Laboratory Phonology
Journal of the Association for Laboratory Phonology

Carignan, C., et al. 2020 Analyzing speech in both time and space: Generalized additive mixed models can uncover systematic patterns of variation in vocal tract shape in real-time MRI. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 11(1): 2, pp. 1–26. DOI: https://doi.org/10.5334/labphon.214

JOURNAL ARTICLE

Analyzing speech in both time and space: Generalized additive mixed models can uncover systematic patterns of variation in vocal tract shape in real-time MRI

Christopher Carignan[1,2,4], Phil Hoole[2], Esther Kunay[2], Marianne Pouplier[2], Arun Joseph[3], Dirk Voit[3], Jens Frahm[3] and Jonathan Harrington[2]

[1] Speech, Hearing and Phonetic Sciences, University College London, UK

[2] Institute of Phonetics and Speech Processing, Ludwig Maximilians Universität Munich, DE

[3] Max Planck Institute for Biophysical Chemistry, Göttingen, DE

[4] The MARCS Institute for Brain, Behaviour and Development, Western Sydney University, AU

Corresponding author: Christopher Carignan (c.carignan@ucl.ac.uk)

We present a method of using generalized additive mixed models (GAMMs) to analyze midsagittal vocal tract data obtained from real-time magnetic resonance imaging (rt-MRI) video of speech production. Applied to rt-MRI data, GAMMs allow for observation of factor effects on vocal tract shape throughout two key dimensions: time (vocal tract change over the temporal course of a speech segment) and space (location of change within the vocal tract). Examples of this method are provided for rt-MRI data collected at a temporal resolution of 20 ms and a spatial resolution of 1.41 mm, for 36 native speakers of German. The rt-MRI data were quantified as 28-point semi-polar-grid aperture functions. Three test cases are provided as a way of observing vocal tract differences between: (1) /aː/ and /iː/, (2) /aː/ and /aɪ/, and (3) accentuated and unstressed /aː/. The results for each GAMM are independently validated using functional linear mixed models (FLMMs) constructed from data obtained at 20% and 80% of the vowel interval. In each case, the two methods yield similar results. In light of the method similarities, we propose that GAMMs are a robust, powerful, and interpretable method of simultaneously analyzing both temporal and spatial effects in rt-MRI video of speech.

Keywords: real-time MRI; GAMM; FLMM; speech dynamics; speech imaging

1. Introduction

One of the primary challenges facing speech articulation researchers is obtaining, quantifying, and interpreting data that capture both the spatial and temporal complexity of speech production. On the one hand, technologies that register flesh point positions (e.g., electromagnetic articulometry; EMA) are excellent for capturing articulatory kinematics, and the resulting data are readily quantifiable. However, interpretation of these data is limited to the kinematics of a set number of flesh points that can physically be accessed by the researcher, thus disregarding a large amount of spatial information about vocal tract action (e.g., pharyngeal constriction/expansion). On the other hand, speech imaging technologies (e.g., real-time magnetic resonance imaging; rt-MRI) yield maximal spatial information about the vocal tract. However, creating metrics from the images that are both phonetically interpretable and statistically testable is less than straightforward. Moreover, even with a suitable technology and analysis method for maximizing spatial information

in place, obtaining this information often comes at the cost of temporal information: Due to limitations on time, resources, and/or computational power, measurements are most commonly carried out at particular 'magic moments' of speech segments, e.g., the temporal midpoint of a vowel, which neglects the often complex dynamic nature of speech (Mücke, Grice, & Cho, 2014).

Almost simultaneously with recent progress in imaging techniques in articulatory research, there have been dramatic improvements in statistical methods which now increasingly allow for the quantification of continuous data while taking into account complex experimental designs with repeated measures on speakers and items. In particular, with the advent of generalized additive mixed models (GAMMs), we are now at a point at which multi-dimensional data can be subjected to statistical modeling. However, model design and the interpretation of the results can be a challenge. In this paper, we explore the application of GAMMs to vocal tract aperture functions over time, gleaned from rt-MRI video of German vowel productions. We thereby use two different versions of GAMMs in order to validate our results and to increase our understanding of how estimated effects obtained from GAMMs can be interpreted with regard to vocal tract dynamics. Three different test cases (in decreasing order of effect size) are investigated, and the GAMM results are independently validated using functional linear mixed models at different time points in the target vowel, in order to determine whether the two methods converge on similar outcomes.

1.1. Real-time magnetic resonance imaging (rt-MRI)

Recent developments in high-quality, high-speed rt-MRI reconstruction techniques (Fu et al., 2012, 2015; Niebergall et al., 2012; Uecker et al., 2010) have made rt-MRI a remarkably suitable tool for capturing data related to vocal tract kinematics in speech (see Lingala, Sutton, Miquel, & Nayak, 2016 for an overview of challenges associated with rt-MRI of speech). Rt-MRI is particularly appealing for studying relatively inaccessible articulatory characteristics, such as velum height (Byrd, Tobin, Bresch, & Narayanan, 2009; Carignan et al., 2019; Martins, Oliveira, Silva, & Teixeira, 2012; Proctor et al., 2013), pharyngeal aperture (Carignan, Shosted, Fu, Z.-P., & Sutton, 2015; Shosted, Sutton, & Benmamoun, 2012; Tiede, 1996), laryngeal position (Demolin, Hassid, Metens, & Soquet, 2002; Honda & Tiede, 1998), and even laryngeal configuration (Ahmad, Dargaud, Morin, & Cotton, 2009; Moisik, Esling, Crevier-Buchman, Amelot, & Halimi, 2015; Moisik, Esling, Crevier-Buchman, & Halimi, 2019). However, unlike articulometry data, rt-MRI video frames must first be quantified in some manner before analysis can be carried out. A variety of quantification methods has been proposed (see Ramanarayanan et al., 2018 for a detailed overview), including (but not limited to) region-of-interest analysis (Lammert, Ramanarayanan, Proctor, & Narayanan, 2013; Teixeira et al., 2012; Tilsen et al., 2016), grid-based area or distance functions (Barlaz, Shosted, Fu, & Sutton, 2018; Proctor, Bone, Katsamanis, & Narayanan, 2010; Zhang et al., 2016), image cross-correlation (Lammert, Proctor, & Narayanan, 2010), region-based principal components analysis (Carignan et al., 2019, 2015), and automated segmentation of individual speech articulators (Eryildirim, M.-O., & Berger, M.-O., 2011; Labrunie et al., 2018; Silva & Teixeira, 2015, 2016).

In the current study, we have chosen to quantify the vocal tract using semi-polar grid functions that represent the aperture (i.e., distance, diameter) of the vocal tract within the midsagittal plane. This particular quantification method was chosen for two reasons, both of which are important for maintaining interpretability in the specific use of GAMMs that we propose in this paper. First, both the number of grid lines and their relative location within the vocal tract remain constant across speakers, items, and conditions. Second, although applying a grid is essentially a manner of discretizing the vocal tract, when

using a relatively large number of grid lines (in our case, 28), the resulting function is a gradient, fine-grained spatial representation of the vocal tract that can be modeled as a continuous variable. In this way, we can subject the dynamic evolution of vocal tract aperture over time to statistical modeling.

1.2. Generalized additive mixed models and functional linear mixed models

In this paper, we will use two different approaches to generalized additive mixed modeling: via smooth-derived random effects (Baayen, Kuperman, & Bertram, 2010; Baayen, Rij, de Cat, & Wood, 2016) and via random effects derived from functional principal components analysis (Cederbaum, Pouplier, Hoole, & Grevens, 2016; Pouplier, Cederbaum, Hoole, Marin, & Greven, 2017).

A generalized additive model (GAM; Hastie & Tibshirani, 1990; Wood, 2006b) is a class of statistical models in which the relationships between the response and predictors are modeled by non-linear smooth functions. Generalized additive mixed models (GAMMs; Wood, 2004, 2006a) are an extension of GAMs as mixed models, in which random effects are estimated from a GAM by computing the variances of the so-called 'wiggly' components of the smooth terms (i.e., the degree of smoothness of the terms). GAMMs have previously been used to investigate speech production over time (Baayen, Vasishth, Kliegl, & Bates, 2017; Kirkham, Nance, Littlewood, Lightfoot, & Groarke, 2019; Mielke, Carignan, & Thomas, 2017; Sóskuthy, 2017; Wieling et al., 2016; Winter & Wieling, 2016) and space (Barlaz et al., 2018; Wieling, 2018), to observe the effects of word frequency and lexical proficiency on articulation (Tomaschek, Tucker, Fasiolo, & Baayen, 2018), and to model spatio-temporal relations in flesh-point kinematics (Tomaschek, Arnold, Bröker, & Baayen, 2018). One distinct advantage of employing GAMMs for speech articulation research is that they can capture the interaction effects of two different continuous variables (such as time and space), using tensor product interaction, which allows the smooth coefficients for one variable to vary in a non-linear fashion depending on the value of the other variable (Wieling, 2018, p. 102). In this way, GAMMs enable speech researchers to investigate how a given articulatory metric (e.g., EMA sensor height; vocal tract aperture) is conditioned by both a temporal dimension (e.g., time within a speech interval; experimental trial) and a spatial dimension (e.g., location of EMA sensor on the tongue; region of the vocal tract).

Functional linear mixed models (FLMMs) are an extension of standard linear mixed modeling, in which both the response and random effects are observed over multiple points in temporal or spatial location. One method of FLM modeling is sparseFLMM (Cederbaum et al., 2016; Pouplier et al., 2017), a non-parametric, spline-based estimation technique for the analysis of correlated functional data which are observed irregularly, or even sparsely. Means are estimated based on penalized splines, and random effects are captured using functional principal component analysis (FPCA), as opposed to deriving random effect structures from the smooth terms (Baayen et al., 2010, 2016). Due to penalized splines being employed, FLMMs imposes no underlying assumptions about the shape and properties of the basis functions apart from an underlying smoothness. The sparseFLMM approach is closely related to GAMMs as applied to two-dimensional data, e.g., time series data (Baayen et al., 2010, 2016; Scheipl, Staicu, & Greven, 2015; Wieling, Montemagni, Nerbonne, & Baayen, 2014).

As previously mentioned, it is possible with GAMMs to model random effects using the degree of factor smoothness. In the sparseFLMM approach, using FPCA bases as a parsimonious representation of the functional random effects provides an interpretable variance decomposition for the random terms in the model, permitting subsequent inspection of these random effects. Importantly, the FPCA basis functions are estimated

from the data as the eigenfunctions of the estimated covariance of the functional random effects (see also Wang, Chiou, & Müller, 2016). GAMMs assume that the error is autocorrelated with a specific parametric first-order autoregressive structure; thus, a fixed correlation parameter (ρ) must be set as a working criterion by the researcher, although it can be estimated directly from the autoregressive structure of the data (see Section 2.4). This may lead to incorrect standard errors and thus incorrect inference, since ρ is assumed to be constant over the dimension of interest. In contrast, sparseFLMM has the advantage of estimating the auto-covariance of the error directly from the data, which allows the error to be heteroscedastic and/or to vary non-parametrically over the dimension of interest, which gives more reliable inference in this respect compared to GAMMs (for discussion, see Pouplier et al., 2017).

In the current study, we thus propose the use of GAMMs for analyzing the conditioning effects of both time and space on vocal tract aperture measured from rt-MRI video. Exploiting the ability of GAMMs to model the effect of two continuous predictors on the response (as proposed by Baayen and colleagues; Baayen et al., 2010, 2016), the use of GAMMs that will be demonstrated here allows us to take advantage of both the temporal and spatial information provided by rt-MR imaging. However, in order to maintain an appropriate degree of circumspection toward the model results, we will employ FLMMs as a way of independently validating the GAMM results, in order to observe whether the two different methods yield similar interpretations of the data. FLMMs currently allow for interpretability of only one of these two factors at a time—e.g., differences in aperture over time at a single location in the vocal tract, or differences in aperture throughout the vocal tract at a single point in time. Therefore, in order to obtain a representative comparison to the GAMM results, two different FLMMs will be created to validate each GAM model: FLMMs created to investigate differences in aperture throughout the vocal tract will be constructed using data from two different time points (20% and 80% of a vowel interval), and the FLMM results will be compared directly to the GAMM results observed at the same two time points. In doing so, it is not our intention to compare GAMM and FLMM as statistical methods; rather, we use them in a complementary fashion in order to gain a comprehensive picture of how rt-MRI data can be analyzed reliably using these relatively new statistical methods.

2. Methodology

2.1. rt-MRI, speakers, stimuli, and segmentation

Rt-MRI data at 50 frames per second were collected using a 3T MRI system (Magnetom Prisma Fit, Siemens Healthineers, Erlangen, Germany) at the Max-Planck-Institute for Biophysical Chemistry (Göttingen, Germany) along with synchronized, noise-suppressed audio. The method relies on highly under-sampled radial gradient-echo acquisitions in combination with serial image reconstruction by regularized non-linear inversion (Uecker et al., 2010). Extending preliminary applications to characterize natural speech at slower speed (Niebergall et al., 2012), the current study employs a temporal resolution of 20 ms[1] (9 radial spokes, repetition time 2.22 ms, echo time 1.47 ms, flip angle 5°). Rt-MRI movies

[1] Methods of reconstructing real-time MRI video sequences at high temporal resolutions have often involved a sliding window technique, wherein shifted reconstructions subdivide a longer acquisition time per frame. In such cases the 'true' resolution is the acquisition time, while the image series has an artificially inflated frame rate dependant on the number of subdivisions of the acquisition. However, no such technique is employed in the reconstruction method used here, yielding a temporal resolution of the image series (50 fps) that is 'self-consistent' with the acquisition time per frame (20 ms) (Frahm et al., 2014; Iltis et al., 2015).

cover a 192 × 192 mm^2 field-of-view at 1.41 mm in-plane resolution in a mid-sagittal plane of 8 mm thickness.

Data are presented here for 36 native speakers of German (22 female), aged between 19 and 35 years (μ = 24.36, SD = 4.22). The corpus consists of ≈300 German lexical items, balanced for coda composition over a wide range of phonetic contexts (e.g., vowel quality, stops versus obstruents, etc.). During the MRI scanning sessions, the words appeared on a computer screen, as reflected on a mirror placed inside the scanner. The words appeared in a variety of carrier phrases constructed to vary the stress placement of the word in three primary conditions: accentuated, de-accentuated, and neutral. The noise-suppressed audio was used for segmentation of the vowel in each word, which was carried out manually in Praat (Boersma & Weenink, 2017) via inspection of the acoustic waveform and corresponding broadband spectrogram.

2.2. Generating vocal tract aperture functions from rt-MRI video

The MATLAB functions used to process the MR images and generate the vocal tract (VT) aperture values are available at: https://github.com/ChristopherCarignan/MRI-analyses. The specific methodological steps that we use to generate the VT aperture functions are not necessary for application of the GAMM method itself—any similar aperture function will do (see, e.g., Narayanan et al., 2014 and Raeesy, Rueda, Udupa, & Coleman, 2013 for alternative solutions)—nor are they central to the goal of our paper, which is to promote and illustrate the application of GAMMs to changing aperture functions over time. However, for the sake of clarity and methodological transparency, we outline in this section the specific processing methods used to create our aperture functions. We refer the reader to the documentation provided in the MATLAB functions for further details. Generation of VT aperture functions was carried out in several steps, each of which is described in the following sections.

2.2.1. Image registration

First, all images are registered in order to control for possible changes in the angle of the speaker's head within the scanner.[2] Image registration is performed by creating a region of interest (RoI) in the upper portion of the head—i.e., the pixel rows extending vertically from ≈ the tip of the nose. Since it can be assumed that any structures in this portion of the head will remain internally stable, any observed movement within the RoI is presumed to be due to movement of the head. Accordingly, each image is aligned to the first image of the recording by estimating rigid transformation matrices of RoI-masked images with the `imregtform` function and applying these geometric transformations to the original images with the `imwarp` function.

2.2.2. Semi-polar grid

After image registration, a semi-polar grid consisting of 28 lines is overlaid from the glottis to the anterior edge of the alveolar ridge (left-most image in **Figure 1**).[3] The choice of 28 grid lines, specifically, is somewhat arbitrary and was reached after manual trial and error, with the final number representing a balance between sufficient coverage throughout the vocal tract and the processing time required for computing the aperture

[2] Changes associated with head movement in our data were relatively rare, were small in magnitude, and usually corresponded to the production of accentuated words.

[3] Labial configuration was not considered for the purposes of this study, since protrusion/retraction of the lips would result in a difference in the number of grid lines that would need to be considered for different items and contexts. Accurate interpretation of the models presented in this study requires that the number of grid lines and their location in the vocal tract be constant.

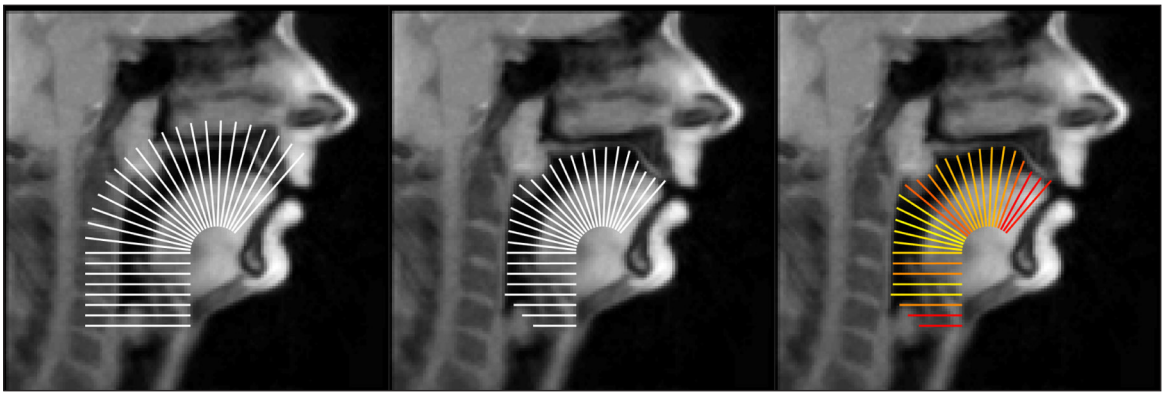

Figure 1: Vocal tract grid line placement (left image), posterior/superior boundary detection (center image), and aperture estimation (right image) from large aperture in yellow to small aperture in red.

values in each MR image.[4] The semi-polar grid is applied to the vocal tract semi-manually, in the following manner. The user selects the location of the glottis, the velopharyngeal port, the anterior edge of the alveolar ridge, and a location of air within the oral cavity. The midpoint of a line extending from the glottis line to the alveolar ridge line—a point located in the *genioglossus* muscle in every case—is then used as the origin of a polar grid of 21 radii extending from the pharynx (parallel to the *genioglossus* origin) to the alveolar ridge. The remaining 7 lines are then set equidistantly from the end of the polar grid to the glottis (left image in **Figure 1**). The grid line closest to the speaker's velopharyngeal port is logged for subsequent spatial normalization across speakers (see Section 2.3). The grid line closest to the air selection is used to compute a threshold of pixel intensity representing the difference in image luminosity between flesh and air; the threshold is defined as 25% of the range of pixel intensities along this line and is used in the estimation of vocal tract aperture along each grid line (see Section 2.2.4).

2.2.3. Air-tissue boundary detection

For each MR image in a speaker's recording, the posterior and/or superior boundary of the vocal tract along each grid line is located semi-automatically and the grid lines are truncated/terminated at this boundary (center image in **Figure 1**). The air-tissue boundary along each grid line is determined based on a weighting of three factors; two representing degree of pixel intensity change and one representing prior assumptions. For each grid line, the first derivative of the pixels falling along the line extending from its central position outward (e.g., from *genioglossus* through the pharynx) is first calculated. The maximum of this differential signal is defined as the most rapid change from low intensity pixels (air) to high intensity pixels (flesh); given the direction/orientation of the line (i.e., extending outward from an anterior and/or inferior location within the vocal tract), the point along the grid line corresponding to this maximum is interpreted as an estimate of the location of the posterior and/or superior air-tissue boundary of the vocal tract (i.e., the point at which air meets flesh along the posterior/superior edge of the vocal tract). Two values associated with this differential peak are logged to be used in the weight calculation: the value of the peak (i.e., the magnitude of change) and the prominence of the peak (i.e., the relative magnitude of the peak in relation to neighboring peaks). In order to generate the prior assumption, the user manually selects

[4] Consideration of the time required to analyze images was a non-negligible factor in our analysis choices, since our data set contains an average of 23,804 MRI video frames per speaker (SD = 2,967), totalling 856,944 images and 24 million grid lines to analyze.

the posterior/superior edge of the vocal tract for each grid line in a representative frame (the 'base assumption').

Finally, for each frame in the MRI video, the air-tissue boundary along each grid line extending from the glottis to the hard palate[5] is calculated automatically by selecting the appropriate peak in the first derivative of the intensity values, as described above, using a weighting that penalizes for small peak magnitude, small peak prominence, and large distance from the base assumption. The resulting 28-point boundary is then smoothed using a Savitzky-Golay second-order polynomial convolutional filter in order to reduce possible errors in the automatic peak selection, under the assumption that the air-tissue boundary is relatively contiguous throughout the vocal tract—i.e., the vocal tract does not contain structures that would introduce an abrupt change in the spatial location of this boundary within the midsagittal plane.

2.2.4. Aperture estimation

After the boundaries are located in each MR image, the aperture of the vocal tract within the boundary-terminated grid lines is estimated using a thresholding technique. The number of pixels along each grid line that have an intensity value below the pre-defined air/flesh boundary threshold (see Section 2.2) is calculated and multiplied by the in-plane voxel resolution (1.41 mm). The result is a 28-point function corresponding to the midsagittal aperture (in mm) of the vocal tract from the glottis to the end of the alveolar process. An illustration of this VT aperture function is shown in the right image in **Figure 1**, in which aperture is denoted by range of color from yellow (large aperture, i.e., VT expansion) to red (small aperture, i.e., VT constriction). Starting from the glottis, we can observe: small aperture at the larynx, followed by increased aperture in the hypo-pharynx just above the larynx, followed by slightly decreased aperture at the epiglottis, followed by expansion at the tongue root in the hyper-pharynx, followed by decreased aperture between the velum and tongue dorsum, followed by intermediate aperture (i.e., orange lines) along both the soft and hard palate, followed finally by very small aperture corresponding to an alveolar constriction.

2.3. Normalization procedures

The current study uses GAMMs to observe how a variety of factors might condition changes in aperture throughout the vocal tract over the time course of the vowel. However, before submitting the data to the GAMMs, both of these dimensions (time and space, i.e., grid line location within the VT) were normalized. Time was normalized in a linear fashion for each token (scale: 0–1). Non-linear spatial normalization was applied to VT grid line locations using spline-based landmark registration, in the following manner. First, the grid line corresponding to the location of the velopharyngeal port (henceforth, velum) was located for each speaker (see Section 2.2.2). Second, for the purposes of interpretability, six major locations in the vocal tract were chosen and considered as equidistant along a 0–1 scale: glottis (0), hypo-pharynx (0.2), hyper-pharynx (0.4), velum (0.6), palate (0.8), and alveolar ridge (1). Finally, the 28 grid line scale was transformed in a non-linear fashion for each speaker by fitting a spline between the sets of coordinates [1, x, 28] (where x = the grid line at the velum) and [1, 16.8, 28] (i.e., 28 * 0.6 = 16.8).[6] The

[5] Since the images are registered to account for head movement, the hard palate can be assumed to remain stable throughout the recording. Thus, the manual boundary selections for the grid lines along the hard palate are held constant throughout the entire recording (rather than estimated algorithmically) to reduce processing time; only the boundaries for grid lines extending from the glottis to the hard palate are allowed to vary, in order to capture differences in vocal tract shape due to, e.g., velum lowering or constriction/expansion of the posterior pharyngeal wall (Carignan et al., 2015).

[6] Since only three coordinate points were used, the spline fit was essentially a second-order polynomial fit in each case.

range of integers 1:28 was then transformed using the coefficients of the fitted spline, resulting in a grid line scale in which 1 = glottis, 16.8 = velum, and 28 = alveolar ridge for each speaker, with speaker-specific non-linear transformations between these three points.[7] Examples of VT aperture functions before and after both the largest and the smallest degrees of this non-linear normalization are shown in **Figure 2**, for two speakers' productions of /a:/ in *bat* "(I) asked."

It is important at this point in time to discuss two caveats associated with the normalization procedures performed on the data presented here. First, as is the case with any similar time normalization, the 0–1 linear temporal scaling performed in this study neglects inherent differences in duration between speech segments. Thus, when investigating reported significant effects in the temporal dimension, the researcher must of course exercise appropriate caution in interpreting the precise nature of these effects. Second, the six equidistant vocal tract locations are not necessarily equidistant in reality. While we believe that they are reasonable approximations—the grid layout displayed in **Figure 1** is a fair representation of the general grid layout across speakers, and these vocal tract locations are each ≈5–6 grid lines apart from one another in the figure—equidistance has been imposed upon these locations for ease of interpretation. We do not claim that they are necessarily accurate estimations of *absolute distance* between regions of the vocal tract (and, thus, they should not be interpreted as such by the reader). Rather, they should be considered as relative landmarks and interpretations of significant effects should be made accordingly.

2.4. GAMM construction

The data and R (R Core Team, 2017) code used to construct the GAMMs and FLMMs, as well as generate the results and figures presented in this paper, are available at: https://github.com/ChristopherCarignan/journal-articles/tree/master/LabPhon_rtMRI-GAMM.

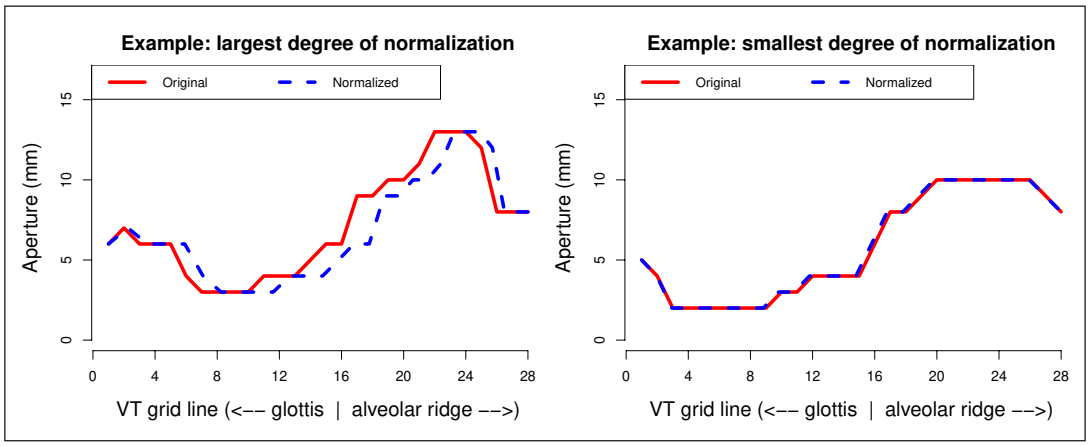

Figure 2: Vocal tract aperture function examples for productions of /a:/, both with (blue, dashed line) and without (red, solid line) normalization. The left plot is a VT aperture function for a speaker who required the largest degree of normalization (velum at grid line 15); the right plot is a VT aperture function for a speaker who required the smallest degree of normalization (velum at grid line 17).

[7] Each of the models presented in this study was also tested without landmark-based transformation of the 28 grid lines. Differences in the results between the two methods were negligible in each case, since the velum grid line was never very different from 16.8, ranging from 15 to 17 for all speakers (μ = 15.81, SD = 0.67). The largest transformation (i.e., velum at grid line 15) yielded a quadratic coefficient of merely –0.0099, indicating only minor deviation from a straight line.

GAMMs were constructed using the `bam` function of the *mgcv* package (Wood, 2019). Autocorrelation in the VT aperture functions is expected for the dimensions of both time (speech articulators move continuously over time) and space (the vocal tract is not composed of discrete structures), resulting in autocorrelation of the model residuals and therefore violating the model assumption of independent errors. The `bam` function includes an autoregression (AR) feature intended to reduce autocorrelation in one dimension, employing a user-supplied ρ parameter. It is suggested that ρ should correspond to the autocorrelation function (ACF) value at lag = 1, i.e., ACF[2], an approach that we have followed here. For the current study, the vocal tract aperture grid lines were chosen as the dimension in which autocorrelation was reduced using this AR feature, using the first VT grid line (i.e., the glottis) of each token as the 'AR.start' commencement point. However, when the data are ordered by time and sub-ordered by grid line at each time point (as is the case for our data), the AR.start parameter set to track the onset of the grid lines at each time point effectively captures autocorrelation of the residuals in *both* dimensions.[8] Models were constructed using the `te`-constructor to fit the non-linear interaction between the temporal and spatial dimensions. Full random effects were included for speaker and random intercepts were included for word. The number of knots (`k` parameter) for all smooths was chosen via model diagnosis using the `gam.check()` function; the marginal basis dimensions in the FLMMs was equal to the number of knots in the random smooths of the GAMMs (i.e., 4), in order to maintain similar model construction. The R code for the full GAMM construction is shown in **Listing 1**.

Listing 1: Structure of generalized additive mixed models used in the current study

```
# Main effect of the predicting factor on VT aperture:
bam(aperture ~ factor
# Tensor product interaction to separate the effect of the predictor
# from the effects of normalized time and space(i.e., VT grid line):
+ te(time, gridline, k=15)
+ te(time, gridline, by=factor, k=15)
# Random smooths(i.e., full random effect) to account for
# non-linear interactions between speaker and time|space:
+ s(time, speaker, bs="fs", m=1, xt="tp", k=4)
+ s(gridline, speaker, bs="fs", m=1, xt="tp", k=4)
# Random intercepts by word:
+ s(word, bs="re", m=1),
# AR.start to control for correlation throughout the vocal tract.
# Since the data are ordered by time(and sub-ordered by gridline),
# AR.start captures autocorrelation of residuals in both dimensions:
AR.start=gridstart, rho=valRho, method="fREML", data=mridata)
```

[8] We would like to thank an anonymous reviewer for bringing this feature to our attention.

This model structure was used to investigate possible articulatory differences in three distinct phonetic contexts, each with different expectations for their effect on vocal tract shaping over time:

1. Difference between monophthongs /aː/ and /iː/: We expect maximal spatial differences (i.e., these vowels place maximally distinct constraints on tongue shape) but minimal temporal differences (i.e., they are both monophthong vowels).
2. Differences between monophthong /aː/ and diphthong /aɪ/: We expect spatial differences and temporal differences to occur concomitantly, since the tongue shapes are expected to be similar earlier in the vowel interval (i.e., [a]–[a]) and to diverge later in the vowel interval (i.e., [a]–[ɪ]).
3. Differences between accentuated and neutrally stressed /aː/: We expect stress to manifest in articulatory differences, but we do not necessarily have any a priori assumptions as to what those differences may be.

In order to validate the GAMM results for each context, separate FLMM models were created at 20% and 80% of the vowel interval using the `sparseFLMM` function of the *sparseFLMM* package (Cederbaum, 2017). These results will be compared to the GAMM results at the same time points to determine whether the two methods converge on similar interpretations of the data.

3. Results

3.1. Monophthongs: Differences between /aː/ and /iː/

In order to investigate differences between the monophthongs /aː/ and /iː/, the following subset of the corpus was created from neutrally stressed (i.e., unaccentuated) lexical items that include these vowels and are preceded by one of /b, d, t, ʀ/ and followed by /t/: *bat* /baːt/, *Rate* /ʀaːtə/, *Tat* /taːt/, *biete* /biːtə/, *Rita* /ʀiːta/, *Dieter* /diːtɐ/. This subset yielded a total of 216 observations and 46,844 data points (29,932 for /aː/, 16,912 for /iː/) across the 36 speakers. GAMM heatmaps of vocal tract aperture for /aː/ and /iː/ are shown in **Figure 3**; these heatmaps were created using the `fvisgam` function of the *itsadug* R package (van Rij, Wieling, Baayen, & van Rijn, 2017). In these heatmaps, small aperture (i.e., VT constriction) is denoted by the red end of the color scale, and

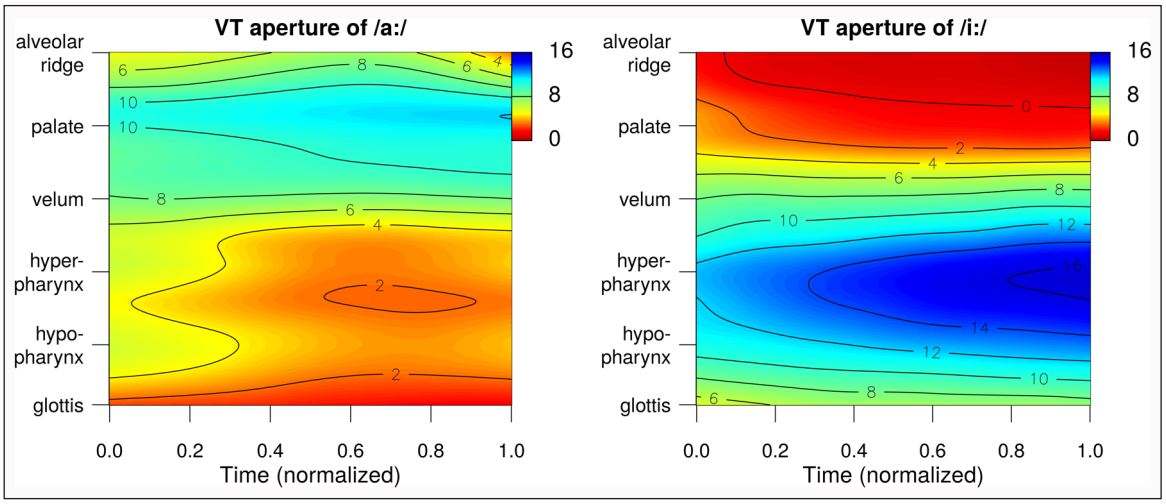

Figure 3: GAMM heatmaps of vocal tract aperture (z-axis) over time (x-axis) throughout the vocal tract (y-axis), for neutrally stressed /aː/ (left) and /iː/ (right). Aperture (mm) is denoted by color grade, and regions of equidistant change (here, Δ2 mm) are denoted by black lines.

large aperture (i.e., VT expansion) is denoted by the blue end of the color scale. The VT shapes implied by these heatmaps are congruent with our general knowledge of these two vowels: /aː/ is produced with expansion along the palate and constriction throughout the pharynx, suggesting a lowered and retracted tongue posture, while /iː/ is produced with constriction along the palate and expansion throughout the pharynx, suggesting a raised and advanced tongue posture. These articulatory configurations are maximally distinct, as predicted. Furthermore, the distinction is enhanced after ~ 50% of the vowel interval: In this latter portion of the vowel, /aː/ becomes even more [aː]-like (greater palatal expansion and pharyngeal constriction), while /iː/ becomes even more [iː]-like (greater palatal constriction and pharyngeal expansion).

While these descriptive heatmaps are useful for understanding the dynamic articulations of these two vowels, it is perhaps more informative (and more advantageous for a phonetic interpretation) to observe the effect of the context itself (here, vowel quality) on VT aperture over time and space. **Figure 4** displays the differences between /aː/ and /iː/, in other words the differences between the two heatmaps shown in **Figure 3**. This difference heatmap of **Figure 4**, as well as all similar difference heatmaps in the current study, were created using the `plot_diff2` function of the *itsadug* R package. In this figure, the differences shown are for /aː/ in comparison to /iː/. Differences that are considered to be significant at $\alpha = 0.05$ are shown in colored areas (i.e., shades of red or blue), while non-significant differences are shown in opaque/dark areas. Contours of equidistant change are denoted by black solid lines (similar to topographic maps), and confidence interval (CI) bands for each contour are denoted by red dotted lines (lower CI) and green dotted lines (upper CI). For example, the red portion that extends from the bottom of the map up to the red CI for the –2 mm contour is considered a region of significant difference, the

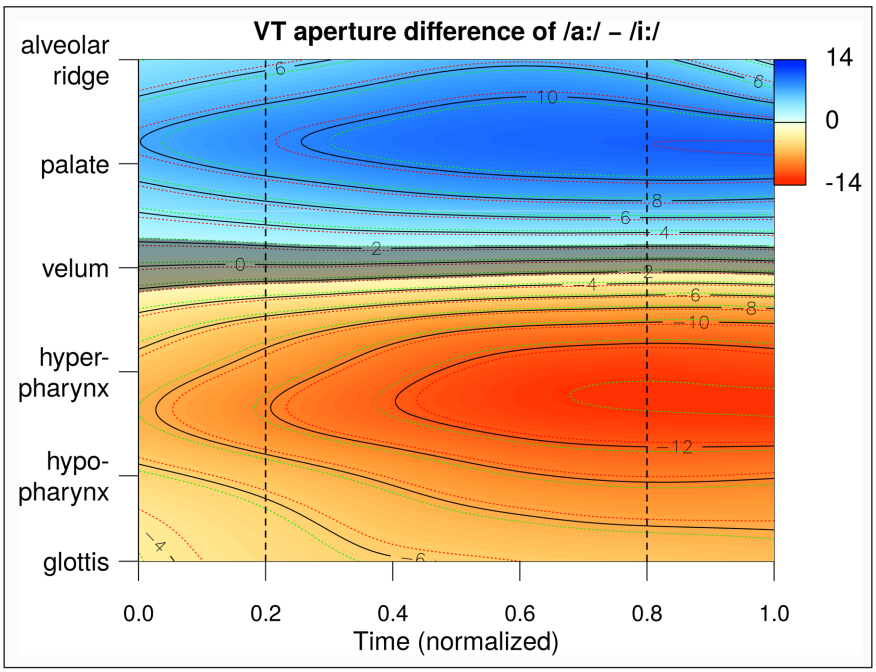

Figure 4: GAMM heatmap of vocal tract aperture (z-axis) over time (x-axis) throughout the vocal tract (y-axis), for the difference between neutrally stressed /aː/ and /iː/. Aperture difference (mm) is denoted by color grade, regions of equidistant change (here, Δ2 mm) are denoted by black solid lines, and 95% confidence interval bands are denoted by red and green dotted lines. Significant differences ($\alpha = 0.05$) are denoted by colored areas (significant) versus opaque areas (non-significant). Vertical dashed lines denote 20% and 80% of the vowel interval, for comparison with FLMMs at the same time points (Figure 6).

opaque region extending from the red CI for the –2 mm contour to the green CI for the 2 mm contour is considered a region of non-significant difference, and the blue portion that extends from the green CI for the 2 mm contour to the top of the map is considered a region of significant difference.

The difference heatmap suggests that, in comparison to /i:/, /a:/ is produced with greater constriction from the glottis to the velum and greater expansion from the velum through the alveolar ridge; the aperture is similar for both vowels at the velum, which is consistent with evidence that the 'pivot' point of lingual variation is around the uvular region (Iskarous, 2005). Moreover, the regions of greatest difference between the two vowels are in the hyper-pharynx (most saturated reds in the heatmap) and in the middle/anterior portion of the palate (most saturated blues in the heatmap). Finally, the differences between the two vowels become more exaggerated in the second half of the vowel interval.

A global summary of the model can be obtained using the standard `summary()` function, as shown in **Table 1**. Additionally, the adjusted R^2 of the model (not shown in **Table 1** but included in the summary output) reveals that 74.5% of the total variance is explained by the model. The model summary provides separate statistics for the parametric coefficients (i.e., linear effects) and the smooth terms (i.e., non-linear effects).[9] In this model, /a:/ was chosen as the reference level; thus, the model intercept shows that the average aperture of /a:/ (throughout the entirety of both the vowel duration and the vocal tract) is 6.11 mm. By comparison, the vowel /i:/ is produced with relatively larger average aperture (1.32 mm more, rendering an estimate of 7.43 mm), although the difference is not significant ($p = 0.25$). With regard to the smooth terms, it is clear that the inclusion of each of the non-linear fixed effects and non-linear random effects is necessary for the model. Thanks to the visualizations provided in **Figures 3** and **4** and the corresponding effects already discussed above, the interpretations of these smooth terms are relatively straightforward: The vowel articulations become enhanced throughout the vowel duration (smooth term 1), with /a:/ becoming more [a:]-like and /i:/ becoming more [i:]-like over time (smooth term 2), with inter-speaker differences over time (smooth term 3) and throughout the vocal tract (smooth term 4), as well as overall differences across words (smooth term 5).

The random smooths can be visualized by selecting the number of the smooth term in the regular `plot()` function, e.g., `plot(m1.rho,select=4)` for the by-speaker vocal tract smooth. Figure 5 displays the by-speaker random smooths over time (left

Table 1: Summary of the GAMM created to compare neutrally stressed /a:/ and /i:/.

A. parametric coefficients	**Estimate**	***SE***	**t-value**	**p-value**
(Intercept)	6.1060	0.8908	6.8544	<0.0001
vowel-i:	1.3240	1.1446	1.1568	0.2474
B. smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
te(time.norm,gridline.norm)	69.5238	82.1682	97.2703	<0.0001
te(time.norm,gridline.norm):vowel-i:	58.1370	68.1674	424.6944	<0.0001
s(time.norm,speaker)	17.3271	143.0000	0.2401	<0.0001
s(gridline.norm,speaker)	119.5418	143.0000	26.7847	<0.0001
s(word)	3.9914	4.0000	367.2035	<0.0001

[9] In the results for the smooth terms, 'Ref.df' is the number of degrees of freedom used in hypothesis testing, while 'edf' is an estimate of the number of parameters required to create the smooth (Wieling, 2018, p. 90).

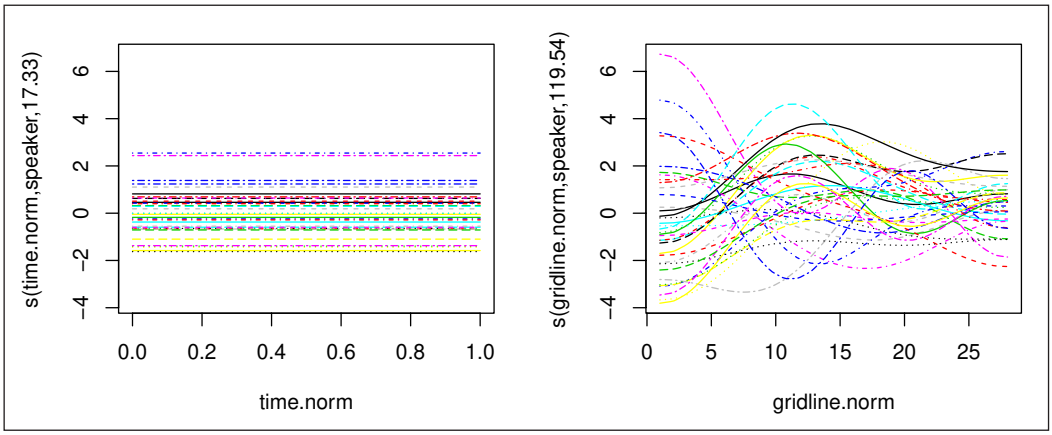

Figure 5: Visualization of by-speaker factor smooths over time (left plot) and space (right plot) in the GAMM created to compare neutrally stressed /a:/ and /i:/.

plot) and throughout the vocal tract (right plot). With respect to the by-speaker random smooth over time, although the smooth term is revealed as significant, its contribution to the model is minimal (evidenced by the small F-value associated with the smooth term); because of this, the by-speaker curves have little to no 'wiggliness.' In contrast, the by-speaker random smooth over space (i.e., throughout the vocal tract) has a much larger contribution to the model, reflected by both the F-value and the difference in curve shapes across speakers (especially around the glottis and in the lower pharyngeal region, i.e., grid lines 1–5). These differences could reflect inter-speaker variation with regard to articulation, physical morphology, changing larynx height, and/or the accuracy of the automatic aperture estimations.

As mentioned in Section 2.4, we now compare the GAMM results to FLMMs created for 20% and 80% of the vowel interval. These static time points are displayed as vertical dashed lines in **Figure 4** (and all similar figures throughout the manuscript). In order for the two methods to converge on similar results, we expect the FLMMs at both of these time points to show evidence of greater pharyngeal constriction for /a:/ compared to /i:/, but greater palatal constriction for /i:/ compared to /a:/, with similar aperture for both vowels at/around the velum. Additionally, we expect the FLMM created with data from 80% of the vowel interval to show greater articulatory differences between the two vowels, in comparison to the FLMM created with data from 20% of the vowel interval.

The results for the FLMMs created to test for differences between /a:/ and /i:/ at these two time points are shown in **Figure 6**. The fitted model for /a:/ is displayed in the red solid line and the fitted model for /i:/ is displayed in the purple dash-dot line; 95% confidence intervals are denoted by ribbons surrounding the fitted means. At the 20% time point, /a:/ is produced with constriction at the glottis, followed by slight expansion in the hypo-pharynx and slight constriction in the hyper-pharynx, followed by increasing expansion up to the palate, followed by increasing constriction up to the alveolar ridge. This aperture profile is in consonance with the profile shown in the aperture heatmap for /a:/ at 20% of the vowel interval, shown above in **Figure 3** (left plot). By contrast, /i:/ is produced with expansion throughout the pharynx with the maximum aperture located between the hypo- and hyper-pharynx, followed by a relatively linear decrease in aperture from this maximal pharyngeal expansion up to the alveolar ridge. This aperture profile is also in consonance with the profile shown in the aperture heatmap for /i:/ at 20% of the vowel interval (**Figure 3**; right plot). With regard to differences between the two vowels: /a:/ displays a smaller aperture than /i:/ throughout the entire pharynx, i.e., from the glottis to the velum, where the aperture for the two vowels converges on ≈7 mm

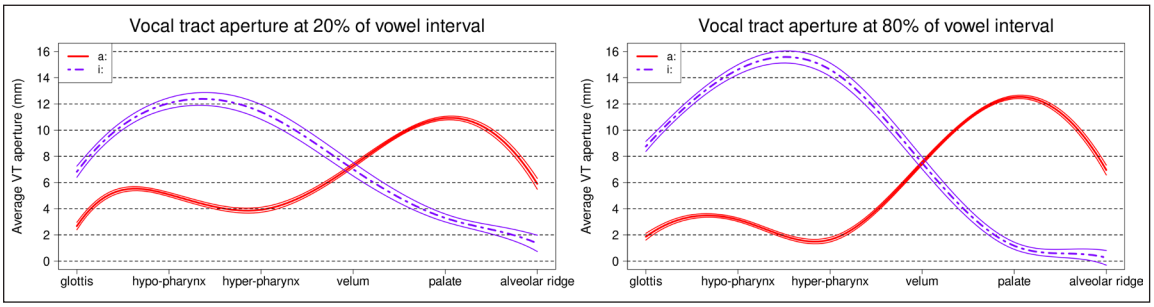

Figure 6: Results for the FLMM created to compare vocal tract aperture (y-axis) throughout the vocal tract (x-axis) between /aː/ (red, solid) and /iː/ (purple, dash-dot) at 20% and 80% of the vowel interval, with 95% confidence interval bands shown for both groups.

(averaged over all speakers). Anterior to the velum, /aː/ displays a larger aperture than /iː/ up to and including the alveolar ridge. The regions of greatest difference between the two vowels are the hyper-pharynx and the palate. Thus, the FLMM results are in agreement with the GAMM results at 20% of the vowel interval, both with regard to the articulatory configurations of the two vowels and with regard to the articulatory differences and similarities between them.

At the 80% time point, the VT aperture profiles for both vowels are similar to those observed at the 20% time point, only more exaggerated: Areas of constriction are more constricted and areas of expansion are more expanded. These results are in agreement with the respective GAMM aperture profiles of these two vowels shown above in **Figure 3**. Given the differences in lingual shape between the two vowels, these exaggerated aperture profiles result in even greater differences between /aː/ and /iː/ at 80% of the vowel interval in comparison to 20% of the vowel interval, with the areas of greatest difference located at the hyper-pharynx and the palate. Thus, the FLMM results are in agreement with the GAMM results at 80% of the vowel interval, both with regard to the articulatory configurations of the two vowels and with regard to the articulatory differences and similarities between them.

3.2. Diphthongs: Differences between /aː/ and /aɪ/

Although the similarities between the GAMM and FLMM results for /aː/ and /iː/ are encouraging, it is expected that these two vowels should yield large effects due to the distinct (and opposing) articulatory constraints that the vowels place on tongue posture. Thus, it is not surprising that both models converge on similar results, since we should expect such large effects to be detected and revealed by both models. Therefore, perhaps a more demanding test is to compare the model results for differences between /aː/ and /aɪ/, for which we expect similar aperture profiles at the beginning of the vowel but distinct aperture profiles at the end of the vowel.

In order to investigate differences between the monophthong /aː/ and the diphthong /aɪ/, the following subset of the corpus was created from neutrally stressed (i.e., unaccentuated) lexical items that include these vowels preceded by a labial /b, v/ and followed by an alveolar /n, t/: *bahne* /baːnə/, *bat* /baːt/, *wate* /vaːtə/, *weine* /vaɪnə/, *weinte* /vaɪntə/, *weihte* /vaɪtə/. This subset yielded a total of 216 observations and 54,628 data points (28,000 for /aː/, 26,628 for /aɪ/) across the 36 speakers. **Figure 7** displays the results for the GAMM created to test for differences between /aː/ and /aɪ/. In this figure, the differences shown are for /aː/ in comparison to /aɪ/. The heatmap suggests that, overall, /aː/ is produced with more constriction throughout the pharynx and more expansion throughout the palate, in comparison to /aɪ/. These articulatory distinctions are evidence of a more retracted and lowered tongue position for /aː/ versus /aɪ/, similar

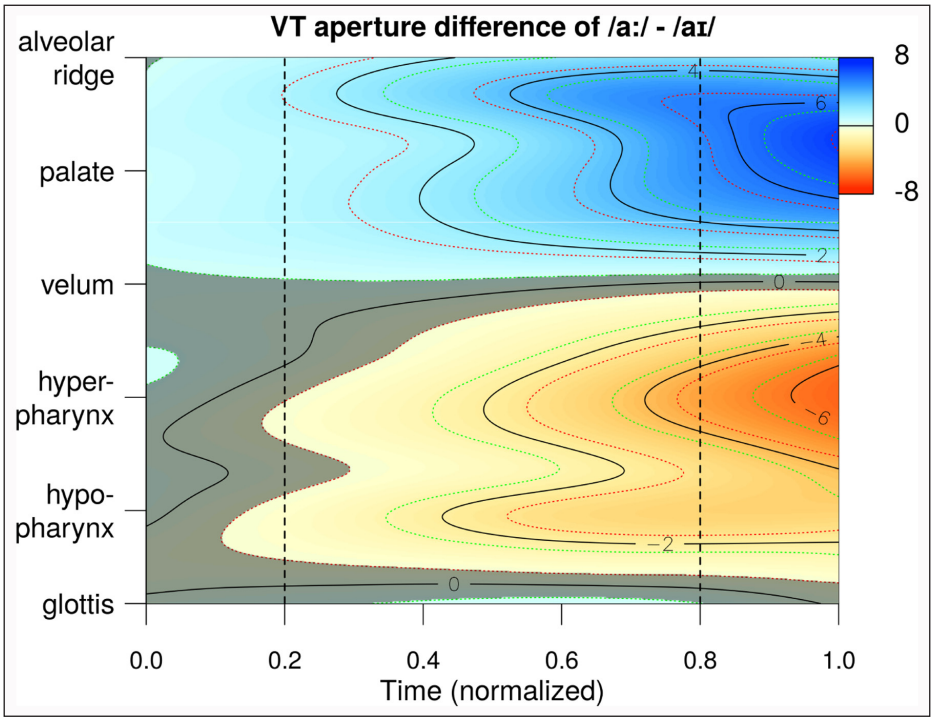

Figure 7: GAMM heatmap of vocal tract aperture (z-axis) over time (x-axis) throughout the vocal tract (y-axis), for the difference between neutrally stressed /a:/ and /aɪ/. Aperture difference (mm) is denoted by color grade, regions of equidistant change (here, Δ2 mm) are denoted by black solid lines, and 95% confidence interval bands are denoted by red and green dotted lines. Significant differences ($\alpha = 0.05$) are denoted by colored areas (significant) versus opaque areas (non-significant). Vertical dashed lines denote 20% and 80% of the vowel interval, for comparison with FLMMs at the same time points (Figure 9).

to those observed for /a:/ versus /i:/ (**Figure 4**), but the magnitude of the distinction is smaller for /a:/–/aɪ/ (absolute difference up to ≈ 7 mm) than for /a:/–/i:/ (absolute difference up to ≈ 13 mm). Although these differences are evidenced from the beginning of the vowel interval, there is a clear dynamic component to the pattern observed in **Figure 7**. The articulatory distinction is relatively minor early in the vowel interval and gradually becomes stronger as time unfolds, as the tongue moves toward the [ɪ] target at the end of the /aɪ/ interval—i.e., the differences increase in a relatively constant manner over time, reaching the most saturated blues and reds at the vowel offset.

The GAMM summary is provided in **Table 2**. 78.6% of the total variance is explained by the model. With regard to the parametric coefficients, the results reveal that /aɪ/ is produced with slightly smaller overall vocal tract aperture ($\mu \approx 6.09$ mm) compared to /a:/ ($\mu \approx 6.54$ mm). With regard to the smooth terms, the results are as expected: The vocal tract aperture changes over time (smooth term 1), but in different ways for the two vowels (smooth term 2), with by-speaker differences over time (smooth term 3) and space (smooth term 4), as well as overall differences across words (smooth term 5). Inspection of the random smooths (**Figure 8**) reveals a greater degree of inter-speaker variation compared to the /a:/–/i:/ GAMM (also evidenced by the larger F-values associated with the by-speaker random smooths in the /a:/–/aɪ/ model compared to the /a:/–/i:/ model). In particular, differences in by-speaker curve shape over time can now be observed (left plot). Interestingly, there is more inter-speaker variation in curve shape for the first half of the vowel interval compared to the second half, suggesting that the articulatory target for the [a] element of /aɪ/ is less consistent/precise across speakers in comparison to the articulatory target for the [ɪ] element.

Table 2: Summary of the GAMM created to compare neutrally stressed /aː/ and /aɪ/.

A. parametric coefficients	**Estimate**	***SE***	**t-value**	**p-value**
(Intercept)	6.5358	0.3690	17.7108	<0.0001
vowel-aI	−0.4451	0.1844	−2.4138	0.0158
B. smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
te(time.norm,gridline.norm)	88.2324	105.2958	74.8627	<0.0001
te(time.norm,gridline.norm):vowel-aI	69.3682	83.9685	73.1307	<0.0001
s(time.norm,speaker)	61.5039	143.0000	1.1402	<0.0001
s(gridline.norm,speaker)	120.6538	143.0000	44.1088	<0.0001
s(word)	3.8240	4.0000	21.3529	<0.0001

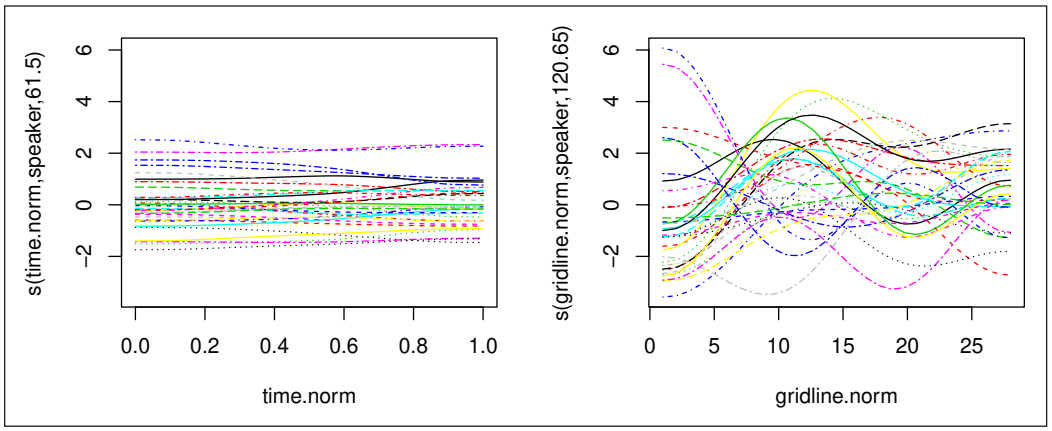

Figure 8: Visualization of by-speaker factor smooths over time (left plot) and space (right plot) in the GAMM created to compare neutrally stressed /aː/ and /aɪ/.

The GAMM results for /aː/–/aɪ/ differences lead to very different predictions for the FLMMs created at 20% and 80% of the vowel interval. At 20%, the GAMM suggests that there is little to no pharyngeal difference between the two vowels—the heatmaps suggest only a marginally significant difference between /aː/ versus /aɪ/ at two locations in the pharynx—and slightly greater expansion in /aː/ from the velum through the alveolar ridge (≈1 mm difference). At 80%, we expect relatively larger differences between the two vowels (i.e., more pharyngeal constriction and palatal expansion for /aː/ versus /aɪ/), but similar VT apertures at the glottis and at the velum, where the GAMM suggests no significant differences between /aː/ and /aɪ/. The results for the FLMMs created to test for differences between /aɪ/ and /aː/ at these two time points are shown in **Figure 9**. At the 20% time point, both vowels are produced with aperture profiles similar to the description provided above for /aː/ in **Figure 6**. The overlapping confidence intervals from the glottis to midway between the hyper-pharynx and the velum reveal no difference in aperture between the two vowels throughout the pharynx, suggesting that the FLMM is perhaps slightly more conservative in this region than the GAMM, which resulted in marginally significant differences in the pharynx at this time point. However, the FLMM confidence intervals diverge posterior to the velum, while the GAMM suggests that the difference between the two vowels is only significant anterior to the velum, suggesting that the GAMM is perhaps slightly more conservative in this region than the FLMM. Nevertheless, both models suggest significant differences in aperture between the two vowels from the velum through the alveolar process, suggesting that /aɪ/ is produced anterior to the velum with slightly more constriction than /aː/ (≈1 mm in both models).

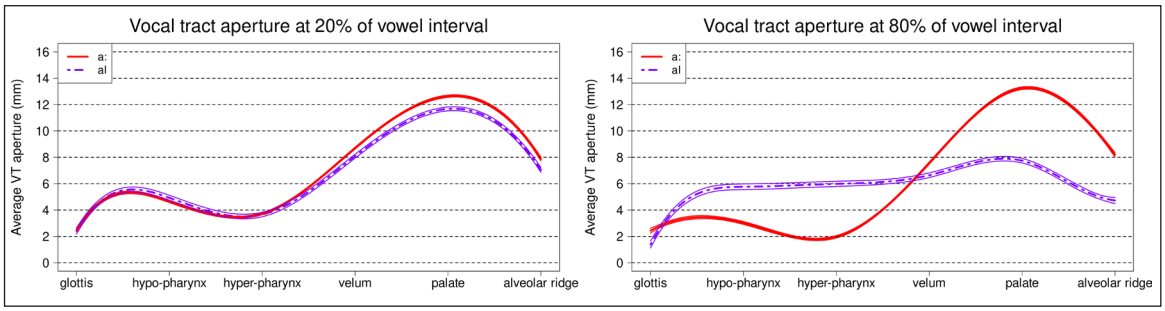

Figure 9: Results for the FLMM created to compare vocal tract aperture (y-axis) throughout the vocal tract (x-axis) between /aː/ (red, solid) and /aɪ/ (purple, dash-dot) at 20% of the vowel interval, with 95% confidence interval bands shown for both groups.

Thus, the FLMM results are generally in agreement with the GAMM results at 20% of the vowel interval, both with regard to areas of significant differences in the vocal tract and with regard to areas of similarity.

At the 80% time point, the shape of the VT aperture profile for /aː/ is similar to the profile seen at 20% of the vowel interval, albeit with slightly more exaggerated characteristics similar to the differences observed for /aː/ in Section 3.1 (**Figure 6**). However, the aperture profile for /aɪ/ at 80% of the vowel interval is substantially different than at 20%: There is more expansion along the pharynx, but less expansion along the palate. The relatively flattened VT aperture profile suggests that the [ɪ] element of the diphthong is somewhat centralized, produced with a fairly similar degree of vocal tract aperture from the pharynx through the palate, ranging from around 6 to 8 mm. This articulatory change in /aɪ/ from 20% to 80% of the vowel interval results in significant differences between /aː/ and /aɪ/ late in the vowel, in which /aɪ/ has less constriction at the pharynx but greater constriction at the palate, in comparison with /aː/. Additionally, there are no significant differences between the two vowels at the glottis and at/around the velum. Thus, the FLMM results are in agreement with the GAMM results at 80% of the vowel interval, both with regard to areas of significant differences in the vocal tract and with regard to areas of similarity.

3.3. Stress: Differences between accentuated and neutral vowels

In Section 3.1 we observed how GAMMs can capture large differences in vocal tract shape; in Section 3.2 we observed how GAMMs can capture both similarities and differences in dynamic vocal tract shaping over time. However, in both of these cases, the expectations were clear with regard to what we should expect the results to look like, and the results indeed confirmed the expectations. In the final test case for using GAMMs to analyze real-time MRI video of speech, we will investigate the possible effect of stress/accentuation on VT aperture of /aː/. As with the previous two test cases, we expect to observe an effect of the condition. However, unlike the previous two cases, we do not necessarily have any clear expectations for what that effect might be.

In order to investigate differences between accentuated and neutrally stressed (i.e., unaccentuated) /aː/, the following subset of the corpus was created from lexical items that include /aː/ followed by an alveolar consonant: *ahnde* /ʔaːndə/, *ahnte* /ʔaːntə/, *sahnst* /zaːnst/, *sahnt* /zaːnt/, *sahst* /zaːst/. This subset yielded a total of 377 observations and 120,568 data points (accentuated: 69,580, neutral: 50,988) across the 36 speakers. **Figure 10** displays the results for the GAMM created to test for differences between accentuated ("A") and neutral ("N") productions of /aː/. In this figure, the differences shown are for accentuated /aː/ in comparison to neutral /aː/. The heatmap suggests that stress does indeed condition differences in VT aperture, but that the

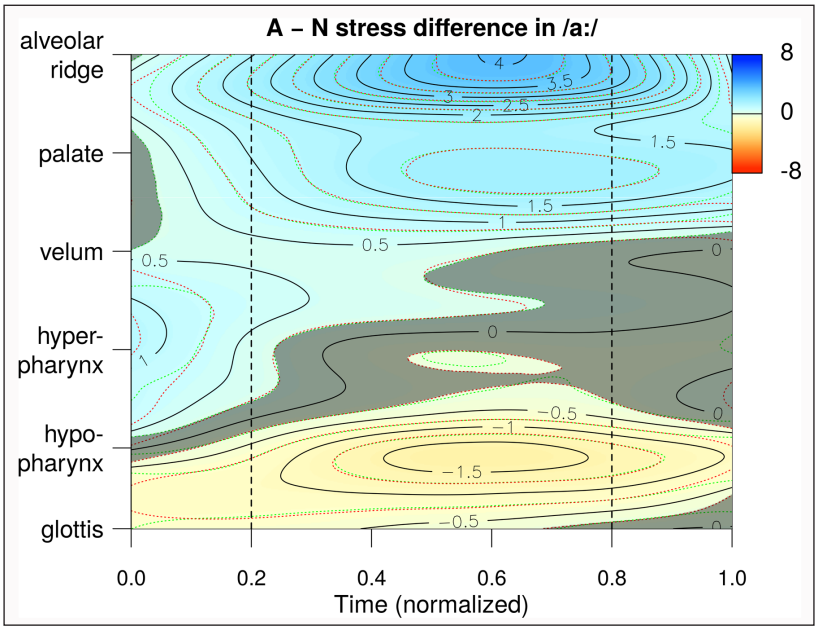

Figure 10: GAMM heatmap of vocal tract aperture (z-axis) over time (x-axis) throughout the vocal tract (y-axis), for the difference between accentuated and neutrally stressed /aː/. Aperture difference (mm) is denoted by color grade, regions of equidistant change (here, Δ0.5 mm) are denoted by black solid lines, and 95% confidence interval bands are denoted by red and green dotted lines. Significant differences (α = 0.05) are denoted by colored areas (significant) versus opaque areas (non-significant). Vertical dashed lines denote 20% and 80% of the vowel interval, for comparison with FLMMs at the same time points (Figure 12).

size of the effect is much smaller compared to the previous two cases. At the vowel onset, accentuated /aː/ is produced with greater constriction in the lower pharynx (as indicated by the light yellow shade with negative values which denote a lesser aperture and hence greater constriction), but greater expansion throughout the rest of the vocal tract, including the hyper-pharynx (as indicated by the light blue regions with positive values). This suggests that stressed /aː/ is produced with a lower and more fronted tongue body position, but with a constriction in the hypo-pharynx, in comparison to neutrally-stressed /aː/.[10] As the time course of the vowel unfolds, these differences increase: The spatial extent of the pharyngeal constriction broadens into the hyper-pharynx, and the magnitude of palatal expansion is enhanced. However, even at the point when these articulatory distinctions are most pronounced—occurring at ≈65% of the vowel interval—the GAMM results suggest that the effect of accentuation on the articulation of /aː/ is substantially smaller than previously observed for /aː/ versus /iː/ (Section 3.1) or for /aː/ versus /aɪ/ (Section 3.2).

The GAMM summary is provided in **Table 3**. 76.8% of the total variance is explained by the model. With regard to the parametric coefficients, the results reveal that neutrally stressed /aː/ is produced with smaller overall vocal tract aperture ($\mu \approx 5.31$ mm) compared to accentuated /aː/ ($\mu \approx 5.88$ mm). In other words, accentuated /aː/ is produced with a more open vocal tract, which is also evidenced by the larger area of blue (i.e., positive difference: expansion) compared to yellow/red (i.e., negative difference: constriction) in **Figure 10**. This difference is likely due to a more open jaw/tongue configuration associated with the accentuation of the (low) vowel /aː/. With regard to the smooth

[10] It is possible that the constriction in the lower pharynx serves the role of enhancing the F1-raising effect of tongue lowering, as has been argued to occur in the production of French nasal vowels (Carignan et al., 2015).

terms, each of the terms is once again significant in contributing to the overall model fit. The by-speaker random smooths are of particular interest here, as they have an even greater contribution compared to the previous two models. This is clearly seen by the differences in curve shapes for the two smooth terms (**Figure 11**); however, the lack of any distinctive patterning in curve variation suggests a relatively large amount of inter-speaker variation both over time (left plot in **Figure 11**) and in overall articulation (right plot in **Figure 11**).

The results for the FLMMs created to test for differences between accentuated and neutral /a:/ at 20% and 80% of the vowel interval are shown in **Figure 12**. The VT aperture profiles suggest that, in comparison with neutral /a:/, accentuated /a:/ is produced with slightly greater constriction at the glottis and slightly greater expansion from the hypopharynx up to the alveolar ridge, where the confidence interval bands meet. Apart from the alveolar ridge, the area of the smallest difference is around the velum. These results are largely in agreement with the GAMM results for the same time point. Although the VT aperture differences at 20% of the vowel interval are rather small, they are nonetheless revealed as significant in both the GAMM and FLMM models. However, whereas the GAMM suggests significantly greater expansion up to 1.5 mm for accentuated versus neutral /a:/ in the anterior portion of the vocal tract, the differences revealed by the FLMM are smaller in magnitude (≤ 0.5 mm), suggesting that the FLMM method is perhaps slightly more conservative at this particular region.

Table 3: Summary of the GAMM created to compare accentuated and neutral /a:/.

A. parametric coefficients	**Estimate**	***SE***	**t-value**	**p-value**
(Intercept)	5.8819	0.4309	13.6488	<0.0001
stress-N	−0.5704	0.0341	−16.7362	<0.0001
B. smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
te(time.norm,gridline.norm)	122.9860	145.5727	163.0388	<0.0001
te(time.norm,gridline.norm):stress-N	78.0490	96.0413	24.6578	<0.0001
s(time.norm,speaker)	74.4270	143.0000	1.6923	<0.0001
s(gridline.norm,speaker)	121.7445	143.0000	77.9771	<0.0001
s(word)	3.9844	4.0000	264.0933	<0.0001

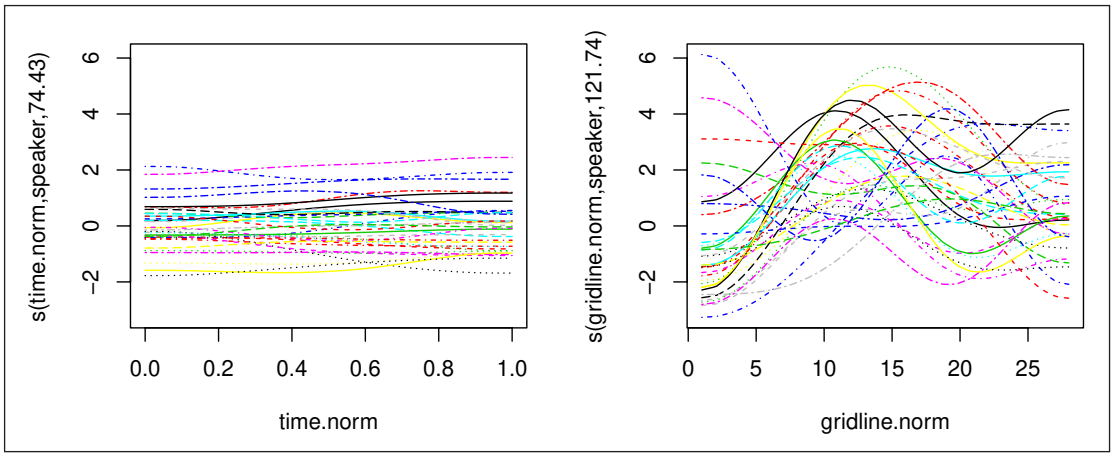

Figure 11: Visualization of by-speaker factor smooths over time (left plot) and space (right plot) in the GAMM created to compare accentuated and neutral /a:/.

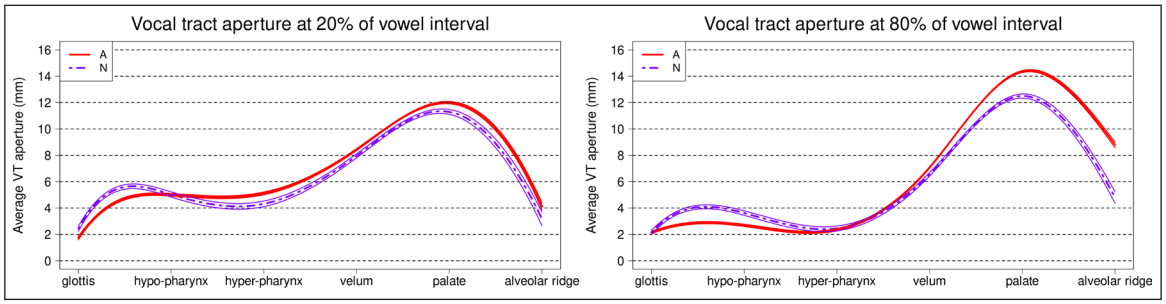

Figure 12: Results for the FLMM created to compare vocal tract aperture (y-axis) throughout the vocal tract (x-axis) between accentuated /a:/ (red, solid) and neutrally stressed /a:/ (purple, dash-dot) at 20% of the vowel interval, with 95% confidence interval bands shown for both groups.

At 80% of the vowel interval, the VT aperture profiles suggest areas of non-significant difference at the glottis and from the hyper-pharynx to the velum, but greater constriction for accentuated versus neutral /a:/ throughout the lower pharynx (≤ 1 mm), and greater expansion for accentuated versus neutral /a:/ beginning at the velum and reaching the greatest differences at the palate (≈ 2 mm) and the alveolar ridge (≈ 3 mm), estimates that mirror the GAMM results at these respective locations in each case. Thus, the FLMM results are in agreement with the GAMM results at 80% of the vowel interval, both with regard to areas of significant differences in the vocal tract and with regard to areas of similarity.

4. Conclusion

In each of the three test cases observed in this study, the context of interest was found to condition changes in rt-MRI vocal tract aperture functions associated with German vowel productions. These changes were observed in the generalized additive mixed models and cross-verified at both the beginning of the vowel (20% of the vowel interval) and the end of the vowel (80% of the vowel interval), using independently constructed functional linear mixed models created with data from these time points. In each of the test cases, the FLMM results were in agreement with the GAMM results, even at different levels of effect size: large differences in VT aperture, minor (but significant) differences in VT aperture, and non-significant differences (i.e., similarities in VT aperture). We conclude that these converging results support the use of both GAMMs and FLMMs in the analysis of real-time MRI video of speech.

Although in some cases the results suggested that the FLMM method is slightly more conservative than the GAMM method, and in other cases the GAMM method more conservative than the FLMM method, any differences between the model estimates were ≤ 1 mm (i.e., smaller than the in-plane resolution of the MR images). Given the similar results achieved by the two methods, we contend that GAMMs are reliable and, in comparison to FLMMs, generally more flexible and more useful for rt-MRI research of this type, which involves not only changes in space (throughout the vocal tract) but also in time (as speech unfolds temporally). A notable disadvantage of using FLMMs with rt-MRI data is that the researcher must prioritize one of these two dimensions at the expense of the other: She must choose either to investigate changing aperture over time at a single point in the vocal tract, or to investigate different degrees of aperture throughout the vocal tract at a single point in time. However, GAMMs allow for observation of aperture *throughout the vocal tract* as it *changes over time*, without needing to sacrifice one dimension for the other. In this way, applying GAMMs to rt-MRI videos of speech

provides a method to automatically identify regions of vocal tract variation in a way that retains both spatial information about the vocal tract and temporal information about speech dynamics.

Although the results observed here are admittedly unremarkable in their informative substance—it is not exactly a ground-breaking revelation to show, for example, that /aː/ and /iː/ differ in articulation—the three test cases examined in this study demonstrate the viability and potential of using GAMMs to investigate dynamic vocal tract characteristics in rt-MRI video. With recent and continuing advances in rt-MRI hardware, scanning techniques, and reconstruction techniques, as well as ever-increasing access to rt-MRI scanning through interdisciplinary research programs, rt-MRI is quickly becoming an attainable and worthwhile method for visualizing speech kinematics. It is our hope that the proofs of concept provided by this study may encourage the extension of rt-MRI GAMMs to more fundamental questions about the nature of human speech sounds and sound systems.

An area of possible future research is to determine whether the method proposed here extends its viability to the analysis of MRI video of spontaneous speech. Given the higher overall speech rate in spontaneous compared to read speech, it would be of particular interest to observe whether the GAMM method can accurately capture differences in: (1) more sparsely sampled temporal data that include (2) a greater degree of kinematic change between subsequent frames. Real-time MRI acquisition of spontaneous speech with high temporal and spatial resolutions is certainly the holy grail for investigating natural speech articulation, a goal that does not come without its methodological challenges. We anticipate that the rt-MRI GAMM method proposed in this study brings us a step closer to meeting those challenges by providing a powerful and interpretable statistical framework for rt-MRI analysis.

Finally, we would like to suggest that the method outlined here can be generalized to other types of speech production data. Similar spatio-temporal GAMMs could be applied to, e.g., ultrasound tongue contours to study changes in tongue shape over time, electropalatography data to study changes in the location of linguo-palatal contact over time, or electromagnetic articulometry data to study changes in the positions of multiple flesh-points over time. The method could even extend to acoustic data, e.g., to study changes in spectral energy over time as captured by either mel-frequency cepstral coefficients or binned frequency energy. When applied to speech production data in these ways, GAMMs offer a flexible, interpretable, and statistically robust method to investigate the organization and structure of speech in both time and space from a laboratory phonology perspective.

Acknowledgements

This research was funded by ERC Advanced Grant 295573 “Human interaction and the evolution of spoken accent” (J. Harrington) and DFG grant HA 3512/15-1 “Nasal coarticulation and sound change: a real-time MRI study” (J. Harrington & J. Frahm). The authors are grateful for comments from audience members at the *New Developments in Speech Sensing and Imaging* satellite event of *16th Conference on Laboratory Phonology* (Portugal, June 2018), where portions of this work were presented. The authors would also like to thank editors Lorenzo Spreafico, Alessandro Vietti, and Mirjam Ernestus, as well as the two anonymous reviewers, for their contribution, help, and insight.

Competing Interests

The authors have no competing interests to declare.

References

Ahmad, M., Dargaud, J., Morin, A., & Cotton, F. (2009). Dynamic MRI of larynx and vocal fold vibrations in normal phonation. *Journal of Voice*, *23*(2), 235–239. DOI: https://doi.org/10.1016/j.jvoice.2007.08.008

Baayen, R. H., Kuperman, V., & Bertram, R. (2010). Frequency effects in compound processing. In S. Scalise & I. Vogel (Eds.), *Cross-disciplinary issues in compounding* (pp. 257–270). Amsterdam: John Benjamins. DOI: https://doi.org/10.1075/cilt.311.20baa

Baayen, R. H., Rij, J. V., de Cat, C., & Wood, S. N. (2016). Autocorrelated errors in experimental data in the language sciences: Some solutions offered by Generalized Additive Mixed Models. In D. Speelman, K. Heylen & D. Geeraerts (Eds.), *Mixed effects regression models in linguistics* (p. arXiv:1601.02043). Berlin: Springer.

Baayen, R. H., Vasishth, S., Kliegl, R., & Bates, D. (2017). The cave of shadows. Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language*, 206–234. DOI: https://doi.org/10.1016/j.jml.2016.11.006

Barlaz, M., Shosted, R., Fu, M., & Sutton, B. (2018). Oropharygneal articulation of phonemic and phonetic nasalization in Brazilian Portuguese. *Journal of Phonetics*, *71*, 81–97. DOI: https://doi.org/10.1016/j.wocn.2018.07.009

Boersma, P., & Weenink, D. (2017). *Praat: Doing phonetics by computer.* Computer software program available from http://www.praat.org/

Byrd, D., Tobin, S., Bresch, E., & Narayanan, S. (2009). Timing effects of syllable structure and stress on nasals: A real-time MRI examination. *Journal of Phonetics*, *47*, 97–110. DOI: https://doi.org/10.1016/j.wocn.2008.10.002

Carignan, C., Hoole, P., Kunay, E., Joseph, A., Voit, D., Frahm, J., & Harrington, J. (2019). The phonetic basis of phonological vowel nasality: Evidence from real-time MRI velum movement in German. In S. Calhoun, P. Escudero, M. Tabain & P. Warren (Eds.), *Proceedings of 19th International Congress of Phonetic Sciences (ICPhS)*. Canberra: Australasian Speech Science and Technology Association.

Carignan, C., Shosted, R., Fu, M., Liang, Z.-P., & Sutton, B. (2015). A real-time MRI investigation of the role of lingual and pharyngeal articulation in the production of the nasal vowel system of French. *Journal of Phonetics*, *50*, 34–51. DOI: https://doi.org/10.1016/j.wocn.2015.01.001

Cederbaum, J. (2017). sparseFLMM: Functional Linear Mixed Models for Irregularly or Sparsely Sampled Data [Computer software manual]. Computer software program available from https://cran.r-project.org/package = sparseFLMM

Cederbaum, J., Pouplier, M., Hoole, P., & Grevens, S. (2016). Functional linear mixed models for irregularly or sparsely sampled data. *Stasistical Modeling*, *16*, 67–88. DOI: https://doi.org/10.1177/1471082X15617594

Demolin, D., Hassid, S., Metens, T., & Soquet, A. (2002). Real-time MRI and articulatory coordination in speech. *Comptes Rendus Biologies*, *325*(4), 547–556. DOI: https://doi.org/10.1016/S1631-0691(02)01458-0

Eryildirim, A., & Berger, M.-O. (2011). A guided approach for automatic segmentation and modeling of the vocal tract in MRI images. In *Proceedings of the nineteenth european conference on signal processing (2011)* (pp. 61–65).

Frahm, J., Schätz, S., Untenberger, M., Zhang, S., Voit, D., Merboldt, K. D., Sohns, J. M., Lotz, J., & Uecker, M. (2014). On the temporal fidelity of nonlinear inverse reconstructions for real-time MRI-the motion challenge. *The Open Medical Imaging Journal*, *8*, 1–7. DOI: https://doi.org/10.2174/1874347101408010001

Fu, M., Christodoulou, A. G., Naber, A., Kuehn, D., Liang, Z. P., & Sutton, B. P. (2012). High-frame-rate multislice speech imaging with sparse sampling of (k,t)-space. In *20th Annual ISMRM Scientific Meeting Exhibition* (p. 12).

Fu, M., Zhao, B., Carignan, C., Shosted, R. K., Perry, J. L., Kuehn, D. P., Liang, Z.-P., & Sutton, B. P. (2015). High-resolution dynamic speech imaging with joint low-rank and sparsity constraints. *Magnetic Resonance in Medicine*, *73*(5), 1820–1832. DOI: https://doi.org/10.1002/mrm.25302

Hastie, T., & Tibshirani, R. (1990). *Generalized additive models*. London: Chapman & Hall.

Honda, K., & Tiede, M. K. (1998). An MRI study on the relationship between oral cavity shape and larynx position. In *Proceedings of the 5th International Conference Spoken Language Processing.*

Iltis, P. W., Frahm, J., Voit, D., Joseph, A. A., Schoonderwaldt, E., & Altenmüller, E. (2015). High-speed real-time magnetic resonance imaging of fast tongue movements in elite horn players. *Quantitative Imaging in Medicine and Surgery*, *5*(3), 374–381.

Iskarous, K. (2005). Patterns of tongue movement. *Journal of Phonetics*, *33*(4), 363–381. DOI: https://doi.org/10.1016/j.wocn.2004.09.001

Kirkham, S., Nance, C., Littlewood, B., Lightfoot, K., & Groarke, E. (2019). Dialect variation in formant dynamics: The acoustics of lateral and vowel sequences in Manchester and Liverpool English. *The Journal of the Acoustical Society of America*, *145*(2), 784–794. DOI: https://doi.org/10.1121/1.5089886

Labrunie, M., Badin, P., Voit, D., Joseph, A. A., Frahm, J., Lamalle, L., Vilain, C., & Boë, L.-J. (2018). Automatic segmentation of speech articulators from real-time midsagittal MRI based on supervised learning. *Speech Communication*, *99*, 27–46. DOI: https://doi.org/10.1016/j.specom.2018.02.004

Lammert, A., Proctor, M., & Narayanan, S. (2010). Data-driven analysis of realtime vocal tract MRI using correlated image regions. In *Proceedings of INTERSPEECH 2010* (pp. 1572–1575).

Lammert, A., Ramanarayanan, V., Proctor, M., & Narayanan, S. (2013). Vocal tract crossdistance estimation from real-time MRI using region-of-interest analysis. In *Proceedings of INTERSPEECH 2013* (pp. 959–962).

Lingala, S. G., Sutton, B. P., Miquel, M. E., & Nayak, K. S. (2016). Recommendations for real-time speech MRI. *Journal of Magnetic Resonance Imaging*, *43*(1), 28–44. DOI: https://doi.org/10.1002/jmri.24997

Martins, P., Oliveira, C., Silva, S., & Teixeira, A. (2012). Velar movement in European Portuguese nasal vowels. In *Proceedings of IberSpeech—VII Jornadas en Tecnología del Habla and II Iberian SLTech Workshop, Madrid, Spain* (pp. 231–240).

Mielke, J., Carignan, C., & Thomas, E. R. (2017). The articulatory dynamics of pre-velar and pre-nasal /æ/-raising in English: An ultrasound study. *Journal of the Acoustical Society of America*, *142*(1), 332–349. DOI: https://doi.org/10.1121/1.4991348

Moisik, S. R., Esling, J. H., Crevier-Buchman, L., Amelot, A., & Halimi, P. (2015). Multimodal imaging of glottal stop and creaky voice: Evaluating the role of epilaryngeal constriction. In *Proceedings of the 18th International Congress of Phonetic Sciences.*

Moisik, S. R., Esling, J. H., Crevier-Buchman, L., & Halimi, P. (2019). Putting the larynx in the vowel space: Studying larynx state across vowel quality using MRI. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of 19th International Congress of Phonetic Sciences (ICPhS)*. Canberra: Australasian Speech Science and Technology Association.

Mücke, D., Grice, M., & Cho, T. (2014). More than a magic moment – Paving the way for dynamics of articulation and prosodic structure. *Journal of Phonetics*, *44*, 1–7. DOI: https://doi.org/10.1016/j.wocn.2014.03.001

Narayanan, S., Toutios, A., Ramanarayanan, V., Lammert, A., Kim, J., Lee, S., Nayak, K., Kim, Y. C., Zhu, Y., Goldstein, L., Byrd, D., Bresch, E., Ghosh, P., Katsamanis, A., & Proctor,

M. (2014). Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC). *The Journal of the Acoustical Society of America*, *136*(3), 1307–1311. DOI: https://doi.org/10.1121/1.4890284

Niebergall, A., Zhang, S., Kunay, E., Keydana, G., Job, M., Uecker, M., & Frahm, J. (2012). Realtime MRI of speaking at a resolution of 33 ms: Undersampled radial FLASH with nonlinear inverse reconstruction. *Magnetic Resonance in Medicine*, *69*, 477–485. DOI: https://doi.org/10.1002/mrm.24276

Pouplier, M., Cederbaum, J., Hoole, P., Marin, S., & Greven, S. (2017). Mixed modeling for irregularly sampled and correlated functional data: Speech science applications. *Journal of the Acoustical Society of America*, *142*(2), 935–946. DOI: https://doi.org/10.1121/1.4998555

Proctor, M., Bone, D., Katsamanis, N., & Narayanan, S. (2010). Rapid semi-automatic segmentation of real-time magnetic resonance images for parametric vocal tract analysis. In *Proceedings of INTERSPEECH 2010* (pp. 1576–1579).

Proctor, M., Goldstein, L., Lammert, A., Byrd, D., Toutios, A., & Narayanan, S. (2013). Velic coordination in French nasals: A real-time magnetic resonance imaging study. In *Proceedings of INTERSPEECH 2013* (pp. 577–581).

Raeesy, Z., Rueda, S., Udupa, J. K., & Coleman, J. (2013). Automatic segmentation of vocal tract MR images. In *Proceedings of the tenth IEEE International Symposium on Biomedical Imaging, (IEEE 2013)* (pp. 1328–1331). DOI: https://doi.org/10.1109/ISBI.2013.6556777

Ramanarayanan, V., Tilsen, S., Proctor, M., Töger, J., Goldstein, L., Nayak, K. S., & Narayanan, S. (2018). Analysis of speech production real-time MRI. *Computer Speech & Language*, *52*, 1–22. DOI: https://doi.org/10.1016/j.csl.2018.04.002

R Core Team. (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria. Computer software program available from http://www.R-project.org

Scheipl, F., Staicu, A.-M., & Greven, S. (2015). Functional additive mixed models. *Journal of Computational and Graphical Statistics*, *24*, 477–501. DOI: https://doi.org/10.1080/10618600.2014.901914

Shosted, R. K., Sutton, B. P., & Benmamoun, A. (2012). Using magnetic resonance to image the pharynx during Arabic speech: Static and dynamic aspects. In *Proceedings of INTERSPEECH 2012* (pp. 2182–2185).

Silva, S., & Teixeira, A. (2015). Unsupervised segmentation of the vocal tract from realtime MRI sequences. *Computer Speech & Language*, *33*(1), 25–46. DOI: https://doi.org/10.1016/j.csl.2014.12.003

Silva, S., & Teixeira, A. (2016). Quantitative systematic analysis of vocal tract data. *Computer Speech & Language*, *36*, 307–329. DOI: https://doi.org/10.1016/j.csl.2015.05.004

Sóskuthy, M. (2017). *Generalised additive mixed models for dynamic analysis in linguistics: A practical introduction.* arXiv:1703.05339 [stat:AP].

Teixeira, A., Martins, P., Oliveira, C., Ferreira, C., Silva, A., & Shosted, R. (2012). Realtime MRI for Portuguese. In *Computational Processing of the Portuguese Language* (pp. 306–317). Springer. DOI: https://doi.org/10.1007/978-3-642-28885-2_35

Tiede, M. K. (1996). An MRI-based study of pharyngeal volume contrasts in Akan and English. *Journal of Phonetics*, *24*(4), 399–421. DOI: https://doi.org/10.1006/jpho.1996.0022

Tilsen, S., Spincemaille, P., Xu, B., Doerschuk, P., Luh, W. M., Feldman, E., & Wang, Y. (2016). Anticipatory posturing of the vocal tract reveals dissociation of speech movement plans from linguistic units. *PloS one*, *11*(1), e0146813. DOI: https://doi.org/10.1371/journal.pone.0146813

Tomaschek, F., Arnold, D., Bröker, F., & Baayen, R. H. (2018). Lexical frequency codetermines the speed-curvature relation in articulation. *Journal of Phonetics*, *68*, 103–116. DOI: https://doi.org/10.1016/j.wocn.2018.02.003

Tomaschek, F., Tucker, B. V., Fasiolo, M., & Baayen, R. H. (2018). Practice makes perfect: The consequences of lexical proficiency for articulation. *Linguistics Vanguard*, *4*(s2). DOI: https://doi.org/10.1515/lingvan-2017-0018

Uecker, M., Zhang, S., Voit, D., Karaus, A., Merboldt, K., & Frahm, J. (2010). Real-time MRI at a resolution of 20 ms. *NMR Biomed.*, *23*, 986–994. DOI: https://doi.org/10.1002/nbm.1585

van Rij, J., Wieling, M., Baayen, R. H., & van Rijn, H. (2017). *itsadug: Interpreting Time Series and Autocorrelated Data Using GAMMs.* Computer software program available from https://cran.r-project.org/package = itsadug

Wang, J.-L., Chiou, J.-M., & Müller, H.-G. (2016). Functional data analysis. *Annual Review of Statistics and Its Application*, *3*, 257–295. DOI: https://doi.org/10.1146/annurev-statistics-041715-033624

Wieling, M. (2018). Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English. *Journal of Phonetics*, *70*, 86–116. DOI: https://doi.org/10.1016/j.wocn.2018.03.002

Wieling, M., Montemagni, S., Nerbonne, J., & Baayen, R. H. (2014). Lexical differences between Tuscan dialects and standard Italian: Accounting for geographical and sociodemographic variation using generalized additive mixed modeling. *Language*, *90*, 669–692. DOI: https://doi.org/10.1353/lan.2014.0064

Wieling, M., Tomaschek, F., Arnold, D., Tiede, M., Bröker, F., Thiele, S., Wood, S. N., & Baayen, R. H. (2016). Investigating dialectal differences using articulography. *Journal of Phonetics*, *59*, 122–143. DOI: https://doi.org/10.1016/j.wocn.2016.09.004

Winter, B., & Wieling, M. (2016). How to analyze linguistic change using mixed models, Growth Curve Analysis and Generalized Additive Modeling. *Journal of Language Evolution*, *1*(1), 7–18. DOI: https://doi.org/10.1093/jole/lzv003

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, *99*, 673–686. DOI: https://doi.org/10.1198/016214504000000980

Wood, S. N. (2006a). *Generalized Additive Models*. New York: Chapman & Hall/CRC.

Wood, S. N. (2006b). Low rank scale invariant tensor product smooths for generalized additive mixed models. *Biometrics*, *62*(4), 1025–1036. DOI: https://doi.org/10.1111/j.1541-0420.2006.00574.x

Wood, S. N. (2019). *mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation.* Computer software program available from https://cran.rproject.org/package = mgcv

Zhang, D., Yang, M., Tao, J., Wang, Y., Liu, B., & Bukhari, D. (2016). Extraction of tongue contour in real-time magnetic resonance imaging sequences. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 937–941). DOI: https://doi.org/10.1109/ICASSP.2016.7471813

How to cite this article: Carignan, C., Hoole, P., Kunay, E., Pouplier, M., Joseph, A., Voit, D., Frahm, J., and Harrington, J. 2020 Analyzing speech in both time and space: Generalized additive mixed models can uncover systematic patterns of variation in vocal tract shape in real-time MRI. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 11(1): 2, pp. 1–26. DOI: https://doi.org/10.5334/labphon.214

Submitted: 18 July 2019 **Accepted:** 22 January 2020 **Published:** 18 March 2020

]u[*Laboratory Phonology: Journal of the Association for Laboratory Phonology* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS