Automatic quantitative analysis of ultrasound tongue contours via wavelet-based functional mixed models

Leonardo Lancia
Department of Linguistics, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany
leonardo_lancia@eva.mpg.de

Philip Rausch[a)]
Berlin School of Mind and Brain, Department of German Studies and Linguistics, Humboldt Universität zu Berlin, Wolfgang Köhler-Haus, Rudower Chaussee 18, 12489 Berlin, Germany
philip.rausch@hu-berlin.de

Jeffrey S. Morris
Department of Biostatistics, The University of Texas, M.D. Anderson Cancer Center, Houston, Texas 77230
jefmorris@mdanderson.org

Abstract: This paper illustrates the application of wavelet-based functional mixed models to automatic quantification of differences between tongue contours obtained through ultrasound imaging. The reliability of this method is demonstrated through the analysis of tongue positions recorded from a female and a male speaker at the onset of the vowels /a/ and /i/ produced in the context of the consonants /t/ and /k/. The proposed method allows detection of significant differences between configurations of the articulators that are visible in ultrasound images during the production of different speech gestures and is compatible with statistical designs containing both fixed and random terms.

[AL]
Date Received: September 15, 2014 **Date Accepted:** December 18, 2014

1. Introduction

Ultrasound imaging is commonly used to display the midsagittal surface contour of the tongue during speech production. By placing an ultrasound (US) probe below the speaker's chin, images of the superior tongue contour can be obtained as a high-intensity shape. Although this technique does not visualize the tip of the tongue, which is concealed by the shadow cast by the jaw, US enables us to study the behavior of the back of the tongue, which is not easily captured with electromagnetic articulography. An important practical limitation of US imaging is related to the extraction of the tongue contour from the noisy images: Automatic algorithms (e.g., that of Li *et al.*, 2005) are prone to errors, and their results therefore need to be corrected manually, one image at a time. In this paper, we propose a fully automatic method of quantifying differences between tongue shapes in recorded US images through the application of wavelet based functional mixed models (WFMM; Morris *et al.*, 2011). Importantly,

[a)]Also at Berlin School of Mind and Brain, Department of Psychology, Humboldt Universität zu Berlin, Wolfgang Köhler-Haus, Rudower Chaussee 18, 12489 Berlin, Germany.

this approach is compatible with complex statistical designs containing both fixed and random factors.

2. Background

WFMM was introduced by Morris and Carroll (2006) to model the effects of one or more factors on the shape of observed trajectories. This approach was extended by Morris *et al.* (2011) to the modeling of grayscale images. The authors propose to represent each image as a vector of coefficients obtained by applying the two-dimensional discrete wavelet transform (DWT) to the data. Because each image is represented by several coefficients, the effects of the experimental factors are estimated through a multivariate mixed model the dependent variables of which are the wavelet coefficients. Once the fixed effects are computed in the wavelet space, these can be back-transformed into the data space (i.e., they can be transformed into differences between intensity values) by means of the two-dimensional inverse discrete wavelet transform (IDWT). The effects of the fixed factors are computed on wavelet coefficients because the covariance matrices of the model in the wavelet space are diagonal, which is motivated by the whitening property of wavelets. This captures local correlation between pixels (Morris and Carroll, 2006) and yet drastically reduces the number of parameters to be computed with respect to a model the dependent variables of which are the intensity values of the images and thus guarantees computational feasibility. The two-dimensional DWT permits modeling of the intensity values of the pixels within an image as a linear combination of functions (or wavelets) located at different coordinates, oriented toward different directions (horizontal, vertical, and diagonal) and varying at different rates (the scales of the wavelets). The coefficients regulating the behaviors of the functions are therefore triple-indexed by (a) wavelet scale (index j), (b) location (index k), and (c) direction of change of the function (index $l = 1$ if row coefficients, $= 2$ if column coefficients, and $= 3$ if diagonal coefficients). The key benefit of using wavelet bases instead of just fitting separate scalar functional mixed models at each pixel position is that the wavelets are multi-scale representations that will allow borrowing of strength from nearby locations in the images that are correlated with each other—i.e., the prediction at a given pixel will also be informed by nearby pixels, thus yielding more efficient estimates and inference.

The number of coefficients required to model a given image depends mainly on the size of the image (cf. Walker, 1999). To further decrease the computation time, the number of coefficients can be reduced by applying compression algorithms to exclude coefficients which are close to 0 across all images while preserving a fixed amount of energy. In the analyses presented below, we retained 99.5% of the image energy, which reduced the number of coefficients by a factor of 12.6 (from 84 618 to 6687) in the model summarized in Fig. 1 and by a factor of 32 (from 84 618 to 2624) in the model summarized in Fig. 2.

Given N observed images and N_c wavelet coefficients per image, the general formulation of the multivariate mixed model in the wavelet space is

$$Y = XB + ZU + E \tag{1}$$

Equation (1) is the functional version of the basic mixed effects model that is often adopted in speech research to test hypotheses about simple scalar data. Each row within the $(N \times N_c)$ matrix Y contains the wavelet coefficients corresponding to one image; X is the $(N \times p)$ fixed effects design matrix of p covariates' values; B is the $(p \times N_c)$ matrix of their effects; Z is the $(M \times m)$ random effects design matrix of m random factors' values; U is the $(M \times N_c)$ matrix of random effects, and E is the $(N \times N_c)$ matrix of residual errors. U follows a matrix normal distribution with $M \times M$ between-row covariance matrix P and $N_c \times N_c$ between-column diagonal covariance matrix $Q = diag(q_{jlk})$ (with j, l, k indexing indexing the location, orientation, and scale of the coefficients, meaning that each coefficient has its own variance). E follows a matrix normal distribution too, with between-row diagonal covariance

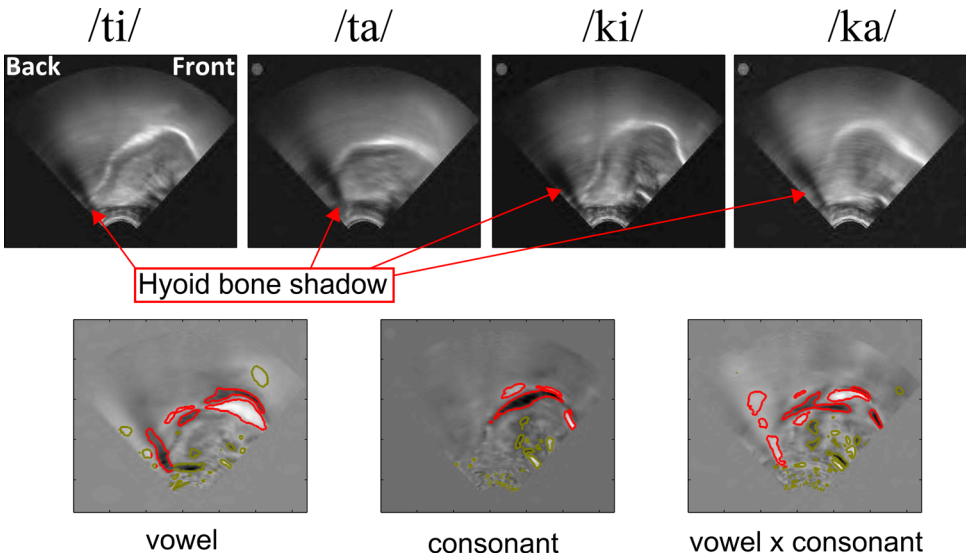

Fig. 1. (Color online) Posterior mean estimates for the cell means (upper panels) correspond to the mean tongue contours in the various conditions as estimated by the model applied to data from the female speaker. The contrast coefficients (lower panels) depict the differences between estimated tongue contours. Marked regions close to the tongue contour indicate portions of images where significant differences in the position of the tongue are observed.

function $S = diag(s_{jlk})$. Because each coefficient has its variance component, the strength of the covariance functions can vary across locations in the images.

Estimation is performed within the Bayesian framework, and the posterior distribution of the model's parameters given a sample of data in the wavelet space is obtained through Markov chain Monte Carlo sampling. Estimates obtained in the wavelet space are back-transformed into the data space coordinates to compute the estimated intensity values and reconstruct the estimated images. Spike-slab priors (Ishwaran and Rao, 2005) are assumed for the fixed effects' wavelet coefficients B_{ijlk}. Priors of this kind (having a spike at zero and medium to heavy tails) help to smooth

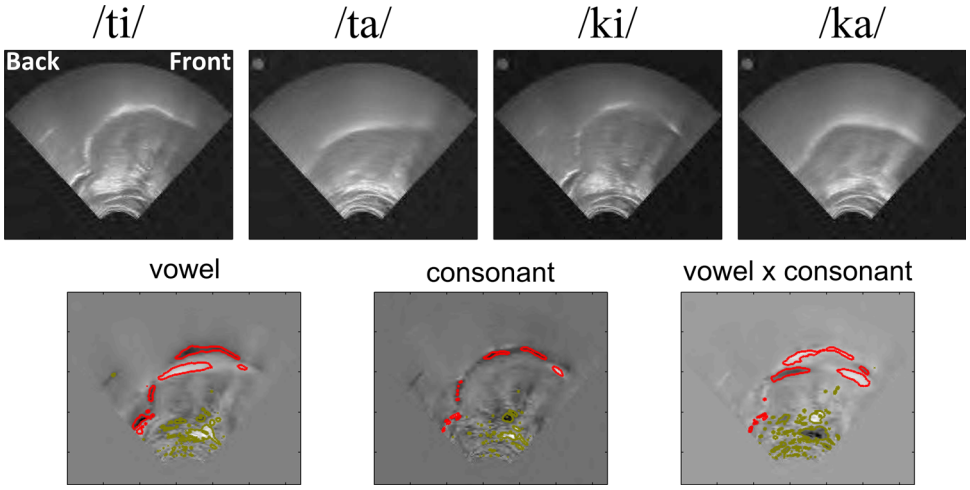

Fig. 2. (Color online) Posterior mean estimates for the cell means (upper panel) and contrast coefficients (lower panel) obtained by applying the model to data from the male speaker (hyoid bone shadow not visible).

the resulting images by adaptively shrinking small coefficients toward zero. Indexing the priors with respect to covariate, scale, location, and orientation permits different degrees of smoothing across predictors, scales, locations, and orientations. Unique to this approach, the smoothing power of the DWT is driven by the hierarchical structure of Bayesian modeling. Finally, vague proper priors are defined for the variance components (the reader is referred to Morris *et al.,* 2011 for details about the automatic choice of the hyper-parameters shaping the prior distributions).

Once the posterior samples of the effects B_{ajlk} with $a = \{1, \ldots, p\}$ have been estimated, they are submitted to IDWT, giving the posterior samples for the parameters $B_a(t_1, t_2)$, where t_1, t_2 vary along the vertical and horizontal directions in the image coordinates system. For each location in t_1, t_2, the probability that the effect of covariate X_a is higher than δ is computed as $p_a^{\delta} = Prob\{|B_a(t_1, t_2)| > \delta\}$. Regions of the image where the effect of the covariate X_a is considered significant are defined as regions where $p_a^{\delta}(t_1, t_2)$ exceeds a threshold θ. For a given δ, the values $1 - p_a^{\delta}(t_1, t_2)$ correspond to false discovery rates because they express the probability of erroneously labeling a location as significant when an effect size δ is selected. Practical considerations can be taken into account to optimize the value of δ and, with that value selected, a threshold θ can be set as the value at which the ratio between the falsely discovered regions and the significant regions is smaller than an arbitrary value. Importantly, due to the continuity of the high intensity regions showing portions of the tongue contour, increasing the values of these parameters does not result in failing to detect significant movements but mainly in a reduction of the extension of the regions that display significant differences.

3. Experiment

We asked one male and one female German speaker to repeatedly utter the following syllables: /ka, ta, ki, ti/. Each syllable was repeated without interruption over 10 s, yielding between 20 and 30 repetitions per speaker and syllable. These syllables were chosen because the production of the composing segments has been extensively studied with various methods including x rays, electromagnetic articulography, and magnetic resonance. Different tongue contour shapes are commonly observed in the production of /i/ and /a/ with the back of the tongue varying mainly along the front-back dimension. The comparison of tongue contours during the production of /t/ and /k/ reveals variation mainly along the top-down dimension. This allows us to formulate solid expectations on the differences across vowels but also on the differences conditioned by coarticulation. We can therefore evaluate the performance of the proposed method by comparing the results of our analyses to the expectations built on previous studies (e.g., Maeda, 1990; Stone *et al.,* 2001). The US probe was fixed to the speakers' chin by means of an Articulate Instruments fixation helmet with an inclination of approximately 30°. Ultrasound video and the corresponding acoustic signal were recorded using a Terason T3000 system and the ULTRASPEECH software (Hueber *et al.*, 2008) at 60 fps.

4. Data preparation and model specification

For each vowel, we collected only the image that was closest in time to the vowel onset (corresponding to the onset of voicing in the acoustic signal) to maximize coarticulation effects. Before modeling the collected images with the WFMM (109 images for the female speaker and 83 images for the male speaker), pixel intensities were normalized for each image with respect to the image-specific mean and standard deviation. The following covariates were tested: Vowel (/a/ vs /i/, reference level: /i/), surrounding consonant (/t/ vs /k/, reference level: /t/), and the interaction between the two. Although this modeling technique allows the inclusion of random factors, this was unnecessary in our case as speakers were modeled separately (due to morphological differences between their vocal tracts). For the DWT, we used Daubechies 4 wavelets (Walker, 1999). Regions were flagged as significantly different across levels if the

probability of finding differences equal to or higher than 1 (the standard deviation of the intensity in each normalized image) was at least 0.95.

5. Results

Figure 1 displays the results for the female speaker. The top panels show the model estimates of the articulatory configurations corresponding to the production of the two vowels in the context of the two consonants. The lower panels depict the differences between the levels of the predictors, i.e., the fixed effects images; the leftmost panel shows the difference between the two vowels in the context of the consonant /t/; the middle panel presents the differences in tongue positions observed in the production of the vowel /i/ in the two consonantal contexts; the rightmost panel depicts the adjustment required to account for the changes conditioned by differences in the consonants in the production of /a/. The contours in the lower panels indicate the boundaries of the regions significantly different between levels of the predictors. The red contours mark the regions where significant differences are observed near to the tongue contour and in areas where the shadow of the hyoid bone is found (i.e., the regions for which we have principled expectations). Green contours indicate regions where significant differences are observed inside the tongue body or outside the tongue contour. Due to the coding scheme adopted for the model covariates, the bright regions close to the tongue contour and marked in red indicate areas of the image containing parts of the tongue contour in the non-reference level of the relative covariate but not in the reference level; conversely, the dark regions marked in red indicate areas of the image containing parts of the tongue only in the reference level of the covariate. For example, in the panel showing the differences between levels of the vowel covariate, bright regions indicate portions of the image containing parts of tongue contour only in the production of vowel /a/, while dark regions indicate portions of the image containing parts of the tongue contour only in the production of the vowel /i/.

We observe that the tongue is more retracted and lower in the production of the vowel /a/ than in the production of /i/. Furthermore, during the production of the vowel /i/ in the context of /k/, the back of the tongue is higher than in the production of the same vowel in the context of /t/. There are clear interactions between vowel quality and consonant context: The back of the tongue is more elevated when /a/ is produced in the context of /k/. Finally, the difference in orientation of the shadow cast by the hyoid bone indicates that this bone either is more elevated or more fronted in the production of the open vowel. This difference, however, is observed only in the context of the coronal consonant. As shown in Fig. 2, the same qualitative differences in tongue contour shape can be observed in both speakers. Importantly, the estimated tongue contours closely match those reported in the literature for comparable articulations (e.g., Stone *et al.*, 2001).

6. Discussion and conclusion

In this paper, we reported the use of the WFMM to estimate significant differences between tongue contours when different articulatory maneuvers are performed. The method proved capable of identifying the expected differences across different vowels and across vowels uttered in the context of different consonants without any manual intervention in tongue contour detection. Moreover, it provides estimates of statistical significance of the results not provided by other fully automatic methods based on whole image analysis (e.g., Fasel and Berry, 2010).

Although in the present papers we analyzed snapshots of the tongue contour recorded at vowel onsets, further possibilities offered by this approach include modeling the evolution of the tongue contour shape over time. In principle, a sequence of images corresponding to a tongue movement can be considered as a three-dimensional (3-D) pattern and submitted to 3-D wavelet transform. The obtained coefficients can then be modeled with the proposed approach. If, due to the limited temporal resolution of most ultrasound recording devices, the 3-D transform cannot be applied, the

temporal dimension can be projected onto an ordered categorical factor. Consecutive images can be sorted into consecutive levels of the factor which can be coded through a successive difference contrast (Venables and Ripley, 2002). This approach allows independent modeling of changes between consecutive images. Therefore the energy at a given location in the images can increase or decrease when moving from one time step to the next, but then it can change in the opposite direction when moving further in time.

The method is currently being extended to use other basis functions, including principal components and innovative combinations of wavelets and principal components, so it is possible for this modeling framework to also incorporate other feature extraction methods such as Eigentongues (Hueber *et al.,* 2007).

Acknowledgments

The work of J.S.M. was supported by NCI Grant Nos. CA-107304 and CA-16672. We would like to thank Georgy Krasovitskiy for proofreading the manuscript.

References and links

Fasel, I., and Berry, J. (**2010**). "Deep belief networks for real-time extraction of tongue contours from ultrasound during speech," in *Proceedings of the 20th International IEEE Conference on Pattern Recognition*, pp. 1493–1496.

Hueber, T., Aversano, G., Chollet, G., Denby, B., Dreyfus, G., Oussar, Y., Russel, P., and Stone, M. (**2007**). "Eigentongue feature extraction for an ultrasound-based silent speech interface," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1245–1248.

Hueber, T., Chollet, G., Denby, B., and Stone, M. (**2008**). "Acquisition of ultrasound, video and acoustic speech data for a silent-speech interface application," in *Proceedings of International Seminar on Speech Production*, pp. 365–369.

Ishwaran, H., and Rao, J. S. (**2005**). "Spike and slab variable selection: Frequentist and Bayesian strategies," Ann. Stat. **33**(2), 730–773.

Li, M., Kambhamettu, C., and Stone, M. (**2005**). "Automatic contour tracking in ultrasound images," Clin. Linguist. Phonet. **19**(6-7), 545–554.

Maeda, S. (**1990**). "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model," in *Speech Production and Speech Modelling*, edited by A. Marchal and W. J. Hardcastle (Kluwer Academic, Dordrecht, The Netherlands), pp. 131–149.

Morris, J. S., Baladandayuthapani, V., Herrick, R. C., Sanna, P., and Gutstein, H. (**2011**). "Automated analysis of quantitative image data using isomorphic functional mixed models with application to proteomics data," Ann. Appl. Stat. **5**(2A), 894–923.

Morris, J. S., and Carroll, R. J. (**2006**). "Wavelet-based functional mixed models," J. R. Stat. Soc. Ser. B **68**(2), 179–199.

Stone, M., Davis, E. P., Douglas, A. S., Aiver, M. N., Gullapalli, R., Levine, W. S., and Lundberg, A. J. (**2001**). "Modeling tongue surface contours from cine-MRI images," J. Speech Lang. Hear. Res. **44**(5), 1026–1040.

Venables, W. N., and Ripley, B. D. (**2002**). *Modern Applied Statistics with S* (Springer, New York).

Walker, J. S. (**1999**). *A Primer on Wavelets and Their Scientific Applications* (Chapman and Hall/CRC, New York).