dimensional continuum between two linguistic categories. For example, participants might hear stimuli along a two-dimensional ban–pan continuum, varying in VOT and fundamental frequency. The relationship between the cue values and participants' categorization decisions are then used to estimate whether, and how strongly, listeners weight the different cues. Here we revisit a common—though often implicit—assumption, that cue weights are static. We show that this assumption is not always warranted and can affect the estimation of cue weights.

In particular, we are interested in understanding how atypical covariations between acoustic and contextual cues affect how listeners weight them during integration. In natural speech, cues are often correlated (e.g., Kingston and Diehl, 1994). However, in perception experiments it is not uncommon to decorrelate the distributions of cues in order to disentangle the effects of the different cues. Decorrelation introduces conflict between the two cues. If listeners are sensitive to these conflicts, this may lead to listeners re-weighting them. Such re-weighting has been observed for the integration of multiple *acoustic* cues: decorrelating or reversing the natural correlation between two acoustic cues can cause listeners to substantially down-weight one of the cues compared to when natural cue correlations are preserved (Idemaru and Holt, 2011; Schertz *et al.*, 2016). Whether similar re-weighting occurs also for the integration of acoustic cues with non-acoustic context is an open question. Here we investigate how the presence or absence of typical correlations between acoustic and subsequent lexical cues effect their integration over the course of an experiment. We focus on this

[a] Author to whom correspondence should be addressed.

particular example because it has played a critical role in research on the limits of information maintenance and integration during speech perception (e.g., Burchill *et al.*, 2018; Connine *et al.*, 1991; McMurray *et al.*, 2009; Szostak and Pitt, 2013; for reviews, see Christiansen and Chater, 2016; Dahan, 2010). Research on this question has typically presented participants with acoustic and contextual cues that were decorrelated. As we show here, this risks *substantially* under-estimating the effect of subsequent lexical context on spoken word recognition.

We present a web-based categorization experiment in which listeners hear sentences that contain an acoustic cue (here, VOT) to a target word and subsequent lexical context to the target. Between participants, we manipulate the correlation between the two cues [Fig. 1(B)]: in the *high conflict* group the two cues are completely uncorrelated (as in previous research); in the *low conflict* group, however, the two cues are correlated and thus tend to support the same categorization decision. We compare the weighting of the two cues and changes therein across the experiment, across the two participant groups. Other than the manipulation of cue conflict, the design, materials, and procedure of the present study closely follows the classic paradigm of Connine *et al.* (1991), also employed in many subsequent studies.

2. Methods

2.1 Participants

A total of 120 participants were recruited from Amazon Mechanical Turk (60 each for the two between-participant conditions). The experiment took approximately 30 min to complete and participants were rewarded $3.00 ($6.00/h prorated). Web-based crowdsourcing paradigms have successfully been used to replicate lab-based experiments on speech perception (e.g., Liu and Jaeger, 2018; Xie *et al.*, 2018).

2.2 Materials

Figure 1(A) shows the general experimental paradigm. We constructed 40 sentence pairs like the following, where the subsequent lexical context biased towards either /t/ (a) or /d/ (b):

(a) When the ?ent in the *forest* was well camouflaged, we began our hike. (*tent-biasing lexical context*).

(b) When the ?ent in the *fender* was well camouflaged, we sold the car. (*dent-biasing lexical context*).

Each sentence pair had identical pre-target material and always consisted of a preposition plus the word “the.” Across the 40 pairs, the distance between the “?”-sound and the first biasing lexical cue (shown in italics above) varied from 3 to 9 syllables.[1] We manipulated the ?-sound to vary between /t/ and /d/ by manipulating its VOT (see supplementary material for stimulus creation procedure).[2] We chose six VOT values based on previous studies using the same stimuli and paradigm in our lab (Bushong and Jaeger, 2017). The resulting VOT steps were 10, 30, 35, 40, 50, and 85 ms.

Between subjects, we manipulated the level of cue conflict between VOT and subsequent lexical context. The high conflict group was exposed to more conflict between VOT and lexical context: tent- and dent-biasing contexts were equally likely to occur regardless of VOT. By contrast, in the low conflict group, VOT and context covaried in a naturalistic way such that context was more likely to occur with VOT steps that were biased in the same direction. This is illustrated in Fig. 1(B).

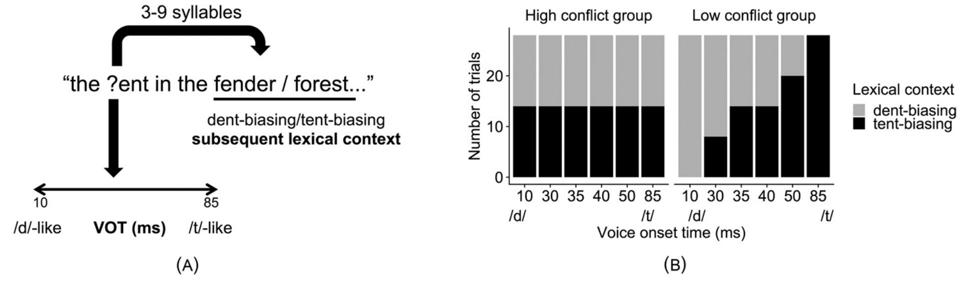

Fig. 1. Experimental design. (A) General experimental paradigm. (B) Between-participant manipulation of the correlation (and thus conflict) between the two cues.

2.3 Procedure

Participants were instructed to listen to the sentence and report whether they heard the word "tent" or "dent." Participants had to listen to the entire sentence, ensuring that they had a chance to process the lexical context before making a decision. To keep the length of the experiment manageable, each participant was exposed to 14 of the 40 sentence frames [the same as in Bushong and Jaeger (2017)]. Each sentence frame was repeated 12 times [6 times for each of the 2 contexts, with VOT distributions as shown in Fig. 1(B)] for a total of 168 trials. So as to maximize statistical power, we balanced how often each sentence frame was used across participants.

2.4 Analysis

Following our previous work (Bicknell *et al.*, in review; Bushong and Jaeger, 2017), we excluded participants who showed no significant effect of VOT on their categorization responses, suggesting that they did not understand the task or had poor audio equipment. This was determined by fitting a logistic regression (Jaeger, 2008) predicting tent-responses from VOT for each individual participant. This exclusion criterion resulted in the removal of 14 subjects from analysis: 9 (15%) from the high conflict group and 5 (8%) from the low conflict group. Including these participants did not change the pattern of results.

Unlike in previous work, we are interested in analyzing changes in the weighting of cues across the experiment, and compare these changes as a function of group. We thus employ a generalized additive mixed model (GAMM; Lin and Zhang, 1999) to detect possible non-linear changes in cue weights across trials. Specifically, we used the bam function of the mgcv library (Wood, 2011) in R (R Core Team, 2016).

We predicted tent-responses as a linear function of VOT (z-scored), lexical context (sum-coded; $1=$tent-biasing vs $-1=$dent-biasing), group (sum-coded; $1=$high conflict vs $-1=$low conflict), and each of their two-way interactions. We included smoothed terms aimed at investigating non-linear effects over the course of the experiment: a smooth of trial (centered and log-transformed); the interaction between the trial smooth and VOT; the three-way interaction between the trial smooth, VOT, and group; and the three-way interaction between the trial smooth, lexical context, and group. Since available GAMM implementations do not accommodate three-way interactions between a smooth and two categorical predictors, we created a dummy-coded context-group variable (with $2\times 2=4$ levels for the combinations of context and group). The interaction between the trial smooth and this context-group variable captures the three-way interaction.[3]

3. Results

Figure 2 shows the overall proportion of tent-responses averaged across the experiment for both groups. In line with Fig. 2, the GAMM analysis found parametric effects of both VOT ($\hat{\beta} = 2.35, z = 23.44, p < 0.001$) and lexical context ($\hat{\beta} = 0.51, z = 8.01, p < 0.001$). The interaction between conflict group and lexical context was significant, such that the context effect was smaller in the high conflict group ($\hat{\beta} = -0.4, z = -9.25, p < 0.001$). There was also an interaction between group and VOT: on average (across trials), the effect of VOT was larger in the high conflict group ($\hat{\beta} = 0.13, z = 3.07, p = 0.002$).[4] There was no main effect of group ($p=0.27$).

We now turn to changes across the experiment. For the low conflict group, the probability of tent-responses in tent-biasing contexts declined significantly throughout the experiment (reference smooth: $edf=3.93$, $\chi^2=16.544$, $p=0.003$); tent-responses in the dent-biasing context followed the same pattern (difference smooth: $p=0.57$). For the high conflict group, tent-responses in tent-biasing contexts did not differ from those in the low conflict group, declining across trials (difference smooth: $p=0.2$). Critically though, tent-responses in the dent-biasing context exhibited a different pattern, such that tent-responses increased over trials ($edf = 1$, $\chi^2 = 4.69$, $p=0.03$). As shown in Fig. 3, participants in the high conflict group—the design used in most previous studies—did not exhibit a context effect throughout most of the experiment. Neither the two-way interaction between trial and VOT, nor the three-way interaction between trial, VOT and group were significant ($ps > 0.19$, see Fig. 4).

4. Discussion

Replicating previous work, we found that VOT and subsequent lexical context together influenced participants' categorization decisions, suggesting that listeners integrate these two cues (e.g., Bushong and Jaeger, 2017; Connine *et al.*, 1991; Szostak and Pitt, 2013). Of primary interest to the present study, the relative weightings of the two cues

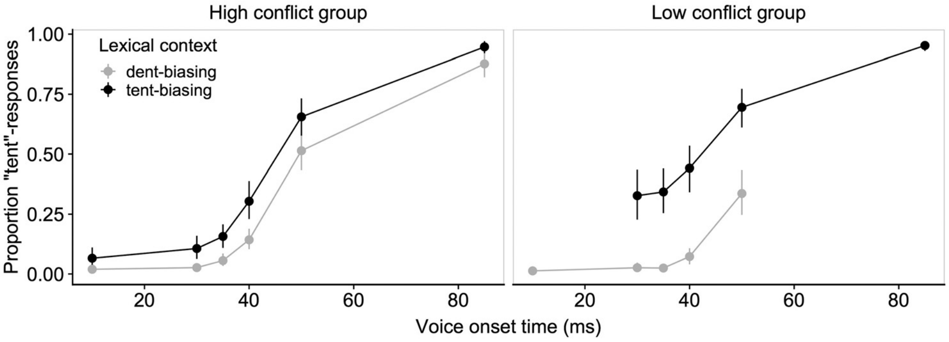

Fig. 2. Proportion of tent-responses by VOT, context, and group, averaged across the experiment. Error bars are bootstrapped 95% confidence intervals (CIs) over subject means.

changed when we manipulated their correlational structure. When VOT and lexical context were completely decorrelated and thus frequently in conflict (as in previous work) the effect of lexical context cues on categorization decisions decreased significantly through the experiment. When the two cues were correlated (as they would be in natural language), the effect of lexical context cues on categorization judgments remained unchanged over the course of the experiment. Figure 3 suggests that these changes occurred rather rapidly, within the first third of the experiment. This suggests that listeners were able to re-weight lexical context as soon as they observed more evidence about the correlational structure.

Decorrelating the two cues also changed the effect of VOT. Participants who were exposed to decorrelated cues had steeper categorization curves, suggesting a higher relative weighting of VOT as compared to participants in the naturalistic condition. The GAMM analysis found no significant change in this effect over the course of the experiment. This is somewhat puzzling as it would suggest the difference is present from the start of the experiment, and thus prior to exposure to our manipulation. One possibility is that we did not have enough power to detect such an effect, or the change occurred so rapidly that we could not detect it with our analysis method. Another potential explanation is that different participants had slightly different *a priori* weightings of VOT and lexical context because of different prior linguistic experience.

The present study has important methodological implications. It is not uncommon that studies on the integration of contextual and acoustic information intentionally decorrelate the cues in the input presented to participants. The present results suggest that this type of design—corresponding to our high conflict group—systematically *underestimates* the effect of contextual cues on spoken word recognition. This has serious consequences for studies where theoretical arguments hinge on whether context affects categorization (as is the case in, e.g., Connine *et al.*, 1991; Szostak and Pitt, 2013). Specifically, the present results suggest that subsequent context might be a more important cue in spoken word recognition than previously suggested.

The present findings further suggest that listeners are able to dynamically re-weight acoustic and contextual cues, perhaps reflecting adaptation to the cues' current statistical structure. Regardless of the specific mechanism, our results resemble

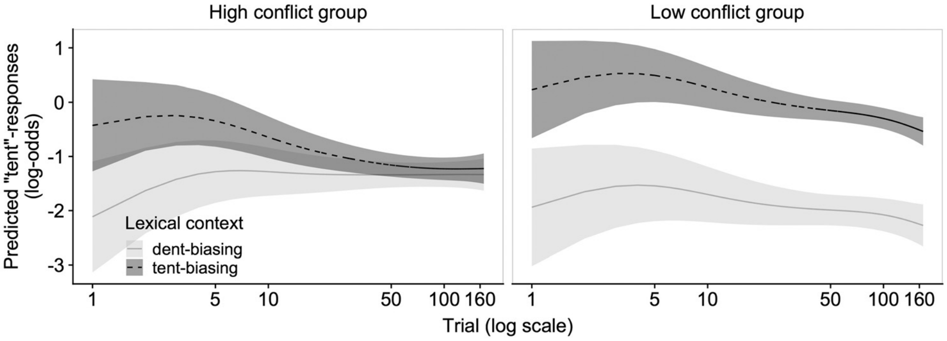

Fig. 3. Effect of lexical context across trials for each conflict group, as determined by GAMM analysis. Bands are 95% GAMM-derived CIs. Predictions were computed at the average VOT value. The wide CIs for the beginning of the experiments are a consequence of log-transforming trial (fewer observations enter those CIs). The estimated variance of participants' responses did not vary across trials (Figures S1 and S2 in the supplementary material) (footnote 2).

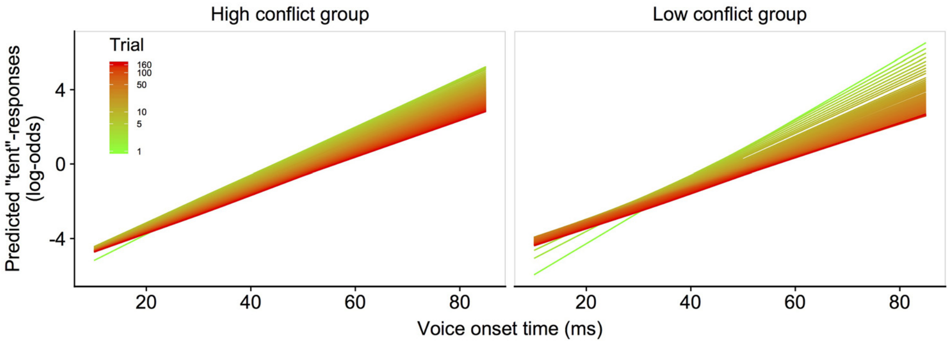

Fig. 4. (Color online) Effect of VOT across trials (note the log scale) for each conflict group, as determined by GAMM analysis. Light gray (green online) corresponds to earlier trials and dark gray (red online) to later trials. While there was a numerical trend for the slope of the categorization curves to decrease over the course of the experiment, this effect was not significant.

those of previous work on multiple *acoustic* cues. These studies might also hold the key to why we observe relative down-weighting of contextual instead of acoustic cues: in English, for example, VOT is the most reliable cue for categorizing voicing while other cues like *F*0 and vowel length are less reliable and listeners typically weight VOT higher in their categorization judgments than other cues (Francis *et al.*, 2008; Idemaru and Holt, 2011; Toscano and McMurray, 2010). In studies of acoustic cue re-weighting, the less reliable cue is the one that tends to be down-weighted in cases of cue conflict (i.e., *F*0 is down-weighted when it conflicts with VOT; Idemaru and Holt, 2011). Our results thus could suggest that under certain conditions, listeners consider lexical context less *a priori* reliable than acoustic cues in spoken word recognition. We note, however, that the acoustic cue tested here corresponds to the primary (most reliable) cue to voicing. Further work is needed to address how listeners estimate these relative reliabilities for different (potentially less reliable) acoustic cues and in different situations (e.g., speaker accent; Schertz and Hawthorne, 2018).

Acknowledgments

This research was partially funded by NIHCD R01 HD075797 (to T.F.J.) and an NSF NRT No. 1449828 fellowship (to W.B.). The views expressed here do not necessarily reflect those of the funding agencies. The authors would like to thank Evan Hamaguchi and Chelsea Marsh for assistance with stimulus sentence creation and recording.

References and links

[1] Here we are not interested in the effect of the distance between the? -sound and subsequent lexical context. Follow-up analyses not reported here confirmed that this distance did not affect categorization (in line with Bushong and Jaeger, 2017).

[2] See supplementary material at https://doi.org/10.1121/1.5119271 for Figs. S1 and S2.

[3] See https://osf.io/924cx/ for the data and analysis scripts. Specifically, we use reference and difference smooths, with the reference level set to the /t/-biasing context in the low-conflict group. This allows us to pinpoint which conditions drive the three-way interaction. In this coding, the "main" trial smooth captures the effect of trial for the reference level of the context-group variable; the interactions capture differences between these trial effects and those observed for the other three levels of the context-group variable.

[4] Figure 3 also seems to suggest a difference in the context effect between groups at the beginning of the experiment. Follow-up analyses found that this difference was not significant.

Burchill, Z., Liu, L., and Jaeger, T. F. (**2018**). "Maintaining information about speech input during accent adaptation," PLoS One **13**(8), e0199358.

Bushong, W., and Jaeger, T. F. (**2017**). "Maintenance of perceptual information in speech perception," *Thirty-Ninth Annual Conference of the Cognitive Science Society*.

Christiansen, M. H., and Chater, N. (**2016**). "The now-or-never bottleneck: A fundamental constraint on language," Behav. Brain Sci. **39**, 1–72.

Connine, C. M., Blasko, D. G., and Hall, M. (**1991**). "Effects of subsequent sentence context in auditory word recognition: Temporal and linguistic constraints," J. Memory Lang. **30**(1), 234–250.

Dahan, D. (**2010**). "The time course of interpretation in speech comprehension," Current Direct. Psychol. Sci. **19**(2), 121–126.

Francis, A. L., Kaganovich, N., and Driscoll-Huber, C. (**2008**). "Cue-specific effects of categorization training on the relative weighting of acoustic cues to consonant voicing in English," J. Acoust. Soc. Am. **124**(2), 1234–1251.

Ganong, W. F. (**1980**). "Phonetic categorization in auditory word perception," J. Exp. Psychol. **6**(1), 110–125.

Idemaru, K., and Holt, L. L. (**2011**). "Word recognition reflects dimension-based statistical learning," J. Exp. Psychol. **37**(6), 1939–1956.

Jaeger, T. F. (**2008**). "Categorical data analysis: Away from anovas (transformation or not) and towards logit mixed models," J. Memory Lang. **59**(4), 434–446.

Kalikow, D. N., Stevens, K. N., and Elliott, L. L. (**1977**). "Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability," J. Acoust. Soc. Am. **61**(5), 1337–1351.

Kingston, J., and Diehl, R. L. (**1994**). "Phonetic knowledge," Language **70**(3), 419–454.

Lin, X., and Zhang, D. (**1999**). "Inference in generalized additive mixed models by using smoothing splines," J. Royal Stat. Soc. Ser. B **61**(2), 381–400.

Lisker, L., and Abramson, A. S. (**1967**). "Some effects of context on voice onset time in English stops," Lang. Speech **10**(1), 1–28.

Liu, L., and Jaeger, T. F. (**2018**). "Inferring causes during speech perception," Cognition **174**, 55–70.

McClelland, J. L., and Elman, J. L. (**1986**). "The trace model of speech perception," Cognitive Psychol. **18**(1), 1–86.

McMurray, B., Tanenhaus, M. K., and Aslin, R. N. (**2009**). "Within-category VOT affects recovery from 'lexical' garden-paths: Evidence against phoneme-level inhibition," J. Memory Lang. **60**(1), 65–91.

Norris, D., and McQueen, J. M. (**2008**). "Shortlist b: A Bayesian model of continuous speech recognition," Psychol. Rev. **115**(2), 357–395.

Oden, G. C., and Massaro, D. W. (**1978**). "Integration of featural information in speech perception," Psychol. Rev. **85**(3), 172–191.

R Core Team (**2016**). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, available at https://www.R-project.org/ (Last viewed 5 April 2019).

Schertz, J., Cho, T., Lotto, A., and Warner, N. (**2016**). "Individual differences in perceptual adaptability of foreign sound categories," Attn., Percept., Psychophys. **78**(1), 355–367.

Schertz, J., and Hawthorne, K. (**2018**). "The effect of sentential context on phonetic categorization is modulated by talker accent and exposure," J. Acoust. Soc. Am. **143**(3), EL231–EL236.

Szostak, C. M., and Pitt, M. A. (**2013**). "The prolonged influence of subsequent context on spoken word recognition," Attn., Percept., Psychophys. **75**(7), 1533–1546.

Toscano, J. C., and McMurray, B. (**2010**). "Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics," Cognitive Sci. **34**(3), 434–464.

Wood, S. N. (**2011**). "Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models," J. Royal Stat. Soc. B **73**(1), 3–36.

Xie, X., Weatherholtz, K., Bainton, L., Rowe, E., Burchill, Z., Liu, L., and Jaeger, T. F. (**2018**). "Rapid adaptation to foreign-accented speech and its transfer to an unfamiliar talker," J. Acoust. Soc. Am. **143**(4), 2013–2031.