Response to reviews

21/01/2019

I updated the title to reflect the suggestion of one reviewer "I recommend the title to include that"longer vowel duration" (only) refers to the "preceding" vowel duration (not necessarily the following). I also think the word 'advancement' will be more suitable than 'displacement'. On the other hand, I wonder if the title purposefully used the word 'displacement' instead of 'advancement'.". The title now reads"Longer vowel duration correlates with greater tongue root advancement at vowel offset: Acoustic and articulatory data from Italian and Polish".

Editor

I have one major concern about your paper. In the main part of your paper, you present results pooled across subjects. These results show some agreement with your hypotheses. However, in Section C. Individual differences, you show that several individuals do not follow the general trends that you have found in the pooled results. So, why not only present the individual results, since they suggest variability across subjects?

> *Based on Figure 8, 14 out of 17 speakers have a positive mean difference in tongue root position at V1 offset, and only three have no difference or a negative mean difference (in accordance with previous work in English), so I think it is fair to say that the pattern of tongue root advancement in voiced stops is generally quite robust across-speakers. Based on Figure 9, 12 speakers agree with the population-level effects of increasing tongue root advancement during the vowel, while only five show either no difference in tongue root or the reversed pattern (more advancement in voiceless stops). Given the size of the sample used in this study (on the smaller side), I think it is best to present and discuss both population-level effects and speaker-level effects, so as to give a more holistic view of the statistical patterns obtained. From a statistical point of view, the effects of the individual speakers inform the way population-level effects are fitted, and uncertainty increases with increased speaker-level differences. This estimation of uncertainty at the population-level can help future studies decide on expected outcomes and statistical priors, so it is valuable to discuss population-level effects together with individual differences. Since the nature of the study is exploratory (hypothesis-generating), rather than confirmatory (hypothesis-testing), it is generally recommended to report all the analyses.*

I also think that you need to say something about the resolution of the ultrasound images, since, as noted by Reviewer 1, some movements are very small. In addition, I would like to see some general comments on tongue movements during speech, in particular when you discuss tongue root movement and vowel duration.

I elaborate on these and other issues below.

Line 17 Fowler, 1992 and Lisker 1974, d not present any data on tongue root movements.

> *They are references to the voicing effect of vowel duration. I've expanded the list as per reviewer's request and moved the references to the next sentence.*

Lines 21 and 24 Here, you make claims about earlier works that need to be referenced.

> *Line 21 refers to this study, so I changed to "In this exploratory study". I added references on line 24.*

Line 72 ...is generally longer than that of voiceless stops..." > ...is generally shorter than that of voiceless stops..."

> *Fixed.*

Page 6, second paragraph Since you bring up possible voicing differences between Italian and Polish, I think it would help if you made clear whether the voiced stops in your material were actually produced with glottal vibrations., If they were not, there would be no need for tongue movements to maintain voicing. Possibly, some of the individual differences may be due to voicing variability.

> *The paragraph mentions differences in the effect of voicing on vowel duration, rather than differences of implementation of consonantal voicing. The EGG data of the study (not discussed here) indicates that virtually all*

voiced tokens were uttered with vocal fold vibration. I added a footnote in Section I.B which mentions this: "Simultaneous electroglottographic data (not discussed here) was also collected during the experiment. This data indicates that virtually all tokens of voiced stops were uttered with vocal fold vibration, with just a few exceptions (4 tokens were voiceless in the speaker PL02)."

Page 11, E, Data processing and statistical analysis. I would suggest saying something about the spatial resolution of the ultrasound system and the reliability of the splines. From Figure 1, it is not immediately clear that you are actually measuring the tongue root.

I have added the following sentence in Section II.A: "The mean pixel size as used by the automatic tracker was 0.47 mm (SD = 0.16), so that differences in tongue position smaller than that would not be captured." As for measuring tongue root displacement, see answer to Reviewer 1 L216-8

Page 17, C. Correlation between tongue rot position and V1 duration Why is this interesting and relevant? Since the tongue is always moving in speech, would it not move longer during a long vowel? The tongue does not appear to have a point target for vowels.

As far as I know, correlation between vowel duration and tongue root position at V1 offset has not been reported before. One possible obvious explanation is that a longer vowel allows for more tongue root advancement (as I argue in the paper). But it could have been that tongue root advancement is initiated at different times during the production of vowels in different vowel duration contexts, so that the same or similar amount of displacement is reached at the end of the vowel independent of its duration. Rather, what the positive correlation interestingly suggests is that the onset of the advancement gesture during the production of the vowel is initiated at a stable temporal distance from the release of C1/onset of V1. Of course, future work will need to assess this (as I recommend in the paper).

Line 302 Can you explain wat you mean with "curvature"?

I have removed this term throughout the paper and used simpler terminology to describe the movement of the tongue root along the duration of the vowel.

Line 333 "...the onset of the forward gesture of the root is timed relative to a landmark preceding the closure, independent of the duration of the vowel." I don't understand what this means. What kind of landmark are you talking about?

Yes, I am sorry the wording was very unclear. I have updated the paragraph to hopefully make it clearer. What I am try to say is that the data suggests the following: If we take the release of C1 as a temporal landmark, then the onset of the tongue root advancement gesture would happen at a fixed temporal distance from C1 release, independent of the duration of the vowel. Of course, C1 release is chosen for convenience, rather than for a theoretical reason (as discussed in Coretta 2018).

Line 391 The duration of 70-90 ms suggested ny Rothenberg (1967) is not based on actual measurements.

Updated the sentence to "Rothenberg 1967 hypothesised, after an informal investigation, that a maximal ballistic expansion movement of the tongue root to increase the size of the lower pharynx would take 70–90 ms" as suggested by one reviewer below.

Figure 6 The labels and the y-axis values are identical in the left and right plots. The two plots appear to be identical.

That was a coding error, now fixed.

References

Ahn. 2015, is available here: http://www.ultrafest2015.hku.hk/docs/S_Ahn_ultrafest.pdf

Ahn & Davidson, 2016, is an abstract, thus not very useful and should be left out.

Halle & Stevens, 1967, is an abstract, thus not very useful and should be left out,. The

Keating, 1984. Page numbers are 35-49, and the text is available here: https://escholarship.org/uc/item/2497n8jq

Line 567 "English"

Lindblom, 1967, is available here: http://www.speech.kth.se/prod/publications/files/qpsr/1967/1967_8_4_001-029.pdf

Maddieson & Gandor, 1976, is availanlde here: https://escholarship.org/uc/item/31f5j8m7

Malisz & Klessa, 2008, is available here: http://www.isle.illinois.edu/sprosig/sp2008/papers/id182.pdf

Line 624 "Phonetics"

Thanks. All changes to references have been applied.

Reviewer 1

Title: Longer vowel duration correlates with greater tongue root displacement: Acoustic and articulatory data from Italian and Polish

Summary – this is an articulatory (ultrasound tongue imaging) study of tongue-root position in the production of CV.CV nonsense words in Italian and Polish, focusing on tongue-root position during V1 and at the onset of C2, while varying whether C2 is a phonemically voiced or voiceless consonant. The main aim of the study is to determine whether tongue root advancement is more extreme during the vowel and at C2 onset when a C2 is voiced than when it is voiceless. It is hypothesized that presence of increased tongue-root advancement when C2 is voiced is a mechanism for increasing the supraglottal cavity and prolonging the transglottal airflow needed to maintain voicing during the C2 closure. Mixed effects modelling is used to identify the impact of C2 voiced/voiceless status (and other factors) on tongue root position at C2 onset. C2 voicing is found to have a significant effect on tongue-root position at C2 onset, with tongue-root onset 0.77mm fronter when C2 is a voiced consonant. A generalised additive mixed model is used to assess the impact of C2 voicing (and other factors) on tongue-root position across V1. The model shows a significant effect of C2 voicing, particularly during the latter half of the vowel. Root advancement is present during the vowel before both voiced and voiceless C2s, but to a greater degree before voiced C2s. A further mixed effects model showed a significant effect of V1 duration on tongue-root advancement, with longer vowels resulting in greater tongue-root advancement. I think it is argued that having a later voiced C2 onset and, concomitantly, a longer preceding vowel, allows for more extreme tongue-root advancement at C2 onset and that therefore the shorter voiced C2 closed phase permits greater expansion of the supralaryngeal cavity to aid transglottal airflow and maintain voicing, while also ensuring that voicing needs to be maintained for a shorter time period.

General comments – the findings of this paper are interesting and seem to show a plausible 2- fold effect of later consonant onset on the maintenance of coronal plosive voicing contrasts. I did find this paper a bit convoluted to read and I found that I didn't have a clear idea of what the author was trying to say. Additionally, raw differences in tongue-root advancement at voiced and voiceless consonant onset, for all they are statistically significant, are extremely subtle <1mm, while UTI images provide a somewhat blurred image of the tongue surface, and there is potential for error during (automatic) spline fitting due to the fuzziness of the image. I also worry about the measurement protocol itself. The ultrasound probe does not seem to have been angled towards the pharynx and yet the probe angles are quite narrow (71-93°). The example given in FIG. 1 does not contain a hyoid shadow, therefore it is difficult to determine whether movement along the selected radial fan axis shown is really capturing root retraction, or some point further up the back of the tongue. Another issue is that an increase in the value of the point at which the tongue spline intersects this chosen radial fan axis is taken to indicate greater tongue-root advancement, whereas tongue-root advancement should result in a lower value – as the tongue root moves forwards, it should move closer to the origin of the radial fan line. It is possible that the author reversed the radial-fan-line intersect values, as in (Kirkham and Nance, 2017), in which case they should state this explicitly. I believe the issues with measurement can be cleared up with some better figures and some assurances about how the measures were taken and transformed, so I suggest revise and resubmit.

These have been addressed below.

Specific comments:

I think some of the content of the section l67-88 should be moved closer to the beginning of the introduction §A, to give context to the ATR focus of the study i.e. there should be an acknowledgement that there are a range of articulatory strategies associated with plosive voicing and then you can say that you are going to focus on one of these features in this study.

This suggestion has been implemented in Section I.A. Now the section starts with a general overview of some of the mechanisms for voicing maintenance and then focusses on tongue root advancement.

1-2 "Voiced stops tend to be preceded by longer vowels and produced with a more advanced tongue root than voiceless stops." Citation missing.

Citations to a long list of publications is given at multiple points within the manuscript, so I think it would not be practical to include them in the abstract.

2 Is "modulated" the right word here? "modified/affected". It doesn't seem to be the right word, given that the voicing you are discussing is phonemic and not necessarily present at the phonetic level, as you point out later.

I changed modulated to "affected".

3-4 "in many languages vowels are longer when followed by voiced stops." Examples and citations missing.

See comment above.

4-6 "Tongue root advancement is known to be an articulatory mechanism which ensures the right pressure conditions for the maintenance of voicing during closure as dictated by the Aerodynamic Voicing Constraint." Citation missing.

See comment above.

7 "enter in a direct statistical relation". This phrase sounds odd. Rephrase as "have a direct statistical relationship"?, or later in lines 25&26, you say "acoustic duration of the vowel is positively correlated with tongue root position". Can you not say that here?

Changed to "have a direct statistical relationship".

8-9 "17 speakers of Italian and Polish (in total)"

Added "in total".

12-13 "in a temporally stable interval". Perhaps it would be better to say "comparatively later closure onset of voiced stops", because it is not clear without context what "stable interval" you mean – C1 release to C2 release.

Changed as suggested.

12-14 It feels as if these lines belong earlier in the paragraph to make the proposed chain of causal relationships clear, i.e. ATR allows maintenance of voicing for voiced stops, ATR correlates with vowel length.

Changed as suggested.

16 "characterised" is not the right word here, given that ATR is covert, "associated with"?

Changed as suggested.

18-19 "a lot of work" citations missing. "aspects", change to "phonetic features"? "phonetic correlates"? Also l19 "relationship", rather than "relation".

A selection of citations from later sections has been included here.

21 "in an exploratory" – change to "in this exploratory".

Changed as suggested.

24 "this replicates previous work on tongue root position" – citation(s) missing.

Citations included.

25 "... indicate that the acoustic duration of the vowel is positively correlated with the tongue root position (at vowel offset/ postvocalic consonant onset)..."

Changed as suggested.

24-8 There is something oddly circular about this section. Compare to the way you talk about the trade-off between vowel and consonant length in lines 121-2. To avoid a circular argument, I would rephrase as: "Furthermore, the results of this study indicate that a comparatively later C2 onset for voiced consonants results in a longer preceding vowel duration which, in turn, results in greater tongue-root advancement during C2 onset. Both the shorter closed phase of the voiced consonant and the more advanced tongue root, which expands the supra-glottal cavity, have the potential to maintain voicing throughout C2 and preserve the voicing contrast."

Changed as suggested.

L32 "relative to the front-back dimension" – "across/in the front-back dimension"?

Fixed.

L33 "This has" – change to "This finding has…"

Fixed.

L43 I think it would be worth mentioning here that there are other articulatory strategies employed to increase supralaryngeal space, e.g. larynx lowering (Rothenberg, 1967). See first comment above about moving content of lines 67-88 to beginning of introductory section.

Changed as suggested.

L52 imply

Fixed.

L57 "closure of a stop" – "closure of a lingual stop".

Fixed.

L157-58 Rephrase as: "Rothenberg (1967) hypothesized after an informal investigation, that a maximal ballistic expansion movement of the tongue root to increase the size of the lower pharynx would take 70-90ms (Rothenberg, 1967: 99).

Fixed.

P65-66 "While the tongue body is more involved in labials". Can you be more specific? What does the tongue body do?

I specified that tongue body lowering is common with labials.

L92 "it can be" -> "it is sometimes"

Fixed.

L104 typo "to study of the"

Fixed.

L105-6 "given their reported differences in magnitude/presence of the effect and the relative ease of comparison." – you need to unpack this a bit.

I've expanded with "Moreover, the segmental phonologies of these languages facilitate the design of sufficiently comparable experimental material (see Coretta 2018 for a more thorough discussion)".

L104-106 I suggest this edit: "Italian and Polish offer an opportunity to study of the articulatory aspects of the voicing effect, given that the former has been consistently reported as a voicing- effect language, while voicing effect in the latter is more complex, with some studies finding an effect and others not."

Changed.

L110-12 What were the mean voicing differences for the two languages? Were they comparable?

I've added details on the raw mean differences of the voicing effect in the two languages and mentioned that such difference is not significant in the linear model.

L112-114 Can you explain further? “The high degree of intra-speaker variation, backed up by statistical modelling, also indicates that these languages possibly behave similarly in regards to the voicing effect.”

I’ve corrected ”intra-speaker” with ”inter-speaker” and expanded on this statment by indicating that, independent of language, speakers show a range of possible magnitudes of the effect.

L124 Can you come up with a better section heading?

Changed to ”Rationale of the current study”.

L126 Citations missing.

I inserted here some of the citations from above for clarity.

L125-131 Suggested rephrasing “Previous research has established that longer preconsonantal vowel durations and greater tongue root advancement are associated with voicing in postvocalic plosives. We know that voicing during plosive closure can be sustained by advancing the tongue root during the production of voiced plosives and that tongue root advancement probably begins before the closure onset (i.e. during the preceding vowel). We also know that vowels followed by voiced plosives tend to be longer than vowels followed by voiceless plosives. Acoustic analysis of the current dataset confirmed an apparent compensatory relationship between the duration of the plosive closure and the duration of the preceding vowel; the shorter the plosive closure, the longer the preceding vowel.”

Rephrased as suggested.

L133 “insights on the link” – “insights into the link”, “between closure” – “between closure duration”?? or “link between closure and vowel durations”.

Changed.

L134 I am not comfortable with the term “modulates”. I think it implies a causal relationship that has not yet been proved. Can you say “correlates”? or “covaries with”?

Changed.

L135-6 Suggested rephrasing “resulting in a three-way relationship between stop consonant duration, vowel duration and tongue-root advancement.”

Changed.

L136-8 “More specifically, the timing of the closure onset within the release-to-release interval decides the duration of the vowel...” Again, is this not a given, see comments l27-28.

Yes. The sentence now says ”More specifically, the timing of the closure onset within the release-to-release interval determines not only the duration of the vowel and that of the closure (as discussed in Coretta 2018), but also the degree of tongue root advancement at V1 offset/C2 onset.”

L159 Can you change to “TELEMED Echo Blaster 128 unit, with a C3.5/20/128Z-3 ultrasonic transducer (20mm radius, 2-4 MHz).” Grouping these items together makes sense. Otherwise you are just writing a random list of items that you used in your recording set up.

Changed.

L160-1 The pre-amp is part of the audio recording set up, not the ultrasound recording system. You need to move it further down the list of equipment, i.e. “a Movo LV4-O2 Lavalier microphone with a FocusRight Scarlett Solo pre-amplifier.”

Changed.

L163 Please gloss “sub-mental triangle”, and in what way was the probe “aligned with the midsagittal plane”?

Now the line reads ”The ultrasonic probe was placed in contact with the flat area below the chin, aligned along the participant’s mid-sagittal plane so that the mid-sagittal profile of the tongue could be imaged.”

L169 ”by means of a synchronisation signal produced by the ultrasound unit and amplified by the P-Stretch unit.

Changed.

P170-2 Does this mean that the ultrasound recording settings varied in every recording session, or between the Italian and Polish recording settings? If the latter, can you give the separate setting by language, rather than a range. Can you also say why you think this variation won't affect what you are measuring? The range of probe angles seem quite narrow.

Yes, they vary for each speaker, since the settings need to be adapted to each speaker's anatomy. I've expanded section II.E to include "Note that, while the data was recorded without taking care that the tongue root was visible, the back part of the tongue just above the hyoid bone shadow (roughly corresponding to the uppermost part of the tongue root) was always imaged." and "The chosen fan-lines across all speakers range between fan-line 25 and 34 (a higher number indicates a more posterior position), and these are always backer in the oral track than the fan-lines along which velar closure is articulated by the respective speaker.". Also see below comment line 216-8.

L184 "explore timing/ articulatory? differences in the closing gesture of voiceless and voiced stops..."

Changed.

L202-9 Can you add a figure showing an example of annotation?

Added.

L216-8 "Tongue root displacement was thus calculated as the displacement of the fitted spline along a selected vector." Based on the figure provided (FIG. 1) I am not convinced that tongue- root advancement is being captured. The probe angle is very narrow. The tongue root does not appear to have been imaged, as there is no visible hyoid shadow. I am not sure that a measure of displacement along a radial fan axis shown would capture tongue-root advancement. Could you provide information on the range of radial axes that were used as measurement axes?

The hyoid shadow is not visible because during UTI set up the angle is adjusted so that the leftmost edge of the ultrasound image corresponds to the edge of the shadow (this is necessary for the automatic tracker to work more consistently). Since the data was collected without tongue root position in mind, I did not make sure to see whole of the actual tongue root all the time. Note though that the measured displacement must be related to tongue root displacement, since the fan-lines from which it was obtained were always more posterior than the fan-lines which corresponded to velar occlusion. This is the best I could do with the data at hand, although I understand it is not ideal.

L218-20 Can you specify the range of radial axes (i.e. their numbers) that were used as measurement axes across all speakers?

Added.

L225 Remove the word "real".

Removed.

L225 I am still confused about why there was no standard frame rate, although I don't think it will make a difference to the results given that the minimum frame rate is 43fps.

I added a footnote "The frame rate is adjusted by the system depending on other settings, so there is no standard frame rate."

L255 "...scaled (z-scored) tongue root position."

Added.

L257-258 I might have missed this part of the data transformation. Fig. 2's caption says that "Higher values indicate advancement." and appears to show greater ATR values at closure onset where a voiced consonant follows. I would have expected the opposite if advanced tongue root is being measured. With increased ATR, the tongue-root section of the tongue spline should intersect the radial fan axis closer to its origin, obtaining a smaller value. Kirkham and Nance (2017) reversed the sign of the z-scored tongue root distance measures, so that greater distance scores were associated with greater ATR (p71). Was that also carried out here? If it was, I think you should be explicit about that, rather than just saying you followed Kirkham and Nance's method.

I've expanded by saying that the sign has been flipped to facilitate interpretation (greater values = greater tongue root advancement).

282 "A second linear mixed regression was fitted to tongue root position ?at V1offset/C2onset?..."

Yes. I have now specified that the time point is V1 offset/C2 onset.

288-289 "V1 duration and tongue root position ?at V1offset/C2onset?..."

Yes. I have now specified that the time point is V1 offset/C2 onset.

378 Change to "It is worth briefly discussing..."

Changed

384 phonemic symbol missing

Fixed.

383-388 Perhaps this needs to be rephrased. It looks as if you are implying that the acoustic difference between /e/ and /ɛ/ is entirely due to differences in tongue-root position. Also phonemic contrast can occur due to quite subtle acoustic cues, so phonemic difference does not necessarily mean a huge difference in acoustics or underlying articulation.

Rephrased as "Given that the articulatory space within which the tongue can move is generally more constrained in stops than in vowels, and given that Kirkham 2017 find a difference of 4 mm in tongue root position in vowels, it makes sense to expect that differences in tongue root position as driven by consonantal factors should be of some magnitude smaller, like the ones found in this study."

390-3 See comments l57-58 Rothenberg's estimated duration for tongue-root advancement.

Rephrased as "If a maximal ballistic forward movement of the tongue root takes between 70 and 90 ms as suggested by the informal investigation by Rothenberg (1967), we can calculate the maximum displacement plausible to be between 4.55 to 5.85 mm (0.065 mm times 70–90 ms).".

393-4 Does this refer to: "Lacking further evidence we might assume it is possible to produce a vertical elongation of the pharynx in plosive production of about 0.5cm." (Rothenberg 1967: 98)?

This refers to a few lines above that: "If we consider the lateral and posterior walls of the pharynx to be held fixed, the anterior-posterior dimension of the pharynx can be increased by a forward motion of the larynx and/or base of the tongue. Forward motions of the base of the tongue of the order of magnitude of 0.5 cm have been reported by PERKELL (1965). The main activation is most likely supplied by the most inferior-posterior fibers of the genioglossus muscle to the base of the tongue, the geniohyoid muscles, and the infrahyoid muscles. The forward motion of the anterior wall of the pharynx might well be more of a pivoting around an axis somewhere near the cricoid cartilage than a uniform forward motion; however, it is difficult to find empirical evidence bearing on this question."

FIG. 6 caption "tornge". Why are there two of each plot?

Fixed. The plot duplication was a coding error and it's now fixed.

L431 "shows", or "these plots show"

Fixed.

L442 "The mean difference in tongue root position (at the onset of?) voiceless vs. voiced stops..."

Yes, added "at the onset of".

L464 and 465 citations missing.

I inserted here some of the citations from above for clarity.

Ll468 "Similarly to what (was) previously found for English (citation)..."

I inserted here some of the citations from above for clarity and I added "was".

L485 “replicate” -> “are replicable”

Changed.

Reviewer 2

The aim of this study is to show the relation between longer vowel duration and tongue root advancement (which is related to the voicing of the following stop). This manuscript contains some very interesting and original/novel data, which was not studied before. Also, it has a potential to be extended to other contexts as well as other languages. However, there are some concerns about the interpretation of the results. Here are my main concerns:

1. The result the author argues to be the main point of the manuscript (how a longer vowel duration corresponds to greater tongue root advancement) is not very well explained. The author shows some results (Figure 5), but does not explain enough why tongue root advancement is related to the longer vowel duration. This study explains why tongue root should be more advanced at the vowel offset of voiced stops, but not necessarily why more tongue root advancement is found with the longer vowel duration. The only part that explains the relation between the vowel duration and tongue root advancement is line 333 (p.21), but more explanations will be necessary—considering the main argument (and the title) of this paper is how longer vowel duration correlates with greater tongue root advancement)—or I wonder if the title purposefully used the word ‘displacement’ instead of ‘advancement’.

   I have reworded the relevant paragraphs of section IV.A (*also see comments below*).

2. Also, I’m concered about inter-speaker variation. In fact, out of 17 speakers, 7 speakers did not show more tongue root advancement at closure onset of voiced stops (Figure 7 and Figure 8). Considering the difficulty of analyzing everyone’s data collectively when doing an ultrasound study (due to the differences in each speaker’s anatomy), the statistical model(s) the author reports in the manuscript can be questionable especially with the high variability found among speakers.

   Figure 8 indicates that one speaker has no difference in tongue root position (*IT11*) *and two have the reversed pattern* (*IT09 and PL02*). *The figure of 7 speakers indicated in the text indicates that these speakers have a more uncertain estimate of tongue root difference, rather than no difference. I have now rephrased this paragraph. In Figure 9, 5 speakers of 17 have no distinction or the reversed pattern.*

3. On p. 19, Figure 5, it shows that when the vowel is longer, the tongue root position at the onset of the vowel is more retracted. The reason for this seems very important, but it’s not explained why that’s the case other than saying ”the trajectory curvature increases with vowel duration (line 302).

   This has changed after having fixed the statistical model (*I realised the number of basis functions was too low according to check.gam*() *so I have increased it*). *Now the initial values increases, except for 200 ms* (*I don’t have an explanation for why this is the case*).

4. The author reports the data on two different languages, Italian and Polish, but it is less clear why these two languages are studied, especially since as the author argues, two languages did not show any difference. Also, since the data were collected in the UK, it’s worth mentioning the length of stay in the UK and English proficiency of each speaker. It is well known that English voiced/voiceless stops behave differently from other languages, so English may have some effects on participants’ native languages. Regarding the materials used for the data collection, I wonder if there was any effect of including the real words. Also, why /b, p/ was not included (at least for a comparison to other lingual consonants?)

   Italian and Polish were chosen because they reportedly have varying degrees of the voicing effect, as discussed in Section I.B. A discussion of the possible influence of English is found in Coretta 2018, so I’ve added a footnote in Section II.B to refer the reader to that paper.

And, here are some parts I found unclear/need more explanations:

p. 3, 2nd paragraph: it’s unclear what the purpose of the last sentence is.

This sentence clarifies the use of "active gesture", to avoid confusion with the more general reading of active gesture as a gestures implemented for a specific purpose.

p. 13, line 235: Generalized additive mixed models (GAMMs). Including more explanations on this model will be very helpful to understand how this model's prediction works in Section III B and Fig. 3.

I have included more details on what each term contributes with to the model fit in Section III.B and III.D.

p. 20, line 330: "Said correlation exists independent of the voicing status of the consonant following the vowel" –> I especially found it difficult to understand the paragraph including this sentence and the following two paragraphs, and I believe these paragraphs are main arguments of this manuscript. More specifically, why do we see a greater tongue root advancement when the vowel is longer? Is it simply because voiced stops show more tongue root advancement and a longer vowel duration? More importantly, why do we see a greater curvature in longer vowels? Why does a longer vowel show a more retracted tongue root at its vowel onset?

I have updated the relevant paragraphs in Section IV.A to reflect changes in Section III.D and expanded them to clarify the point being made.

p. 21, 3rd paragraph: this paragraph suddenly mentions speech rate and its paradox, and it is not very clear what the argument here is. Considering inter-speaker variation. I think at least with speech rate and tongue root advancement of vowel onset mentioned here, it may be worth looking at individual speaker's results instead of just use speech rate as a factor in the statistical model.

I have expanded and moved the relevant paragraphs to after the discussion of speakers' variation.

p. 23-24, Section IV B: I think this section can be either condensed or deleted since estimates of tongue root displacement will be highly variable depending on speaker. Since (1) Speaker variation was reported in this study, (2) only two languages are studied, and (3) the author mentions that "the correlation between tongue root position and vowel duration needs to be replicated by expanding the enquired contexts to other types of consonants and vowels, and with other languages" (p. 24, line 410), I recommend to condense this section.

It is generally desirable to discuss in some details the estimates of the sought effects, as recommended by recent literature on statistical power. It is true that the estimates will be highly variable depending on speaker, but since such estimates refer to a population of speakers it can still be useful, for example, for the determination of Bayesian priors in future work. The discussion of the estimates also links the results of this study back to theoretical argumentations from previous work, and it thus constitutes a grounding point.

There are some parts where the tables and figures may be necessary and/or more explanations on figures may be helpful:

First of all, in the results, tables with these acoustic data are necessary: e.g. mean V1 duration, mean speech rate, mean closure-to-closure interval

p. 11, line 206: what is "higher formant"? Either give a very specific example, or provide a sample spectrogram.

I have specified what is meant by "higher formants structure", taken from Machač 2009.

Figure 1: please indicate the front and back of the tongue

Added indication of back/front.

Figure 2: How was the data of each speaker scaled? Also, it will be helpful if this figure is in color.

I've added that those are z-scores and I made the figure in colour.

Figure 4: Maybe using color and indicating voiced/voiceless stops may be helpful.

I've added colour by vowel in the plot. The model from which the regression lines are obtained does not contain C2 voicing. This is because a separate model indicated that voicing nor its interaction with vowel duration is significant, so it was dropped from the final model. I've added a mention to this in Section III.C.

Figure 5: Why 145ms was used instead of 150ms? (while other V1 duration values used are 50, 100, and 200). Also, why 145 ms and 200 ms starts from more retracted tongue root position? It's not explained in the text.

The figure has now 150 ms rather than 145 ms. I have also increased the number of basis functions in the smooth for vowel duration, since I realised it was too low according to "gam.check(). This has partially changed the starting values for the different vowel durations, so that now there is an increase in starting tongue root position with the exception of 200 ms which decreases. I do not know why this is the case. I've added in the text: "I have no explanation for why the advancement of the root seemingly increases with increasing vowel duration except when the duration goes from 150 to 200 ms."

The following other minor suggestions are noted:

p. 2, line 18: "a lot of work has been done on each of these aspects separately,..." but actual references are not mentioned; also, line 24—references

I have added here some of the references from later sections for clarity.

p. 3, line 39: "Aerodynamic Voicing Constraint": since this concept is mentioned as a very important concept of this manuscript, more detailed information about it will be helpful

The constraint simply states that there must be a positive trans-glottal air pressure differential, as discussed in lines 33-42.

p. 4, line 58: ...root "advancement" about....

The sentence has been rephrased as suggested by one of the other reviewers and moved at the beginning of the paragraph.

p. 4, line 72: typo–'longer' –> 'shorter'

Fixed.

p. 4, line72-73: please indicate what language(s) each paper worked on

Added.

p. 5, line 79: (Ahn, 2018)–> Ahn (2018)

Fixed.

p. 5, line 98: here as well, please specify what language(s) each paper worked on

Added.

p. 6, line 106: "relative ease of comparison" –> not sure what exactly it means

I've explained this now "Moreover, the segmental phonologies of these languages facilitate the design of sufficiently comparable experimental material (see Coretta 2018 for a more thorough discussion)."

p. 6, line 111: "the stressed vowels of disyllabic words"–> indicate which vowel is stressed

Stress is indicated with a diacritic although it is not properly rendered due to restrictions to the LaTeX submission system, but it should be fixed in the final proof. I have now specified that the stressed vowel is the first.

p. 8, line 147: missing '.' After Verbania (Italy)

Fixed.

p. 10, line 182: "high and front vowels usually produce less tongue displacement from and to a stop consonant" –> reference(s)?

This fact hasn't been systematically studied, but since the tongue is already quite high because of the production of high-front vowels, the displacement between the position of the tongue during high-front vowels and that of the tongue during stop closure is generally quite small and it makes it difficult to detect gestural landmarks from UTI displacement data.

p. 13, line 231: the calculated speech rate values—here, please explain how the values were calculated

I have included there a condensed version of the formula from above for clarity.

p. 13, line 235: Generalised additive mixed models—since it's first time mentioning this model, show the abbreviation here (GAMMs?); line 238: GAMs–> GAMMs??

Added abbreviation and fixed line 238.

P. 21, line 349 and 351: Two "howevers" are confusing. maybe change the second however to "nevertheless"?

I have fixed the repetition of "however" by rephrasing.

P. 22, line 357: 'previous work' > reference(s)

I have added a cross-reference to the patterns described in the Introduction section.

P. 23, line 384: a +ATR vowel, a -ATR vowel,

Fixed.

P. 23, line 384: also missing IPA symbol in the second / /

Fixed.

P. 25, Figure 6: typo (torngue–> tongue)

Fixed.

P. 26, line 447: 11 speakers–> 10 speakers?

Fixed.