Review: “Temporal (in)stability in English monosyllabic and disyllabic words: Insights on the effect of voicing on vowel duration”

1. Summary

This paper investigates the effect of voicing on the duration of the release-to-release interval in English. It is argued that, based on the presented evidence, that voicing effects the second consonant in disyllabic words, but not in monosyllabic words. A biomechanical explanation is put forward, arguing that this pattern is a reflex of the gestural phasing of vocalic gestures.

2. Verdict

The reviewer self-identifies as Timo Roettger (timo.b.roettger@gmail.com). This paper is a timely and well-executed contribution to the literature and can shed light on possible asymmetries in the production of words containing voiced sounds.

As opposed to many other empirical studies in linguistics, the author makes their data and scripts available, giving the reader the opportunity to reproduce their analysis, reanalyze their data, and critically evaluate their empirical claims. This is in itself a scholarly contribution. Moreover, the study was preregistered, thus clearly drawing a visible line between confirmatory and exploratory data analysis. The analysis utilizes state-of-the-art Bayesian regression models and embraces the estimation of uncertainty. Taking these points together, this paper has the potential to be a prime example of how experimental phonetics can safely navigate the inherent noisiness of its subject (speech) without overconfidently overstating observed patterns. Having said all that, I have several issues with the analysis and the interpretation that need to be resolved before I can recommend this paper for publication (see a summary and demonstration below). I will append an R script to this review in which I offer solutions to these issues and I’d highly welcome it if the author wants to get in touch to discuss these matters. Beyond these analytical questions, I think there is some room for improving the narrative of the paper. On the one hand, I as a reader did not immediately understand the relevance and scope of the presented findings. I recommend a clearer outline of the anticipated impact of the findings in the introduction (including some signposting of it very early one). Moreover, I found the link between formulated hypotheses and the presentation of the data analysis a little bit out of touch. I recommend streamlining both sections and aligning them more closely. This issue also affects the actual description of the data analysis. Given that Bayesian inference is still a rather new method in the linguistic tool box, I recommend better guidance of the reader when walking through the analysis. I made extensive use of the comment function within the pdf to make concrete suggestions about how to do this.

3. **Analytical issues**

Since the author provided his data and scripts, I was able to critically evaluate the choices made. Note that I did not thoroughly check how well the preregistration maps onto the presented analysis, but since I recommend overhauling the analysis, the analysis will diverge from the planned analysis anyway.
I identify three families of issues.

1. The interpretation and presentation of dummy coded regression coefficients.
2. The use of informative priors when testing hypotheses.
3. The use of anti-conservative random effect structures.

(re: 1) The author uses dummy-coded predictors. For simplicity sake, let's simply look at a model that involves voicing (voiced vs. voiceless), syllable number (monosyllabic vs. disyllabic) and their 2-way interaction. If dummy coded, the regression will estimate a reference level as the intercept (in his analysis voiceless + disyllabic); the effect of voicing on that reference level (e.g. the effect of voicing in disyllabics); the effect of syllable number on that reference level (e.g. the effect of syllable number on voiceless); and an interaction term that describes the additional change in the DV if both predictor levels need to be changed relative to the reference levels. This results in the following additions if we want to arrive at predicted values for all four categories:

Voiceless disyllabic = Intercept
Voiceless monosyllabic = Intercept + syllable.number
Voiced disyllabic = Intercept + voicing
Voiced monosyllabic = Intercept + syllable.number + voicing + Interaction(syllable.number x voicing)

Thus, the interaction coefficient is numerically not interpretable in a straightforward way but must rather be related to the other coefficients. There is nothing wrong with this coding scheme, however, great care needs to be taken when interpreting and reporting these effects. There are some rare formulations in the manuscript that indicate that the author indeed interprets these coefficients correctly, but the way the analysis is reported often suggests a misinterpretation (or at least will lead to many misinterpretations on the side of the reader). I think what the author actually wants to look at (and discuss) is the voicing effect in monosyllabics, the voicing effect in disyllabics, and the difference between these voicing effects. This can either be achieved with contrast coding or, since we are in Bayesian land, these effects can simply be extracted from the posterior distributions. I added some code to the script to offer a (non-elegant) way to extract said posteriors and will present them accordingly below. In short, I highly recommend to present the results in a more intuitive way. This will tremendously help the reader and make the comparison to the literature much easier. It might actually change the interpretation quite drastically. Again, this is an easy fix which requires some work on the analysis and rewriting the results.

(re: 2) The second point is conceptually a more severe problem in my opinion (but can also be easily fixed). The author wants to test the effect of voicing on the acoustic measures discussed. Using Bayesian statistics offers a whole new family of degrees of freedom to any analysis: Priors. For any Bayesian analysis, the researcher needs to formulate prior beliefs about plausible values. In hypothesis-testing scenarios (such as the present one), we commonly specify so-called weakly informative priors, i.e. priors that are agnostic about a relative effect, but constrain the parameter space reasonably well.

An agnostic researcher departs from the assumption that voicing does not affect the dependent variable, so there is no difference between voiced and voiceless. Moreover, we should be uncertain about which values are plausible values. Our best prior guess is that there is, on average, no voicing effect (mean = 0) but we allow for this difference to vary quite a bit by adding noise to this guess (e.g. a normal distribution centered on 0 with a certain spread). Now the author does two things, that I don't agree with and that might be problematic:

First, the author tests a hypothesized relationship between syllable number and voicing. Thus, the author should be agnostic about the presence of such an interaction and use a weakly informative prior centered on zero. But the author instead uses an informative prior (mean = 50, SD = 25), basically building in the expected effect. When there are enough data, the prior won't matter much, but as I will show below it does add more uncertainty and decreases the effect estimates. I recommend using weakly informative priors for all relevant parameters that are part of the tested hypotheses.

Second, the author lengthily discusses the important issue of effect magnitudes in the literature and points to concerningly low sample sizes in previous studies (i.e. small sample sizes and publication bias can lead to systematic overestimation of underlying effects in the published literature.). Thus, the present study sets out to assess these magnitudes with a larger sample. Good. But if this is one of the goals of the analysis, the priors should give this assessment a fair shot. The author sets rather narrow priors for the voicing effect with a SD of 25 ms in light of the fact that some of the previous studies report effects of up to 150 ms. I recommend using more liberal priors in order to avoid down-weighting of possibly large values.

(re: 3) Finally, the author uses anticonservative random effect structures, artificially decreasing estimated variance and thus possible increasing overconfident certainty estimations. Concretely, the interaction of voicing and syllable number is central to some of the proposed hypotheses, but the interaction is not added as a random slope for speakers (allowing the model to consider speaker-specific variation regarding said interaction). Luckily, this might actually not make a large difference (at least for model 1). I thus highly recommend adding these slopes. It is more convincing to more analysis-savvy readers and potentially does not even increase the estimated uncertainty.

I reran the 1st model and present the results of (i) the original analysis, (ii) an alternative analysis with appropriate priors (based on my own judgement), and (iii) an alternative model with appropriate priors and appropriate random effects structure. The table displays the posterior means and 95% Credible Intervals for seven different parameters. Absolute estimates for each voicing x syllable number combination (1-4), the difference between voiced and voiceless in monosyllabics (5) and disyllabics (6), and the difference between the voicing effects (7, corresponding to a traditional interaction term).

First some good news. It becomes apparent that the posteriors are rather similar. So, none of my remarks changes things categorically. However, since decisions are made based on magnitude of the effects and based on its associated uncertainty, these changes might matter and point me even more to an interpretation that assumes that there is nothing compelling to observe. The voicing effect for monosyllabic words is more uncertain and numerically smaller (-12 vs. -8, thus. in fact below the JND). This becomes even clearer for the interaction effect which decreases from 17 in the original analysis to 10 in my proposed analysis.

From what I can see in these results is the following: There is no compelling evidence for neither a voicing effect in monosyllabic words (0 is a very plausible value in all posterior distributions), nor in disyllabic words, and there is no compelling evidence for an interaction effect. Since the author's research questions and discussion surrounds these effects, I am not sure whether a more appropriate narrative here would be the following: Given the data set (which is much larger than what other people have done), and the model (which is likely to be more appropriate than previous analyses in the literature), there is no compelling evidence for any of these effects. It is of course possible that these effects are too small and variable, that a much larger sample is necessary to find compelling evidence for them, but that really raises the question whether the potentially pay-off of showing these patterns are worth the immense resources required for their investigation. I think such a narrative would be well-worth publishing in *Laboratory Phonology*.

Parameter	**(i) original analysis**	**(ii) with appropriate priors**	**(iii) appropriate slopes + (ii)**
voiceless disyllabic	264 [245,283]	264 [244,286]	265 [246,283]
voiceless monosyllabic	281 [262,301]	281 [261,303]	281 [260,302]
voiced disyllabic	259 [240,279]	263 [241,284]	262 [242,281]
voiced monosyllabic	293 [273,313]	289 [269,311]	289 [267,312]
voicing effect monosyllabic	-12 [-33,9]	-8 [-30,15]	-8 [-31,15]
voicing effect disyllabic	4 [-14,25]	2 [-21,26]	2 [-19,24]
diff of voicing effects	-17 [-42,8]	-10 [-43,20]	-10 [-42,20]

All in all, I recommend overhauling the analysis according to my suggestions, potentially taking my R code as a departure point. I also highly recommend reconsidering the interpretation of the results in light of my remarks. These changes might have two consequences: Given that the effect estimates will probably be much more uncertain, the relevance of the ROPE for the presentation of the results will be less useful (as most effects overlap with it). That said, a reference to JNDs is still useful and desirable. Second, the initial separation between confirmation and exploration via the preregistration might become less compelling given these changes to the analysis. However, given that these changes are all more conservative than the preregistered analyses, this does not take away anything from the value of the results.

Please find attached an annotated .pdf and a zipped data analysis folder with my own script names as *analysis_TR.R.*