Mixed modeling for irregularly sampled and correlated functional data: Speech science applications

Marianne Pouplier[a)]
Institute of Phonetics and Speech Processing, Ludwig Maximilians University, Munich, Germany

Jona Cederbaum
Department of Statistics, Ludwig Maximilians University, Munich, Germany

Philip Hoole and Stefania Marin
Institute of Phonetics and Speech Processing, Ludwig Maximilians University, Munich, Germany

Sonja Greven
Department of Statistics, Ludwig Maximilians University, Munich, Germany

(Received 30 March 2016; revised 22 May 2017; accepted 20 July 2017; published online 16 August 2017)

The speech sciences often employ complex experimental designs requiring models with multiple covariates and crossed random effects. For curve-like data such as time-varying signals, single-time-point feature extraction is commonly used as data reduction technique to make the data amenable to statistical hypothesis testing, thereby discarding a wealth of information. The present paper discusses the application of functional linear mixed models, a functional analogue to linear mixed models. This type of model allows for the holistic evaluation of curve dynamics for data with complex correlation structures due to repeated measures on subjects and stimulus items. The nonparametric, spline-based estimation technique allows for correlated functional data to be observed irregularly, or even sparsely. This means that information on variation in the temporal domain is preserved. Functional principal component analysis is used for parsimonious data representation and variance decomposition. The basic functionality and usage of the model is illustrated based on several case studies with different data types and experimental designs. The statistical method is broadly applicable to any types of data that consist of groups of curves, whether they are articulatory or acoustic time series data, or generally any types of data suitably modeled based on penalized splines. [http://dx.doi.org/10.1121/1.4998555]

[MAH-J] Pages: 935–946

I. INTRODUCTION

In empirical speech and language research, it is very common for data to consist of time-varying signals, such as fundamental frequency contours, formant, or articulator time series. Speakers typically record multiple repetitions of the same stimulus item. The data thus contain multiple sources of correlation, since curves belonging to the same speaker as well as curves belonging to the same item are inherently correlated, yet it is important to be able to generalize across both speakers and items. The temporal signal evolution itself is usually not directly subjected to statistical analysis due to a lack of statistical methods which allow taking the complex experimental designs, in particular crossed random factors, into account. Statistical methods in the field have undergone a revolution with the advent of linear mixed modeling, precisely because linear mixed models are ideally suited for the most typical design in the speech sciences with crossed random factors, in that they circumvent many of the problematic assumptions of repeated measures analysis of variance (ANOVA) (Baayen *et al.*, 2008; Quené and van den Bergh, 2008; Barr *et al.*, 2013). To be able to use statistical analyses which can mathematically accommodate such complex correlation structures, data reduction methods are commonly employed. For instance, observations are analyzed at one or several "magic moment(s)" (Mücke *et al.*, 2014); i.e., measurement values are extracted at a single, often normalized time point (e.g., formant values at vowel-midpoint, or maximal articulator position), or curves are reduced to a single value by evaluating, e.g., the slope of two connected points, thereby reducing the signal to a straight line. The unfortunate corollary of this is that a large portion of the information contained in the signal is discarded. Analysis time points have to be chosen as an arbitrary working criterion. Often the same statistical model is applied repeatedly on the same data at different measurement time points in order to understand whether observations of, e.g., 25% time point measurements generalize to say, 50 or 75% time points. Also, qualitative observations on signal dynamics of individual speakers, items, or repetitions are used to gauge the degree of generalization of any given single time point analysis. This has to remain unsatisfactory, and qualitative observations have become increasingly unfeasible with the ever increasing amount of data recorded for a single study, enabled by technological advances in recording techniques, storage space, and computation times.

Functional data analysis (FDA) (Ramsay and Silverman, 2005) has provided a significant methodological advance for a global analysis of signal dynamics. Generally, FDA is a data analysis technique that quantifies shape variation

[a)]Electronic mail: pouplier@phonetik.uni-muenchen.de

between curves by operating over functions instead of scalars or vectors as classical multivariate statistical methods do. The basic assumption is that the curves resulting from a finite set of empirical measurements arise from an unknown underlying smooth function plus some observational error. FDA makes no *a priori* assumption about the shape of the underlying function, although often a smoothness restriction is imposed. The empirical observations are either converted to functional data for example via a linear combination of spline basis functions (commonly penalized B-splines), or similar basis expansions are used for each term in a model for the functional data. For this conversion to functional data, the number of basis functions and their locations (knots) have to be specified. A high number of basis functions will be more truthful to the variability in the curve, but a penalty term is used to avoid overfitting (overly wiggly estimation). For more detail on FDA, the reader is referred to, among others, Ramsay *et al.* (2009) and Wang *et al.* (2016); tutorials on the application of FDA to speech data specifically can be found in a series of papers by Gubian and colleagues (Gubian *et al.*, 2009; Gubian *et al.*, 2015). FDA has found broad application in phonetics, yet most methods for functional data assume that curves are independent and thus do not provide a way to account for the complex correlation structure due to repeated observations for speakers and items. Moreover, FDA based methods commonly assume that all curves of a given dataset are observed on a regular grid that is common to all curves, which is often not true for studies in the speech sciences, requiring resampling. We present a tutorial in this paper to functional mixed models for irregularly or sparsely sampled data (sparseFLMM), a recent extension of functional linear mixed models (FLMM; first introduced by Guo (2002)) to irregularly sampled and correlated data (Cederbaum *et al.*, 2016a; Cederbaum *et al.*, 2016b). The main idea of sparseFLMM is to estimate mean functions based on penalized splines, and to use functional principal component analysis (FPCA) to model the functional random effects. The approach is further embedded in the framework of functional additive mixed models [FAMMs; Scheipl *et al.* (2015)] which allows for the computation of confidence bands. Recently, generalized additive mixed models (GAMs) have been used to analyse time series data (Baayen *et al.*, 2010; Wieling *et al.*, 2014; Scheipl *et al.*, 2015; Baayen *et al.*, 2016; Wieling *et al.*, 2016). Our approach can be formulated as a particular kind of generalized additive mixed model by reformulating the functional linear mixed model using appropriate (spline-based) basis expansions for all model terms. We use this fact during estimation. The main difference is that we use functional principal components (FPCs) instead of spline basis functions to expand the functional random effects. We will discuss this further below.

The purpose of the present paper is to illustrate the use of the sparseFLMM method based on several example datasets with different data types [acoustic, electropalatography (EPG), and articulography (EMA) data]. The example datasets all evaluate signal characteristics over time for one or several covariates and require either crossed or simple random effects. For the case studies, only limited detail on the experimental methods for each dataset is given since the main interest is to demonstrate the application of the statistical procedure to different types of data and for different model specifications. The technical details of the statistical approach and methodological validation via simulations are presented in Cederbaum *et al.* (2016a) and Cederbaum *et al.* (2016b). The supplementary online material provided with our paper provides all of the example datasets along with step-by-step instructions for conducting the analyses presented in this paper using the sparseFLMM R package (Cederbaum, 2017) which is available on CRAN (https://CRAN.R-project.org/package=sparseFLMM).[1]

II. FUNCTIONAL MIXED MODELING FOR SPARSELY SAMPLED AND CORRELATED FUNCTIONAL DATA

This section of the paper gives a general overview of the model and the estimation approach, which is then applied to three different datasets with different experimental designs. The model allows for both discrete and continuous covariates. We illustrate both cases in our case studies (Secs. III–V). Note that in the sparseFLMM package, covariate interactions have only been implemented for discrete covariates. The current implementation of the model allows for crossed random intercepts, simple intercepts (one random factor only), and independent curves, but not for random slopes (a point we will return to in Sec. VI).

A. Random effect and covariance estimation

At the heart of the sparseFLMM approach to the estimation of functional random effects is dimension reduction via functional principal component analysis [see, e.g., Ramsay and Silverman (2005)], and it is this aspect that differentiates sparse FLMM from other generalized additive mixed model approaches. Principal component analysis (PCA) is a nonparametric dimension reduction technique designed to extract patterns from high-dimensional multivariate data by finding orthogonal axes of variation in terms of eigenvectors and their associated eigenvalues. The data are projected into a new (lower dimensional) coordinate space with each new coordinate accounting for, in decreasing order, the next greatest amount of variance. The eigenvectors (principal components) of the covariance matrix correspond to the directions of greatest variance. The eigenvalues give the amount of variance which is accounted for by the corresponding eigenvector. Principal components are linear combinations of the original variables weighted by their contribution to explaining the variance in a particular dimension. FPCA analysis is an extension of PCA to functions, therefore the eigenvectors of PCA become eigenfunctions in FPCA. It is then the first eigenfunction of the covariance function that points into the direction of the largest variance of the data, and the variance of the weights for this eigenfunction in the observed curves equals the corresponding eigenvalue. Correspondingly, the second largest eigenfunction is orthogonal to the first eigenfunction, and points in the direction of the second largest spread of the data.

In the sparseFLMM model, the random effects functions are expressed as weighted sums of the eigenfunctions (FPCs) of their covariance operators (in the functional case,

the covariance matrix becomes a covariance operator). This contrasts with other approaches such as Scheipl *et al.* (2015), which use a spline basis to model random effects. [Scheipl *et al.* (2015) also provide an estimation approach based on FPCs if these FPCs are provided by the user, and it is this variant that we later use during estimation after we have estimated the FPCs from the data.] Using FPCs to expand the functional random effects, as done in sparseFLMM, has several advantages over spline basis functions. For one, FPCs have an optimal approximation property. Specifically, for a given basis size and e.g., random intercepts, the best approximation is obtained using an FPC basis expansion. This will be in particular better than a spline basis expansion with the same number of basis functions. This leads to small and parsimonious bases and thus much reduced computation times and memory footprints. Cederbaum *et al.* (2016a) compared the FPC-based and spline based estimation and found reduced computation times, e.g., 7–8 h vs 10 days in an example. The FPC approach also yields better estimation quality. Cederbaum *et al.* (2016a) found between 2.8 and 6 times lower root relative mean squared errors for the covariate effects for our approach compared to a spline based additive model, and also lower errors for other model components. Second, FPCs provide an interpretable decomposition of the variance, and the FPC weights can be used for further analysis or exploration of the data: the model allows us to assess how much variability is explained by which model component, and how many FPCs are needed for each random effect to explain a given amount of variability in the data, as will be illustrated in our case studies below. A variance threshold is used to choose the number of FPCs to describe the random functions. By weighting a given principal component function with the coefficients for the observations, the variance or spread of the observations in a particular dimension can be observed. The weights can in principle be used for further analysis, such as to probe the data for groupings in the random effects (e.g., speakers and items). It has to be kept in mind though that FPCA calculates orthogonal axes of variation in terms of eigenfunctions and their associated eigenvalues from overall distributional properties of the data independently of the variables used to construct the dataset. This means that it is not guaranteed that any grouping effects of, say, speakers in the weights of a given FPC, has an experimentally relevant interpretation in terms of speaker properties.

As criterion for the selection of the number of FPCs, the cumulative percentage of explained variance is used. The explained variance is calculated for all functional random effects in the model and the FPCs are ranked across all functional random effects according to the percent variance they account for. The FPCs are then selected one-by-one in rank order until the specified level of variance (e.g., 95%) is accounted for. The threshold is usually set such that the main FPCs accounting for the biggest amount of variance are included in the model and may typically range between 95 and 99%. In contrast to the random effects, sparseFLMM estimates the fixed effects on the basis of splines. Note that the effects of covariates are solely estimated as fixed effects in the current implementation of the model.

Another distinctive advantage of sparseFLMM is that it not only estimates the effects of the covariates and their interactions in crossed designs, but also allows for irregular spacing of the (possibly very few) observation points. This is achieved by estimating all curves simultaneously; thus, estimation for any given curve is informed by the overall data distribution. As a corollary, the number and the location of the observation points can differ between curves, and the observation points of a given curve do not have to be equally spaced. Point-wise confidence bands enable the evaluation of significant differences between groups of curves. Due to penalized splines being employed for mean and covariance estimation, there are no underlying assumptions about the shape and properties of the functions apart from an underlying smoothness. This means that the model is generally applicable for all types of data that can be suitably approximated using splines. The possibility of irregular grids for the temporal part of the model and the possibility of modeling crossed functional random effects are in their combination the features differentiating the current model from related developments, such as wavelet-based models on equal grids (Morris and Carol, 2006; Lancia *et al.*, 2015) or FDA approaches for independent curves or non-crossed designs (e.g., Di *et al.*, 2014; Gubian *et al.*, 2015; Guo, 2002). As already mentioned, the sparseFLMM approach is closely related to GAMs which have been applied to time series data (Baayen *et al.*, 2010; Wieling *et al.*, 2014; Scheipl *et al.*, 2015; Baayen *et al.*, 2016). In GAMs, it is possible to model individual variation by speaker and by item using factor smooths in the *bam* function of the R mgcv package (Wood, 2011; Wood *et al.*, 2015). Using, as done by the sparseFLMM approach, FPC bases as a parsimonious representation of the functional random effects instead of spline bases, as done by these other approaches, provides an interpretable variance decomposition for the random terms in the model and increases computational efficiency. Moreover, using FPCA the basis functions are estimated from the data as the eigenfunctions of the estimated covariance of the functional random effects (see also Wang *et al.*, 2016). GAMs as proposed by Baayen *et al.* (2010) and Baayen *et al.* (2016) assume that the error is autocorrelated with a specific parametric first order autoregressive [AR(1)] structure with a fixed correlation parameter (rho), which has to be set as an arbitrary working criterion by the researcher. This may lead to incorrect standard errors and thus incorrect inference. In contrast to this, sparseFLMM has the distinct advantage of estimating the auto-covariance of the error from the data, which allows the error to be heteroscedastic and/or vary nonparametrically over the time interval, giving more reliable inference in this respect. While we also use penalized splines for covariance estimation, we can increase the number of spline basis functions in this step to allow for more flexibility than is usually possible (for computational reasons) in approaches directly expanding the random effects in splines. For a further discussion on and comparison of differences between the current model and other approaches, see Cederbaum *et al.* (2016a).

The general form of the model with crossed random effects for Speaker and Item is given in Eq. (1),

$$Y_{ijh}(t) = \mu(t, \boldsymbol{x}_{ijh}) + B_i(t) + C_j(t) + E_{ijh}(t) + \varepsilon_{ijh}(t), \quad (1)$$

with $Y_{ijh}(t)$ being the index curve for speaker i, item j, and repetition h observed at time $t \in T = [0,1]$. $\mu(t, \boldsymbol{x}_{ijh})$ is a curve specific smooth mean function, $\boldsymbol{x}_{ijh}$ are known covariates and possible interactions of covariates. $B_i(t)$ and $C_j(t)$ are random functional intercepts for Speaker and Item, respectively. $E_{ijh}(t)$ is a Speaker-, Item-, and Repetition-specific smooth random deviation and also includes the interaction between Speaker and Item. $\varepsilon_{ijh}(t)$ is white noise measurement error and captures random uncorrelated variation within each curve. The mean function $\mu(t, \boldsymbol{x}_{ijh})$ includes, in the case of p covariates, $p+1$ effects in the form of the reference mean ($f_0(t)$), p covariate fixed effects ($f_1(t),\ldots, f_p(t)$), and their interactions. Equation (2) gives a case with four effects, i.e., two item covariates and their interaction

$$\mu(t, \boldsymbol{x}_{ijh}) = f_0(t) + f_1(t) \cdot covariate1_j + f_2(t) \cdot covariate2_j + f_3(t) \cdot covariate1_j \cdot covariate2_j, \quad (2)$$

where $f_0(t), f_1(t),\ldots, f_3(t)$ are unknown fixed functions. A more complex case with eight effects is provided in Cederbaum *et al.* (2016a). Note that the covariate effects are linear at each time point t and thus can be interpreted in the usual way when fixing a time point. In particular, if we fix a set of time points, the linear covariate effect estimates and confidence intervals for those time points correspond to the kind of result one would obtain from a "magic moment" analysis, but without the need to restrict the analysis to those time points or to specify them in advance.

B. Estimation procedure

Estimation is conducted in several steps. First, the mean effects are estimated using penalized splines under a working independence assumption of the curves and time points, and the curves are centered with the estimated mean. To avoid the choice of the number of basis functions, we choose a penalized spline basis approach for our covariate effects. Then, the number of basis functions should not make a great difference to the estimation as long as a sufficiently high number is used, see, e.g., Ruppert (2002). A higher number of basis functions should not lead to overfitting due to the penalization parameters, but it will mostly affect computation times. Using mgcv, the model estimates the smoothing parameter using restricted maximum likelihood (Wood *et al.*, 2015), i.e., the optimal amount of smoothness is estimated from the data using a maximum likelihood based method. At the next step, the auto-covariances of the functional random effects (and the error variance) are obtained, again using penalized spline smoothing. For our case studies, we use the third order derivative as order of the penalty for covariance estimation of the functional random intercepts. This can be motivated in general for speech production data by the fact that jerk (which, as the derivative of acceleration, corresponds to the third derivative with respect to position) minimization is generally taken to be an important parameter in movement optimization (Nelson, 1983; Wada *et al.*, 1995). Eigen decompositions of the covariances are conducted to obtain the FPCs for each functional random effect. The FPCs are then selected with respect to the user-specified variance threshold. Finally, the complete model [Eq. (1)] can be estimated, using these FPC bases to expand the random effects, and thus accounting for the correlation structure (between subjects, items, and near-by time points) in the inference for the mean; that is, incorporating the functional random effects in the estimation of the covariate effects and computation of the confidence bands. The effects of the covariates are solely estimated as fixed effects. The model outputs estimated covariate-specific changes to the mean function; i.e., prototypical curves over time which represent the covariate effects. Point-wise (not simultaneous) confidence bands are used to evaluate significance. Point-wise significance means that a test for the curve value being equal to zero at a given time point would be rejected at the specified alpha level (e.g., 0.05), but that multiple testing across time points is not taken into consideration. Significance is claimed when the confidence band does not include zero at a given time-point. Confidence bands do not take into account the variability in the estimated FPCs (they are "conditional" on the estimated FPCA), but simulations in Cederbaum *et al.* (2016a) and Cederbaum *et al.* (2016b) show that coverage of the true value is still close to the nominal level (e.g., 95%).

The case studies in the remainder of the paper illustrate how to interpret estimated mean functions for the covariates and their interactions, as well as the decomposition of variance.

III. FORMANT DATA OF ROMANIAN VOWEL SEQUENCES

The first case illustrates the evaluation of formant dynamics based on Romanian vowel sequences. These data require a single covariate only but crossed random effects of Subject and Item. Romanian features a typologically uncommon diphthong–hiatus contrast for the vowel sequences /ea/ and /oa/ [see Gubian *et al.* (2015) for the investigation of a similar contrast in Spanish using FDA]. Previous modelling work using articulatory synthesis has proposed that the hiatus /e.a/ and the diphthong /ea/ differ in their temporal coarticulation pattern with /e.a/ showing a lesser degree of coarticulation between the two members of the vowel sequence compared to the diphthong (Marin and Goldstein, 2012). We compare here the formant dynamics of the /ea/ – /e.a/ contrast for five Romanian speakers. The data were acquired as part of the work reported in Marin (2014), which compared mid and high vowel and diphthong sequences articulatorily by measuring five different time points in each vowel sequence for each comparison with a corresponding number of statistical tests. The publication includes an articulatory analysis for two out of the three diphthong–hiatus pairs analyzed here. The full experimental protocol and more information on the diphthong–hiatus inventory of Romanian can also be found therein. Of the experimental participants, three speakers were female, and two were male. The stimulus material consisted of the diphthong-hiatus pairs *teama* vs *te am* (clitic+auxiliary sequence); *seara* vs *seat*; *cafea fină* vs *gafe afine* with the first member of each pair being the diphthong. Each speaker repeated the stimulus items

six times over the course of the experiment. The stimulus items were embedded in a neutral carrier phrase. The targeted token total was 6 words × 5 speakers × 6 repetitions = 180. Due to recording failure or undetected mispronunciations, 17 repetitions are missing leaving $N = 163$ tokens. For these, we obtained, via LPC analysis, formant time series for the first two formants ($F1$, $F2$) for the /ea/ part of each token. Formants were tracked using the AUDAPTER software package developed by Cai *et al.* (2008). LPC order (and fundamental frequency to control cepstral smoothing) were specified individually for each speaker. Figure 1 gives the average formant curves with 1 standard deviation (SD) bands. SDs were computed point-wise across all curves for $F1$, $F2$ separately.

For the dataset, the number of equidistant points per curve varies between 28 and 135 (median: 65). Using sparseFLMM, it was possible to enter the formant curves for all speakers into the statistics without correction for vocal tract size differences, nor were the formant time series smoothed or resampled prior to statistical analysis. Since our main point is a methodological one, we restrict our analysis to $F2$ here. The model only had a single covariate (Status) since we wanted to know whether there is a difference in $F2$ dynamics between diphthong and hiatus. For categorical covariates, the code requires dummy coding; here, zero was given to the diphthong. This means that the reference group mean corresponds to Status Diphthong (stimulus items *teama*, *seara*, *cafea finǎ*).

The model has a random effects structure as specified in Eq. (1) in Sec. II with $i = 1,\ldots, 5$, $j = 1,\ldots, 6$, and $h = 1,\ldots, 6$. The mean function $\mu(t, \boldsymbol{x}_{ijh})$, given in Eq. (3), includes the reference mean and one covariate (Status)

$$\mu(t, \boldsymbol{x}_{\mathrm{ijh}}) = f_0(t) + f_1(t) \cdot Status_j, \tag{3}$$

where $f_0(t)$ and $f_1(t)$ are unknown fixed functions.

The model estimates are evaluated via plots of the covariate effects and are for interpretation best visualized in a summed effects plot as given in Fig. 2. Figure 2(a) gives the reference mean $f_0(t)$ (here, the diphthong condition), and

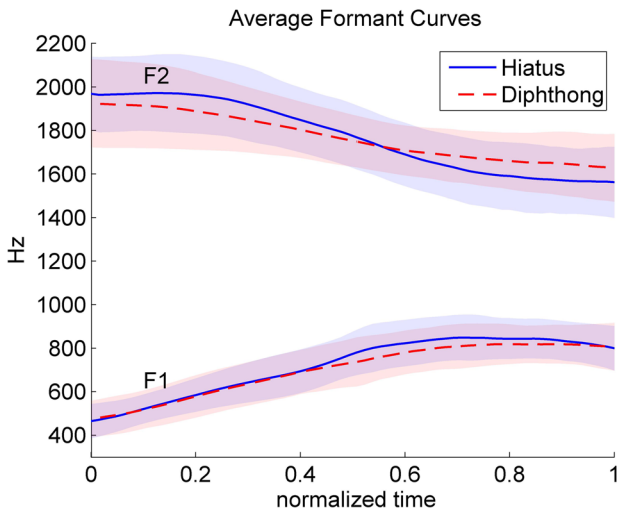

FIG. 1. (Color online) Average $F1$, $F2$ curves for the Romanian vowel sequences by Status (diphthong, hiatus) with 1 standard deviation bands. Note that the averages are across all speakers, male and female.

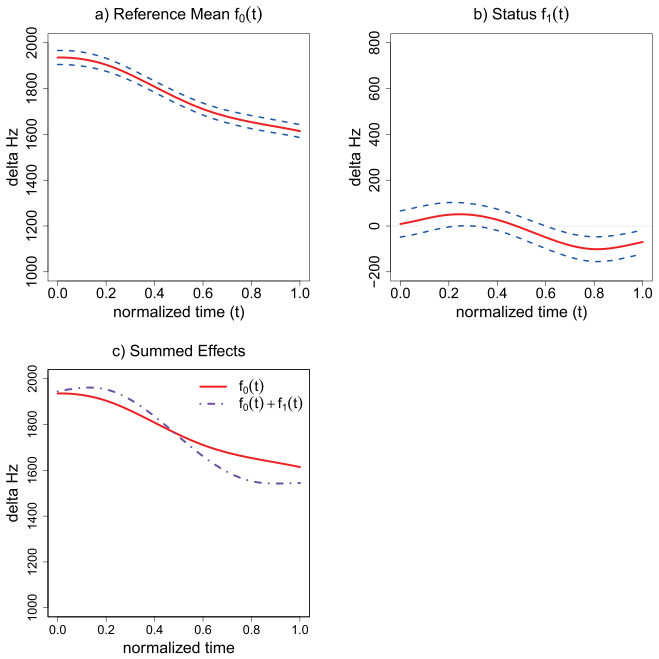

FIG. 2. (Color online) Results for $F2$. Covariate mean effect (solid line) with conditional point-wise confidence bands (dashed line). (a) Reference mean ($f_0(t)$), (b) covariate effect Status ($f_1(t)$), (c) summed estimated effects curves (solid line $f_0(t)$: diphthong; dashed-dotted line $f_0(t) + f_1(t)$: hiatus).

the effect of the covariate $f_1(t)$ with conditional point-wise confidence bands is in Fig. 2(b). The effect on the formant curve is denoted by delta Hz. The reference mean displays a clear $F2$ falling dynamic as to be expected for an /ea/ diphthong [Fig. 2(a)]. The effect of a covariate on the mean is obtained by multiplying the covariate effect with the dummy coding (1 or 0) of that covariate and adding it to the reference mean. This is precisely what is shown in a summed effects plot [Fig. 2(c)]. Recall that significance is claimed when the confidence band of the covariate effect plot [Fig. 2(b)] does not include zero at a given time-point. In $F2$, we observe a significant difference between diphthong and hiatus: the confidence band for the covariate effect [Fig. 2(b)] does not cover zero from about 0.65 normalized time onwards, that is, during the /a/ target of the diphthong. At around 0.25, the confidence band is just above the zero line; the lower limit of the confidence band at timepoint 0.25 is 0.3. We might call this borderline significance. Note that confidence bands are narrower than the 1 SD bands in Fig. 1 due to accounting for the correlation structure in the data.

The hiatus has a more extreme (lower) $F2$ target for /a/ compared to the diphthong. The /e/ targets of diphthong and hiatus do differ significantly from each other at trend level (there is, as mentioned above, a trend for significance at about 0.25) with a tendency for a more extreme /e/ target for the hiatus. Overall, the results mean that there is less carryover (e-to-a) coarticulation in the hiatus compared to the diphthong. Note in this context in particular how the formants have no static portion at all, but rather continuously rising and falling curves. While these curve dynamics can be investigated by measuring, e.g., five different time points in each vowel sequence as done for articulatory data by Marin (2014), this entails running five statistical models per comparison and

carefully choosing the time-points *a priori* to fall into informative time regions.

We now turn to the decomposition of variance of the random effects. Functional random effect Speaker accounts for 43% of the variance (3 FPCs selected by the model), Item for only 3.71% (1 FPC selected), the smooth error term E, which subsumes both a speaker-by-item interaction and a repetition-specific deviation, accounts for 46% (3 FPCs selected). A way to evaluate the functional random effects further is given in Fig. 3 for the speaker specific deviation. Figure 3(a) plots the global mean and how positive (+) and negative (–) weights on FPC1 capture speaker-specific differences from the global mean. Note that these plots give the global mean, which corresponds to all covariates being set to the empirical average, *not* the reference mean. Speakers with a positive weight for FPC1 show an $F2$ curve below the global mean, while the $F2$ curve for speakers with a negative weight for FPC1 lies above the global mean. Figure 3(b) plots the estimated means per speaker. Of the five participants, speakers 3 and 4 were male, the others female. There is some tendency for $F2$ curves for the male participants to be below the global mean, while the female participants are above the mean as one would expect from typical vocal-tract size differences between male and female speakers. Speaker 4 is below the global mean initially, but then has a rather flat curve compared to the other participants. This may be related to this participant speaking noticeably faster than the other speakers, meaning his data may show undershoot. The first FPC selected by the model for random effect speaker thus seems to capture some of the sex related vocal-tract size differences but not unequivocally so; possibly also differences in speech rate (related to undershoot) factor into this FPC.

We now apply the same plotting procedures to the random effect Item. Our experiment included three different diphthong–hiatus pairs, and thus three items per condition (three diphthong, three hiatus items). Figure 4 visualizes the weights of the FPC associated with the Item effect [Fig. 4(a)] and the estimated item-specific means [Fig. 4(b)]. It is noticeable in Fig. 4(b) that items 5 and 6 have somewhat steeper $F2$ trajectories mainly due to a higher initial starting point of the curves, whereas items 1–4 are very similar. Items 5 and 6 are the two-word stimuli consisting of a noun + adjective phrase (5: *cafea fină*, 6: *gafe afine*). Interestingly both diphthong (5) and hiatus (6) diverge similarly from both the global mean

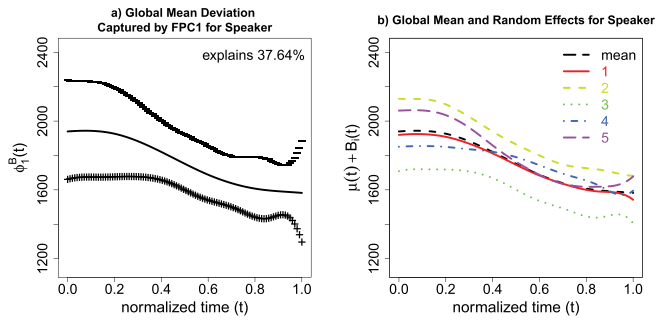

FIG. 3. (Color online) Variance decomposition, random effect Speaker. (a) Global mean function and effect of adding (+) and subtracting (−) a suitable multiple (twice the square root of the eigenvalue) of FPC1 for random effect Speaker. (b) Global mean and estimated mean curves for the five speakers.

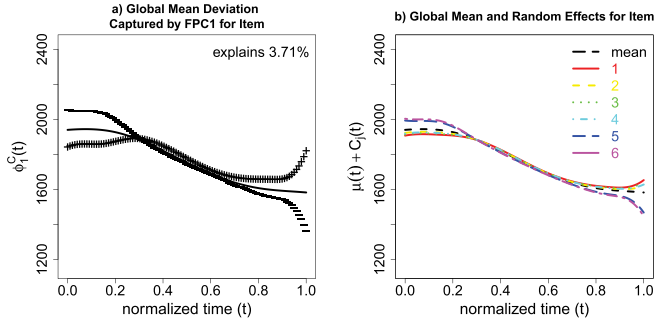

FIG. 4. (Color online) Variance decomposition, FPC1 of random effect Item. (a) Global mean function and effect of adding (+) and subtracting (−) a suitable multiple (twice the square root of the eigenvalue) of FPC1 for random effect Item. (b) Global mean and estimated mean curves for the six stimulus items (from 1 to 6: *teama, te am, seara, seat, cafea fină, gafe afine*).

and the other stimulus items, suggesting that it is either the difference in syntactic structure or the higher number of syllables that may be captured by this FPC.

We now turn to the second case study, which illustrates the application of sparseFLMM to EPG data and the use of a model with a covariate interaction.

IV. EPG DATA OF GERMAN FRICATIVE SEQUENCES

The study we discuss here consists of EPG data of German fricative sequences, which was recorded as part of a larger study looking at fricative sequences in German (Pouplier and Hoole, 2016). We illustrate how a more complex experimental design with crossed random effects, two covariates, and their interaction can be evaluated using sparseFLMM. The details of data acquisition and treatment can be found in Pouplier and Hoole (2016), where magic-moment analyses of these data have been published. For present purposes, we contrast the fricative sequences /f#ʃ/ and /s#ʃ/ and ask whether the extent of anticipatory coarticulation varies as a function of $C1$ (/f/ vs /s/) and lexical stress.

The dataset comprises EPG data from nine native speakers of German. Stimuli consisted of (semantically nonsensical) noun-noun compound phrases with abutting medial fricatives sequences /f#ʃ/ and /s#ʃ/; the compounds were embedded in a neutral carrier sentence. The stimuli varied according to whether the V_1C_1 syllable of the $V_1C_1\#C_2V_2$ sequence bore lexical stress or not. There were four items per condition. Palatograms were extracted for each token during the acoustically identified $C1\#C2$ interval. The EPG data were sampled at 200 Hz. For EPG data, each sample consists of a binary on/off contact pattern for 62 electrodes. Prior to statistical modelling, these palatograms underwent a normalization procedure which mapped the 62 contact patterns of each sample onto a single valued, time-varying similarity index such that any given sample can take on a value ranging from −1 to 1, indicating its Euclidean distance to prototypical /f, s, ʃ/. The prototypes were computed in a speaker and condition specific manner based on homorganic control conditions (/f#f/, /s#s/, /ʃ#ʃ/). The similarity index was computed such that 1 represents a prototypical realization of $C1$ (/f/ for the /f#ʃ/ condition, /s/ for the /s#ʃ/ condition), a value of −1

represents a prototypical realization of $C2$ (/ʃ/). This means that for any given sample, the similarity index quantifies, for the /s#ʃ/ condition, how close this sample is to /s/ (ideal value of 1) or /ʃ/ (ideal value of −1) or, for the /f#ʃ/ condition, how close any given sample is to /f/ (ideal value of 1) or /ʃ/ (ideal value of −1). A value of zero represents an EPG pattern intermediate between the two prototypical reference patterns (see Fig. 5). To the extent that there is anticipatory coarticulation of /ʃ/ during the $C1$, the index will be smaller than 1 during the initial parts of the fricative interval. For further detail on the normalization procedure, the reader is referred to Pouplier and Hoole (2016).

In sum, the data consist of curves representing a time-varying similarity index computed from EPG data. The question we pursue here is whether there is an effect of $C1$ in the coarticulatory patterns of $V_1C_1\#C_2V_2$ fricative sequences in interaction with prosody (lexical stress of the first syllable). The expectation is that there is more anticipatory coarticulation in $C1$ = /f/ and in the unstressed condition. This is so because the tongue is free to anticipate $C2$ during a labial $C1$ /f/ but not during a lingual $C1$ /s/ consonant and because unstressed positions are known to be articulatorily "weaker" compared to stressed positions. The average curves given in Fig. 5 are consistent with there being more anticipatory coarticulation for the labial $C1$ with the /f#ʃ/ curve showing negative (/ʃ/-like) index values earlier in time compared to the /s#ʃ/ curve.

The statistical model has two covariates: $C1$ (/f/, /s/) and Stress of $C1$ (stressed, unstressed), plus their interaction, and crossed random effects for Subjects and Items. The model equation and mean function thus correspond to Eqs. (1) and (2) given in Sec. II. Each $C1$-Stress combination is represented by four stimulus items, and there are five repetitions per stimulus. Targeted token total was 9 speakers × 2 $C1$ conditions × 2 stress conditions × 4 items × 5 repetitions = 720, minus token loss due to technical failure resulting in a dataset of $n = 709$ curves. All trials were mapped onto a time interval of [0,1] without resampling to a regular grid. The data consist of 24–65 equidistant measurements (median = 34) per curve with different spacings across curves.

The two independent variables were dummy coded into two covariates: stress of the first syllable: 0 = stressed, 1 = unstressed, and $C1$: 0 = /s/ and 1 = /f/. The reference condition for the following analyses corresponds to all dummy codings being zero, i.e., /s#ʃ/ stressed. Figure 6 shows the mean effects of the covariates and their interaction with conditional point-wise confidence bands. The effect on the index curve is denoted by delta index. As explained before, the effect of a covariate on the mean can be obtained by multiplying the covariate effect with the dummy coding (1 or 0) and adding it to the reference group mean. To evaluate the effect of an interaction, the effect curves of the two covariates of interest and their interaction have to be added to the reference mean curve. Relevant comparisons are then of the sum of all four curves (reference mean + covariate1 + covariate2 + covariate1 · covariate2 effect, meaning here: reference mean + $C1$ + Stress + $C1$ · Stress effect) to the sum of the reference mean and only the covariate1 ($C1$), or only the covariate2 (Stress) effect. The relevant confidence band is the one of the interaction curve [here, $f_3(t)$ in Eq. (2)]. Significance is claimed if the confidence bands do not include zero: in this case (Fig. 6), $C1$ identity is significant in the middle part of the interval [Fig. 6(b)], Stress shows a tendency for significance at the very beginning of the time interval [Fig. 6(c)], and the interaction is significant in the second part of the interval [Fig. 6(d)]. The summed effects curves are illustrated in Fig. 7.

It can be seen from the effects plot in Fig. 6(b) that the difference between $C1$ = /s/ and $C1$ = /f/ is significant

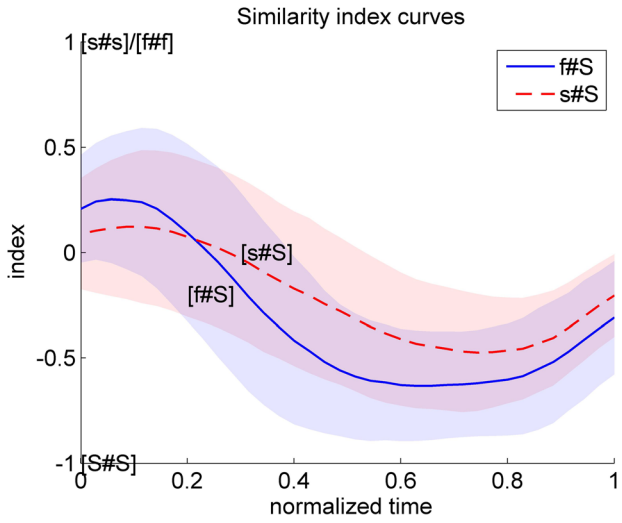

FIG. 5. (Color online) Mean index curves with 1 SD bands for /s#ʃ/ (dashed line) and /f#ʃ/ (solid line), averaged across all speakers and items. The SD was computed point-wise across all curves separately for /s#ʃ/ and /f#ʃ/. Capital ⟨S⟩ is used to denote /ʃ/.

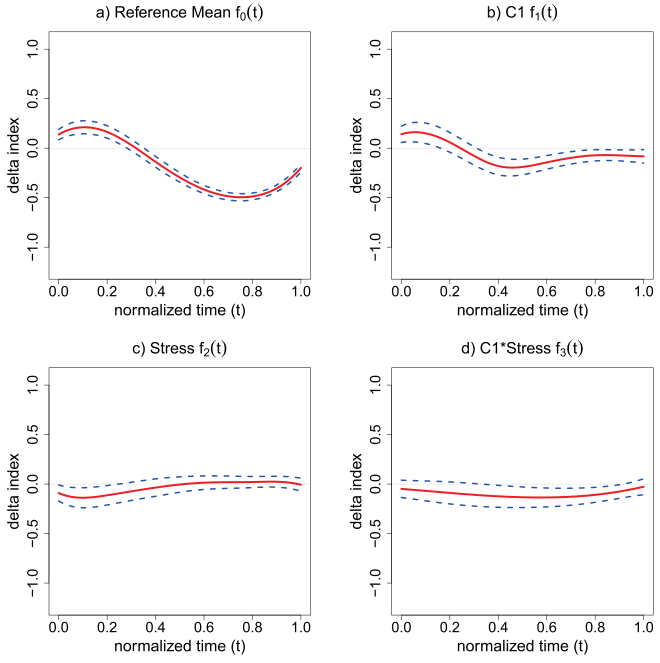

FIG. 6. (Color online) Covariate mean effects (red solid lines) with conditional point-wise confidence bands (dashed lines). (a) Reference mean ($f_0(t)$); (b) covariate effects $C1$ ($f_1(t)$); (c) Stress ($f_2(t)$), and (d) interaction effect for $C1$ · Stress ($f_3(t)$).

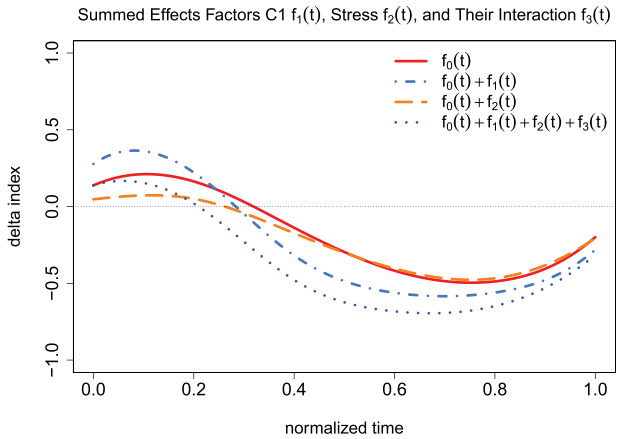

FIG. 7. (Color online) Summed effects curves for the covariate combinations $C1$ and Stress. Solid line ($f_0(t)$): /s#ʃ/ stressed; dotted-dashed line ($f_0(t) + f_1(t)$): /f#ʃ/ stressed; dashed line ($f_0(t) + f_2(t)$): /s#ʃ/ unstressed; dotted line ($f_0(t) + f_1(t) + f_2(t) + f_3(t)$): /f#ʃ/ unstressed.

between roughly 0.4–0.7 normalized time. In Fig. 7, it is evident that this is due to an earlier stronger constriction formation for /ʃ/ in /f#ʃ/ compared to /s#ʃ/. At the same time, the significant difference at the beginning of the $C1$ effects curve shows that /s#ʃ/ has a lower index value compared to /f#ʃ/, which indicates that /ʃ/ exerts a subtle influence on the spatial properies of /s/ early on in the constriction, as to be expected from the interaction of two lingual consonants.

Adding the effect curve of Stress [Fig. 6(c)] to the reference (/s#ʃ/) mean reveals that the unstressed condition has an index lowering effect for /s#ʃ/: the $f_0(t)$ curve has a higher index value compared to $f_0 + f_2(t)$ initially, and the confidence band confirms this as significant in Fig. 6(c). The interaction effect can be visualized by comparing the space between relevant pairs of summed effects curves in Fig. 7. Throughout the entire time interval, the distance between the solid (red) line /s#ʃ/ reference mean curve ($f_0(t)$), and the dashed (orange) line $f_0(t) + \text{Stress}$ ($f_2(t)$) curve is smaller compared to the distance between the (light blue) dashed-dotted /f#ʃ/ ($f_0(t) + C1$ ($f_1(t)$) curve and the (dark blue) dotted curve $f_0(t) + C1$ ($f_1(t)$) $+ \text{Stress}$ ($f_2(t)$) $+ \text{interaction}$ ($f_3(t)$). The interaction is (barely) significant roughly after 0.4 normalized time [Fig. 6(d)]. Changing the lexical stress level of the first syllable from stressed to unstressed reduces the index values for /f#ʃ/ but not for /s#ʃ/. For /f#ʃ/, this means that the index/EPG pattern is pulled towards the /ʃ/ reference pattern throughout, indicating that there is significantly stronger anticipatory coarticulation in unstressed /f#ʃ/ compared to stressed /f#ʃ/.

In terms of variance decomposition for the random effects, two FPCs were selected for Speaker (accounting for 11.7% of the variance), one for Item (2.8% variance accounted for), and five for the smooth error term E (Speaker $\times$ Item $\times$ Repetition, 76.7% variance accounted for). This means that the item effects are rather weak in the present dataset even though each $C1 \cdot$ Stress combination was represented by four different stimulus items. The lion's share of the variability is accounted for by error term E, and here it is the first of the four chosen FPCs (FPC1) that accounts for as much as 51.19% of the variance. Figure 8 plots the effect of

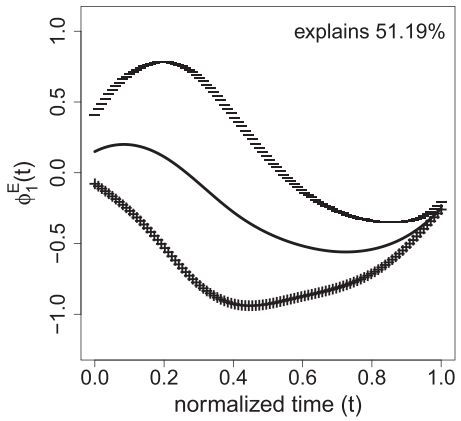

FIG. 8. Global mean function and effect of adding (+) and subtracting (−) a suitable multiple (twice the square root of the eigenvalue) of FPC1 for functional random error term E.

adding and subtracting a suitable multiple of this FPC to the mean. This FPC suggests very strong idiosyncratic speaker-item-repetition specific coarticulation patterns in both directions (anticipatory, carryover).

In cases in which more than one FPC is selected for a functional random effect, it can be useful to look at a scatterplot for the FPC weights for two of the FPCs in order to gauge the interpretability of the FPCs in two dimensions of variability. To illustrate this, we now look at the FPC weights for Speaker in more detail. Figures 9(a) and 9(b) illustrate the effect of adding and subtracting a suitable multiple of FPCs 1 and 2 selected for functional random effect Speaker to the mean and additionally gives the weights of these two FPCs selected for Speaker as a scatterplot [Fig. 9(c)]. The estimated mean curves for some of the speakers with different weights

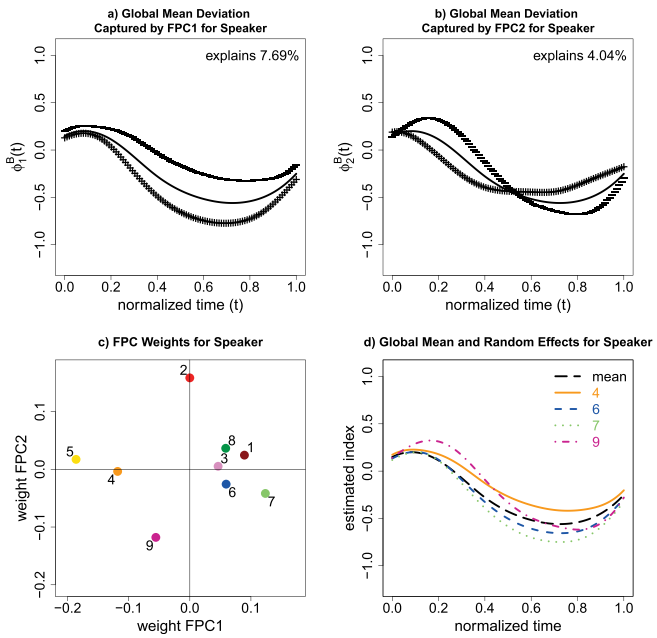

FIG. 9. (Color online) (a), (b) Global mean function and effect of adding (+) and subtracting (−) a suitable multiple (twice the square root of the eigenvalue) of FPC1 (a) and FPC2 (b) for random effect Speaker, (c) scatterplot of the weights for FPC1 and FPC2 for random effect Speaker; (b) estimated mean curves for speakers 4, 6, 7, 9, and the global mean.

on the two FPCs are also shown [Fig. 9(d)]. For instance, speakers 6 and 7 [positive weight on FPC1 in Fig. 9(c)], show more anticipatory coarticulation than, e.g., speakers 4 and 9 (positive weight on FPC1). Gender and dialect come to mind as potential grouping factors, but they do not seem to be very relevant for our sample at hand. Speaker 5 was the only male in our sample. Regionally 5, 8 were of Northern German origin, speakers 1, 3, 6 from Swabian dialect areas, 2, 7, 9 from dialectal Bavaria, and speaker 4 from Hesse. While the variance decomposition for random effect Speaker in case of the Romanian vowels seemed to have a certain degree of interpretability in terms of (presumed) vocal tract size and speech rate differences, for the current dataset there is no obvious interpretation of the grouping in terms of speaker characteristics.

V. FREQUENCY EFFECTS IN ARTICULOGRAPHY DATA

In this third case study, we illustrate the application of the sparseFLMM model to EMA data for the case of a continuous covariate. We evaluate variation in the relative timing of the two consonants of Russian *C*1*C*2*V* onset clusters as a function of cluster frequency. The data consist of EMA data from 11 Russian speakers. For details of data acquisition and treatment, the reader is referred to Pouplier *et al.* (2015) and Pouplier *et al.* (2017). The speakers recorded *C*1*C*2*V* syllables embedded in a constant carrier phrase. Five repetitions were recorded at two speech rates ("slow," "fast"). We include only the fast condition here, since the current sparseFLMM implementation does not allow for interactions involving continuous covariates. Table I lists the stimuli, which consisted of 12 different syllables of differing token frequency. The numbers in parentheses first show the log frequency determined based on the Russian Internet Corpus (Sharoff, 2005), and second the number of tokens available. Due to technical failure during data recording, the number of tokens per cluster can differ.

For this type of data, a common research question concerns the relative timing between the successive consonants. Typically, so-called landmark segmentation is employed: based on the temporal evolution of the velocity profile, time points such as movement onset or target achievement are determined (see, among many others, e.g., Marin and Pouplier, 2014). Relative timing is then evaluated by simple landmark subtraction, for example, by calculating the temporal interval between release of *C*1 and target achievement of *C*2. This entails that spatial variation, which may be conditioned by the temporal overlap, is completely left aside and can only be assessed independently of the temporal measurements; also the general temporal evolution dynamics cannot adequately be taken into consideration.

TABLE I. Stimuli in order of decreasing frequency from top left to bottom right. Numbers in parentheses give log frequency and available number of tokens per stimulus.

gla (12.7, $n=52$)	mla (9.5, $n=54$)	ʃpa (7.2, $n=55$)
mno (12.4, $n=53$)	tka (8.9, $n=48$)	lga (7.0, $n=52$)
kto (12.1, $n=46$)	lba (8.8, $n=55$)	ʃma (6.5, $n=55$)
bla (11.5, $n=53$)	xma (7.9, $n=31$)	mxa (6.0, $n=22$)

To evaluate relative timing patterns of independent articulators, curve statistics have, to our knowledge, usually not been employed since relative timing analyses require to evaluate the evolution of temporally overlapping curves with quite different spatial coordinates (due to the anatomical differences in articulator location and movement range). For our present dataset, the idea is to circumvent this problem by focusing on the movement kinematics of *C*2 only, but in the time interval in which *C*1 and *C*2 potentially co-exist. This approach essentially quantifies how *C*2 evolves within the movement cycle of *C*1.

For data extraction, we employed the following segmentation procedure: based on the velocity profile of each consonant, we algorithmically identified the peak velocity of constriction formation of *C*1 and the time point of constriction release of *C*2 (20% threshold of peak velocity of the release). We then extracted the articulator time series of *C*2 (lip aperture for labials, tongue tip for coronals, and tongue back for dorsals) for that time interval (Fig. 10). The kinematic data were sampled at 1.25 kHz and for present purposes downsampled by a factor of 3 to reduce the computational power required to run the model (the number of points per curve is a major determinant of memory demands). The current dataset comprises 576 curves. The number of equidistant points per curve varies between 37 and 125 (median = 62). Again, all curves are mapped onto a [0,1] time interval without resampling. Since place of articulation for *C*2 varies across labial, coronal, and dorsal and the data therefore have very different spatial coordinates across subjects and articulators, we z-scored (standardized) the time series by articulator by subject.

For the current data, we specify in sparseFLMM a single covariate, (log transformed) Frequency, and random effects of Speaker $B_i(t)$ as well as of Speaker-by-Repetition $E_{ij}(t)$. No item effect is specified, since there is only a single cluster

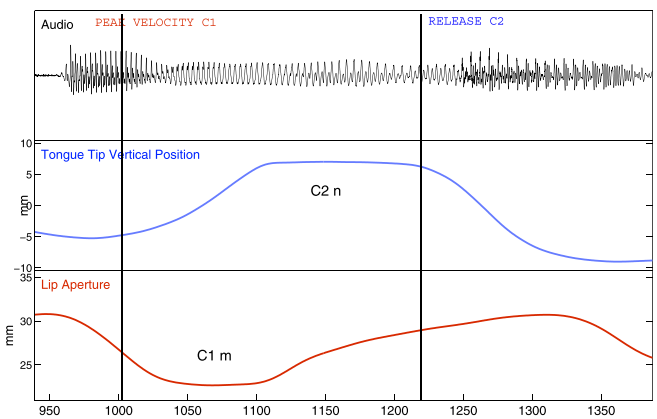

FIG. 10. (Color online) Illustration of data segmentation conventions for one token of the syllable /mno/. Panels show from top to bottom: oscillogram of the audio recording, tongue tip vertical position, lip aperture (Euclidean distance between lip sensors). Time on the *x*-axis is in milliseconds. The constrictions of *C*1 /m/ and *C*2 /n/ are informally indicated. The vertical lines across all three panels mark the time interval over which the time series for *C*2 (tongue tip vertical position, middle panel) was extracted for analysis. Only the vertical position curve of *C*2 entered into the analysis.

per frequency. Prior to running the model, the Frequency covariate was centered to the mean so that the effect can be interpreted relative to this mean. The model only has one functional random intercept and can be specified as

$$Y_{ij}(t) = \mu(t, \boldsymbol{x}_{ij}) + B_i(t) + E_{ij}(t) + \varepsilon_{ij}(t), \quad (4)$$

with $\mu(t, \boldsymbol{x}_{ij}) = f_0(t) + f_1(t) \cdot Frequency_j$ implying a linear effect of Frequency on articulator position for each time point. Significance is evaluated as in the previous examples by generating the effects plots with confidence bands, as done in Fig. 11. Figure 11(b), in which the effect curve represents the effect on articulator position when increasing the variable Frequency by 1, shows no significant effect. The reference mean here represents the mean position for the mean of covariate Frequency. For understanding the impact of a higher or lower frequency relative to the reference mean, the summed effects plot can be generated by adding to the reference mean the estimated curve of covariate 1 (Frequency) multiplied by the given centered frequency value of interest. For the summed effects plot in Fig. 11(c), we plot the reference mean and five selected frequencies, in rank order 1 (lowest), 4, 8, 9, and 12 (highest).

The curves for the lower frequency clusters show a later movement onset of *C*2 compared to higher frequency clusters, as well as a spatially reduced movement [compare the dotted orange curve of freq1 against the dashed-dotted pink curve of freq12 in Fig. 11(c)]. Since the effect is not significant, we abstain from further interpretation here. In terms of variance decomposition, no FPC was selected for Speaker. This is not surprising given the data were z-scored by speaker by articulator. For $E_{ij}(t)$, four FPCs were selected which together account for 96.12% of the variance.

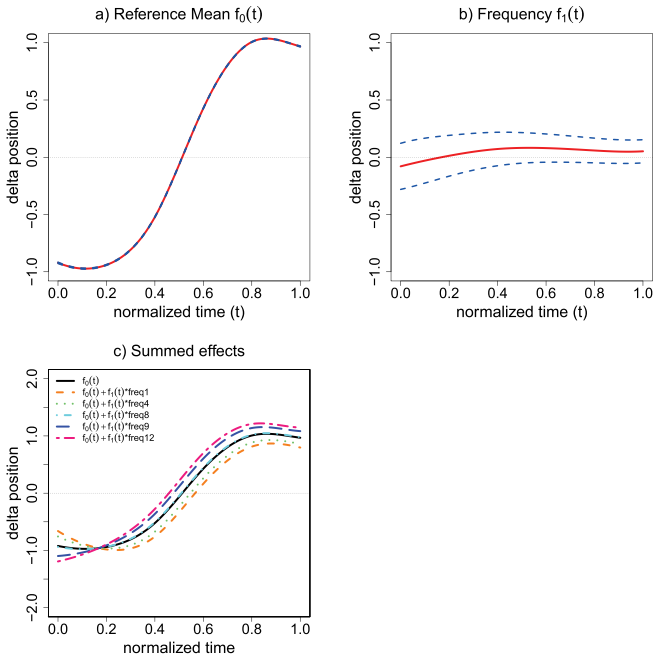

FIG. 11. (Color online) Covariate mean effects (red solid lines) with conditional point-wise confidence bands (dashed lines). (a) Reference mean ($f_0(t)$), (b) continuous covariate effect Frequency ($f_1(t)$), and (c) summed effects curves for five selected frequencies (1 = lowest, 12 = highest; 1: /mxa/, 4: /ʃpa/, 8: /mla/, 9: /bla/, 12: /gla/, see Table I).

VI. DISCUSSION

Our current case studies illustrated how data in which each observation is a curve—whether formant, EPG, or articulography data—can be analyzed as functional data for experimental designs requiring several (continuous or discrete) covariates and crossed random or simple random effects. One innovation of the current approach to the estimation of sparseFLMM is that it does not require data points to be observed at identical, equidistant grid points and allows for sparsely sampled curves, which in the extreme may be measured at only a single point [given a sufficient number of curves with a sufficient number of points from which strength can be borrowed across curves; see also Yao *et al.* (2005)]. Simulations in Cederbaum *et al.* (2016a) have shown that the model can be estimated well when 3–10 points per curve are observed, although the number of speakers and items then needs to be much larger than in the current studies (on the order of 40 or more). Estimation uses information across observations within a curve as well as across curves. Any given estimated curve is therefore informed by the distribution of all curves of the entire dataset. Another important feature of the approach is that the curves are not smoothed prior to entering the model. Instead, smoothing is performed as part of covariance and mean estimation. This ensures that the original observations enter the overall model, no variation is removed beforehand, and the variation is accounted for in subsequent inference.

For the final step of the estimation procedure, the current model implementation offers two ways to predict the weights of the eigenfunctions in the expansion of the random effects. In the computationally more efficient version (computation time being on the order of seconds), the weights are computed directly as best linear unbiased predictors, but no confidence bands for the mean and covariate effects can be computed as the mean and covariate effects are estimated under the working assumption that the curves and time points are independent from each other. In order to obtain valid statistical inference, the second option re-estimates the whole model within one framework [Scheipl *et al.* (2015), based on the estimated FPCs] accounting in the mean effects estimation for the crossed random effects structure via FPC expansions of the functional random intercepts, and constructs confidence bands. This is the option we focused on here. One drawback of that method is that it is computationally time intensive for large datasets. Nonetheless, it is faster than spline-based alternatives [Scheipl *et al.* (2015), when using splines as bases] which do not make use of parsimonious representation via FPCA [see Cederbaum *et al.* (2016a) for explicit comparisons and further discussion]. Additionally, we have recently proposed ways to speed up the covariance estimation part of the algorithm (Cederbaum *et al.*, 2016b) which have also been implemented in the sparseFLMM R-package. In a further development of the method compared to Cederbaum *et al.* (2016a), sparseFLMM now includes the option to approximate the covariate effect estimation using the "discrete" option of the R mgcv package [as of version 1.8.7. (Wood *et al.*, 2017)]. This considerably speeds up computation times and can be used for large datasets. However, this is at the cost

of estimation accuracy. The discrete option bins observations (i.e., time points) into intervals and bases the computation on the number of observations in each bin rather than the value of the observation itself. This is equivalent to a rounding effect on the covariate. It is thus advantageous to not use the discrete option unless required by computational constraints. We did not use the discrete option for any of the examples in this paper. In terms of the confidence bands, it has to be kept in mind that these are point-wise rather than simultaneous confidence bands. Moreover, the uncertainty associated with estimating the FPCs is not taken into account. Nonetheless, simulations have shown that the confidence bands provide coverage close to the nominal level for the actual effects and thus are indeed appropriate to evaluate (point-wise) significance (Cederbaum *et al.*, 2016a).

A further point we would like to address is the issue of random slopes and the possibility of modeling by-subject variability in the main effects (Baayen *et al.*, 2008). Currently, the mgcv package, which we are building on, does not allow for specifying correlated intercepts and slopes (or correlated intercepts). Random slopes could in principle be added to a FLMM model, see for instance the denseFLMM package on CRAN (Greven and Cederbaum, 2017), which estimates FPCs for more complex models, including functional random slopes for functional data observed on the same dense grid. The extension for sparsely or irregularly sampled functional data, however, is far from trivial. (For dense data, the model can basically be estimated point-wise.) Importantly, even if the software were extended to estimate FPCs for random slopes, it is currently not possible to estimate models with correlated random intercepts and slopes in mgcv, the R package on which sparseFLMM builds internally once we have estimated the FPCs (and which is also used for estimating GAMs; i.e., GAMs likewise do not allow for estimating correlated intercepts and slopes or correlated intercepts). Such a model can thus unfortunately not be estimated with the current state of the art, but is an important avenue for future research.

Finally, with mixed-effects models, model selection is often performed in order to test for the necessity of a predictor in building a complex model. With the current implementation of sparseFLMM, comparisons of model quality based on the Akaike information criterion (AIC) are not possible. Given the current debate in statistics on statistical inference after model selection [which is not valid unless post-selection inference is used and valid procedures for this are still in their infancy; see, among others, Fithian *et al.* (2017)], we are in favor of avoiding model selection if possible when statistical inference is of interest. The strategy should be to maximally include all predictors chosen *a priori* by virtue of the experimental design. The effect curves with confidence bands indicate whether a given covariate has a significant effect or not, with a pointwise confidence band not containing zero for some time point corresponding to a significantly different from zero effect at that time point. If there is a small region of significance or only a small distance to the zero line it is, as always, up to the researcher to interpret the statistical result.

The details of data treatment presented here are specific to each case study. The model, however, is generally suitable for any kind of time series data or data with time series-like observations, be they acoustic, articulatory, eyetracking curves, or ERPs, but also more generally the method is in principle applicable to the comparison of groups of curves/functional data, such as they commonly occur, e.g., in ultrasound research. For ultrasound research, in which usually tongue contours are tracked over time and each sample represents a curve, another development will have to involve the analysis of curves over time. Hoole and Pouplier (2017) have recently shown how PCA can successfully be applied to raw ultrasound images to obtain time series of scalar values similar to the ones presented here; these can then in principle be subject to sparseFLMM modelling.

With technological advances in speech production methods, it has become increasingly feasible to collect large datasets for many speakers, making it all the more important to develop methods which allow for adequate statistical modeling of functional data. While methods for linear mixed modeling of individual measurement points have become well established, data in the speech sciences usually consist of time series in the form of signals or signal-derived analysis parameters, recorded with multiple repetitions per item for each speaker. sparseFLMM obviates the need for data reduction by means of feature extraction and magic moment analyses. At first blush, it may seem a high price to pay that effects have to be interpreted graphically. But in fact, analysis is not strictly dependent to graphical evaluation: While we allow for smooth effects over time—anything linear would be way too simplistic—the sparseFLMM model is in fact linear in the covariates for each time point. Thus, the linear covariate effect at any given time point, which can be visually seen in our figures (e.g., Fig. 2), could also be extracted with its confidence intervals from the model output and interpreted in the usual way. In that sense, if one chose, say three or five time points along the time interval and extracted the linear covariate effects and confidence intervals for those time points from the model output, one could in fact simulate the result of a "magic moment" analysis from the output (as we did when discussing extracting the lower limit of the confidence band at time point 0.25 for the Romanian diphthong case study in Sec. III). Moreover, as we show in our case studies, the variance decomposition of the sparseFLMM method renders information which can be interpretable and in principle is available for further analysis. The advantage of sparseFLMM is that analysis is not restricted to a given set of time points, nor do they have to be specified in advance. More importantly, there may be considerable benefit from using more complex models which allow for curve shape quantification of correlated data. As we have pointed out in Sec. I, it has to be a great concern if data reduction techniques either discard a wealth of information in order to be able to take the complex correlation structure of crossed designs into account or successfully quantify curve shape at the price of disregarding the complex correlation structures in the data. We therefore believe that statistical models allowing us to quantify curve evolution over time will provide an important tool for the speech and language sciences.

ACKNOWLEDGMENTS

Work supported by the ERC under the EU's 7th Framework Programme (FP/2007-2013)/Grant Agreement No. 283349-SCSPL to M.P. and by the German Research Foundation (DFG) through Emmy Noether Grant GR 3793/1-1 to S.G.

[1]See supplementary material at http://dx.doi.org/10.1121/1.4998555 for step-by-step instructions for analyzing the datasets for all case studies discussed in this paper, including R code.

Baayen, R. H., Davidson, D. J., and Bates, D. M. (**2008**). "Mixed-effects modeling with crossed random effects for subjects and items," J. Memory Lang. **59**, 390–412.

Baayen, R. H., Kuperman, V., and Bertram, R. (**2010**). "Frequency effects in compound processing," in *Compounding*, edited by S. Scalise and I. Vogel (Benjamins, Amsterdam), pp. 257–270.

Baayen, R. H., van Rij, J., de Cat, C., and Wood, S. N. (**2016**). "Autocorrelated errors in experimental data in the language sciences: Some solutions offered by Generalized Additive Mixed Models," in *Mixed Effects Regression Models in Linguistics*, edited by D. Speelman, K. Heylen, and D. Geeraerts (Springer, Berlin), arXiv:1601.02043.

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (**2013**). "Random effects structure for confirmatory hypothesis testing: Keep it maximal," J. Memory Lang. **68**, 255–278.

Cai, S., Boucek, M., Ghosh, S. S., Guenther, F. H., and Perkell, J. S. (**2008**). "A system for online dynamic perturbation of formant frequencies and results from perturbation of the Mandarin triphthong /iau/," in *Proceedings of the 8th International Seminar on Speech Production*, December 8–12, Strasbourg, France, pp. 65–68.

Cederbaum, J. (**2017**). "sparseFLMM: Functional linear mixed models for irregularly or sparsely sampled data. R package version 0.1.0," https://CRAN.R-project.org/package=sparseFLMM (Last viewed May 19, 2017).

Cederbaum, J., Pouplier, M., Hoole, P., and Greven, S. (**2016a**). "Functional linear mixed models for irregularly or sparsely sampled data," Stat. Modell. **16**, 67–88.

Cederbaum, J., Scheipl, F., and Greven, S. (**2016b**). "Fast symmetric additive covariance smoothing," arXiv:1609.07007.

Di, C.-Z., Crainiceanu, C., and Jank, W. S. (**2014**). "Multilevel sparse functional principal component analysis," Stat **3**, 126–143.

Fithian, W., Sun, D., and Taylor, J. (**2017**). "Optimal inference after model selection," arXiv:1410.2597v4, 1-39.

Greven, S., and Cederbaum, J. (**2017**). "denseFLMM: Functional Linear Mixed Models for Densely Sampled Data," R package version 0.1.0.

Gubian, M., Torreira, F., and Boves, L. (**2015**). "Using functional data analysis for investigating multidimensional dynamic phonetic contrasts," J. Phon. **49**, 16–40.

Gubian, M., Torreira, F., Strik, H., and Boves, L. (**2009**). "FDA as a tool for analyzing speech dynamics. A case study on the French word c'était," in *Proceedings of Interspeech 2009*, September 6–10, Brighton, UK, pp. 2199–2202.

Guo, W. (**2002**). "Functional mixed effects models," Biometrics **58**, 121–128.

Hoole, P., and Pouplier, M. (**2017**). "Öhman returns: New horizons in the collection and analysis of imaging data in speech production research," Comput. Speech Lang. **45**, 253–277.

Lancia, L., Rausch, P., and Morris, J. S. (**2015**). "Automatic quantitative analysis of ultrasound tongue contours via wavelet-based functional mixed models," J. Acoust. Soc. Am. **137**, EL178–EL183.

Marin, S. (**2014**). "Romanian diphthongs /ea/ and /oa/: An articulatory comparison with /ja/ - /wa/ and with hiatus sequences," Rev. Filología Román. **31**, 83–97.

Marin, S., and Goldstein, L. (**2012**). "A gestural model of the temporal organization of vowel clusters in Romanian," in *Consonant Clusters and Structural Complexity*, edited by P. Hoole, L. Bombien, M. Pouplier, C. Mooshammer, and B. Kühnert (Mouton de Gruyter, Berlin), pp. 177–203.

Marin, S., and Pouplier, M. (**2014**). "Articulatory synergies in the temporal organization of liquid clusters in Romanian," J. Phon. **42**, 24–36.

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

Mücke, D., Grice, M., and Cho, T. (**2014**). "More than a magic moment—Paving the way for dynamics of articulation and prosodic structure," J. Phon. **44**, 1–7.

Nelson, W. L. (**1983**). "Physical principles for economy of skilled movements," Biol. Cybernet. **46**, 135–147.

Pouplier, M., and Hoole, P. (**2016**). "Articulatory and acoustic characteristics of German fricative clusters," Phonetica **73**, 52–78.

Pouplier, M., Marin, S., Hoole, P., and Kochetov, A. (**2017**). "Speech rate effects in Russian onset clusters are modulated by frequency, but not auditory cue robustness," J. Phon. **64**, 108–126.

Pouplier, M., Marin, S., and Kochetov, A. (**2015**). "Durational characteristics and timing patterns of Russian onset clusters at two speaking rates," in *Proceedings of Interspeech 2015*, September 6–10, Dresden, Germany, pp. 2679–2683.

Quené, H., and van den Bergh, H. (**2008**). "Examples of mixed-effects modeling with crossed random effects and with binomial data," J. Memory Lang. **59**, 413–425.

Ramsay, J. O., Hooker, G., and Graves, S. (**2009**). *Functional Data Analysis with R and Matlab* (Springer, Dordrecht, the Netherlands).

Ramsay, J. O., and Silverman, B. W. (**2005**). *Functional Data Analysis* (Springer, New York).

Ruppert, D. (**2002**). "Selecting the number of knots for penalized splines," J. Comput. Graph. Stat. **11**, 735–757.

Scheipl, F., Staicu, A.-M., and Greven, S. (**2015**). "Functional additive mixed models," J. Comput. Graph. Stat. **24**, 477–501.

Sharoff, S. (**2005**). "Russian internet corpus," http://corpus.leeds.ac.uk/list.html (Last viewed June 30, 2014).

Wada, Y., Koike, Y., Vatikiotis-Bateson, E., and Kawato, M. (**1995**). "A computational theory for movement pattern recognition based on optimal movement pattern generation," Biol. Cybernet. **73**, 15–25.

Wang, J.-L., Chiou, J.-M., and Müller, H.-G. (**2016**). "Functional data analysis," Ann. Rev. Stat. Appl. **3**, 257–295.

Wieling, M., Montemagni, S., Nerbonne, J., and Baayen, R. H. (**2014**). "Lexical differences between Tuscan dialects and standard Italian: Accounting for geographic and sociodemographic variation using generalized additive mixed modeling," Language **90**, 669–692.

Wieling, M., Tomaschek, F., Arnold, D., Tiede, M., Bröker, F., Thiele, S., Wood, S. N., and Baayen, R. H. (**2016**). "Investigating dialectal differences using articulography," J. Phon. **59**, 122–143.

Wood, S. N. (**2011**). "Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models," J. R. Stat. Soc. Ser. B. **73**, 3–36.

Wood, S. N., Goude, Y., and Shaw, S. (**2015**). "Generalized additive models for large datasets," J. R. Stat. Soc. Ser. C. **64**, 139–155.

Wood, S. N., Zeheyuan, L., Shaddick, G., and Augustin, N. (**2017**). "Generalized additive models for gigadata: Modelling the UK Black Smoke Network daily data," J. Am. Stat. Assoc. (published online).

Yao, F., Müller, H.-G., and Wang, J.-L. (**2005**). "Functional data analysis for sparse longitudinal data," J. Am. Stat. Assoc. **100**, 577–590.