Assessing midsaggital tongue contours in polar coordinates using generalised additive (mixed) models

Stefano Coretta

1 Introduction

Since the publication of the seminal paper by Davidson (2006), statistical modelling of whole tongue contours obtained with ultrasound imaging has been dominated by the use of Smoothing Splines Analysis of Variance (SSANOVA, Gu 2013). These models have greatly advanced our understanding of tongue articulation and speech modelling. On the other hand, a variety of research disciplines is witnessing an increased use of Generalised Additive Models (GAMs) and their mixed-effects counterpart (GAMMs, Wood 2006), especially when dealing with complex data. GAMs have being increasingly adopted in linguistics as a means to model dynamic speech data. This paper introduces an implementation of GAMs with tongue contours using polar coordinates. The use of polar GAMs is illustrated with ultrasound tongue imaging data comparing voiceless and voiced stops. The R package `rticulate` has been developed to facilitate the use of the model, and it is briefly introduced here.

1.1 Ultrasound tongue imaging

Ultrasound imaging is a non-invasive technique for obtaining an image of internal organs and other body tissues. 2D ultrasound imaging has been successfully used for imaging sections of the tongue surface (for a review and applications in field settings, see Gick 2002 and Lulich et al. 2018). An image of the (2D) tongue surface can be obtained by placing the transducer in contact with the sub-mental triangle (the area under the chin), aligned either with the mid-sagittal or the coronal plane. The ultrasonic waves propagate from the transducer in a radial fashion through the aperture of the mandible and get reflected when they hit the air above the tongue surface. This 'echo' is captured by the transducer and translated into an image like the one shown in Figure 1.

1.2 Generalised Additive models

Generalised additive models, or GAMs, are a more general form of non-parametric modelling that allows fitting non-linear as well as linear effects, and combine properties of linear and additive modelling (Hastie & Tibshirani, 1986). GAMs are built with smoothing splines (like SSANOVA, see Helwig & Ma 2016), which are defined piecewise with a set (the *basis*) of polynomial functions (the *basis functions*). When fitting GAMs, the smoothing splines try to maximise the fit to the data while being constrained by a smoothing penalty (usually estimated from the data itself). Such penalisation constitutes a guard against overfitting. GAMs

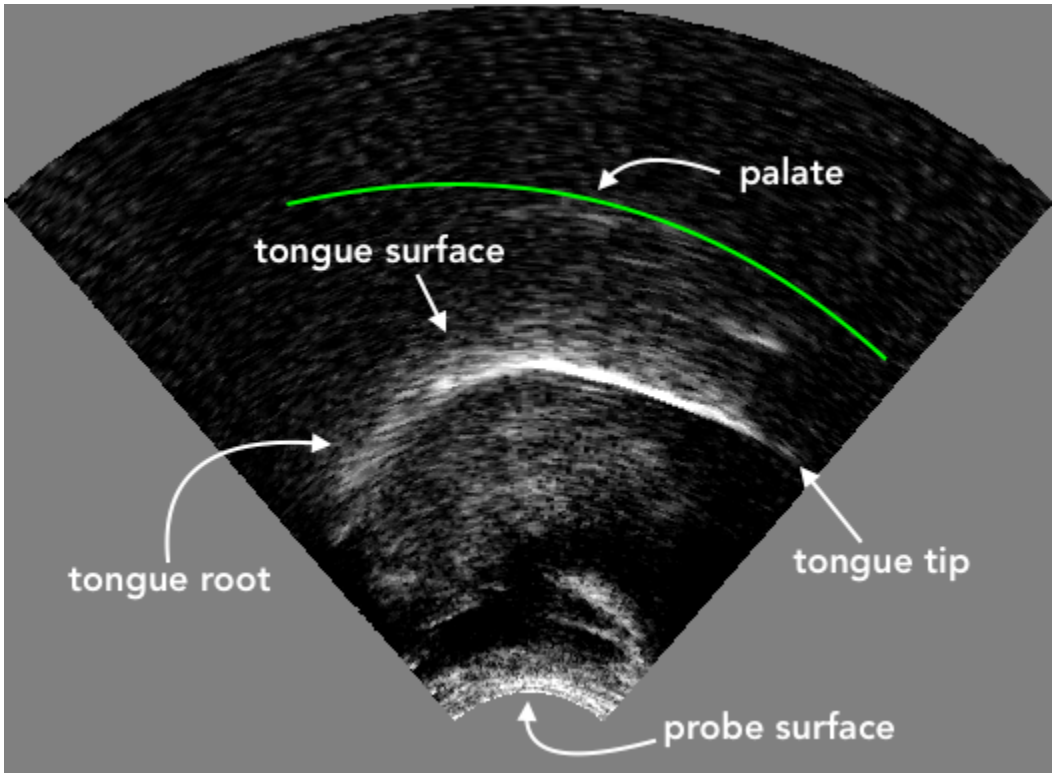

Figure 1: An ultrasound image showing a mid-sagittal view of the tongue. The white curved stripe in the image indicates where the ultrasonic waves have been reflected by the air above the tongue. The tongue surface corresponds to the lower edge of the white stripe. In this image, the tongue tip is located on the right. The green curve approximates the location of the palate.

are thus powerful and flexible models that can deal with non-linear data efficiently. Moreover, GAMs have a mixed-effect counterpart, Generalised Additive Mixed Models (GAMMs), in which random effects can be included (for a technical introduction to GAM(M)s, see Zuur (2012) and Wood (2017)). GAMs can offer relief from issues of autocorrelation between points of the contour (given that points close to each other are not independent from one another). For example, GAMs can fit separate smooths to individual contours, or a first-order autoregression model can be included which tries to account for the autocorrelation between each point in the contour and the one following it. Tongue contours obtained from ultrasound imaging lend themselves to be efficiently modelled using GAM(M)s.

1.3 Polar coordinates

Mielke (2015) and Heyne & Derrick (2015b,a) have shown that using polar coordinates of tongue contours rather than cartesian coordinates brings several benefits, among which reduced variance at the contour edges. Points in a polar coordinate system are defined by pairs of radial and angular values. The point is describes with a radius, which corresponds to the radial distance from the origin, and the angle from the reference radius. Tongue contours, due to their shape, tend to have increasing slope at the left and right edges, in certain cases tending to become almost completely vertical. The almost verticality of the contours has the effect of increasing the variance of the fitted contours (and hence increased confidence intervals), and in some cases it can even generate uninterpretable curves.

This issue is illustrated in Figure 2. The x and y axes are the x and y cartesian coordinates in millimeters. The plot shows LOESS smooths superimposed on the points of the individual tongue contours of an Italian speaker (IT01, see Section 2.1). These contours refer to the mid-sagittal shape of the tongue during the closure of four consonants (/t, d, k, g/) preceded by one of three vowels (/a, o, u/). The tip of the tongue is on the right-end side of each panel. Focussing on the smooths, it can be noticed that the smooths in the contexts of the vowel /u/ diverge substantially from the true contours (as inferred by the points). In the contexts of velar consonants and the other two vowels, one could say that the back/root of the tongue is somewhat flatted out relative to the actual contours. These artefacts of smoothing happen because, especially in the right-edge of these particular contours, the slope of the curve increases in such a way that at times the curve bends under itself (see for example the context /ug/, when x is between -30 and -20). Since those points on the bend share the x value, the smooth just averages across the y values of those points.

Figure 3 shows a better way of representing individual tongue contours. In these plots, the points of each contour are connected sequentially by a line, rather than smoothed over. The parts in which the contours bends over are kept and visualised correctly

These figures illustrate that using cartesian coordinates for modelling can introduce smoothing artefacts which in turn can negatively affect the model output. When tongue contours are expressed with polar coordinates, on the other hand, the variance is reduced and the fitted contours generally reflect more closely the underlying tongue shape. Mielke has implemented a series of R (R Core Team, 2018) functions for fitting polar SSANOVAs to tongue contours. While model fitting is achieved using polar coordinates, plotting of the model output is subsequently obtained by reconverting the coordinates to cartesian. The method introduce in this paper is based on Mielke's implementation applied to GAMs.

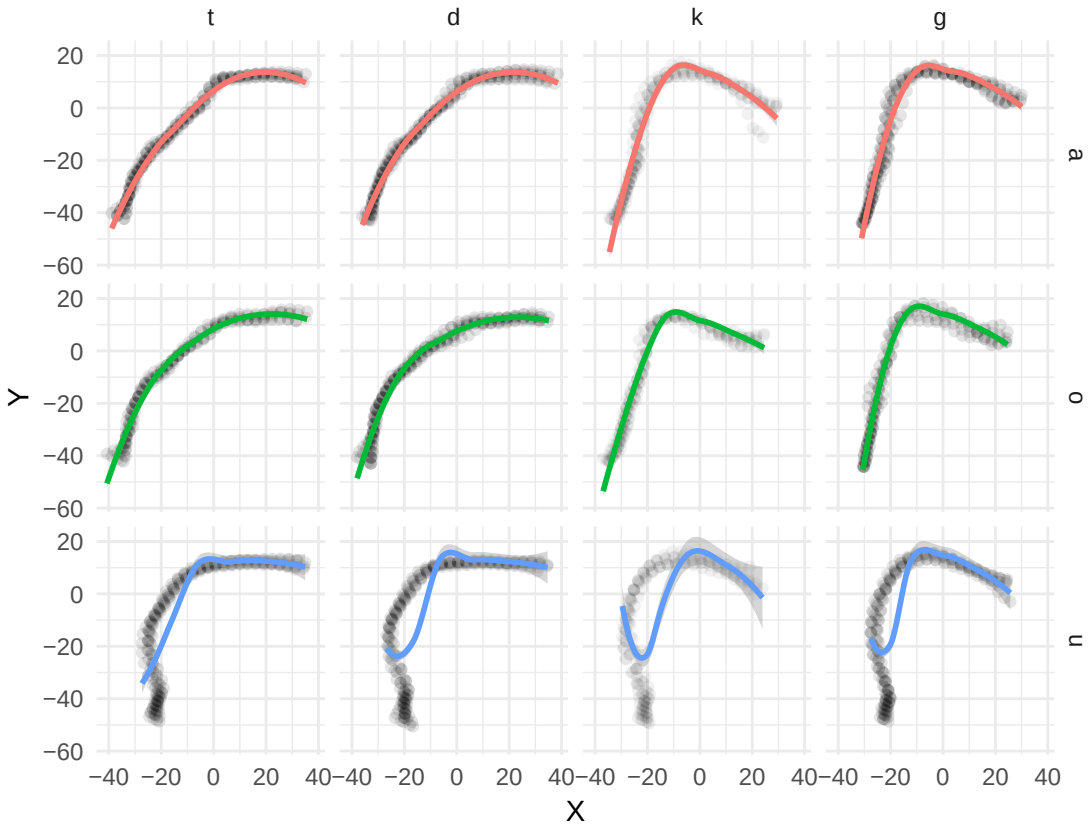

Figure 2: Estimated tongue contours of IT01 depending on C2 place, vowel and C2 voicing.

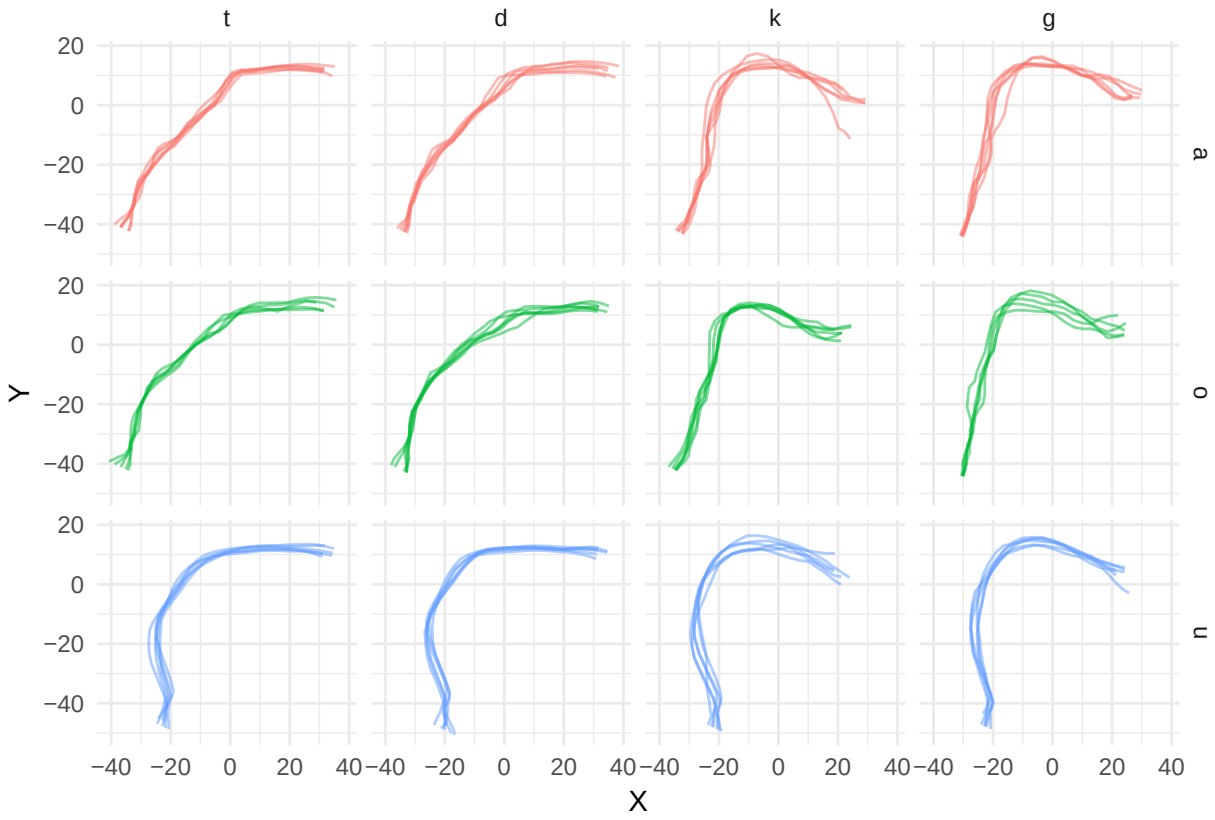

Figure 3: Estimated tongue contours of IT01 depending on C2 place, vowel and C2 voicing.

2 Polar GAM(M)s

GAMs fitted to tongue contours in polar coordinates are introduced here. A polar GAM is constructed as follows. The outcome variable of the model are the radial coordinates, while a smooth term over the angular coordinates is the predictor which takes care of modelling the curved shape of the contour. Other predictors, such as consonant or vowel type, speech rate, or random effects, can be also included. The model returns fitted smooths with polar coordinates as units. The predicted polar coordinates can be derived from the fitted smooths and converted into a cartesian coordinate system (centred on the origin of the polar system) for plotting. A simple example with data from one speaker will illustrate how to fit polar GAMs with the R package `rticulate`. The following section gives information on the ultrasonic system used for data collection and on how the data has been processed, before moving onto model fitting itself.

2.1 Data collection and processing

Synchronised audio and ultrasound tongue imaging data have been recorded from 4 speakers of Italian while reading a series of controlled sentences. An Articulate Instruments Ltd™ set-up was used for this study. The ultrasonic data was collected through a TELEMED Echo Blaster 128 unit with a TELEMED C3.5/20/128Z-3 ultrasonic transducer (20mm radius, 2-4 MHz). A synchronisation unit (P-Stretch) was plugged into the Echo Blaster unit and used for automatic audio/ultrasound synchronisation. A FocusRight Scarlett Solo pre-amplifier and a Movo LV4-O2 Lavalier microphone were used for audio recording. The acquisition of the ultrasonic and audio signals was achieved with the software Articulate Assistant Advanced (AAA, v2.17.2) running on a Hawlett-Packard ProBook 6750b laptop with Microsoft Windows 7. Stabilisation of the ultrasonic transducer was ensured by using the metallic headset produced by Articulate Instruments Ltd™ (2008).

Before the reading task, the participant's occlusal plane was obtained using a bite plate (Scobbie et al., 2011). The participants read nonce words embedded in the frame sentence *Dico* ____ *lentamente* 'I say ____ slowly'. The words follow the structure $C_1V_1C_2V_2$, where C_1 = /p/, V_1 = /a, o, u/, C_2 = /t, d, k, g/, and $V_2 = V_1$. Each speaker repeated the stimuli six times.

Spline curves were fitted to the visible tongue contours using the AAA automatic tracking function. Manual correction was applied in those cases that showed clear tracking errors. The time of maximum tongue displacement within consonant closure was then calculated in AAA following the method in Strycharczuk & Scobbie (2015). A fan-like frame consisting of 42 equidistant radial lines was used as the coordinate system. The origin of the 42 fan-lines coincides with the centre of the ultrasonic probe, such that each fan-line is parallel to the direction of the ultrasonic signal. Tongue displacement was thus calculated as the displacement of the fitted splines along the fan-line vectors. The time of maximum tongue displacement was the time of greater displacement along the fan-line vector that showed the greatest standard deviation (as assessed manually). The vector standard deviation search area was restricted to the portion of the contour corresponding to the tongue tip for coronal consonants, and to the portion corresponding to the tongue dorsum for velar consonants.

The cartesian coordinates of the tongue contours were extracted from the ultrasonic data

at the time of maximum tongue displacement (always within C2 closure). The contours were subsequently normalised within speaker by applying offsetting and rotation relative to the participant's occlusal plane (Scobbie et al., 2011). Each participants' dataset is thus constituted by x and y coordinates of the tongue contours that define respectively the horizontal and vertical axes. The horizontal plane is parallel to the speaker's occlusal plane.

2.2 Fitting a polar GAM

GAMs can be fitted in R with the `gam()` function from package `mgcv` (Wood, 2011, 2017). `bam()` is a more efficient function when the datasets has several hundreds observations. The package `rticulate` has been developed as a wrapper of the `bam()` function to be used with tongue contours. The special function `polar_gam()` can fit any specified GAM model to tongue contours coordinates, using the same syntax of `mgcv`. The function accepts tongue contours either in cartesian or polar coordinates. In the first case, the coordinates can be transformed into polar before fitting. If the data is in the AAA fan-like coordinate system, the origin is automatically estimated with the method in Heyne & Derrick (2015b). If the data is not exported from AAA, the user can specify the known coordinates of probe origin. The function `plot_polar_smooths()`, used for plotting the estimated contours, converts the coordinates back into cartesian using the same origin as with GAM fitting.

A GAM in R can be specified with a formula that uses the same syntax of `lme4`, a commonly used package for linear mixed-effects models (Bates et al., 2015). The `mgcv` package allows to specify smoothing spline terms with the function `s()`. This function takes the term along which a spline is created (for example, time in a time series, or x-coordinates in a cartesian system). Among the arguments of `s()`, the user can select the type of spline (the `bs` argument) and the grouping factor used for comparison (the `by` argument). For a more in-depth introduction to GAMs in R for linguistics, see Sóskuthy (2017) and Wieling (2017).

As means of illustration, the following paragraphs will show how to fit a polar GAM with data from one of the 4 Italian speakers. Due to differences in the placement of the probe and in the speakers' anatomy, different portions of the tongue are likely to be imaged across speakers, so that scaling might not be possible (or wise). For this reason, it is recommended to fit separate models for each participant, rather than aggregate all of the data in a single model.

We can start from a simple model in which we test the effect of C2 place, vowel, and voicing on tongue contours. `vc_voicing` is an ordered factor that specifies the combination of C2 place, vowel, and voicing. Modelling different contours for each combination of the three predictors can be achieved by using `vc_voicing` with the `by` argument of the difference smooth, and by including `vc_voicing` as a parametric term. The following code fits the specified model to the contour data of IT01. When running the code, the coordinates of the estimated origin used for the conversion to polar coordinates are returned. The model is fitted by Maximum Likelihood (ML) here to allow model comparison below).

```
it01_gam <- polar_gam(
  Y ~
    vc_voicing +            # parametric term
```

```
    s(X) +                    # reference smooth
    s(X, by = vc_voicing),    # difference smooth
  data = tongue_it01,
  method = "ML"
)
```

```
## The origin is x = 14.3900999664996, y = -65.2314226131983.
```

The function `plot_polar_smooths()` can be used to plot the estimated contours. The shaded areas around the estimated contours are 95% confidence intervals. Note that, differently from SSANOVA, statistical significance can't be assessed from the overlapping (or lack thereof) of the confidence intervals.

```
plot_polar_smooths(
  it01_gam,
  X,
  voicing,
  facet_terms = c2_place + vowel,
  # the following splits the factor interaction in the individual terms,
  # so that they can called in the plotting arguments
  split = list(vc_voicing = c("vowel", "c2_place", "voicing"))
) +
  coord_fixed() +
  theme(legend.position = "top")
```

One way to assess significance of model terms is to compare the ML score of the full model against one without the relevant predictor, using the function `compareML()` from the `itsadug` package. Both the parametric term and the difference smooth need to be removed in the null model.

```
it01_gam_0 <- polar_gam(
  Y ~
    # vc_voicing +            # remove parametric term
    s(X),                     # keep reference smooth
    # s(X, by = vc_voicing),  # remove difference smooth
  data = tongue_it01,
  method = "ML"
)
```

```
## The origin is x = 14.3900999664996, y = -65.2314226131983.
```

```
compareML(it01_gam_0, it01_gam)
```

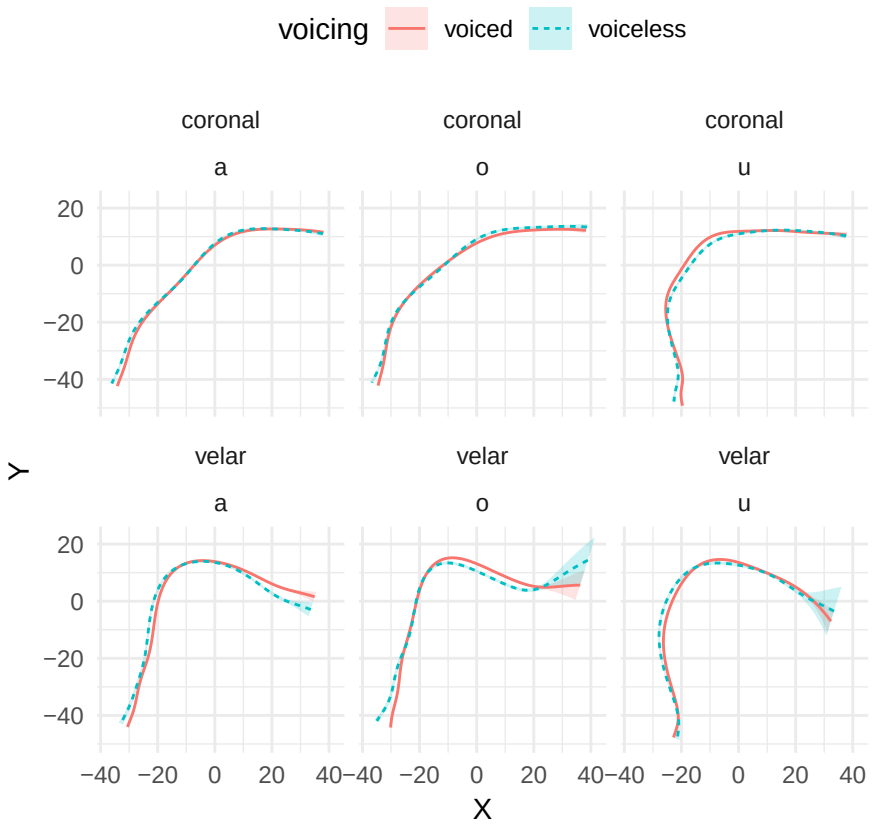

Figure 4: Estimated tongue contours of IT01 depending on C2 place, vowel and C2 voicing.

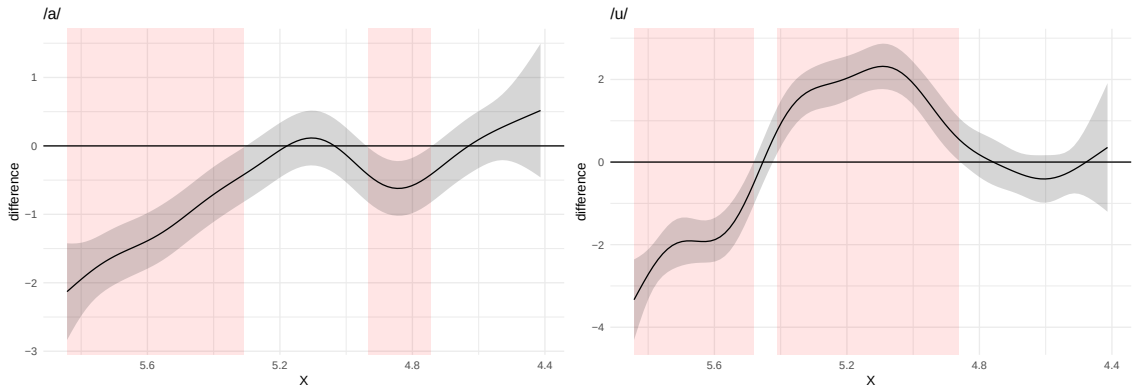

Figure 5: Difference smooth of voiceless vs. voiced stops in the context of /a/ (left) and /u/ (right).

```
## it01_gam_0: Y ~ s(X)
##
## it01_gam: Y ~ vc_voicing + s(X) + s(X, by = vc_voicing)
##
## Chi-square test of ML scores
## -----
##        Model     Score Edf Difference     Df  p.value Sig.
## 1 it01_gam_0 12395.227   3
## 2   it01_gam  7423.356  36   4971.871 33.000  < 2e-16  ***
##
## AIC difference: 10258.65, model it01_gam has lower AIC.
```

To check which part of the contour differs among conditions, the method recommended in Sóskuthy (2017) is to plot the difference smooth and check the confidence interval. The parts of confidence interval that don't include 0 indicate that the difference between contours in that part is significant. Figure 5 illustrates the use of difference smooths with the difference smooths of voiceless vs. voiced coronal stops when the vocalic context is /a/ or /u/. As per usual, the tongue tip is on the right-end side of each plot. The difference smooths indicate that there is a significant difference along the most anterior part of the tongue (the root and dorsum). Based on the predicted smooths sown in Figure 4, we can argue that, in the context of coronal consonants, the root is more advanced in voiced relateive to voiceless stops (when the vowel is either /a/ or /u/), and that the dorsum is also somewhat retracted in voiced stops if the vowel is /u/.

As mentioned in the introduction, autocorrelation in the data can produce unwanted patterns in the residuals, which in turn can affect the estimated smooths (and falsely increase certainty about them). A first-order autoregressive (AR1) model can be included to reduce autocorrelation at lag 1. Figure 6 show the autocorrelations in the residuals without and with an AR1 model. The GAM model with the AR1 correction has lower values of autocorrelations. In this case, it is thus advisable to perform ML comparison and smooths plotting with models in which an AR1 model has been included. For a more in-depth treatment of issues related to autocorrelation, see Sóskuthy (2017).

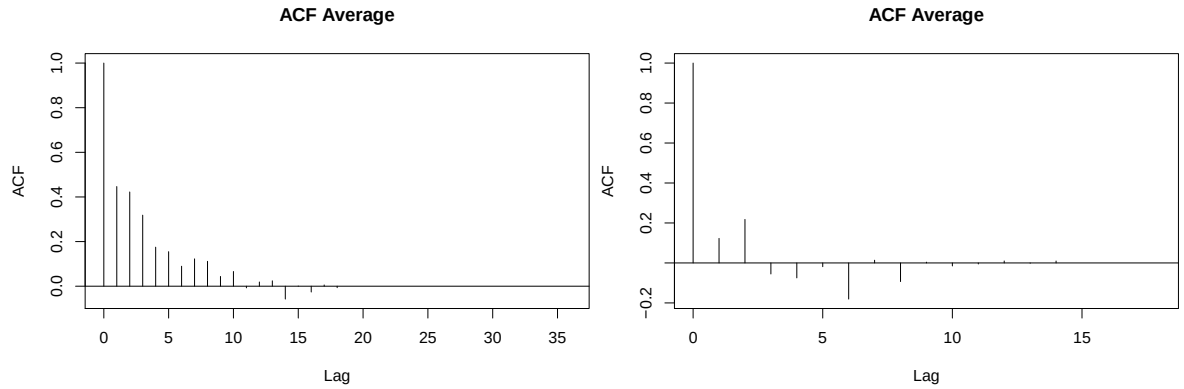

Figure 6: Autocorrelation plots of a model fitted without (left) and with (right) a first-order autoregressive model (AR1).

3 Comparing tongue contours in voiceless and voiced stops

Mid-sagittal tongue contours at maximum tongue displacement of voiceless and voiced stops have been compared using polar GAMs. Figure 7 to Figure 10 show an appreciable degree of variation across speakers and phonological contexts in relation to the differences in tongue shapes between voiceless and voiced stops. In some speakers and contexts, the tongue root (the left of the tongue contours) is more advanced in voiced stops than in voiceless stops, especially in the context of a coronal C2 and /a/. Tongue root advancement is a well known mechanism employed to maintain intra-oral pressure below the threshold required for voicing (Ohala, 2011; Kent & Moll, 1969; Perkell, 1969; Westbury, 1983; Ahn, 2018).

The magnitude of the difference in tongue root position in the data reported here is about 2 millimetres. Kirkham & Nance (2017) find that the tongue root in +ATR vowels is on average 4 millimetres more advanced than the respective −ATR vowels. Rothenberg (1967) argues, based on modelling, that the tongue root can move forward by a maximum of about 5 mm mid-sagittally. This movement corresponds to an average volume increase of 18 cm^2. Given these estimates, it can be argued that a 2 mm change reasonably contributes to an appreciable volume increase, also considering that other volume expansion mechanisms can operate along with the advancement of the tongue root (like larynx lowering, slack oral walls, etc.).

4 Conclusions

Generalised additive (mixed) models (GAMs) can be efficiently used to statistically assess differences in tongue contour shapes as obtained from ultrasound tongue imaging. This paper showed how GAMs can be fitted to tongue contours in polar coordinates in R with the specialised package `rticulate`. An example of how GAMs can help modelling differences in tongue contours has been illustrated with data from 4 speakers of Italian in which the mid-sagittal tongue contours of voiceless and voiced stops where compared. The same general advantages and issues noted in Davidson (2006) for SSANOVA apply to polar GAMs. In

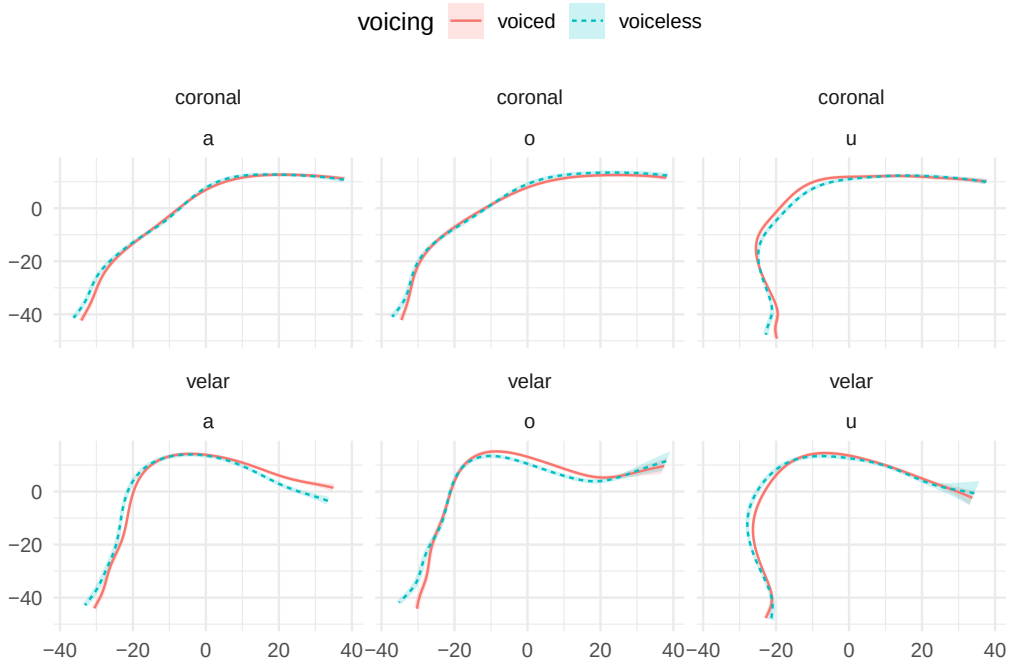

Figure 7: Tongue contours of voiceless and voiced stops in IT01.

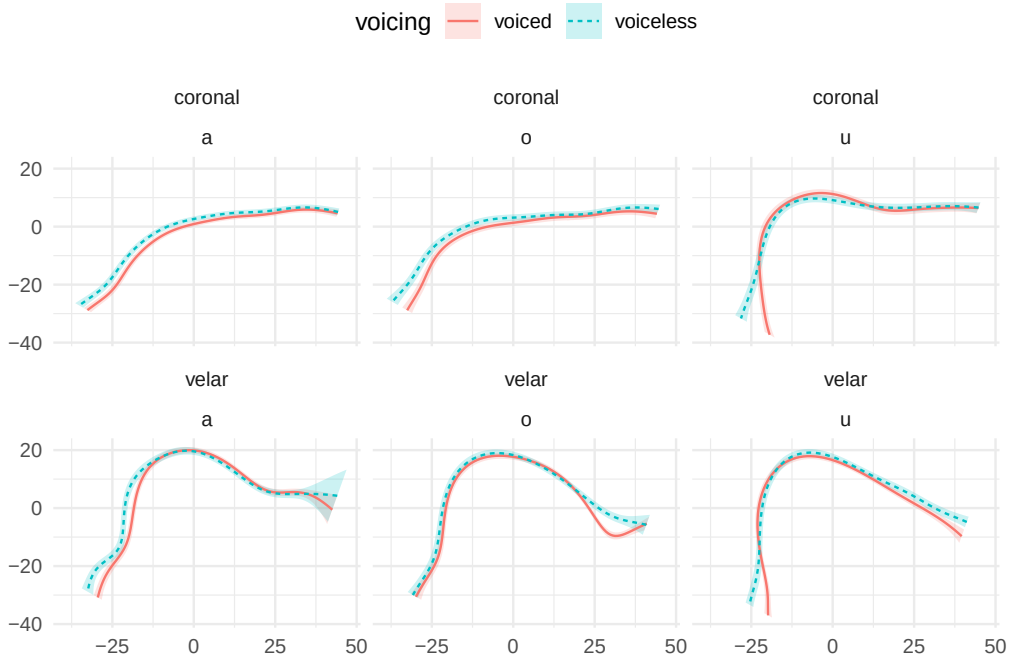

Figure 8: Tongue contours of voiceless and voiced stops in IT02.

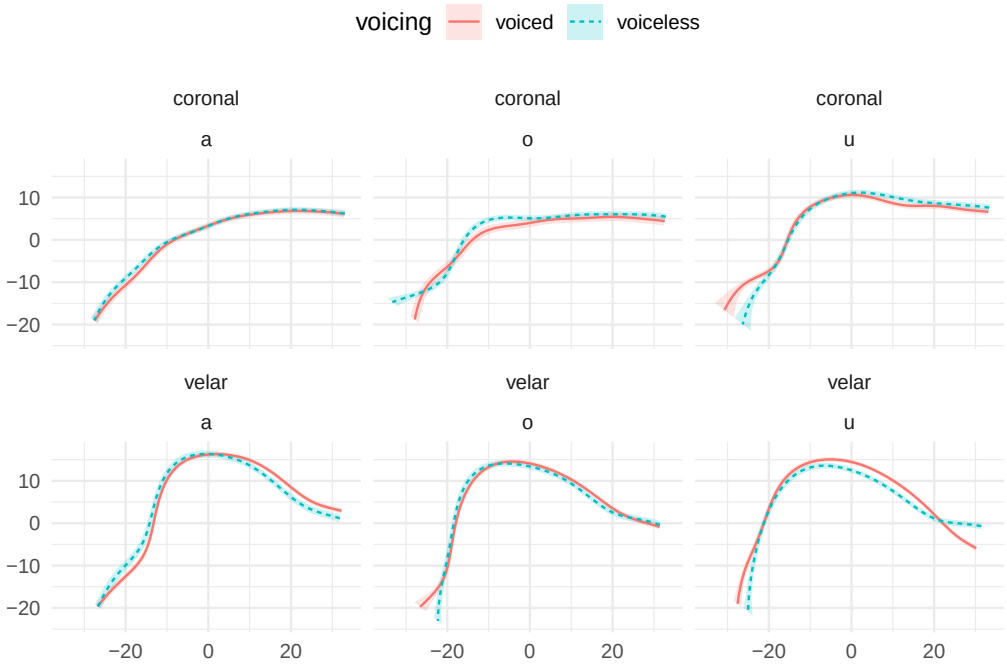

Figure 9: Tongue contours of voiceless and voiced stops in IT03.

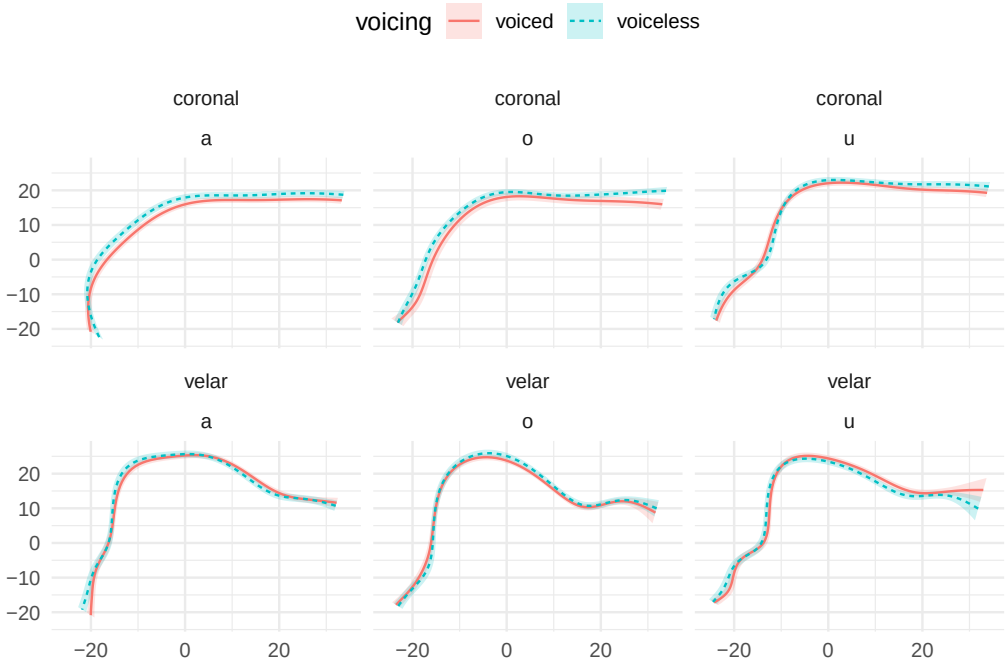

Figure 10: Tongue contours of voiceless and voiced stops in IT04.

particular, while within-speaker normalisation can be achieved by rotation and offsetting of the data relative to a bite plate (as done here), across-speaker normalisation represents a bigger challenge. Since it can't be deduced with sufficient certainty from the ultrasonic image which part of the tongue is being actually imaged, it is not possible to define fixed anatomical landmarks across speakers that can be used in normalisation. For this reason it has been recommended here to fit separate models for each speaker. Future work will explore ways of allowing the user to use data aggregated from multiple speakers while accounting for the uncertainty in which parts of the tongue are imaged. To conclude, polar GAMs can also be extended to model whole tongue contours differences over time (in other words, how the sectional shape of the tongue changes over time) and 3D tongue surfaces.

References

Ahn, Suzy. 2018. The role of tongue position in laryngeal contrasts: An ultrasound study of english and brazilian portuguese. *Journal of Phonetics* 71. 451–467.

Articulate Instruments Ltd™. 2008. Ultrasound stabilisation headset users manual: Revision 1.4. Edinburgh, UK: Articulate Instruments Ltd.

Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48. doi:10.18637/jss.v067.i01.

Davidson, Lisa. 2006. Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance. *The Journal of the Acoustical Society of America* 120(1). 407–415. doi:10.1121/1.2205133.

Gick, Bryan. 2002. The use of ultrasound for linguistic phonetic fieldwork. *Journal of the International Phonetic Association* 32(02). 113–121. doi:10.1017/S0025100302001007.

Gu, Chong. 2013. *Smoothing spline ANOVA models*, vol. 297. Springer Science & Business Media.

Hastie, Trevor & Robert Tibshirani. 1986. Generalized additive models. *Statistical Science* 1(3). 297–310.

Helwig, Nathaniel E & Ping Ma. 2016. Smoothing spline ANOVA for super-large samples: Scalable computation via rounding parameters. arXiv.org preprint, arXiv:1602.05208 [stat.CO].

Heyne, Matthias & Donald Derrick. 2015a. Benefits of using polar coordinates for working with ultrasound midsagittal tongue contours. *The Journal of the Acoustical Society of America* 137(4). 2302–2302. doi:10.1121/1.4920405.

Heyne, Matthias & Donald Derrick. 2015b. Using a radial ultrasound probe's virtual origin to compute midsagittal smoothing splines in polar coordinates. *The Journal of the Acoustical Society of America* 138(6). EL509–EL514. doi:10.1121/1.4937168.

Kent, Raymond D. & Kenneth L. Moll. 1969. Vocal-tract characteristics of the stop cognates. *Journal of the Acoustical Society of America* 46(6B). 1549–1555.

Kirkham, Sam & Claire Nance. 2017. An acoustic-articulatory study of bilingual vowel production: Advanced tongue root vowels in Twi and tense/lax vowels in Ghanaian english. *Journal of Phonetics* 62. 65–81. doi:10.1016/j.wocn.2017.03.004.

Lulich, Steven M., Kelly H. Berkson & Kenneth de Jong. 2018. Acquiring and visualizing 3D/4D ultrasound recordings of tongue motion. *Journal of Phonetics* 71. 410–424. doi:10.1016/j.wocn.2018.10.001.

Mielke, Jeff. 2015. An ultrasound study of Canadian French rhotic vowels with polar smoothing spline comparisons. *The Journal of the Acoustical Society of America* 137(5). 2858–2869. doi:10.1121/1.4919346.

Ohala, John J. 2011. Accommodation to the aerodynamic voicing constraint and its phonological relevance. In *Proceedings of the 17th International Congress of Phonetic Sciences*, 64–67.

Perkell, Joseph S. 1969. *Physiology of speech production: Results and implication of quantitative cineradiographic study.* Cambridge, MA: MIT Press.

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

Rothenberg, Martin. 1967. *The breath-stream dynamics of simple-released-plosive production*, vol. 6. Basel: Biblioteca Phonetica.

Scobbie, James M., Eleanor Lawson, Steve Cowen, Joanne Cleland & Alan A. Wrench. 2011. A common co-ordinate system for mid-sagittal articulatory measurement. In *QMU CASL Working Papers*, 1–4.

Sóskuthy, Márton. 2017. Generalised additive mixed models for dynamic analysis in linguistics: A practical introduction. arXiv.org preprint, arXiv:1703.05339.

Strycharczuk, Patrycja & James M. Scobbie. 2015. Velocity measures in ultrasound data. Gestural timing of post-vocalic /l/ in English. In *Proceedings of the 18th International Congress of Phonetic Sciences*, 1–5.

Westbury, John R. 1983. Enlargement of the supraglottal cavity and its relation to stop consonant voicing. *The Journal of the Acoustical Society of America* 73(4). 1322–1336.

Wieling, Martijn. 2017. Generalized additive modeling to analyze dynamic phonetic data: a tutorial focusing on articulatory differences between l1 and l2 speakers of english. *The Mind Research Repository (beta)* (1). doi:10.1016/j.wocn.2018.03.002.

Wood, Simon. 2006. *Generalized additive models: An introduction with R.* CRC Press.

Wood, Simon. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)* 73(1). 3–36.

Wood, Simon. 2017. *Generalized additive models: An introduction with R.* Chapman and Hall/CRC 2nd edn.

Zuur, Alain F. 2012. *A beginner's guide to generalized additive models with R.* Highland Statistics Limited: Newburgh.