Compensatory aspects of the effect of voicing on vowel duration in English

Stefano Coretta

1 Introduction

Consonants and vowels are known to exert a reciprocal influence on each other in a variety of ways. One such way is the well-established tendency for vowels to have shorter durations when followed by voiceless stops and longer durations when followed by voiced stops. This so-called 'voicing effect' has been long recognised in a plethora of languages across different linguistic families. English is possibly by far the most investigated language in relation to the voicing effect. Several hypotheses have been proposed as to the origin of this phenomenon, however no one particular hypothesis has gained unequivocal consensus.

Coretta (2018) proposes to seek the source of the voicing effect in a compensatory mechanism between vowel and consonant closure duration. In an exploratory study of acoustic durations in 17 Italian and Polish speakers, Coretta finds that the duration of the interval between two consecutive consonant releases is not affected by the voicing status of the second consonants. In particular, Coretta investigated disyllabic words of the form CV́CV (with lexical stress on the first syllable), where the second consonant was either voiceless or voiced. The results of that study suggest that, for example, the duration of the release-to-release interval in a word like /pata/ is non-significantly different from that in /pada/.

Coretta (2018) argues that the temporal stability of the release-to-release interval is compatible with a compensatory temporal adjustment account of the voicing effect, which states that vowels are shorter when followed by voiceless stops because the latter have long closure durations, and vice versa with voiced stops, which have shorter closure durations (Lindblom 1967; Slis & Cohen 1969b,a; Lehiste 1970b,a). Coretta also reviews some of the shortcomings of the compensatory account and concludes that the release-to-release invariance offers a resolution to those. In particular, previous versions of the account are not clear on within which speech interval the compensation is implemented. Both the syllable (Lindblom 1967; Farnetani & Kori 1986) and the word (Slis & Cohen 1969b,a; Lehiste 1970b,a) have been proposed but subsequently criticised on empirical grounds (Chen 1970; Jacewicz et al. 2009; Maddieson & Gandour 1976).

Given the temporal stability of the release-to-release interval and the differential closure duration in voiceless vs. voiced stops (Lisker 1957; Van Summers 1987; Davis & Van Summers 1989; de Jong 1991), it follows that the timing of the consonant closure onset within that interval will decide on the respective durations of vowel and consonant closure. Coretta (2018) thus proposes that the source of the voicing effect can be seen in the temporal stability of the release-to-release interval in relation to the voicing of the second consonant and the effect of voicing on closure duration. Other properties of speech production and perception can of course further contribute to the emergence and enhancement of the voicing effect (see Beguš 2017 for an overview).

English is generally the language given as an example in which the voicing effect has a big magnitude. This fact is normally attributed to phonologisation of the voicing effect in this language Sharf (1964); de Jong (2004). Indeed, previous studies on English report a difference in vowel duration before voiceless vs. voiced stops which ranges between 20 and 150 ms. A Bayesian meta-analysis of the English voicing effect (See Supplemental materials) indicates that 95% credible interval for the effect of voicing in monosyllabic words is between 55 and 95 ms. This means that, based on the data,

the true effect lies within that interval at a probability of 95%, on in other words we can be 95% sure that the effect is between 55-95 ms.

1.1 Research hypotheses

The principal aim of this study is to test whether the same temporal stability observed for the release-to-release interval in Italian and Polish disyllabic words can be observed in English. Jacewicz et al. (2009) report that monosyllabic words in American English are longer when the second consonant is voiced. Based on this finding, it is expected that the duration of the release-to-release interval will differ in monosyllabic words depending on C2 voicing. More specifically, the release-to-release duration should be longer when C2 is voiced. Jacewicz et al. (2009) attribute the difference in word duration to the difference in vowel duration before voiceless vs. voiced stops. Thus, we can expect the magnitude of the difference in release-to-release duration to be close to the difference in vowel duration.

The data in Coretta (2018) suggest that the intrinsic duration of vowels and consonants can contribute to the duration of the release-to-release interval. In particular, release-to-release intervals containing a high vowel have shorter durations than those with a low vowel in Coretta (2018). This is not surprising, given that a well known cross-linguistic tendency is that high vowels are shorter than low vowels (Hertrich & Ackermann 1997; Esposito 2002; Mortensen & Tøndering 2013; Toivonen et al. 2015; Kawahara et al. 2017). As for the consonant place of articulation, the interval is shorter in Italian and Polish if the second consonant is velar compared to when it is coronal. The closure of velar stops is shorter than that of other stops. For example, Sharf's data 1962 on closure duration in English suggests that the closure of labial stops (60-90 ms) is about 10 ms longer than that of velar stops (55-75 ms). It can thus be expected that intervals with a velar stop in English will be about 10 ms shorter than intervals with labial consonants.

To summarise, the following research questions and respective hypotheses were formulated:

- Q1: Is the duration of the interval between two consecutive stop releases (the release to release interval) in monosyllabic and disyllabic words affected by the voicing of C2 in English?
  - H1a: The duration of the release to release interval is not affected by C2 voicing in disyllabic words.
  - H1b: The release to release interval is longer in monosyllabic words with a voiced C2 than in monosyllabic words with a voiceless C2.
- Q2: Is the duration of the release to release interval affected by (a) the number of syllables of the word, (b) the quality of V1, and (c) the place of C2?
  - H2a: The release to release interval is longer in monosyllabic than in disyllabic words.
  - H2b: The duration of the release to release interval decreases according to the hierarchy /ɑ:/, /ɜ:/, /i:/.
  - H2c: The release to release interval is shorter when C2 is velar.
- Q3: What is the estimated difference in the effect of voicing on vowel and stop closure duration between monosyllabic and disyllabic words?
  - H3: The effect of voicing on vowel duration is greater in monosyllabic than in disyllabic words.

2 Methods

The research design and data analyses of this study has been pre-registered at the Open Science Framework.

2.1 Participants

The participants of this study were 15 native speakers of British English, who were born and brought up in the Greater Manchester area. All the speakers were undergraduate students at the University of Manchester with no reported hearing or speaking disorders, and with normal or corrected to normal vision. The participants signed a written consent form and received £5 for participation.

Sample size and stopping rule were determined prior to data collection by the Region Of Practical Equivalence (ROPE) method (Kruschke 2015; Vasishth et al. 2018b). A 'no-effect' region of values around 0 is first identified. The null region (the ROPE) can be thought of as the Bayesian 95% credible interval from a distribution, the values within which can be interpreted as a null effect. For this study, a ROPE between −10 and +10 ms has been chosen. The width of 20 ms is based on the estimates of the just noticeable difference in Huggins (1972) and Nooteboom & Doodeman (1980).

Once a ROPE width is set, the goal is to collect data until the width of the 95% credible interval of the tested effect is equal to or less than the ROPE width (in this study, 20 ms). Due to resource and time constraints specific to this particular study, a second condition had to be included in the stopping rule such that data collection would be have to stop on April 5th 2019, independent of the the ROPE condition.

2.2 Equipment

Audio recordings were obtained in a sound-attenuated room in the Phonetics Laboratory of the University of Manchester, with a Zoom H4n Pro recorder and a RØDE Lavalier microphone, at a sample rate of 44100 Hz (16-bit, downsampled to 22050 Hz for analysis). The Lavalier microphone was clipped on the participants cloths, about 20 cm under their mouth, displaced a few centimetres on one side.

2.3 Materials

The test words were $C_1V_1C_2$(VC) words, where C_1 = /t/, V_1 = /iː, əː, ɑː/, C_2 = /p, b, k, g/, and (VC) = /əs/. This structure specification generates 24 test words, shown in **??**. Each word was embedded in the following frame sentences: *I'll say X this Thursday*, *You'll say X this Monday*, *She'll say X this Sunday*, *We'll say X this Friday*, *They'll say X this Tuesday*. Each word + frame combination was included once in the stimuli list, so that each speaker would read 120 sentence stimuli (24 words × 5 frames).

2.4 Procedure

The participants were first debriefed on the experimental procedure. Prior to recording, the participants familiarised themselves with the materials by reading them aloud and were instructed not to insert pauses anywhere within the sentence stimuli and to try and keep a similar intonation contour for the total duration of the experiment. They were also given the change to take any number of breaks at any point during recording. Misreadings or speech errors were corrected by letting the participant repeat the stimulus. The reading task took around 6 to 10 minutes.

2.5 Data processing and measurements

The audio recordings were downsampled to 22050 Hz for analysis. A forced-aligned transcription was obtained with the SPeech Phonetisation Alignment and Syllabification software (SPPAS, Bigi 2015). The automatic annotation was corrected by the author according to the principles of phonetic segmentation detailed in Machač & Skarnitzl (2009). A custom Praat script was written to automatically detect the burst onset of the consonants in the test words, using the algorithm in Ananthapadmanabha et al. (2014). The output was checked and manually corrected by the author when necessary.

The following measures were obtained:

- Duration of the release-to-release interval: from the release of C1 to the release of C2.
- V1 duration: from appearance to disappearance of higher formant structure in the spectrogram in correspondence of V1 (Machač & Skarnitzl 2009).
- C2 closure duration: from disappearance of higher formant structure in the V1C2 sequence to the release of C2 (Machač & Skarnitzl 2009).
- Speech rate: calculated as number of syllables per second (number of syllables in the sentence divided by the sentence duration in seconds, Plug & Smith 2018).

2.6 Statistical analysis

All statistical analyses were performed in R v3.5.2 (R Core Team 2018). Bayesian regression models were fit with brms (Bürkner 2017, 2018). Each model was run with four MCMC chains and 2000 iterations per chain, of which 1000 for warmup. A Gaussian (normal) distribution was used in all the models as the response distribution. The posterior predictive check plots indicate that the observed distributions are slightly positively skewed so that a log-normal distribution would have been more appropriate. However, the deviations from a Gaussian distribution are minimal so that using a Gaussian distribution is still acceptable. It is desirable that future studies would specifically investigate the distributional properties of speech segment durations.

All factors were coded using treatment contrasts with the first level as the reference level: number of syllables (disyllabic, monosyllabic), vowel (/ɑː/, /ɜː/, /iː/), C2 voicing (voiceless, voiced), C2 place of articulation (velar, labial).[1] Speech rate has been centred when included in the models so that the intercept could be interpreted as the intercept at mean speech rate. Model convergence was reached in all the models reported here ($\hat{R} = 1$) and no major divergences in the MCMC chains were observed.

The choice of Bayesian over frequentist statistics stems from the misplaced reliance on *p*-values in a context of commonly low-powered studies (Munafò et al. 2017; Roettger 2019). For an introduction to Bayesian statistics in phonetics, see Vasishth et al. (2018a); Nicenboim et al. (2018), while for a general introduction see Etz et al. (2018); McElreath (2015); Kruschke (2015), and reference therein. While a thorough discussion of Bayesian methods would be beyond the scope of this paper, it is relevant to give the less familiar reader the tools for interpreting analyses and results.

More weight will be given here on the estimated distributions of the sought effects, rather than on point estimates (as in frequentist regression models). The estimated distribution of an effect (or parameter) is the posterior distribution of that effect (or parameter). The posterior distribution is an approximation of the parameter distribution, and it takes into account the specified prior, i.e. the theoretical probability of the parameter as known or derived by the researcher. This characteristics is at the heart of Bayesian modelling, which includes prior knowledge into the estimation of parameter values. For each relevant term in the models, the 95% credible intervals (CI) should be taken as a

[1] Note that the order of the levels in the vowel factor are reversed compared to that in the pre-registration, and this was done to match the height order in Coretta (2018), from low to high. Changing the order of the levels of course does not affect the results.

summary of the posterior distribution, and inference should be based on the posterior rather than on the point estimate (the posterior mean, represented here with $\bar{\theta}$). A 95% CI can be interpreted as the 95% probability that a parameter lies within the interval range. For example, if the 95% CI is between 10 and 30 ms, there is a 95% probability that the true parameter value is between 10 and 30 ms.

In each model, a prior is specified for each of the parameters to be estimated. The priors are in the form of particular distributions, like the normal or the Cauchy distributions. A prior defines the prior knowledge of where the parameter might lie within a range of values. For example, a prior as a normal distribution with mean 200 ms and standard deviation 50 indicates the researcher's belief that the parameter lies between 100 and 300 ms with 95% probability (i.e., the mean minus twice the standard deviation, and the mean plus twice the standard deviation).

3 Results

3.1 Release-to-release duration

A Bayesian regression model was fit to model the duration of the release-to-release interval. The following terms were included as fixed effects: C2 voicing (voiceless, voiced), number of syllables (disyllabic, monosyllabic), centred speech rate, an interaction between C2 voicing and number of syllables. A by-speaker and by-word random intercept, and a by-speaker random coefficient for C2 voicing were entered as random effects. The following priors have been used. Two weakly informative priors based on the results from Coretta (2018) were chosen for the intercept and the effect of C2 voicing. The former prior was a normal distribution with mean 200 ms and SD = 50, while the latter a normal distribution with mean 0 ms and SD = 25. The same prior was specified for the effect of number of syllables and the interaction between C2 voicing and number of syllables. This weakly informative prior was a normal distribution with mean 50 ms and SD = 25, and it is based on differences in vowel duration in mono- vs. disyllabic words and in monosyllabic words before voiceless vs. voiced stops, which range in both cases between 30 and 100 ms (see Sharf 1962 and Klatt 1973 for the difference in vowel duration in mono- vs. disyllabic words). The prior for the effect of centred speech rate was a normal distribution with mean -25 ms and SD = 10, and was based again on results from Coretta (2018). For the random effects, a half Cauchy distribution (location = 0, scale = 25) was used for the standard deviation (also used for the residual standard deviation) and a LKJ(2) distribution for the correlation.

The posterior distribution of the estimated effect of C2 voicing on the release-to-release duration has a 95% credible interval (95% CI) between -23.86 and 15.45 ms (the mean is -4.43 ms, estimated error = 10.03). The 95% CI of the estimated interaction between C2 voicing and number of syllables is between -8.41 and 41.41 ms ($\bar{\theta}$ = 16.53 ms, SD = 12.72). The difference in duration of the release-to-release interval between monosyllabic and disyllabic words is between -1.58 and 36.53 ms (95% CI, $\bar{\theta}$ = 17.34, SD = 9.76). Speech rate has a strong negative effect on the release-to-release duration with 95% CI [-40.14, -32.13].

A second Bayesian regression was fitted with the release-to-release duration as the outcome variable to test the effects of vowel and C2 place of articulation, which were entered as terms in the model without interactions. Centred speech rate was also included. The relevant priors from the first model were kept. For the effects of vowel (/ɜ:/, /i:/) and place of articulation (labial), the very weakly informative prior was a normal distribution with mean = 0 ms and SD = 30. This prior was based on duration differences depending on vowel height (Heffner 1937; House & Fairbanks 1953; Hertrich & Ackermann 1997) and labial place Sharf (1962), which range between 10 and 30 ms.

The random effects structure was the same as with the first model. The posterior distribution of the effect of the vowel /ɜ:/ shows that this vowel tends to a somewhat negative effect, with a 95% CI

between −21.90 and 4.87 ms ($\bar{\theta}$ = −8.58 ms, SD = 6.9). The vowel /iː/ has a more robust negative effect on release-to-release duration, with a 95% CI between −50.10 and −22.26 ($\bar{\theta}$ = −36.94 ms, SD = 6.96). Less clear is the effect of C2 place of articulation (labial stop): the mean of the posterior is 2.46 ms (SD = 5.68), and the 95% CI is [−9.15, 13.28].

The credible intervals of the effects in the models reported above have widths which are greater than the chosen ROPE width of 20 ms. The large credible intervals indicate that the estimated posterior distributions of the effects have a somewhat high degree of uncertainty with them. The uncertainty has been possibly brought about by not controlling for vowel and number of syllables in respectively the first and second model. An exploratory model was thus fitted to the data, in which all the terms from the two models above were included together. The same priors of the two separate models were used in the combined model.

Including all the relevant terms in the model (C2 voicing and place, vowel, number of syllables in interaction with C2 voicing) has indeed reduced the credible intervals substantially. Note that inferences from this model should be treated with caution, since the model was not pre-registered. Figure 1 shows a variety of crebile intervals for the model terms. The posterior distribution of the C2 voicing effect on release-to-release duration is tighter than that of model 1 (95% CI = [−10.45, 5.65]) while the mean (−2.43 ms, SD = 4.06) is virtually unchanged (−4.43 ms, only a 2 ms difference). The estimated effect of syllable number is now more robustly positive (95% CI = [9.17, 22.48]), with a mean (16.03 ms, SD = 3.32) similar to that in model 1. The posterior distribution of the interaction between number of syllables and C2 voicing (95% CI = [2.65, 20.98]) suggests a positive and medium-sized coefficient ($\bar{\theta}$ = 11.67 ms, SD = 4.71). This result indicates that the duration of the release-to-release is greater in monosyllabic words with voiced C2 than in monosyllabic words with voiceless C2. The effects of vowel and place or articulation have similar means, but the credible intervals are smaller. The release-to-release is on average 10.05 ms (SD = 2.95, 95% CI = [−15.92, −4.24]) shorter if the vowel is /ɜː/ and 39.3 ms (SD = 2.99, 95% CI = [−45.03, −32.76]) shorter if the vowel is /iː/. C2 place of articulation (labial) has a negligible positive mean effect (2.6 ms, SD = 2.39, 95% CI = [−2.29, 7.28]).

3.2 Vowel duration

A Bayesian regression model was fitted to test vowel duration (the outcome variable in the model). The following terms were entered: C2 voicing (voiceless vs. voiced), vowel ((/ɑː/, /ɜː/, /iː/)), number of syllables (disyllabic, monosyllabic), centred speech rate, all possible interactions between C2 voicing, vowel, and number of syllables. The same random structure as in the previous models was used (a by-speaker and by-word random intercept, and a by-speaker random coefficient for C2 voicing).

For the prior of the intercept of vowel duration we used a normal distribution with mean 145 ms and standard deviation 30 (Heffner 1937; House & Fairbanks 1953; Peterson & Lehiste 1960; Sharf 1962; Chen 1970; Klatt 1973; Davis & Van Summers 1989; Laeufer 1992; Ko 2018). A normal distribution with mean 50 ms and standard deviation 20 was used as the prior for the effect of voicing on vowel duration (based on the above studies). A normal prior with mean 50 and standard deviation 25 was chosen instead for the effect of number of syllables and the interaction C2 voicing/number of syllables. For the effects of vowel, vowel/number of syllables interaction, and the three-way interaction vowel/number of syllables/C2 voicing, a normal distribution with mean 0 and standard deviation 30, based on differences reported in the previous studies. A slightly more informative prior was used for the interaction between C2 voicing and vowel (mean = 0, SD = 20). The same priors as in the previous model were included for the random structure.

The 95% CI of the posterior distribution of the duration of /ɑː/ is included in the range 113.15–137.06 ms ($\bar{\theta}$ = 125.16 ms, SD = 6.02). The vowel /ɜː/ is 9.19 ms shorter (SD = 5.29) with CI =

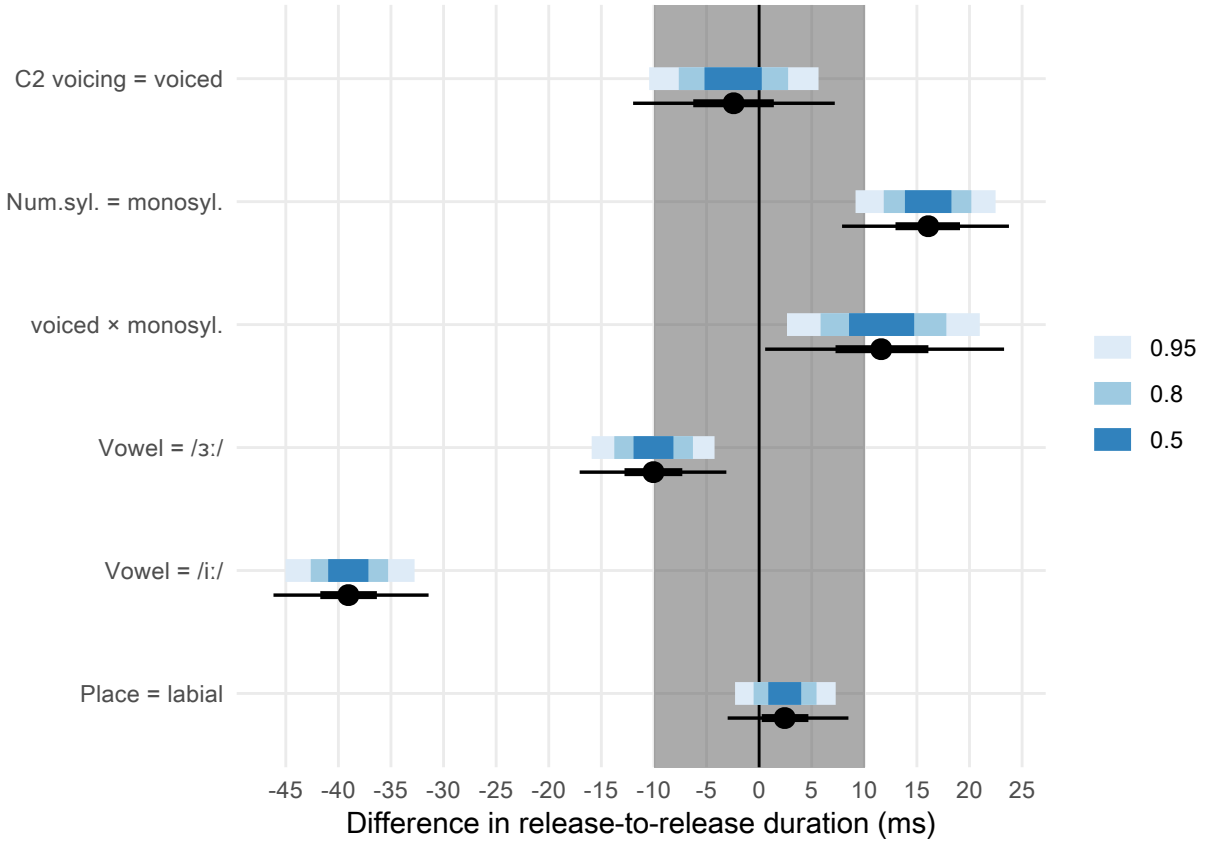

Figure 1: Bayesian credible intervals of the effects on release-to-release duration (model 3). For each effect, the thick blue-coloured bars indicate (from darker to lighter) the 50%, 80%, and 95% CI. The black point with bars are the posterior median (the point), the 98% (thin bar) and 66% (thicker bar) CI

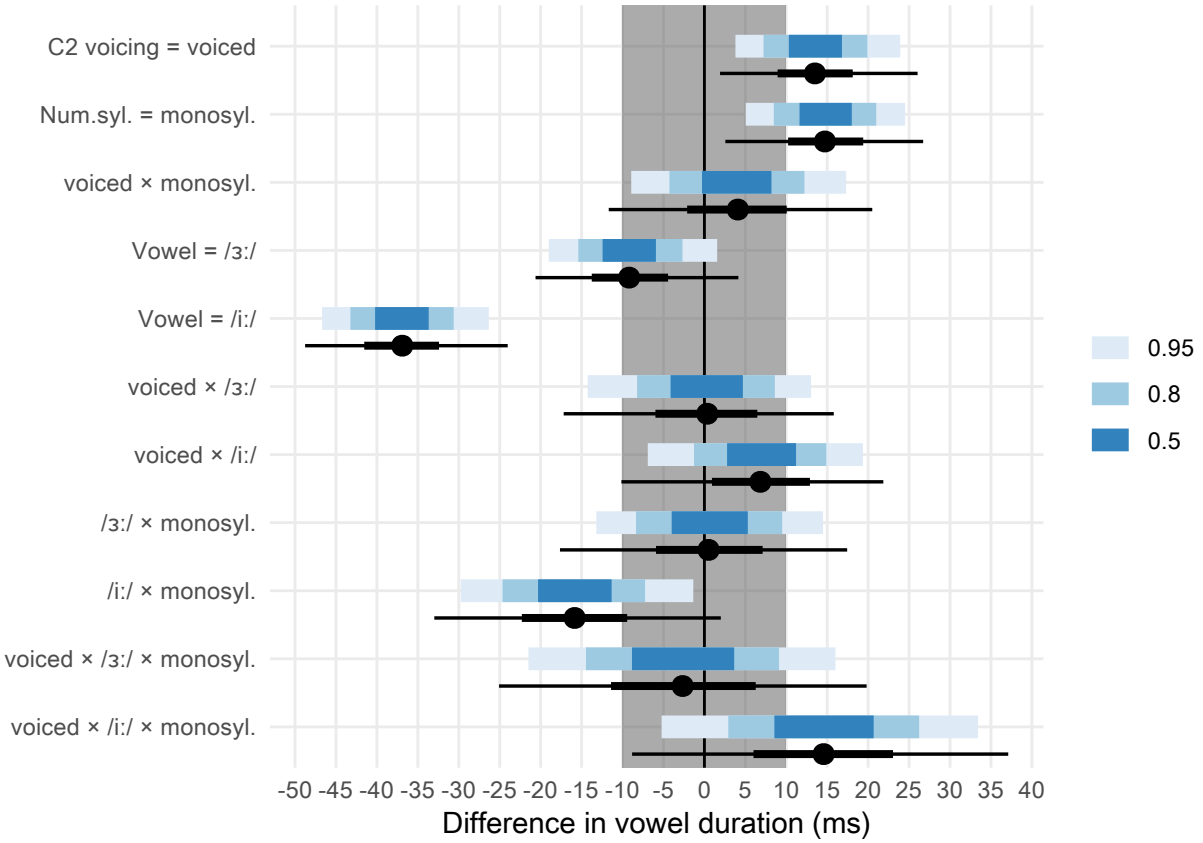

Figure 2: Bayesian credible intervals of the effects on vowel duration (model 4). For each effect, the thick blue-coloured bars indicate (from darker to lighter) the 50%, 80%, and 95% CI. The black point with bars are the posterior median (the point), the 98% (thin bar) and 66% (thicker bar) CI

[−19.24, 1.49], while /i:/ is 36.97 ms shorter (SD = 5.14, 95% CI = [−46.99, −26.76]). C2 voicing as a smaller but robust positive effect on vowel duration in disyllabic words. The posterior distribution of the effect of voicing on /ɑ:/ has mean 13.47 ms (SD = 5.14) and 95% CI = [3.80, 24.36]. The posterior of the interaction of voicing with vowel when the vowel is /ɜ:/ is quite spread out, with 95% CI between −13.77 and 13.40 ms. This can indicate that /ɑ:/ and /ɜ:/ behave similarly for what concern voicing-driven difference in duration, or that, if non-null, the interaction coefficient is very small. On the other hand, the effect of voicing is on average 6.75 ms greater (SD = 6.76, 95% CI = [−7.45, 19.55]) with the vowel /i:/. The magnitude of the voicing effect in disyllabic vs. monosyllabic words is modulated by the identity of the vowel. The posterior distribution for the interaction C2 voicing/number of syllables when the vowel is /ɑ:/ has mean 4.07 ms (SD = 6.6) and 95% CI [−8.94, 17.30], indicating the possibility for a very small increase of the effect from disyllabic to monosyllabic /ɑ:/-words. The three-way interaction C2 voicing/vowel/number of syllables indicates that the effect of voicing with monosyllabic /ɜ:/-words is very similar to that of monosyllabic /ɑ:/-words ($\bar{\theta}$ = −2.66, SD = 9,44, 95% CI = [−21.57, 16.23]), while in monosyllabic /i:/-words it increases by 14.5 ms (SD = 9.43, CI = [−4.27, 33.41]). Note that the credible intervals of some of the effect are quite large, so that a large range of values are possible. The posterior means of such effects should be treated with more caution.

3.3 Consonant closure duration

To test various effects on C2 closure duration, a model was fit with closure duration as the outcome variable and the following predictors: C2 voicing (voiceless, voiced), C2 place of articulation (velar, labial), number of syllables (disyllabic, monosyllabic), all interactions between these predictor terms, and centred speech rate. The random effects were again a by-speaker and by-word random intercept, and a by-speaker random coefficient for C2 voicing.

As priors, a normal distribution with mean 90 ms (SD = 20) was used for the intercept, based on Sharf (1962) and Luce & Charles-Luce (1985). The means reported in Sharf (1962) and Luce & Charles-Luce (1985) also indicate that the closure of the stop in monosyllabic words is 10-30 ms shorter when the stop is voiced. A normal distribution with mean -20 ms (SD = 10) was chosen as the prior of the effect of C2 voicing on closure duration. The same studies indicate that labial stops have a closure which is 10-20 ms longer than the closure of velar stops. For the effect of C2 place, a normal distribution with mean 15 ms (SD = 10) was used.

The posterior distribution of the intercept for closure duration (corresponding to the duration of voiceless velar stops in disyllabic words) has mean 74.62 ms (SD = 2.85) and 95% CI = [69.17, 80.42]. The effect of C2 voicing on closure duration is estimated to be between −26.76 and −14.57 ms (95% CI). The posterior mean of this effect is −20.83 ms (SD = 3.09). A very small positive effect of place of articulation (labial) is suggested by the 95% CI from −0.30 to 10.63 ms ($\bar{\theta}$ = 5.1 ms, SD = 2.73). A possibly even smaller effect of number of syllables or no effect at all can be inferred from the posterior distribution which has mean 2.81 ms and estimated error 2.84 (95% CI = [−2.70, 8.19]). See Figure 3 for the credible intervals of the effects on closure duration.

4 Discussion

4.1 Release-to-release interval

The first question asked whether the voicing of C2 in disyllabic and monosyllabic words in English influence the duration of the release-to-release interval. Coretta (2018) showed that the release-to-release interval duration is not affected by C2 voicing in disyllabic words of Italian and Polish. The hypotheses were that, in English, the interval is not affected in disyllabic words, like in Italian and Polish, but it is in monosyllabic words. The results of this study indicate that the release-to-release duration of disyllabic words in English is very similar whether C2 is voiceless (like *tarpus*) or voiced (*tarbus*).

A Bayesian regression model was fitted to the release-to-release duration (model 3). In this model, the stipulated precision based on the ROPE width of 20 ms was reached for all the relevant terms. The results of model 3 suggest a null effect of C2 voicing on the interval duration in disyllabic words (hypothesis 1a), with a 95% probability that the true effect is between −10 and +5 ms (within the ROPE). At lower levels of probability, the posterior distribution of the effect indicates an effect between −5 and 0 ms (50% probability). If the voicing of C2 is conditioning the duration of the release-to-release interval, this effect is very small and negative (around −2.5 ms).

The possible very small negative effect of C2 voicing in disyllabic words could be related to an annotation bias which affects the identification of stop releases. English voiceless stops are generally followed by aspiration, and the related glottal friction could mask the exact location of the release burst. If the release of the post-vocalic voiceless stops is annotated later than the actual release (by mistaking peaks in aspiration for the release burst), this could lead to longer release-to-release durations when C2 is voiceless compared to when it is voiced. Such annotation bias could explain the (indeed quite small) negative effect of voicing on the interval duration, and why it is in the opposite direction of the

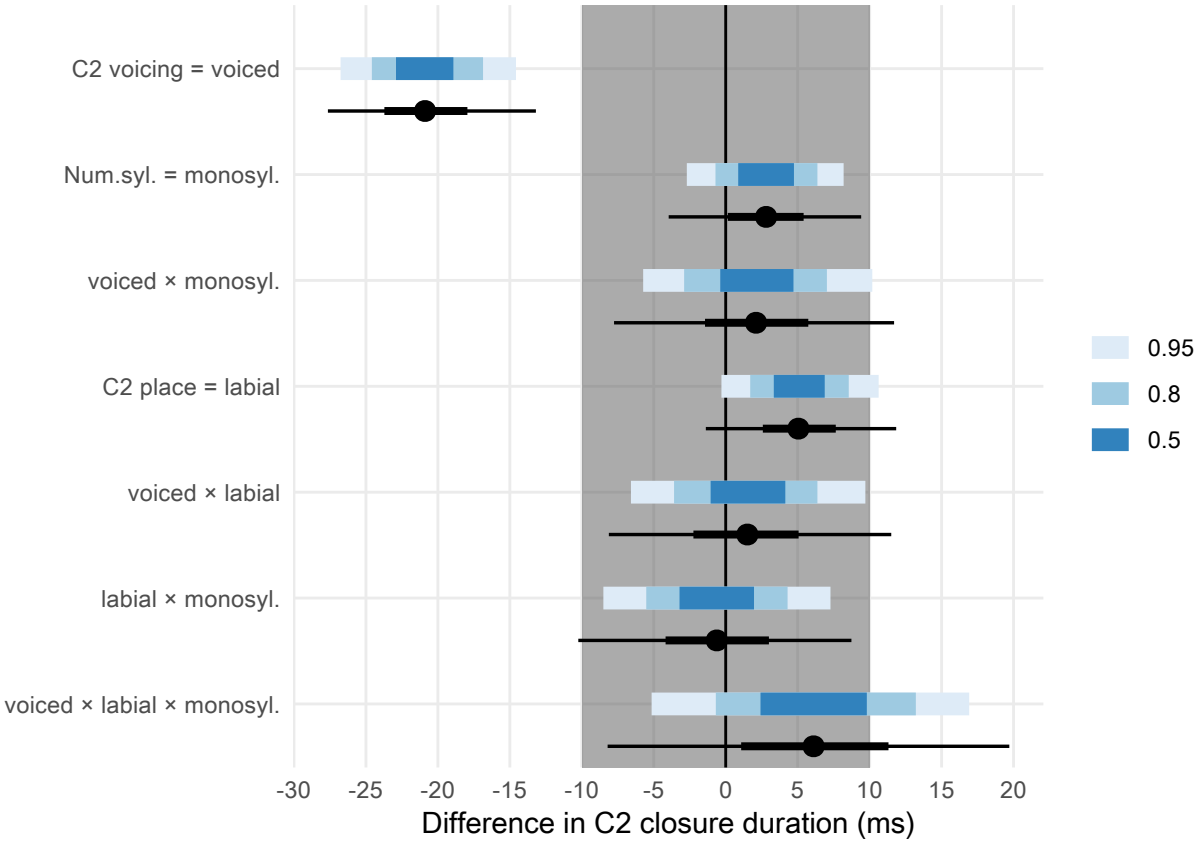

Figure 3: Bayesian credible intervals of the effects on closure duration (model 4). For each effect, the thick blue-coloured bars indicate (from darker to lighter) the 50%, 80%, and 95% CI. The black point with bars are the posterior median (the point), the 98% (thin bar) and 66% (thicker bar) CI

one predicted for monosyllabic words (i.e. *longer* release-to-release when C2 is voiced).

On the other hand, the interval in monosyllabic words is longer when C2 is voiced (for example, *tarb*) vs. when it is voiceless (*tarp*). The interaction term between number of syllables in the word and C2 voicing is positive, between +2.5 and +21 ms (at 95% probability), which means that the effect of C2 voicing increases by 2.5 to 21 ms in monosyllabic words relative to the effect in disyllabic words. This result is compatible with Hypothesis 1b that the release-to-release interval is longer in monosyllabic words with a voiced C2.

The second question was about other effects on the release-to-release duration. As expected by hypothesis 2a, the release to release interval is longer in monosyllabic than in disyllabic words. At 95% probability, the effect of number of syllables (from di- to mono-syllabic) is between 9 and 22.5 ms. As for hypothesis 2b, the results are more robust for /iː/ than for /ɜː/. When the vowel is /iː/, the release-to-release interval is 33 to 45 ms shorter compared with an interval with /ɑː/. The posterior distribution of the effect when the vowel is /ɜː/ substantially overlaps with the ROPE, although it tends towards the negative side. If there is an effect with this vowel compared to /ɑː/, it is negative and possibly around −10 ms. Finally, hypothesis 2c is not unequivocally corroborated. The posterior distribution of the effect of C2 place of articulation (labial) has very high precision (9.5 ms) and it is between 0 and 5 ms (at somewhat less than 80% probability). However, it lies within the ROPE and it is very close to 0.

4.2 Vowel and closure duration

Question 3 addresses the effect of voicing on vowel and closure duration, and the possible differences between disyllabic and monosyllabic words. Starting from vowel duration, the effect of voicing is estimated to lie between 4 and 25 ms. This range is very similar to that reported by Coretta (2018) for Italian and Polish disyllabic words (95% confidence intervals [8, 25], monosyllabic words were not tested). Note however that the posterior mean (13.5 ms) is lower than the values reported by Sharf (1962), Klatt (1973), and Davis & Van Summers (1989) for disyllabic words, which range around 25 ms. The number of syllables produce a similar effect to that of voicing, whereby vowels increase in duration by 5 to 25 ms in monosyllabic words. This relation correspond to what previously reported in the literature.

It was expected that the voicing effect on vowels would be stronger in monosyllabic than in disyllabic words (hypothesis 3). The wide credible intervals of the posterior distributions from the Bayesian model 4, which are larger than the ROPE, make interpretation difficult. It can be said that, at 80% probability, the difference in voicing effect between mono- and disyllabic words is between −5 and +12.5 ms. The distribution is skewed towards the positive side, and this is compatible with results from previous studies. Based on what the posterior distribution indicates, and with the caveat that more data is needed to reach a sensible estimate precision and reduce uncertainty, we can argue for an effect increase in monosyllabic words by a factor of 1.5 to 2.

The three-way interaction between C2 voicing, vowel, and number of syllables reveals that in monosyllabic words with the vowel /iː/ the effect is indeed larger. The model estimates an increase in the effect of about 14.5 ms ([−4.27, 33.41]). Note that, although the credible interval is again very wide (38 ms), it overlaps less extensively with the ROPE around 0, thus somewhat more clearly suggesting a positive effect. However, the vowel /iː/ followed by a voiceless stop is also about 16 ms shorter in monosyllabic words than the same vowel in disyllabic words. While it is not clear why the vowel is that much shorter in that context, it is possible that the simultaneous increase (by the voicing effect) and decrease (by vowel identity) of vowel duration in that context is a statistical product. As for the vowel /ɜː/, it can be inferred from the posterior of the interaction that this vowel behaves similarly to /ɑː/. Research on the duration of English tense vowels and on a possible process of /iː/ shortening is

needed to shed light on the observed patterns.

There was no specific hypothesis concerning closure durations. C2 voicing has a robust negative effect on closure duration, so that voiced closures are 14.6-26.8 ms shorter than voiceless closures. The effects of number of syllables, place, and interactions all have credible intervals that are narrower than 20 ms (the ROPE width) but they lie entirely within the ROPE around 0. If these variables do have an effect on closure duration, the present analysis suggests that the means of these effects are between 0 and 5 ms. These values are smaller than what the results in Sharf (1962), which indicate a difference of 15 ms between velar and labial closure durations.

4.3 General discussion

Coretta (2018) proposes that the voicing-related adjustments in the relative timing of the closure onset within an isochronous speech interval (acoustically identified as the release-to-release interval) is the diachronic precursor of the widespread effect of voicing on vowel duration.[2] Given that the duration of the release-to-release interval in Italian and Polish disyllabic words is not affected by the voicing of the post-vocalic consonant, the relative durations of vowel and closure depend on when the closure is achieved within that interval. A later closure onset implies a longer vowel and a shorter closure, while, vice versa, an earlier closure onset produces a shorter vowel and a longer closure.

The current study shows that also the release-to-release interval of English disyllabic words with either a voiceless or a voiced stop is isochronous. While future research will have to investigate the source of the isochrony of the interval (which probably lies in the temporal organisation of articulatory gestures), the same pathway proposed in the context of the Italian and Polish data can be envisaged for English, and possibly cross-linguistically in general.

A complication arises from the pattern observed with English monosyllabic words, in which the release-to-release interval is not isochronous and is instead affected by C2 voicing. Words with a voiced C2 have longer release-to-release intervals (by about 5-17.5 ms at 80% probability). If compensatory mechanism expounded in this paper is based on the premise that the interval within which compensation happens must be isochronous. Since English monosyllabic words don't show isochrony, we either have to abandon a compensatory account or explain a mechanism by which the release-to-release interval in monosyllabic words are characterised by both compensation and absence of isochrony.

A possible solution could be cautiously put forward from principles of Evolutionary Phonology Blevins (2004, 2006). Most of present-day English monosyllabic words harken back to disyllabic words in Old English and Proto-Germanic. It is possible that a mechanism of compensation was in action in what were disyllabic words which, through vowel reduction and loss which affected different diachronic stages of Germanic and English, became monosyllabic. Speculative, a voicing effect could thus be reconstructed for Proto-Germanic itself. This hypothesis is compliant with a general principle of historical linguistics by which if a feature is present in a majority of daughter languages (which is the case for the voicing effect), it is sensible to reconstruct such feature for the ancestor proto-language. Furthermore, given that in almost all the investigated Indo-European languages the duration of vowels is modulated by the voicing of a following stop, we could even stretch this principle to argue that the voicing effect can be reconstructed even from Proto-Indo-European.

Once the second vowel has gone from the production of disyllabic words and isochrony can be dispensed, perceptual biases (as proposed by perceptual accounts of the voicing effect like Javkin 1976 and Kluender et al. 1988) can operate. The duration of the remaining vowel is free to be modified

[2] Note that isochrony here is intended as pertaining the context of voiceless vs. voices stops only. It is not implied that the interval is isochronous in absolute terms.

further to enhance the perceptual difference of voiceless vs. voiced stops by cues of vowel duration Lisker (1974, 1986); Stevens & Keyser (1989).

Why, then, monosyllabic words have lost isochrony? The following section offers an articulatory account which could answer this question.

4.3.1 Release-to-release isochorony as a consequence of gestural phasing

As already argued in Coretta (2018), the release-to-release interval in itself is not special. The compensatory temporal adjustment account can be understood as an account of vowel acoustic duration, hence the scope of compensation can be defined in terms of acoustic intervals. The interval that was found to be temporally stable across voicing context in Coretta (2018) is the release-to-release interval. However, it is desirable to derive the isochrony of this acoustic interval from properties of articulatory coordination. While a fully fledged theory of gestural phasing is beyond the scope of this study, I offer here a tentative account of the underlying gestural coordination from which the release-to-release isochrony can be derived.

Fowler (1983) builds on Öhman's 1966; 1966 idea that the speech stream is composed by a series of continuous vocalic gestures interspersed with oral constriction (consonantal) gestures. Fowler proposes that the vocalic gestures of a VCV sequence are characterised by a cyclic pattern of production, so that the temporal distance between the two vowels is constant and it is not affected by the number of intervening consonants. The task dynamic model (Saltzman et al. 2008) of Articulatory Phonology (Ohala et al. 1986; Browman & Goldstein 1988, 1992), based on the coupled oscillators model (O'Dell & Nieminen 2008), also argues that onset consonant gestures are generally produced in-phase with the following vowel (meaning that the initiation of the vocalic and consonantal gesture is synchronous). This mechanism generates the so-called C-centre effects, by which the acoustic duration of the in-phase vowel depends on the number of onset consonants, while coda consonants, which are anti-phase with the preceding vowel, do not have an effect on vowel duration.

Furthermore, results which are compatible with a vowel-based rhythmic gestural implementation come from work by Farnetani & Kori (1986) and Celata & Mairano (2014), which investigate vowel duration and syllable structure in Italian. In the former study, it was found that vowels followed by a singleton stop (for example /la.ta/) are longer than vowels followed by a tautosyllabic cluster (/la.dra/). This pattern can be easily derived if the cluster /dr/ follows a C-centre alignment, and if we assume that the distance between the vowels is the same in the two contexts (/la.ta/ and /la.dra/). Celata & Mairano (2014) show that the duration of consonant or consonant cluster is negatively correlated with the preceding vowel (although the magnitude of the correlation is low to moderate, see Beguš 2017 and Coretta 2018 on how speech rate could mask statistical relations).

Van Summers (1987) and de Jong (1991) show that the consonantal closing gesture of voiceless stops has greater velocity than that of voiced stops. If the closing gesture of both voiceless and voiced stops is initiated in synchrony with that of the following vowel and the latter is at a stable temporal distance from the preceding vowel, a later complete oral closure will be achieved in the latter case without affecting the time of the consonantal release.

We can now turn back to the issue of the absence of isochrony in monosyllabic words. When the second vowel in disyllabic CVCV words is deleted via diachronic change, the second consonant loses its anchor vowel (V2) and shifts its affiliation to the preceding vowel instead, in other words, it becomes a coda consonant. The shift could be brought about by described in terms of Tilsen's 2013; 2016 selection–coordination theory. Coda consonants are anti-phase with the preceding vocalic nucleus and among themselves, meaning that each gesture is executed in sequence. The cyclicity of vowel production does not apply in this case, since that there is only one vowel, and the vocalic gesture is free to be stretched or compressed.

References

Ananthapadmanabha, T. V., A. P. Prathosh & A. G. Ramakrishnan. 2014. Detection of the closure-burst transitions of stops and affricates in continuous speech using the plosion index. *The Journal of the Acoustical Society of America* 135(1). 460–471. doi:10.1121/1.4836055.

Beguš, Gašper. 2017. Effects of ejective stops on preceding vowel duration. *The Journal of the Acoustical Society of America* 142(4). 2168–2184. doi:10.1121/1.5007728.

Bigi, Brigitte. 2015. SPPAS - Multi-lingual approaches to the automatic annotation of speech. *The Phonetician* 111–112. 54–69.

Blevins, Juliette. 2004. *Evolutionary phonology: The emergence of sound patterns*. Cambridge University Press.

Blevins, Juliette. 2006. A theoretical synopsis of Evolutionary Phonology. *Theoretical linguistics* 32(2). 117–166.

Browman, Catherine P. & Louis Goldstein. 1988. Some notes on syllable structure in articulatory phonology. *Phonetica* 45(2-4). 140–155.

Browman, Catherine P. & Louis Goldstein. 1992. Articulatory phonology: An overview. *Phonetica* 49. 155–180.

Bürkner, Paul-Christian. 2017. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software* 80(1). 1–28. doi:10.18637/jss.v080.i01.

Bürkner, Paul-Christian. 2018. Advanced Bayesian multilevel modeling with the r package brms. *The R Journal* 10(1). 395–411. doi:10.32614/RJ-2018-017.

Celata, Chiara & Paolo Mairano. 2014. On the timing of V-to-V intervals in Italian: a review, and some new hypotheses. *Revista de Filología Románica* 31. 37. doi:10.5209/rev_RFRM.2014.v31.n1.51022.

Chen, Matthew. 1970. Vowel length variation as a function of the voicing of the consonant environment. *Phonetica* 22(3). 129–159.

Coretta, Stefano. 2018. An exploratory study of voicing-related differences in vowel duration as compensatory temporal adjustment in Italian and Polish. Submitted.

Davis, Stuart & W. Van Summers. 1989. Vowel length and closure duration in word-medial VC sequences. *Journal of Phonetics* 17. 339–353.

Esposito, Anna. 2002. On vowel height and consonantal voicing effects: Data from Italian. *Phonetica* 59(4). 197–231. doi:10.1159/000068347.

Etz, Alexander, Quentin F. Gronau, Fabian Dablander, Peter A. Edelsbrunner & Beth Baribault. 2018. How to become a Bayesian in eight easy steps: An annotated reading list. *Psychonomic Bulletin & Review* 25(1). 219–234. doi:10.3758/s13423-017-1317-5.

Farnetani, Edda & Shiro Kori. 1986. Effects of syllable and word structure on segmental durations in spoken Italian. *Speech communication* 5(1). 17–34. doi:10.1016/0167-6393(86)90027-0.

Fowler, Carol A. 1983. Converging sources of evidence on spoken and perceived rhythms of speech: Cyclic production of vowels in monosyllabic stress feet. *Journal of Experimental Psychology: General* 112(3). 386. doi:10.1037/0096-3445.112.3.386.

Heffner, R.-M.S. 1937. Notes on the length of vowels. *American Speech* 12. 128–134. doi:10.2307/452621.

Hertrich, Ingo & Hermann Ackermann. 1997. Articulatory control of phonological vowel length contrasts: Kinematic analysis of labial gestures. *The Journal of the Acoustical Society of America* 102(1). 523–536. doi:10.1121/1.419725.

House, Arthur S. & Grant Fairbanks. 1953. The influence of consonant environment upon the secondary acoustical characteristics of vowels. *The Journal of the Acoustical Society of America* 25(1). 105–113. doi:10.1121/1.1906982.

Huggins, A. William F. 1972. Just noticeable differences for segment duration in natural speech. *The Journal of the Acoustical Society of America* 51(4B). 1270–1278. doi:10.1121/1.1912971.

Jacewicz, Ewa, Robert Allen Fox & Samantha Lyle. 2009. Variation in stop consonant voicing in two regional varieties of American English. *Journal of the International Phonetic Association* 39(3). 313–334. doi:10.1017/S0025100309990156.

Javkin, Hector R. 1976. The perceptual basis of vowel duration differences associated with the voiced/voiceless distinction. *Report of the Phonology Laboratory, UC Berkeley* 1. 78–92.

de Jong, Kenneth. 1991. An articulatory study of consonant-induced vowel duration changes in English. *Phonetica* 48(1). 1–17. doi:10.1121/1.2028316.

de Jong, Kenneth. 2004. Stress, lexical focus, and segmental focus in English: patterns of variation in vowel duration. *Journal of Phonetics* 32(4). 493–516. doi:10.1016/j.wocn.2004.05.002.

Kawahara, Shigeto, Donna Erickson & Atsuo Suemitsu. 2017. The phonetics of jaw displacement in Japanese vowels. *Acoustical Science and Technology* 38(2). 99–107. doi:10.1250/ast.38.99.

Klatt, Dennis H. 1973. Interaction between two factors that influence vowel duration. *The Journal of the Acoustical Society of America* 54(4). 1102–1104. doi:10.1121/1.1914322.

Kluender, Keith R., Randy L. Diehl & Beverly A. Wright. 1988. Vowel-length differences before voiced and voiceless consonants: An auditory explanation. *Journal of Phonetics* 16. 153–169.

Ko, Eon-Suk. 2018. Asymmetric effects of speaking rate on the vowel/consonant ratio conditioned by coda voicing in English. *Phonetics and Speech Sciences* 10(2). 45–50. doi:10.13064/KSSS.2018.10.2.045.

Kruschke, John. 2015. *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan (2nd edition)*. Amsterdam, The Netherlands: Academic Press.

Laeufer, Christiane. 1992. Patterns of voicing-conditioned vowel duration in French and English. *Journal of Phonetics* 20(4). 411–440.

Lehiste, Ilse. 1970a. Temporal organization of higher-level linguistic units. *The Journal of the Acoustical Society of America* 48(1A). 111–111. doi:10.1121/1.1974906.

Lehiste, Ilse. 1970b. Temporal organization of spoken language. In *Working papers in linguistics*, vol. 4, 96–114. doi:10.1121/1.1974906.

Lindblom, Björn. 1967. Vowel duration and a model of lip mandible coordination. *Speech Transmission Laboratory Quarterly Progress Status Report* 4. 1–29.

Lisker, Leigh. 1957. Closure duration and the intervocalic voiced-voiceless distinction in English. *Language* 33(1). 42–49. doi:10.2307/410949.

Lisker, Leigh. 1974. On "explaining" vowel duration variation. In *Proceedings of the Linguistic Society of America*, 225–232.

Lisker, Leigh. 1986. "Voicing" in English: a catalogue of acoustic features signaling /b/ versus /p/ in trochees. *Language and Speech* 29(1). 3–11.

Luce, Paul A & Jan Charles-Luce. 1985. Contextual effects on vowel duration, closure duration, and the consonant/vowel ratio in speech production. *The Journal of the Acoustical Society of America* 78(6). 1949–1957.

Machač, Pavel & Radek Skarnitzl. 2009. *Principles of phonetic segmentation*. Epocha.

Maddieson, Ian & Jack Gandour. 1976. Vowel length before aspirated consonants. In *UCLA Working papers in Phonetics*, vol. 31, 46–52.

McElreath, Richard. 2015. *Statistical rethinking: A bayesian course with examples in R and Stan*. CRC Press.

Mortensen, Johannes & John Tøndering. 2013. The effect of vowel height on Voice Onset Time in stop consonants in CV sequences in spontaneous Danish. In *Proceedings of Fonetik 2013*, Linköping, Sweden: Linköping University.

Munafò, Marcus R., Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie Du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware & John P. A. Ioannidis. 2017. A manifesto for reproducible science. *Nature Human Behaviour* 1(1). 0021. doi:10.1038/s41562-016-0021.

Nicenboim, Bruno, Timo B. Roettger & Shravan Vasishth. 2018. Using meta-analysis for evidence synthesis: The case of incomplete neutralization in german. *Journal of Phonetics* 70. 39–55. doi:10.1016/j.wocn.2018.06.001.

Nooteboom, Sieb G. & Gert J. N. Doodeman. 1980. Production and perception of vowel length in spoken sentences. *The Journal of the Acoustical Society of America* 67(1). 276–287. doi:10.1121/1.383737.

O'Dell, Michael L. & Tommi Nieminen. 2008. Coupled oscillator model for speech timing: Overview and examples. In *Nordic prosody: Proceedings of the xth conference*, 179–190.

Ohala, John J, Catherine P Browman & Louis M Goldstein. 1986. Towards an articulatory phonology. *Phonology* 3. 219–252.

Öhman, Sven E. G. 1966. Coarticulation in VCV utterances: Spectrographic measurements. *The Journal of the Acoustical Society of America* 39(1). 151–168. doi:10.1121/1.1909864.

Peterson, Gordon E. & Ilse Lehiste. 1960. Duration of syllable nuclei in English. *The Journal of the Acoustical Society of America* 32(6). 693–703. doi:10.1121/1.1908183.

Plug, Leendert & Rachel Smith. 2018. Segments, syllables and speech tempo perception. In *Proceedings of the 9th international conference on speech prosody 2018*, 279–283. doi:10.21437/SpeechProsody.2018-57.

R Core Team. 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org.

Roettger, Timo B. 2019. Researcher degrees of freedom in phonetic sciences. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 10(1). 1–27. doi:10.5334/labphon.147.

Saltzman, Elliot, Hosung Nam, Jelena Krivokapic & Louis Goldstein. 2008. A task-dynamic toolkit for modeling the effects of prosodic structure on articulation. In *Proceedings of the 4th international conference on speech prosody (speech prosody 2008), campinas, brazil*, 175–184.

Sharf, Donald J. 1962. Duration of post-stress intervocalic stops and preceding vowels. *Language and speech* 5(1). 26–30.

Sharf, Donald J. 1964. Vowel duration in whispered and in normal speech. *Language and speech* 7(2). 89–97.

Slis, Iman H. & Antonie Cohen. 1969a. On the complex regulating the voiced-voiceless distinction II. *Language and speech* 12(3). 137–155. doi:10.1177/002383096901200301.

Slis, Iman Hans & Antonie Cohen. 1969b. On the complex regulating the voiced-voiceless distinction I. *Language and speech* 12(2). 80–102. doi:10.1177/002383096901200202.

Stevens, Kenneth N. & Samuel Jay Keyser. 1989. Primary features and their enhancement in consonants. *Language* 81–106.

Tilsen, Sam. 2013. A dynamical model of hierarchical selection and coordination in speech planning. *PLoS ONE* 8(4). e62800. doi:10.1371/journal.pone.0062800.

Tilsen, Sam. 2016. Selection and coordination: The articulatory basis for the emergence of phonological structure. *Journal of Phonetics* 55. 53–77. doi:10.1016/j.wocn.2015.11.005.

Toivonen, Ida, Lev Blumenfeld, Andrea Gormley, Leah Hoiting, John Logan, Nalini Ramlakhan & Adam Stone. 2015. Vowel height and duration. In Ulrike Steindl, Thomas Borer, Huilin Fang, Alfredo García Pardo, Peter Guekguezian, Brian Hsu, Charlie O'Hara & Iris Chuoying Ouyang (eds.), *Proceedings of the 32nd west coast conference on formal linguistics*, vol. 32, 64–71. Somerville, MA: Cascadilla Proceedings Project.

Van Summers, W. 1987. Effects of stress and final-consonant voicing on vowel production: Articulatory and acoustic analyses. *The Journal of the Acoustical Society of America* 82(3). 847–863. doi:10.1121/1.395284.

Vasishth, Shravan, M. Beckman, B. Nicenboim, Fangfang Li & Eun Jong Kong. 2018a. Bayesian data analysis in the phonetic sciences: A tutorial introduction. *Journal of Phonetics* 71. 147–161. doi:10.1016/j.wocn.2018.07.008.

Vasishth, Shravan, Daniela Mertzen, Lena A. Jäger & Andrew Gelman. 2018b. The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language* 103. 151–175.