INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Zooming in on Spatiotemporal V-to-C Coarticulation with Functional PCA

Michele Gubian, Manfred Pastätter, Marianne Pouplier

Institute of Phonetics and Speech Processing, LMU Munich, Germany

m.gubian,manfred,pouplier@phonetik.uni-muenchen.de

Abstract

It has long been proposed in speech production research that in CV sequences, the movement for consonant and vowel are initiated synchronously. However, mostly due to limitations on the statistical analysis of articulator motion over time, this could only be shown in a limited fashion, based on positional differences at a single time point during consonantal constriction formation. It is unknown to which extent this observation generalizes to earlier timepoints. In this paper, we illustrate the use of functional principal component analysis (FPCA) for the statistical analysis of articulator motion over time. Using articulography data, we quantify CV coarticulation during constriction formation of [k] in two vowel contexts. We show how FPCA enables us to analyse both horizontal and vertical movement components over time in a single model while preserving information on temporal variability. We combine FPCA with linear mixed modelling to obtain estimated mean trajectories and confidence bands for [k] in the two vowel contexts. Results show that well before the timepoint of peak velocity the vowel causes a substantial spatial separation of the consonantal trajectories, estimated to be at least 3 mm at peak velocity. This lends support to the hypothesis that vowel and consonant are initiated synchronously.

Index Terms: Functional data analysis, mixed models, time series analysis, CV coarticulation, articulography

1. Introduction

The concept of coarticulation has been a long-standing topic in phonetic research; it describes the interaction of adjacent speech sounds when the respective articulatory gestures overlap in space and time (e.g. [1, 2, 3]). This spatiotemporal notion of coarticulation implies that coarticulation takes place in multiple dimensions. However, continuous time functions of, for instance, two-dimensional (X, Y) articulatory data have long been problematic for statistical analyses. This is why many studies sampled measurements at discrete *magic moments* at which spatial V-to-C coarticulation was statistically evaluated, separately for horizontal (X) and vertical (Y) movement components. While these studies have contributed to our fundamental understanding of how coarticulation arises from overlapping gestures, magic moment analyses discard a wealth of information about signal dynamics. In this paper, we demonstrate how functional principal component analysis (FPCA) in conjunction with Linear Mixed-Effects models (LME) can be used for the analysis of continuous multi-dimensional articulatory data in order to obtain a more comprehensive and fine-grained picture of the spatiotemporal characteristics of V-to-C coarticulation.

Functional Data Analysis (FDA) [4] and Generalized Additive Mixed Models (GAMMs) [5], have previously been applied to time series data in the speech sciences (e.g. [6, 7, 8, 9, 10, 11, 12, 13]). However, many studies using FDA are limited to a qualitative evaluation of time deformation curves (e.g. [10]) or functional PCs (e.g. [11]), and usually do not simultaneously analyse multiple spatial dimensions over time. GAMMs on the other hand cannot straightforwardly be applied to multi-dimensional time series analysis (though see [14]) and do not incorporate non-linear relative timing information.

In our current work we apply a variant of FDA based on FPCA and non-linear time warping developed in [15] that allows us to (i) analyse multi-dimensional trajectories (ii) include temporal distortion as an extra dimension which makes it easy to derive explicit relations between spatial and temporal variation across trajectory paths. LME models allows us to select the PC score that best captures the variation associated with a given independent factor (in our case, vowel context) while at the same time taking repeated measures on speakers and items into account. The potential of this method will be illustrated with a case study that concerns V-to-C coarticulation in [kV] sequences. Previous research on labial-vowel sequences [16] observed that articulator paths diverge as a function of vowel context as early as 50 ms into the consonant [16]. By jointly investigating horizontal (X) and vertical (Y) tongue dorsum movement of [k] in two different vowel contexts ([i, a]) we can estimate the dynamic evolution of the articulator trajectory from movement onset and also ask whether this effect holds beyond labial consonants. Many articulatory studies focus on independently measurable articulators (often tongue and lips) particularly when interested in temporal effects. Our methodological approach illustrated here opens the possibility for the statistical evaluation of the spatiotemporal dynamics of CV coarticulation within the same articulator. Data analysis thereby proceeds in several steps: First, the data are registered in space and time, with temporal information being encoded in non-linear time-warping functions which are added as an extra dimension to the data. FPCA is then applied in a second step, followed by the construction of LME models that identify the PC that captures the variation associated with the covariate of interest (here: vowel context) while taking into account speaker as a random effect. Lastly, we use the estimated means from the LME analysis for the PC scores to reconstruct the articulatory trajectories for the vowel contexts with by-speaker variability removed.

2. Methods

2.1. Materials

For our study we used available electromagnetic articulography (EMA, AG501) data from five speakers of Polish (for details see [17]). The recordings contained several repetitions of the target words *kim* and *kap*, pronounced in a constant carrier phrase. Based on the filtered velocity profile of vertical tongue dorsum movement, we identified during the constriction formation of [k] the following articulatory *landmarks*: *Gesture onset* (t_0), *Peak velocity of closing movement* (t_1), *Attainment of articulatory plateau* (t_2), and *Articulatory target* (t_3) (Figure 1). While t_1 and t_3 refer to the timepoints of maximum and minimum ve-

http://dx.doi.org/10.21437/Interspeech.2019-2143

locity, respectively, t_0 and t_2 define where velocity exceed or fall below a 20% threshold of the local velocity range. We then extracted the (X, Y) trajectories of the tongue dorsum between landmarks t_0 and t_3.

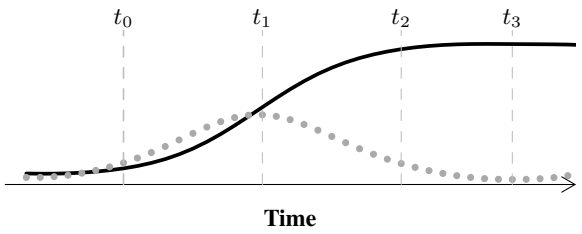


Figure 1: *Schematised tongue dorsum raising gesture for the [k] (solid line) plus velocity profile (dotted line) over time. t_0: Gesture onset (time zero), t_1: Peak velocity, t_2: Plateau attainment, t_3: Articulatory target.*

2.2. Space and time normalisation

A total of 73 sampled sensor trajectories (X, Y) underwent a number of pre-processing steps, the aim of which was to deal with variability among speakers and among realisations in the space and in the time domain. First, the spatial coordinates (X, Y) were speaker-normalised using z-score normalisation, in order to mitigate oral cavity size differences among subjects, which are greater than the size of the effects under study. The second step was to refer the spatial origin $(0, 0)$ to the gesture onset, i.e. $t = t_0 = 0$ implies $X = 0$ and $Y = 0$ in each trajectory, i.e. all trajectories start at the same point on the sagittal plane. This made it possible to inspect the degree of separation between [ka] and [ki] trajectories independently of any possible (small) systematic spatial offset.

Temporal variability was tackled as follows. Each (X, Y) trajectory was segmented into the landmark intervals as described in Section 2.1 (Figure 1). Generally, there are two complementary ways for (X, Y) trajectories to differ from each other. One is that they travel in different regions of the sagittal plane, the other is that they travel at different speed along the same path. In order to distinguish these two modalities we applied *landmark registration* [4], a non-linear warping of the time axis that synchronises a set of trajectories on a specified number of time points, i.e. the landmarks. The result is a set of transformed trajectories that reach the same corresponding landmarks exactly at the same time, i.e. t_1 in Figure 1 is the same value in all trajectories, and the same holds for t_2 and t_3, the latter implying that also total durations are identical. Formally, from a set of trajectories $(X_n(\tau), Y_n(\tau))$, where τ is the original (physical) time variable and n indexes the trajectories from 1 to 73, we obtained a corresponding set of time-normalised trajectories $(X_n(t), Y_n(t))$, where t is the normalised time axis. While this procedure allows comparison between (X, Y) trajectories, it comes at a cost because the original information about interval durations is lost. For the present study we adapted the procedure developed in [15] so that such information could be included in the analysis of the articulatory data that follows (see [15] and Appendix A in [18] for more details). The result of landmark registration is encoded in the time warping functions $\tau = h_n(t)$, which map the normalised time t to the original time axis τ for every trajectory. These functions are defined on the same normalised time axis as the trajectories $(X_n(t), Y_n(t))$. Hence $h_n(t)$, or any invertible transformation thereof, can be added as extra dimension to the normalised trajectories in order to preserve the original information on duration. For mathematical convenience, it is better to transform $h(t)$ as follows:

$$r(t) = -\log \frac{dh(t)}{dt}, \tag{1}$$

where $r(t)$ can be interpreted as relative log rate [15]. The result is that each (previously spatially normalised) 2-dimensional function $(X_n(\tau), Y_n(\tau))$ is mapped to a 3-dimensional function $(X_n(t), Y_n(t), r_n(t))$, where $r_n(t)$ defined in (1) is the log rate function associated to the n-th time-normalised function $(X_n(t), Y_n(t))$. The 2- and the 3-dimensional representations encode exactly the same information, each one can be obtained from the other. The advantage of the latter is that space and time variation have been decoupled from each other, which makes it possible to infer explicit relations between them, as it will be shown below.

Finally, trajectories were interpolated using a B-splines basis following the procedure illustrated in [19], this last step being necessary in order to feed articulatory trajectories as input to FPCA.

2.3. Functional PCA

FPCA [4] provides a data-driven parametrisation of a set of input curves or trajectories, the latter represented by continuous functions defined on the same time interval, which in our case are the space- and time-normalised 3-dimensional functions $(X_n(t), Y_n(t), r_n(t))$ defined in Section 2.2. The FPCA parametrisation is expressed by the following equations:

$$X_n(t) \approx \mu_X(t) + \sum_{k=1}^{K} s_{k,n} \cdot PCk_X(t) \tag{2a}$$

$$Y_n(t) \approx \mu_Y(t) + \sum_{k=1}^{K} s_{k,n} \cdot PCk_Y(t) \tag{2b}$$

$$r_n(t) \approx \mu_r(t) + \sum_{k=1}^{K} s_{k,n} \cdot PCk_r(t) \tag{2c}$$

where $(\mu_X(t), \mu_Y(t), \mu_r(t))$ is the mean trajectory, $(PCk_X(t), PCk_Y(t), PCk_r(t))$ are K Principal Component trajectories, $k = 1, \ldots, K$, which are based on the entire trajectory data set, and each $s_{k,n}$ is a weight or *score* modulating PCk for the n-th trajectory. Formally, Eq. (2) follow the same structure of ordinary PCA, namely any input is approximately decomposed into a linear combination of K PCs added to the data set mean. What is different from ordinary PCA is that input, mean and PCs are functions of time as opposed to vectors of real numbers, and in this particular case these functions are multi-dimensional, as they take values in X, Y and r at each point t in time. Crucially, PC scores $s_1, s_2, \ldots, s_K$ modulate all PC dimensions at once, e.g. the same coefficient s_1 multiplies $PC1_X(t)$, $PC1_Y(t)$ and $PC1_r(t)$ in Eq. (2), which is the key to find relations across space and time dimensions.

We computed the first $K = 3$ PCs[1], which combined explain 93.7% of the trajectory variance. The result is that each of the 73 trajectories is associated with a triplet of scores

[1]To avoid confusion, note that the decision to consider the first 3 PCs has nothing to do with the fact that the functions are 3-dimensional. In fact, each PC influences all 3 dimensions.

(s_1, s_2, s_3), which substituted into (2) reconstruct the original trajectory shape, to some approximation. In a next step, we use LME models in order to identify the PC which separates the data according to our variable of interest, vowel context.

2.4. Linear Mixed-Effects models

We aim at constructing models that characterise the tongue dorsum trajectories of [k̲a] and [k̲i] in terms of shape as well as speed differences between the two. To achieve this we made use of the trajectory parametrisation obtained from FPCA in Section 2.3. In particular, we constructed three independent LME models [20], one for each PC score, where the score is predicted by one binary fixed factor encoding vowel context, plus a random intercept term accounting for speaker. The rationale is that any systematic difference between PC scores predicted by the vowel following [k] translates into a spatial (X, Y) *and* temporal (r) dynamic characterisation of the corresponding trajectories, thanks to the FPCA parametrisation in Eq. (2).

3. Results

The fixed factor Vowel Context is highly significant in the LME model predicting s_1 ($\chi^2 = 293.3$, $p \ll 0.001$) and explains alone 65.4% of the variance of s_1. The same fixed factor is not significant in the model predicting s_2 and explains only 0.1% of its variance. In the model predicting s_3, the fixed factor is significant ($\chi^2 = 14.2$, $p < 0.001$) and explains 12.0% of the variance. The explained variance was estimated computing Marginal Pseudo-R^2 [21, 22]. Figure 2 displays the estimated marginal mean values for the three PC scores, together with their confidence intervals.

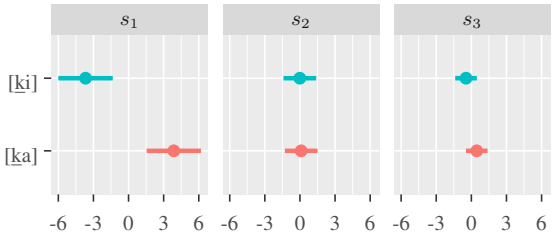


Figure 2: *Estimated marginal PC scores and associated confidence intervals.*

Figures 3a and 3b translate Figure 2 into (X, Y) trajectories and interval durations respectively by virtue of Eq. (2). In Figure 3a solid trajectories for [k̲a, k̲i] are obtained by substituting the estimated marginal mean PC scores, i.e. the dots in Figure 2, in Eq. (2a) and (2b), while dashed and dotted trajectories are constructed in the same way but using the lower and upper boundaries of their respective confidence intervals shown in Figure 2. For example, the solid trajectory for [k̲a] in panel PC1 of Figure 3a is obtained from Eq. (2a) and (2b) by substituting $s_1 = 3.87$, i.e. the value marked by the corresponding dot in panel s_1 of Figure 2, and setting the other scores to zero. Arrows at the upper end of trajectories indicate the movement direction, dots indicate the positions corresponding to time landmarks, i.e. the interval boundaries in Figure 1. The same principles were applied to obtain Figure 3b, in this case using Eq. (2c). Instead of displaying the (hard to interpret) relative log rate curves $r(t)$, the durations of the three intervals between landmarks were reconstructed by inverting Eq. (1) [18]. For example, the PC1 × [k̲a] panel (top left) of Figure 3b shows three horizontal bars representing durations, each bar divided in three adjacent between-landmarks intervals. Durations of each interval are written on the bars, and the cumulative durations can be read off the horizontal axis. The middle bar (solid contour) shows the durations corresponding to the expected marginal s_1 for [k̲a], the lower and upper bars (dashed and dotted contours) the durations corresponding to the lower and upper ends of the confidence interval of s_1 for [k̲a] (Fig. 2). Note that both Figures 3a and 3b depend on the same mean and confidence intervals shown in Figure 2, i.e. when trajectory shapes vary, durations vary too.

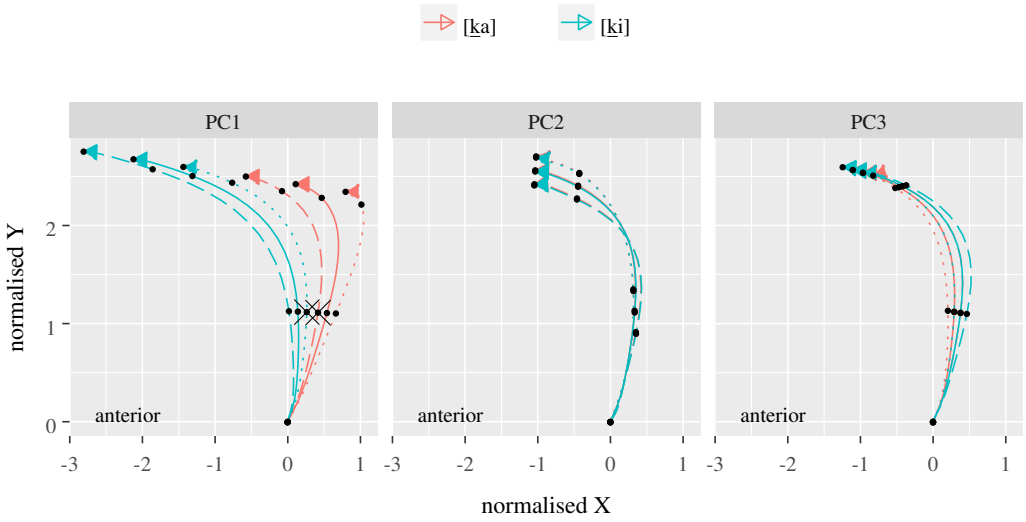
Considering trajectory shapes first, Figure 3a illustrates three independent variations in the trajectory paths along which velar constriction formation in the two vowel contexts may or may not differ from each other. Only PC1 and PC3 matter in distinguishing vowel context, the former much more than the latter, which is why we will mainly discuss PC1. While the velar trajectories start at the same $(0, 0)$ point for both vowel contexts (due to elimination of initial spatial offsets, Section 2.2), their respective boundary trajectories span non-overlapping regions of the (X, Y) plane way before they reach the first time landmark (t_1, i.e. peak velocity of closing movement). With respect to our predictions this confirms that V-to-C coarticulation conditions a very early separation of the tongue dorsum trajectories of [k], close to t_0 (cf. [16, 23]). This separation with the movement in the [i] context taking a more anterior path relative to the [a] context increases until the articulatory target (t_3) is reached, consistent with the existing literature on velar fronting (e.g. [24, 2, 25, 1]).

Considering the spatial differences we computed the separation index as the distance between the rightmost [k̲i] and the leftmost [k̲a] trajectories at time t_1, i.e. the two "×" points in Figure 3a panel PC1. As (X, Y) coordinates have been normalised by speakers' standard deviations (Section 2.2), we unnormalised (X, Y) multiplying by the respective smallest standard deviation, i.e. we took a conservative stand. Even so, the distance is about 3.0 mm, which is above the spatial resolution of the EMA instrumentation (<1 mm), while the distance between the mean trajectories at the same landmark is about 7.7 mm.

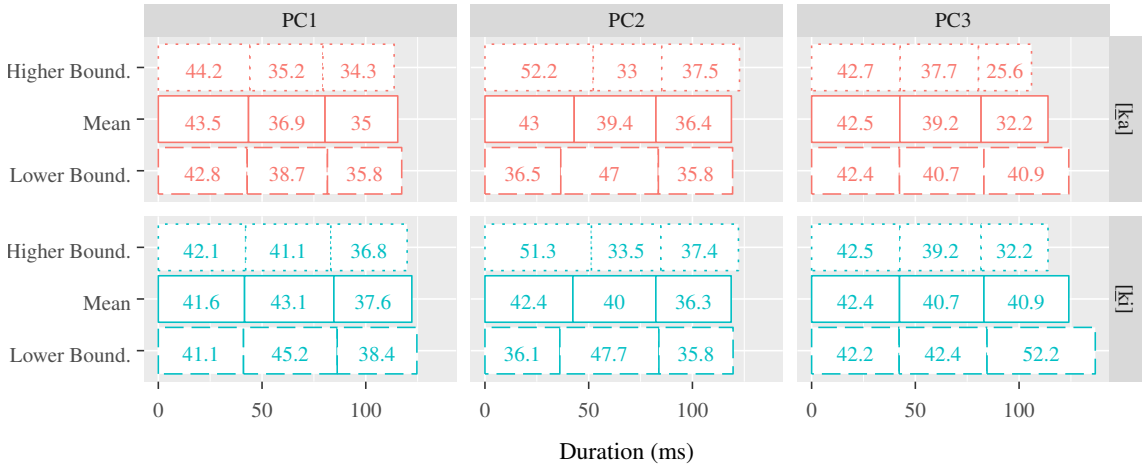
Considering duration differences, panel PC1 in Figure 3b shows that the first landmark is reached after between 41.1 and 42.1 ms for the [i]-context and between 42.8 and 44.2 ms for the [a]-context respectively, which allows us draw two conclusions. First, even the longest of the predicted durations is shorter than 50 ms, which is the benchmark reported in [16] as hypothesised duration at which vowel-induced variability should be apparent. In fact, Figure 3a shows that such differentiation has already occurred *before* that point. Second, by inspecting interval variations in Figure 3b we note that those involving the first interval in PC1 are rather small compared to e.g. the same interval in PC2, which for [k̲a] varies between 36.5 and 52.2 ms. This suggests that the trajectory differentiation induced by the vowel during the first interval of constriction formation is only very mildly correlated to a difference in duration, i.e. speed cannot be a confounding effect.

4. Discussion

In this paper, we demonstrated the joint use of FPCA and LME modelling to characterize tongue dorsum trajectories during the constriction formation of [k] in two vowel contexts. Our analyses confirm previous research which pointed, for independent articulators, to a very early onset of vowel-induced separation in



(a) *Spatial variation.*



(b) *Temporal variation.*

Figure 3: *Mean and boundary trajectories for velar constriction formation in (a) [ka] (red) and [ki] (green), and corresponding interval durations (b) as predicted by PC scores. Solid lines correspond to means (dots) in Figure 2, dashed and dotted lines to lower and higher confidence interval boundaries (left and right ends of segments) in Figure 2, respectively.*

articulator paths. Our results suggest that this separation likely preceeds the 50 ms time point reported by [16], with the estimated effect at peak velocity being at least 3 mm separation. Moreover, we successfully estimated dynamic vowel effects in a single articulator while separating out speaker variability. These results were based on PC1 only. Minor corrections could be applied by considering the shape and duration fluctuations modelled by PC2 and PC3. The two "×" points in Figure 3a can be pulled apart in the Y direction by variations along PC2 (note that the [ki] and [ka] trajectories are almost on top of each other for this PC) and pulled closer in the X direction (in the worst case) by variations along PC3. As these two fluctuations tend to compensate each other, we still maintain 3 mm as a (very) conservative minimum distance at peak velocity. Another point of consideration is the substantial durational fluctuation in the first interval of PC2 (Figure 3b). This fluctuation is the same for both vowel contexts but it suggests that the peak velocity time point may have 50 ms [16] as an upper bound.

5. Conclusions

Tongue dorsum trajectories of [k] in two vowel contexts were characterised using a combination of FPCA and ordinary LME models. This technique proved capable of extracting salient global spatiotemporal trends, as well as of zooming into a particular point of interest (e.g. t_1, peak velocity of closing movement) to produce statistically justified estimates of both location (distance) and timing across speakers. We believe that the potential of this approach lies in its flexibility in considering several spatial and temporal dimensions at once while separating out variability associated with random effects. In future work we will extend this method to higher-dimensional data (e.g. more than one EMA sensor).

6. References

[1] D. Recasens and A. Espinosa, "An articulatory investigation of lingual coarticulatory resistance and aggressiveness for consonants and vowels in Catalan," *Journal of the Acoustical Society of America*, vol. 125, no. 4, pp. 2288–2298, 2009.

[2] S. E. Öhman, "Coarticulation in VCV utterances: Spectrographic measurements," *Journal of the Acoustical Society of America*, vol. 39, no. 1, pp. 151–168, 1966.

[3] W. Chen, Y. Chang, and K. Iskarous, "Vowel coarticulation: Landmark statistics measure vowel aggression," *Journal of the Acoustical Society of America*, vol. 138, no. 2, pp. 1221–1232, 2015.

[4] J. O. Ramsay and B. W. Silverman, *Functional Data Analysis*. Springer Series in Statistics, 2005.

[5] S. N. Wood, *Generalized Additive Models: An Introduction with R*, 2nd ed. Boca Raton, FL: Chapman and Hall/CRC, 2017.

[6] J. O. Ramsay, K. G. Munhall, V. L. Gracco, and D. J. Ostry, "Functional data analyses of lip motion," *Journal of the Acoustical Society of America*, vol. 99, no. 6, pp. 3718–3727, 1996.

[7] J. Lucero and A. Löfqvist, "Measures of articulatory variability in VCV sequences," *Acoustic Research Letters Online*, vol. 6, no. 2, pp. 80–84, 2005.

[8] B. Parrell, S. Lee, and D. Byrd, "Evaluation of prosodic juncture strength using functional data analysis," *Journal of Phonetics*, vol. 41, pp. 442–452, 2013.

[9] C. Mooshammer, L. Bombien, and J. Krivokapic, "Prosodic effects on speech gestures: A shape analysis based on functional data analysis," *Journal of the Acoustical Society of America*, vol. 133, no. 5, p. 3565, 2013.

[10] S. Lee, D. Byrd, and J. Krivokapić, "Functional data analysis of prosodic effects on articulatory timing," *Journal of the Acoustical Society of America*, vol. 119, no. 3, pp. 1666–1671, 2006.

[11] S. Lee, E. Bresch, and S. Narayanan, "An exploratory study of emotional speech production using functional data analysis techniques," *Proc. ISSP*, pp. 11–17, 2006.

[12] S. Schötz, J. Frid, L. Gustafsson, and A. Löfqvist, "Functional Data Analysis of tongue articulation in palatal vowels: Gothenburg and malmöhus Swedish /i:, y:, u-:/," in *Proceedings of Interspeech*, vol. 2013, 2013.

[13] M. Pouplier, J. Cederbaum, P. Hoole, S. Marin, and S. Greven, "Mixed modeling for irregularly sampled and correlated functional data: speech science applications," *Journal of the Acoustical Society of America*, vol. 142, no. 2, pp. 935–946, 2017.

[14] M. Wieling, F. Tomaschek, D. Arnold, M. Tiede, F. Bröker, S. Thiele, S. N. Wood, and R. H. Baayen, "Investigating dialectal differences using articulography," *Journal of Phonetics*, vol. 59, pp. 122–143, 2016.

[15] M. Gubian, L. Boves, and F. Cangemi, "Joint analysis of f0 and speech rate with functional data analysis," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 4972–4975.

[16] A. Löfqvist and V. L. Gracco, "Interarticulator programming in VCV sequences: Lips and tongue movements," *Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1864–1876, 1999.

[17] M. Pastätter and M. Pouplier, "Articulatory mechanisms underlying onset-vowel organization," *Journal of Phonetics*, vol. 65, pp. 1–14, 2017.

[18] Y. Asano and M. Gubian, ""Excuse meeee!!":(Mis) coordination of lexical and paralinguistic prosody in L2 hyperarticulation," *Speech Communication*, vol. 99, pp. 183–200, 2018.

[19] M. Gubian, F. Torreira, and L. Boves, "Using functional data analysis for investigating multidimensional dynamic phonetic contrasts," *Journal of Phonetics*, vol. 49, pp. 16–40, 2015.

[20] B. T. West, K. B. Welch, and A. T. Galecki, *Linear mixed models: a practical guide using statistical software*. Chapman and Hall/CRC, 2014.

[21] S. Nakagawa and H. Schielzeth, "A general and simple method for obtaining R2 from generalized linear mixed-effects models," *Methods in Ecology and Evolution*, vol. 4, no. 2, pp. 133–142, 2013.

[22] P. C. Johnson, "Extension of Nakagawa & Schielzeth's R2GLMM to random slopes models," *Methods in Ecology and Evolution*, vol. 5, no. 9, pp. 944–946, 2014.

[23] H. Nam, L. Goldstein, and E. Saltzman, "Self-organization of syllable structure: A coupled oscillator model," in *Approaches to phonological complexity*, F. Pellegrino, E. Marsico, I. Chitoran, and C. C., Eds. Berlin/New York: Mouton de Gruyter, 2009, pp. 299–328.

[24] D. Abercrombie, *Elements of general phonetics*. Edinburgh: University Press, 1961.

[25] C. A. Fowler and L. Brancazio, "Coarticulatory resistance of American English consonants and its effects on transconsonantal Vowel-to-Vowel coarticulation," *Language and Speech*, vol. 43, pp. 1–41, 2000.