Characterizing fundamental frequency in Mandarin: A functional principal component approach utilizing mixed effect models

Pantelis Z. Hadjipantelis
Centre for Complexity Science and Department of Statistics, University of Warwick, Coventry CV4 7AL, United Kingdom

John A. D. Aston[a)]
Department of Statistics, University of Warwick, Coventry CV4 7AL, United Kingdom

Jonathan P. Evans
Institute of Linguistics, Academia Sinica, 128 Academia Road Sec 2, Taipei 115, Taiwan Republic of China

(Received 6 February 2011; revised 6 January 2012; accepted 18 April 2012)

A model for fundamental frequency (F0, or commonly pitch) employing a functional principal component (FPC) analysis framework is presented. The model is applied to Mandarin Chinese; this Sino-Tibetan language is rich in pitch-related information as the relative pitch curve is specified for most syllables in the lexicon. The approach yields a quantification of the influence carried by each identified component in relation to original tonal content, without formulating any assumptions on the shape of the tonal components. The original five speaker corpus is preprocessed using a locally weighted least squares smoother to produce F0 curves. These smoothed curves are then utilized as input for the computation of FPC scores and their corresponding eigenfunctions. These scores are analyzed in a series of penalized mixed effect models, through which meaningful categorical prototypes are built. The prototypes appear to confirm known tonal characteristics of the language, as well as suggest the presence of a sinusoid tonal component that is previously undocumented.
 [http://dx.doi.org/10.1121/1.4714345]

PACS number(s): 43.60.Cg, 43.60.Uv, 43.66.Hg [MAH] Pages: 4651–4664

I. INTRODUCTION

A. Theoretical background

Speech sounds consist of complex periodic waves characterized by their frequency and amplitude. Phonetic sound properties of research interest include the pulse, intensity, sound wave components, spectrum, and/or duration of the examined sound segment, as well as fundamental frequency (F0), the focus of this paper. F0 as a speech phenomenon is the major component of what a human listener identifies as pitch and relates to how fast the vocal folds of the speaker vibrate during speech.[1,2]

While in many languages pitch differences are mostly detected in matters of intonation or semantic alterations (such as expression of sarcasm), in tonal languages, such as Taiwanese Mandarin, pitch (and the closely related F0) plays a crucial role in the actual lexical entry of the word. As such, má (↗) said with a mid rising tone means *hemp*, while articulated with a high falling tone, mà (↘), means *to scold*. In the past, linguistic studies treated F0 as a single point by utilizing target values[3,4] or obtained estimates of the F0 contour by treating it as a bounded rigid curve through processes of averaging.[5] Such approaches, by necessity, impose simplifying assumptions which make interpretation difficult when considering a complete corpus of data from a more natural language experiment.

In this paper, a different approach is adopted. We propose a model where the F0 curve is characterized as the realization of a stochastic Gaussian process.[6] A Gaussian process is essentially a generalization of a multivariate Gaussian random variable to an infinite index set.[7] As a consequence, our methodology treats the fundamental frequency of each rhyme as a bounded continuous curve, rather than a time-indexed vector of readings.

Functional data analysis offers tools for analyzing data that consist of functions—often but not always, smooth curves.[8] In the current study, a functional principal component analysis (FPCA) is first performed on the data set's F0 measurements to extract the principal curves, those curves which explain the most variation in the data. Similar approaches might utilize Legendre polynomials,[9] quadratic splines,[10] or Fourier analysis to derive lower and higher ranking basis functions that would correspond to slower and faster varying components of the utterance. However, these functions are fixed in advance rather than derived directly from the data and are not guaranteed to be optimal in terms of the minimal number required to explain a certain percentage of the variation in the data as in the case of principal component functions.[11]

Building on the FPCA findings, the functional principal component (FPC) scores are used as the dependent values in a series of linear mixed effect (LME) models, allowing the scores to act as proxy data for the complete curves. The

[a)]Author to whom correspondence should be addressed. Electronic mail: J.A.D.Aston@warwick.ac.uk

scores essentially quantify the weight each FPC carries in the final F0 curve formation, and will be described in further detail in Sec. II. LME models allow the inclusion of both fixed and random effects to achieve a flexible modeling of the data. In the current case, the difference between individual speakers due to genetic, environmental[12] or even chance factors[13] are modeled as a series of random additive effects acting on the F0 contours.[14,15] In order to compute the FPC's, the sample mean is subtracted from the data, and the covariance of the data then calculated, in a similar way to standard multivariate principal component analysis. Another possible approach would be to subtract the speaker specific mean from each syllable, prior to further analysis. We chose not to follow this direction as determining the effect of speaker on each of the components is of interest. However, as might be expected, the two approaches yield very similar results (see supplementary material,[16] Sec. III).

The methodology presented here addresses the issue that, while it has been widely accepted and documented that F0 undergoes variations due to phonetic processes in speech production that are attributed to fixed effects (e.g., the sex of the speaker), unmeasurable variables such as the length of the speaker's vocal folds or the state of their health, also affect the final F0 utterance. This measurability problem is countered by considering such covariates as random effects. This theoretical perspective is not *ad hoc*; it corresponds directly with the linguistic, para-linguistic and non-linguistic parameters presented in the work of Fujisaki.[17,18] The Fujisaki model implementations have been extended by Mixdorff[19] to account for micro-prosodic effects by taking advantage of the MOMEL algorithm.[10] Other approaches utilize the automatic intonation modeling approach as offered by the INTSINT (Refs. 20, 21) and/or the TILT algorithmic implementations.[22] Furthermore the qTA model[23] also builds on Fujisaki's assumption, proposing a description of the physiological mechanisms behind F0 production, a goal somewhat different from the one in this paper. In the present framework and analogous to the Fujisaki rationale, F0 is the dependent variable of interest with standard fixed effects such as the vowel in the rhyme corresponding to linguistic effects, sentence variations and break points within the utterance corresponding to para-linguistic effects, and speaker variations corresponding to non-linguistic effects.

As Evans *et al*. have already presented[24] and Aston *et al*. have further extended,[15] the explanatory power that can be yielded from the application of LME models for F0 is insightful in cases of tonal languages. In the current study, the F0 track of each rhyme in the utterance is used; as a result, while the two previously mentioned works focused on one position in a frame sentence, in this project a large number of read texts of varying lengths are investigated, adding new dimensions of complexity and further enhancing the generality of the approach by analyzing complete corpus data. In addition, while the previous studies utilized two phonologically level tones, Mandarin has both level and contour tones as well as toneless syllables and thus poses a significantly more complex analytical challenge.

As a starting point, a smoothing and interpolation procedure is utilized to change the measurement from real-time into that of normalized syllable time, building partially on the assumption of syllable-synchronization.[23] Next, regression models are introduced to help identify significant covariates of speech production. Afterwards, a penalized system of model selection is put forward to obtain the final models. Given the amount of data present in the study, over-fitting is a concern, and therefore a penalty on the number of regressors in the model is imposed through an AIC approach (as outlined by Faraway[25]) and jackknifing is also implemented to further enhance and test the robustness of the findings. This use of FPCA and mixed effects modeling offers a generalized semi-parametric approach to the linguistic modeling of Mandarin Chinese F0.

The application of FDA (functional data analysis) in relation to linguistics is not without precedence. The early work of Ramsay *et al*.[26] used FDA to model the coordinates of lip motion in order to infer basic principles of lip coordination. Since then a number of speech production related questions associated with articulatory issues,[27–29] as well as with issues of physiological interests,[30–32] have been addressed with FDA. The current work differs from the above mentioned projects by employing an entire corpus as raw data. Rather than using a small linguistic sample by a single speaker,[18] employing monosyllabic utterances and a small number of sentences[33] and/or frames within the utterances[5,15,27,30] to minimize possible confounds at the data collection level, a large corpus is analyzed and the confounds explicitly modeled. In contrast to existing intonation synthesis algorithms, the current methodology's primary goal is to offer insights into how linguistic and non-linguistic factors combine in the estimation of F0 and presents an auxiliary approach for existing speech synthesis algorithms in terms of modeling the acoustic shapes of tones.

B. Dataset presentation

The Sinica Continuous Speech Prosody Corpora 1 (COSPRO-1) is a large-scale comprehensive data-set consisting of recordings of Taiwanese Mandarin read speech.[34] Five participants each uttered a total of 599 predetermined sentences. After pre-processing and annotation, their utterances, having a median length of 20 syllables, resulted in a total of 54 707 frequency curves. Each F0 curve corresponds to the rhyme portion of one syllable. The three female and two male participants were native Taiwanese Mandarin speakers. The recordings themselves were conducted by the Institute of Linguistics, Academia Sinica in 1994. Using the in-house developed speech processing software package COSPRO TOOLKIT,[34,35] the fundamental frequency (F0) of each rhyme utterance was extracted at 10 ms intervals, a duration under which the speech waveform can be regarded as a stationary signal.[36] Associated with the recordings were characterizations of tone, rhyme, initial consonant as well as speech break or pause; the presented corpus is a real language corpus, designed to include all tonal combinations but still have semantic meaning. The syllables are labeled with the four lexically specified tones as well as encoding that some syllables are phonologically toneless (tone 5), and

additional contextual information is also associated with each curve (see Table IV for a list of covariates included). This data set has been previously analyzed using a Fujisaki approach.[37]

II. STATISTICAL METHODOLOGY

A. Functional data analysis

Ferraty and Vieu provide the following definition: "A random variable x is called a functional variable if it takes values in an infinite dimensional space (or functional space)."[38] Here we interpret the F0 trajectory as observed functional data. Based on the Ferraty and Vieu definition and given that the examined dataset is indeed in curve-form, the current study adopts the notion put forward by Chiou *et al.*, that "each observed curve is a (independent) realization of a stochastic process reflecting the random nature of the individual curves."[39] As a logical result, given a stochastic process $Y(t)$, $t\epsilon$, [0,1] the sample curves can be thought of as having a mean $E[Y(t)] = \mu(t)$ and a covariance $\mathrm{cov}[Y(s), Y(t)] = C(s, t)$. Taking advantage of the symmetric nature of C ($C(s, t) = C(t, s)$) the following spectral decomposition follows by Mercer's theorem[40] for $C(s, t)$:

$$C(s, t) = \sum_{\nu=1}^{\infty} \lambda_\nu \phi_\nu(s)\phi_\nu(t), \tag{1}$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ are ordered eigenvalues of the operator C and ϕ_ν's are the corresponding eigenfunctions.

Going back and reviewing the notion of PCA, it is worth noting that PCA is not only a convenient transformation for dimensionality reduction; the principal components (PCs) themselves serve as characterizations of the sample's trajectories around an overall mean trend function,[41] in other words each PC gives a representation of the F0 contour components for our data. As Castro *et al*. briefly summarized in their seminal work on continuous sample curves,[42] given a vector process $Y = (y_1, y_2, \ldots, y_p)^T$, where $y_1, y_2, \ldots, y_p$ are scalar vectors, an expression of the form

$$Z = M + \sum_{\nu=1}^{m} \alpha_\nu Z_\nu(t) \tag{2}$$

is called a m-dimensional model of Y, where M denotes the mean vector of the process, $Z_1, Z_2, \ldots, Z_m$ are fixed unit length p vectors and $\alpha_1, \alpha_2, \ldots, \alpha_k$ are scalar variates dependent on Y. Proposing now that a process $Y(t)$ is observed at p distinctive times $t_1, t_2, \ldots, t_p$ it yields the analogous random vectors $y(t)$, describing the stochastic process $Y = (y(t_1), y(t_2), \ldots, y(t_p))^T$, fitting perfectly with the theoretical notions of longitudinal data being a variation of repeated measurements. Therefore, coming back to the original notion of a stochastic process $Y(t)$, the m-dimensional linear model for such process is

$$Y_j(t) = \mu(t) + \sum_{\nu=1}^{m} \alpha_{\nu j}\phi_\nu(t), \tag{3}$$

where α_ν are once more the uncorrelated random variables with zero mean and refer to the νth principal component score of the jth subject and ϕ_ν are linear independent basis-functions, of the random trajectories Y_j. This expansion (3) is referred to as the Karhunen–Loève or FPC expansion of the stochastic process Y.[43]

It must be noted here, that as Rice and Silverman emphasized, the mean curve and the first few eigenfunctions are smooth and the eigenvalues λ_ν tend to zero rapidly so that the variability is predominantly of large scale.[44] In physical terms, smoothness of data is critical so that the discrete sample data can be considered functional.[11] A number of smoothing techniques have been proposed over the years concerning FPCA; linear smoothing, basis function methods such as wavelet or regression splines bases, or smoothing by local weighting using local polynomial smoothing or kernel smoothing, being some of the most frequently encountered. Kernel smoothing, considered to be the optimal choice in the case of local weighting,[38] is the one applied here due to its simplicity and computational ease, yielding smooth sample F0 curves.

Utilizing the methodology proposed by Chiou *et al.*[39] a locally weighted least squares smoother, denoted by S_L, is implemented, so local lines are fitted to the data. A point t is used as the center of a smoothing window or interval $[t - b, t + b]$ where b is the fixed parameter commonly known as bandwidth. The formal definition of the smoother itself is

$$S_L\left\{t; b, (t_i, y(t_i))_{i=1,\ldots,s}\right\} = \underset{a0}{\mathrm{argmin}}\left\{\min_{a1}\left(\sum_{i=1}^{s} K\left(\frac{t - t_i}{b}\right)[y(t_i) - \{a_0 + a_1(t - t_i)\}]^2\right)\right\}, \tag{4}$$

where K is the kernel function selected, t is the argument of the smoother S_L, b is the smoothing parameter (how big the window of the smoother will be in relation to actual available data-points), and $(t_i, y(t_i))_{i=1,\ldots,s}$ is the actual data scatter-plot consisting of s points. Using cross-validation the optimal bandwidth b was found to be 3% of the total rhyme duration signifying the fact that even the initial sample was quite smooth. It must also be mentioned that for the data to be suitable for FPCA, besides the smoothing, time normalization is of importance. As the main focus of the present work is on the phonetic F0 contour shapes, the results it yields are implicitly time normalized on a [0, 1] interval. Therefore all the data-curves were not only smoothed but concurrently interpolated in a [0, 1] interval in order to be directly comparable with each other, resulting in F0 curves on a rhyme time scale rather than in real time. The reader will note that interpolation itself does impose

a certain degree of smoothing, as well an *ad hoc* choice of the number of points over which the interpolation takes place. The actual readings in our study, after disregarding missing values, had on average (15.38 ≈) 16 points per syllable, and based on this estimate, the basis of 16 points is chosen. The analysis was also conducted using 12- and 20-point interpolation so that the impact of the smoothing could be more easily identified but this produced negligible differences. Furthermore, to ensure the beginning and ends of each syllable are not subjected to substantial smoothing errors due to limited data, the beginning and the end point of the curve are not smoothed.

As a final remark on the smoother implementation, the function K, denoting a non-negative kernel function, was chosen to be a Gaussian basis function $K(x) = e^{-x^2/2}$, being the most standard weight function and also ensuring that its product is never negative.

Having established the smoothness of the data, the next step in the actual implementation of the K-dimensional linear model of Eq. (2) is the estimation of the mean function. Given that we have an equispaced design, the overall mean function is estimated as

$$\hat{\mu}(t_i) = \frac{1}{n}\sum_{i=1}^{n} y_j(t_i), \;\; i = 1, \ldots, s, \tag{5}$$

where n is the number of sample F0 curves available and s is the number of points in each curve (in the current data, 54 707 and 16, respectively).

The final step to calculate the FPCA scores is actually the most straightforward. Following the same methodology as Aston *et al.*,[15] the eigenfunctions are calculated by the spectral analysis of the estimated covariance matrix

$$\hat{C}(t_k, t_l) = \frac{1}{n}\sum_{j=1}^{n}\{y_j(t_k) - \hat{\mu}(t_k)\}\{y_j(t_l) - \hat{\mu}(t_l)\}, \quad k, l \in \{1, \ldots, s\}. \tag{6}$$

As a result, we can estimate the eigenfunctions ϕ_ν, which as shown in Eq. (1), correspond to solutions of

$$\hat{C}(t_k, t_l) = \sum_{\nu=1}^{m} \lambda_\nu \hat{\phi}_\nu(t_k)\hat{\phi}_\nu(t_l), \tag{7}$$

where $\lambda_1, \lambda_2, \ldots, \lambda_m$ are the ordered eigenvalues of the system. Finally the FPCA $A_{j,\nu}$ scores are estimated as

$$\hat{A}_{j,\nu} = \sum_{j=1}^{s}\{y_j(t_i) - \hat{\mu}(t_i)\}\hat{\phi}_\nu(t_i)\Delta_i, \tag{8}$$

where $\Delta_i = t_i - t_{i-1}$. These scores, $A_{j,\nu}$, are the ones finally used for the estimation analysis by the LME. The choice and number of FPC's used is related to the amount of variation that each of these components reflect. Given the large number of available sample utterances, a relatively high number of FPC's is required in order to account for phonetic effects that might occur in just a relatively small number of sample instances. Despite the need for statistical accuracy, it should be mentioned that the actual information content found in the FPC scores is of importance. Thus, only the FPC's reflecting variation that is audible are selected. In reality, only pure tone F0 fluctuations above a 2 Hz threshold can definitely be clearly perceived by the human auditory system (just noticeable difference—JND);[45] in the presence of noise, JND is at a minimum of 10 Hz. As advocated by Kochanski,[46] in the case of human speech, the JND for pitch motions seems to be rather larger. Black and Hunt[4] show that a 9.9 Hz RMS error is not detrimental to the model's success. This threshold will be used throughout the paper; however our approach is flexible enough for other practitioners to utilize it with different cut-off thresholds.

B. Linear mixed effects models

Having determined the eigenfunctions and corresponding FPCs from the data, the next step involves the LME model construction and selection. LME models are models in which both random and fixed effects occur linearly in the model's implementation. As Pinheiro and Bates[47] presented: "(LME models) extend linear models by incorporating random effects which can be regarded as additional error terms, to account for correlation among observations within the group." More formally, and using the classical linear mixed effect model notation proposed by West *et al.*,[48] combined with the distributions notion presented by Faraway,[25] a standard fixed effect model with normal errors:

$$A_\nu = X_\nu\beta + \epsilon_\nu \text{ or } A_\nu \sim N(X_\nu\beta, \sigma^2 I) \tag{9}$$

can be extended to account for random effects in the following form:

$$A_\nu = X_\nu\beta + Z_\nu\gamma + \epsilon_\nu \;\; \text{or} \;\; A_\nu|\gamma \sim N(X_\nu\beta + Z_\nu\gamma, \sigma^2 I), \tag{10}$$

where in the presented case A_ν is the vector of length $n \times 1$ of FPC scores associated with the νth FPC, X_ν is the $n \times p$ model matrix, the vector ϵ_ν of length n encapsulates the random variables representing the error in the relation, and β is a vector of length p that contains the linear (fixed) regression coefficients, where p is the number of those coefficients. The extension of this model now to account for mixed effects is such that Z_ν is a model matrix $n \times r$ (Ref. 49) associated with a vector γ of random effects. It needs to be stressed that random effects are by definition random variables themselves.[14] As such, the γ vector will follow a multivariate Gaussian distribution $\gamma \sim N(0, D)$, where D represents the covariance matrix of the elements in vector γ. In a similar manner, the error residual vector ϵ also follows a multivariate Gaussian distribution where $\epsilon \sim N(0, R)$ and R is the covariance matrix for residuals in vector ϵ.

Having established that $\gamma \sim N(0, D)$ and $\epsilon \sim N(0, \sigma^2 I)$ the variance of a is subsequently written as

$$Var(A_\nu) = Var(Z_\nu\gamma) + Var(\epsilon) = Z_\nu D Z_\nu^T + \sigma^2 I \tag{11}$$

resulting in the unconditional distribution:

$$A_\nu \sim N\left(X_\nu\beta,\ \sigma^2 I + Z_\nu D Z_\nu^T\right). \quad (12)$$

Model construction requires the use of a definition for the goodness of fit achieved by the model estimated. Existing literature suggests the log-likelihood function as a standard choice. Nevertheless, a number of issues have to be highlighted. An important problem arising when estimating the log-likelihood function of the data is that the unrestricted maximum likelihood estimator (MLE) might involve a negative variance, which is clearly unacceptable. Moreover the MLEs are biased. Given that the number of samples in the random vector might be quite small, as in the case of speakers, the difference between a biased and an unbiased MLE can be significant. Therefore when estimating the final parameters, the restricted maximum likelihood (ReML) is used. ReML tries in essence to find linear combinations of the responses, k, such that $k^T X = 0$ and thus to exclude any fixed terms parameters from the likelihood function. However, ML is used for the model selection procedure as the theory for model comparisons is based on ML estimation. As ReML will try to transform the fixed effect response in the manner described above, this would lead to a series of different transformations for each model setting, making them incomparable. Therefore it is essential to use ML estimators if likelihood ratio tests are to be implemented.

For each FPC's scores, the LME modelling procedure was initiated by a model containing the maximal number of linguistically plausible covariates. By employing an Akaike information criterion (AIC) selection of the models examined, models with both important covariates and also parsimony were identified.[50,51] AIC for each model is defined as

$$AIC = 2(-loglik + q), \quad (13)$$

where q is the number of parameters in the model examined and loglik the maximum value of the log-likelihood function of the model. AIC chooses a model that is adequately detailed to capture the variation exhibited in the data but concurrently attaches a penalty as the number of included covariates increases. This is achieved by the terms loglik and q, respectively.

To assess significance and give confidence intervals for the model's estimates, highest posterior density intervals were found[14] using MCMC sampling for the chosen models. In addition, in order to check the robustness of the results obtained, jackknifing was performed by constructing partitions of the data into five random sets and then examining 180 such randomizations, comparing the models produced. In the case of a discrepancy between jackknifing and AIC, the more parsimonious model was chosen. More details of the AIC scores and jackknifing are given in the supplementary information.[16]

Overall, the complete procedure to obtain the F0 estimate, once the components (in the example four chosen components are used) and models are found, can be summarized by Fig. 1.

III. DATA ANALYSIS AND RESULTS

We must emphasize that while the statistical robustness of the methods employed is crucial, the actual targets of this project are the phonetic significance and interpretation of its results. The analysis requires high-specificity as some tonal combinations and other covariate interactions of interest are relatively sparse within the data. Therefore, initially at least 99.99% of the total variation in the original data has to be accounted for. This figure results from the need to ensure effects that might only systematically alter a small number of sample curves are not missed in the analysis. Thus, the first 12 FPC's were selected as necessary to incorporate in the modeling procedure. This unusually large number of FPC's was also dictated by the fact that significant regression-related effects might actually appear in a small percentage of the sample variation. These 12 FPCs account for the 99.992% of the total variation in the sample (Table I). Nevertheless, in a worst case scenario, even by accounting for such high variation, relevant characteristics that may occur in five syllables or fewer within the corpus could be filtered away (based on the residual variation of the discounted FPC's).

Moreover, given the large number of samples, by taking the upper model percentile (99%) of the FPC scores and multiplying it by the maximum absolute value of each eigenfunction, we can effectively derive an upper limit of the actual variation attributed to each component in Hz, the unit that was originally used for measurement. This is of interest because any actual variation found to be below the minimum threshold assumed (9.9 Hz in this case) is likely to remain unnoticed. This cut-off threshold in essence excludes all FPC's with rank equal to or higher than 5, which were previously deemed as of possible importance (see Table II). Statistically, it should be emphasized that our estimates of the maximum actual auditory variation per FPC are quite conservative as they are based on a 99% quantile. As shown in Table II, if a 95% quantile were used, it would suggest that we actually exclude the components that are below 4.2 Hz, a significantly narrower range.

The eigenfunctions of each principal component are computed and used to compute the FPC scores relating to each curve. As mentioned in the previous section, not only the smoothness of the covariance function of this transformation is essential, but also the smoothness of the eigenfunctions themselves. A visual inspection of our results confirms that the kernel smoothing undertaken was successful, with the data being smooth enough for the notions of FDA to be applicable (even though only a minimal smoothing was performed). The covariance function appears smooth throughout its values (see supplementary material,[16] Sec. II) as do the mean and FPC curves (Fig. 2). It must be noted that the fifth and sixth FPC's seem somewhat less smooth in appearance, further signifying that the transformation starts to reach an explanatory threshold and these components start to exhibit the characteristics of noise. It is also noticeable that the eigenfunctions appear to exhibit a distinctive polynomial pattern, with each successive FPC's eigenfunction reflecting the component rank in the eigenfunction's curvature (Fig. 2). This result concurs with the assumed contour shapes of Grabe *et al.*[9] where Legendre polynomials $L0$ to $L3$ were utilized for the contour basis of F0 to examine intonation. In principal, given our statistical findings and the well attested shapes of Mandarin tones in the literature, the basic tone curve of the syllable can essentially be

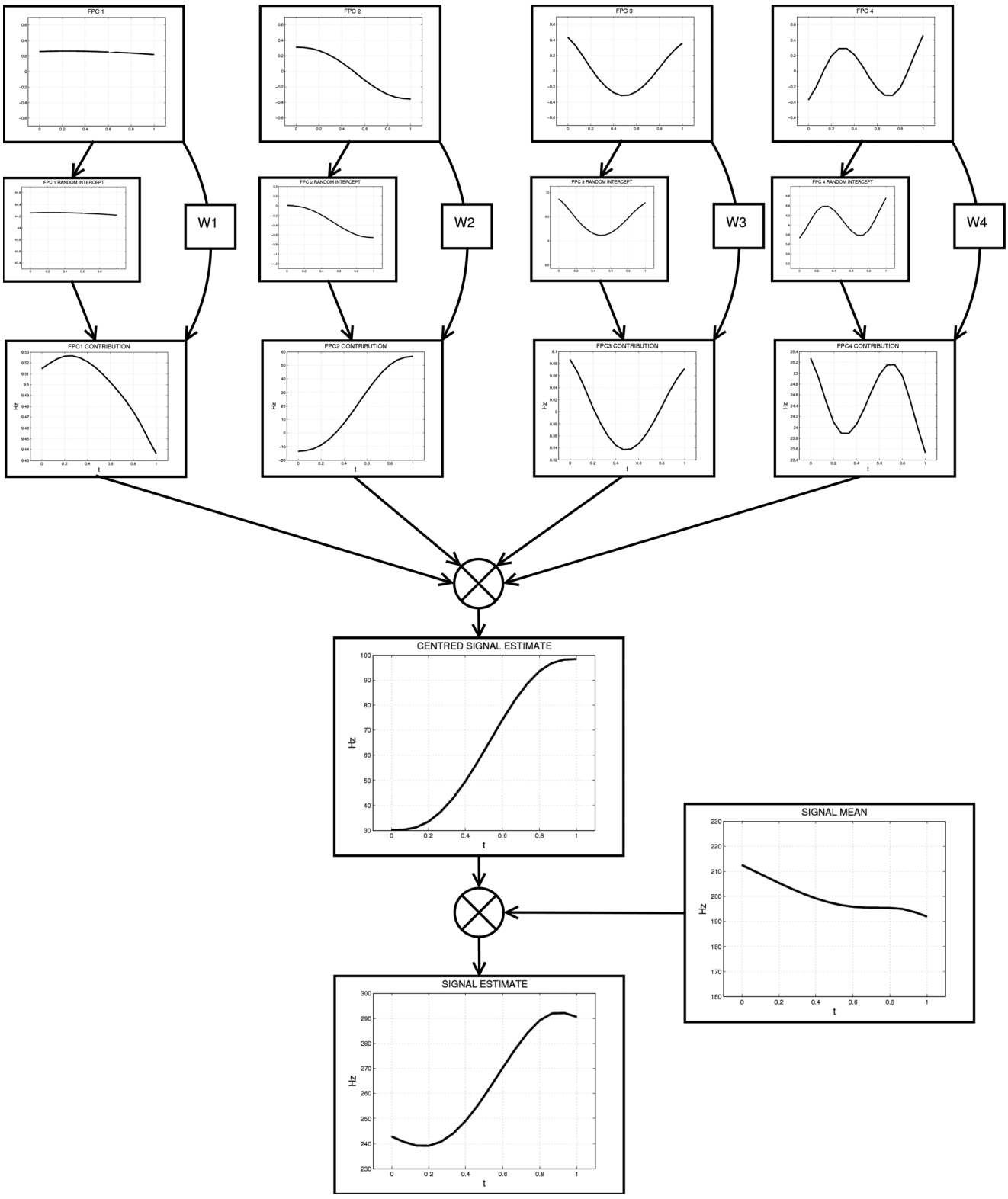

FIG. 1. The first four eigencomponents (top row) are used to construct the final syllable estimate of F0 (bottom row). The individual component magnitude (third row) is calculated by using the weight estimates (w_i) obtained as the sum of the relevant utterance covariates from the LME model and the component specific random intercept (second row). Subsequently, these components added together produce the centralized syllable estimate (row 4). Finally, the addition of the sample mean (row 5) produces the final syllable estimate of F0 (bottom row).

reconstructed by using FPC1, FPC2, and FPC3, as can be seen from the actual shape of those components, with FPC4 allowing contextual movement between tones.

While the kernel smoothing and interpolation was implemented by a custom built C++ program written by the first author, utilizing the GSL package,[52] the calculation for the eigenfunction decomposition and the production of the FPC scores was conducted using standard built-in MATLAB procedures.[53] The rest of the analysis was carried out in the statistical environment R.[54] Except for the obvious standard R

TABLE I. Individual and cumulative variation percentage per FPC.

FPC No.	Individual variation	Cumulative variation
FPC1	88.23	88.23
FPC2	9.78	98.01
FPC3	1.42	99.43
FPC4	0.32	99.75
FPC5	0.11	99.86
FPC6	0.05	99.91
FPC7	0.03	99.94
FPC8	0.02	99.96
FPC9	0.01	99.97
FPC10	0.01	99.98
FPC11	0.01	99.99
FPC12	0.01	99.99

methods used (qqplot(), lm(), etc.) the major body of the analysis was done using methods from the statistical package LME4 (Ref. 55) (for the LME model estimation and prediction) and LanguageR (Ref. 56) (for the MCMC sampling required for the construction of confidence intervals relating to the model's estimators). As mentioned earlier, we examined the robustness of the selected models through jackknifing, and extensive sub-sampling was implemented using 180 5-sub-sample partitions of our original samples, yielding a total of 900 subsamples. (For a detailed discussion and relevant histograms refer to the supplementary material,[16] Sec. VII.)

The model selection procedure was initiated by selecting a large but still linguistically plausible model and then de-constructing it using AIC; excluding covariates that were viewed as statistically redundant or insignificant. The following equation presents the original basis equation:

$$\begin{aligned} FPCx = \Big\{ & \left[tn_{previous} * tn_{current} * tn_{next}\right]. \\ & + \left[cn_{previous} * tn_{current} * cn_{next}\right] \\ & + \Big[(B2) + (B2)^2 + (B2)^3 \\ & + (B3) + (B3)^2 + (B3)^3 \\ & + (B4) + (B4)^2 + (B4)^3 \\ & + (B5) + (B5)^2 + (B5)^3\Big] * Sex + [rhyme_t]\Big\}\beta \\ & + \{[Sentence] + [Spkr\,ID]\}\gamma + \epsilon. \end{aligned} \quad (14)$$

TABLE II. Actual auditory variation per FPC (in Hz) (human speech auditory sensitivity threshold ≈ 10 Hz).

FPC No.	Hz (99%)	Hz (95%)
FPC1	133.3	101.3
FPC2	55.3	38.3
FPC3	35.8	20.7
FPC4	19.1	9.1
FPC5	8.9	4.2
FPC6	5.7	2.5
FPC7	3.6	1.7
FPC8	2.9	1.2
FPC9	2.4	1.1
FPC10	1.8	0.85
FPC11	1.7	0.68
FPC12	1.3	0.45

The standard R notation is used here for simplicity regarding the interaction effects; [K*L] represents a shorthand notation for [K + L + K:L] where the colon specifies the interaction of the covariates to its left and right.[57] Table IV offers a list of each covariate and its definition. It must be pointed out that, from the set of fixed effects, all fixed covariates, with the exception of break counts, are in factor form. Break (or pause) counts represent the number of syllables between successive breaks of a particular type and are initialized in the beginning of the sentence and are subsequently reset every time a corresponding or higher order break occurs. They represent the perceived degree of disjunction between any two words, as defined in the ToBi annotations.[3] B2 break types correspond to smaller breaks occurring usually at the end of words, while B5 types occur exclusively at a full stop at the end of each utterance; essentially signifying an utterance boundary pause. Breaks B3 and B4 represent intermediate or intonational phrase stops, respectively. B1 breaks were not used as these are coincident with our data observation unit (i.e., each syllable). Table III offers a comprehensive list of what each break represents. Break annotation is of great importance because physiologically a break has a possible resetting effect on the vocal folds' vibrations and thus its duration and strength significantly affects the shape of the F0 contour, not just within a rhyme but across phrases. During data generation, each speaker read the text in his/her natural manner, and these recordings were then hand annotated with break information. Allowing the break indexes to form interactions with the speaker's sex, the model can associate different rates of curvature declination among male and female speakers. This effect found to be usually associated with lower order breaks (faster variational components). Furthermore, the ability to allow different curvature declinations between speakers of different genders enables the modeling of more complex down-drift patterns. This approach allows an analogy to be drawn with the phrase component used in the Fujisaki modeling approach.[17] The different tones of each syllable may be associated with the accent component as proposed by Mixdorff. The linguistic data were transcribed using ASCII symbols[58] to encode the nine vowels [ə, ɚ, a, e, i, ɛ, y, o, u]. Combinations of these vowels, with and without final [−n, −ŋ], add up to 37 rhymes, which are listed in the supplementary material[16] (Sec. V).

As shown in Table IV, 13 possible covariates (not counting their interactions) were included in the model. Eleven of them account for fixed effects and two for random effects. The initial model incorporates three-way interactions and their embedded two- and one-way interactions. Three-way interactions have been known to be present in Taiwanese Mandarin and therefore were deemed as significant effects to incorporate[5,15,59] both in the form of previous_tone: current_tone: next_tone interaction as well as a previous_consonant: current_tone: next_consonant interaction. Consonant refers only to the consonant's voicing status, not the identity of the sound. Four levels were present in the consonant covariate. It is well attested that syllables with no initial consonant in Chinese can have an epenthetic glottal stop before the rhyme, as in the second syllable of [tɕiāuʔaù] "proud" [e.g., as in Lin[60] (pp. 113–115, 173–174)]. The glottal stop [ʔ] is defined as a

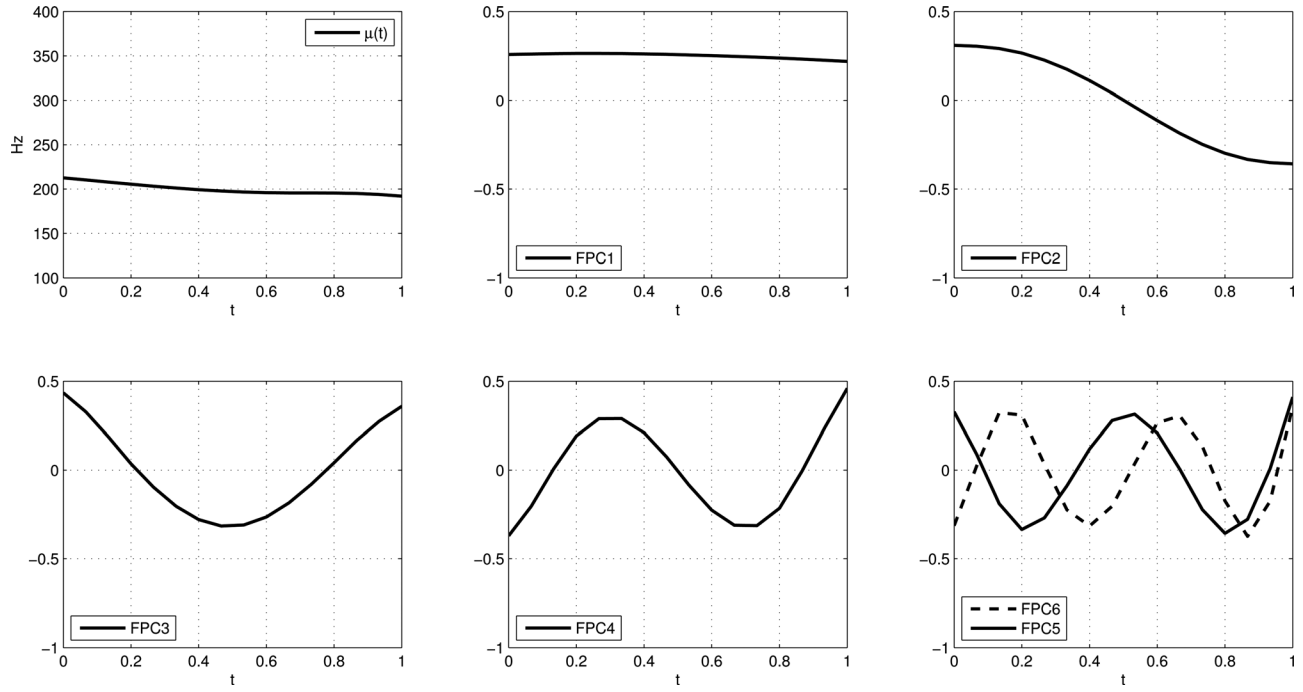

FIG. 2. Mean function and first, second, third, fourth, fifth, and sixth functional principal components. Together these account for 99.994% of the sample variance, but only the first four have linguistic meaning (99.933 % of samples variation) and as such the fifth and sixth were not used in the subsequent analysis.

voiceless sound, as the glottis cannot be simultaneously closed and vibrating. However, there are two reasons why we did not simply label all such syllables as beginning with a voiceless consonant. First, glottal stop is not always inserted in this context, being most likely after a higher order break, such as B4 or B5. Second, recent research on this topic [for example, Borroff[61] (p. 82)] has shown that voicing is often continuous through a perceived glottal stop. Thus, glottal stop is neither predictably present, nor always voiceless. For these reasons, we have labeled zero-initial as neither voiced nor voiceless but its own category. Furthermore, break counts were allowed to assume squared and cubic values, as this would allow up to a cubic form of down-drift in the final model. In addition to the inclusion of speaker identity as a random effect, which was included for reasons such as age, sex, health, and emotional condition among others, utterance instance was incorporated as a random effect, since it is known that pitch variation is associated with the utterance context (e.g., commands have a different F0 trajectory than questions).

The initial analysis shows that in all cases, the random effects of speaker and sentence were found to be significant, in spite of the fact that certain effects (especially sentence) appeared to be rather smaller than the actual model residuals (Table V).

Furthermore, it is shown that while third order interactions are not present in the analysis of the first FPC (this being partially expected as the first FPC appears to specify curve placement) third order interactions are present on the modeling of the second and third FPCs, those that appear to represent phonological rather than physiological features. In addition, the second eigenfunction reflects a considerable proportion (9.78%) of the total sample variation; thus significantly affecting the beginning and the end of the curve, dictating the syllable's overall trend.

We now outline the role that each individual eigenfunction plays in the F0 curve formation. As mentioned, the first eigenfunction appears to have a shifting effect on the F0 curve itself, raising or lowering the overall F0. In contrast, the second, third and fourth eigenfunctions have an average effect on the F0 curve quite close to 0 over the entire trajectory (as can easily be seen on the plots themselves). Therefore FPC-2, -3, and -4 do not have an overall shifting effect on the curve, but rather only dictate properties of the curve's shape, essentially bending it.

Finally, it should be pointed out that FPC-4 findings were rather interesting linguistically in the sense that the sinusoid-like suggested F0 formation does not correspond to any known/formal individual Mandarin tones. Nevertheless,

TABLE III. COSPRO break annotation.

Break type	Meaning
Break 1	Normal syllable boundary. In written Chinese, this corresponds to one character. (As this is our experimental data unit, B1 is equivalent to the mean value in the regressions and thus not included separately).
Break 2	Prosodic word boundary. Syllables group together into a word, which may or may not correspond to a lexical word.
Break 3	Prosodic phrase boundary. This break is marked by an audible pause.
Break 4	Breath group boundary. The speaker inhales.
Break 5	Prosodic group boundary. A complete speech paragraph.

TABLE IV. Covariates examined in relation to F0 production in Taiwanese Mandarin. Tone variables in a five-point scale representing tonal characterization, 5 indicating a toneless syllable, with 0 representing the fact that no rhyme precedes the current one (such as at the sentence start).

Effects	Values	Meaning	Notation mark
Fixed effects			
Previous tone	0:5	Tone of previous syllable, 0 no previous tone present	tn_{previous}
Current tone	1:5	Tone of syllable	tn_{current}
Following tone	0:5	Tone of following syllable, 0 no following tone present	tn_{next}
Previous consonant	0:3	0 is voiceless, 1 is voiced, 2 not present, 3 sil/short pause	cn_{previous}
Next consonant	0:3	0 is voiceless, 1 is voiced, 2 not present, 3 sil/short pause	cn_{next}
B2	linear	Position of the B2 break in sentence	B2
B3	linear	Position of the B3 break in sentence	B3
B4	linear	Position of the B4 break in sentence	B4
B5	linear	Position of the B5 break in sentence	B5
Sex	0:1	1 for male, 0 for female	sex
Rhyme type	1:37	Rhyme of syllable	rhyme_t
Random effects			
Speaker	$N(0, \sigma^2_{speaker})$	Speaker effect	spkrID
Sentence	$N(0, \sigma^2_{sentence})$	Sentence effect	sentence

it appears native speakers do indeed exhibit components of sinusoidal-shape in their production of F0, as FPC-4 accounts for 19 Hz variation, hence represents an audible signal. It is likely that this F0 curve component is needed to move between different tones in certain tonal configurations, as will be discussed below.

Reviewing each model eigenfunction in an individual manner it is important to stress the main qualitative features that each model suggests. We must also note that during the modeling procedure the fixed effects do not incorporate an intercept as such. Tone 1, the presence of a voiceless next consonant, the absence of a next or a previous tone and the vowel_type ə (schwa) served as intercepts in the cases of tones, consonants, next or previous tone and vowel type covariates, respectively. (For a detailed listing of the relevant covariates and the jackknifing results please refer to the supplementary material,[16] Secs. VI–VIII.) Taking into account the results from AIC and jackknifing, the following model for FPC_1 was chosen:

$$\begin{aligned} FPC_1 = \{ & [tn_{previous} * tn_{current}] + [tn_{current} * tn_{next}] \\ & + [cn_{previous} * tn_{current}] + [tn_{current} * cn_{next}] \\ & + [cn_{previous} * tn_{next}] \\ & + [(B2) + (B2)^2 + (B2)^3 + (B3) + (B3)^2 \\ & + (B3)^3 + (B4) + (B4)^2 + (B4)^3 \\ & + (B5) + (B5)^2 + (B5)^3] * Sex + [rhyme_t]\}\beta \\ & + \{[Sentence] + [Spkr\,ID]\}\gamma + \epsilon. \end{aligned} \quad (15)$$

The first eigenfunction is almost exclusively associated with the speaker's F0 curve placement. Complex third order tonal interactions were not present. The speaker-identify random effect is significantly high despite the inclusion of speakers' sex as a covariate. Thus, this random effect captures speaker related variance that cannot be accounted for by indexing the sex of the speaker alone. Tones-2, -3, and -4 register lower in F0 than tone 1. Also, a number of rhymes appear to have significant associations with the first eigenfunction, indicating that a number of rhymes have a characteristic influence or shift on F0 (see supplementary material,[16] Sec. XI). These results are all relatively well known, but it is reassuring to find them all present in the model.

The type of voicing of the rhyme's neighboring consonants is of significance for all tone types. Specifically, the voicing of the preceding consonant resulted in a statistically significant lower overall F0 placement, when compared to the F0 placement associated with a preceding voiceless consonant. Overall, voiced neighboring/initial consonants (including epenthetic glottal stop) resulted in lower F0 placements, although the value of the effect depended on the tone type.

Break types B2, B3, and B4 associated both with males and females are statistically significant emphasizing the role of speech units larger than the word (but smaller than the utterance) on the formation of F0. In contrast, B5 breaks, in effect syllable index within the utterance, did not appear significant individually in terms of p-values; however, AIC deemed them worthy of incorporating as a group, yielding a cubic curve, thus demonstrating that while one covariate

TABLE V. Random effects and 95% highest posterior density confidence intervals for the first, second, third, and fourth FPC scores models as produced by using 10 000 samples.

	FPC1 estimate (95 lower, 95 upper)	FPC2 estimate (95 lower, 95 upper)	FPC3 estimate (95 lower, 95 upper)	FPC4 estimate (95 lower, 95 upper)
Speaker	71.7510 (37.479, 152.281)	4.6889 (2.333, 22.050)	6.9769 (3.824, 16.035)	3.0921 (1.492, 6.543)
Sentence	30.8339 (26.875, 31.330)	3.4920 (2.782, 4.037)	1.9485 (1.637, 2.180)	0.5968 (0.350, 0.752)
Residual	118.9193 (118.306, 119.712)	45.0915 (44.833, 45.380)	21.2468 (21.126, 21.382)	12.1229 (12.053, 12.196)

value might exhibit insignificant effects, the group might be quite important. A more detailed examination of the break term coefficients reveals more information about the down-drift effects in the samples. These suggest that, as the speaker progresses, while F0 might exhibit short jumps because of the generally additive effect of B2, the negative effects of B3 and B4 start to carry more weight and the down-drift becomes more prominent forcing the F0 estimate to be lower. Furthermore, the break interactions with speaker's sex suggest that, male speakers do not exhibit B2-related effects to such an extent but due to their B3 and B4-related interaction their F0 track drifts to lower frequency levels more smoothly as the additive lowering effects of B3 and B4 influences become more prominent. These types of features are reminiscent of the kinds of features that can be explored using a Fujisaki approach to the data.

The model for FPC_2 was chosen as

$$\begin{aligned} FPC_2 = \{ & [tn_{previous} * tn_{current} * tn_{next}] \\ & + [cn_{previous} * tn_{current} * cn_{next}] \\ & + \Big[(B2) + (B2)^2 + (B2)^3 + (B3) + (B3)^2 \\ & + (B3)^3 + (B4) + (B4)^2 + (B4)^3\Big] * Sex \\ & + \Big[(B5) + (B5)^2 + (B5)^3\Big] + [rhyme_t]\}\beta \\ & + \{[Sentence] + [Spkr\,ID]\}\gamma + \epsilon. \end{aligned} \tag{16}$$

The second eigenfunction scores exhibit third order interactions incorporating both triplet types tested, previous_tone: current_tone: next_tone and previous_consonant: current_tone: next_consonant. These kinds of interactions are of importance as they reflect not only physiological but also linguistic relations in the language corpus. At first glance, only uncommon triples (such as the tone triple 1-4-3 or 1-3-2 and the consonant-vowel-consonant triplets where the tones-2 and -3 occur in-between voiced consonants) appear statistically significant. Nevertheless, the effects that both third order interactions groups have in the final modeling outcome were found to enhance the whole model in a statistically significant way by AIC. It is noteworthy that both the speaker's identity and the sentence random effects carry almost equal weighting in the eigenfunction's final formation, but their individual impacts are a whole scale of magnitude smaller than the model's residual (see Table V). Thus, while they are not excluded by the model during our selection procedure, it is clear that their effect (or rather lack of it) suggests that non-linguistic covariates play a lesser role in the formation of this FPC. As expected from the shape of FPC-2, tones 2 and 4 appear significantly affected by the second eigenfunction, as the slopes of these two tones are phonological mirror-images. As a consequence, the two have actual parameter values of opposite signs (-73 and 95 for tones 2 and 4, respectively). Analogous with the known Mandarin tones, the negative parameter effect in tone 2 will cause tone 2 curves to have an upward curvature, while a positive parameter effect in tone 4 will cause downwards bending of the syllable's curve. Fewer rhymes appear to be associated with FPC-2 and thus with the shaping of its contour. Breaks do come through as significant covariates, despite not having significant interactions with the speaker's sex, showing that the overall down-drift effect in an utterance is a sex-independent phenomenon for this FPC. Finally, the voicing nature of the adjacent neighboring consonants proved of importance both individually and in association with the syllable's tone. The influence of a voiced initial consonant was negative overall, resulting in lowering the start and raising the end of the F0 curve. However, the following consonant's voicing effect depended mostly on the associated tone.

The scores associated with FPC_3 had the following model chosen:

$$\begin{aligned} FPC_3 = \{ & [tn_{previous} * tn_{current}] + [tn_{current} * tn_{next}] \\ & + [tn_{previous} * tn_{next}] \\ & + [cn_{previous} * tn_{current} * cn_{next}] \\ & + \Big[(B2) + (B2)^2 + (B2)^3 + (B3) + (B3)^2 \\ & + (B3)^3\Big] * Sex + [rhyme_t]\}\beta \\ & + \{[Sentence] + [Spkr\,ID]\}\gamma + \epsilon. \end{aligned} \tag{17}$$

The third eigenfunction possibly plays a dual role. Firstly, it is most associated with tone 3 in terms of its covariate value, which is unsurprising given its shape. It also appears to have strong effects on many tonal and voicing interactions, indicating that it is being used to transition between syllables. In addition, the speaker's identity random effect appears to play a statistically significant role to the eigencomponent's final weighting, especially when compared to the sentence effect. FPC-3 appears to carry statistically significant associations with the majority of different rhymes considered; suggesting that a hill, valley or a flattening in the curvature of the rhyme of the vowel is a prominent feature. Furthermore emphasizing the linguistic and local relevance of FPC-3, B2 and B3 break types appear to have the highest association both as individual covariates and in interaction with sex.

As in the case of FPC2, the voicing nature of the surrounding consonants interacting with the current rhyme tone influences the final curvature. This effect was most prominent in the cases where the rhyme occurred immediately after a short pause or another rhyme (i.e., there was no preceding consonant) and resulted in the curvature exhibiting a clear hill-top tendency. Also noteworthy is that this eigenfunction appears to have significant interactions when modeling adjacent pairs of the same tone, its positive influence easily seen in the cases of tones-2 and -3.

The model for the fourth FPC was chosen as

$$\begin{aligned} FPC_4 = \{ & [tn_{previous} * tn_{current}] + [tn_{next}] \\ & + [cn_{previous} * tn_{current}] + [tn_{current} * cn_{next}] \\ & + [cn_{previous} * cn_{next}] \\ & + \Big[(B2) + (B2)^2 + (B2)^3 + (B3) + (B3)^2 \\ & + (B3)^3\Big] * Sex + (B4) + (B4)^2 + (B4)^3 \\ & + [rhyme_t]\}\beta + \{[Sentence] \\ & + [Spkr\,ID]\}\gamma + \epsilon. \end{aligned} \tag{18}$$

This fourth eigenfunction, which does not display the shape characteristics of a single Mandarin tone, shows strong association with the voicing of the next initial consonant. This eigenfunction appears to reflect strongly localized effects mostly associated with the transition from one tonal segment to another. As expected, specific tones do not exhibit correlation with this eigenfunction, however, the interaction between current_tone and next_consonant appears statistically significant in all cases; suggesting a phonetic functionality that is associated with linguistic characteristics of the following syllable. While only a handful of rhymes appeared to have statistical significance in terms of p-values, AIC does not exclude them, showing that at least part of the eigenfunction's shape is indeed reflected in the rhyme shaping. Another important issue is that breaks 2 and 3 (prosodic word and phrase) have much influence on the F0 contour through this eigenfunction. B4 (breath group) has a very small influence, and B5 (paragraph) was not deemed statistically significant enough to even incorporate. Thus, this eigenfunction reflects the influence of prosodic units no larger than the prosodic phrase. It can be suggested that such a small percentage of F0 variance approaches the limit of the explanatory power of our modeling rationale. Therefore, fluctuations smaller than this (small) magnitude are due to articulatory and/or phonetic effects that are beyond the mostly linguistic covariates the current model entails.

Choosing the relevant covariates from each FPC for the syllable of interest, summing them up and using this sum as a factor to weight the influence of each respective eigenfunction to the original sample mean yields the final F0 estimate (see Figs. 1 and 4). Here the estimates correspond to generic speakers and to estimations of the behavior of the underlying Gaussian process. The estimates do not specify individual speakers; therefore the random effects are set to 0 across all FPC's as random effects always have mean 0. As can be seen, the example tone estimates (Fig. 3) generated by the model exhibit qualitatively similar characteristics with those of the YR Chao tone chart.[62]

Table VI gives a brief overview of each eigencomponent model's performance in terms of adjusted R_a^2 with and without the incorporation of random effects.[25] It is immediately seen that the overall adjusted R_a^2 score is declining as the models try to capture the highly variable nature of each higher order individual eigencomponent. Nevertheless, in all cases the inclusion of random effects seems beneficial and was not rejected by the full sample AIC model comparison or the jackknifing model selection procedure. While the third and fourth components' R_a^2 are very low, this likely results from the inherent variability in the sample data being captured by these components, beyond the explanatory factors available to model the data (such as speaker mood through the experiment, changes in attention, etc.).

Given the break information in the model, it is also possible to construct the F0 track for rhymes over time. As can be seen in Fig. 5, the curves estimated from the models are not only fairly good fits to the data on a rhyme by rhyme basis (with of course the expected estimation error), but the overall time normalized track from rhyme to rhyme is captured through the break covariate estimation. (See supplementary material,[16] Table X, for a detailed listing of relevant covariates.) Thus, in a similar manner to the Fujisaki framework, estimation can be achieved for tracks both associated with single rhyme curves and also longer phrasal (multiple rhyme) instances.

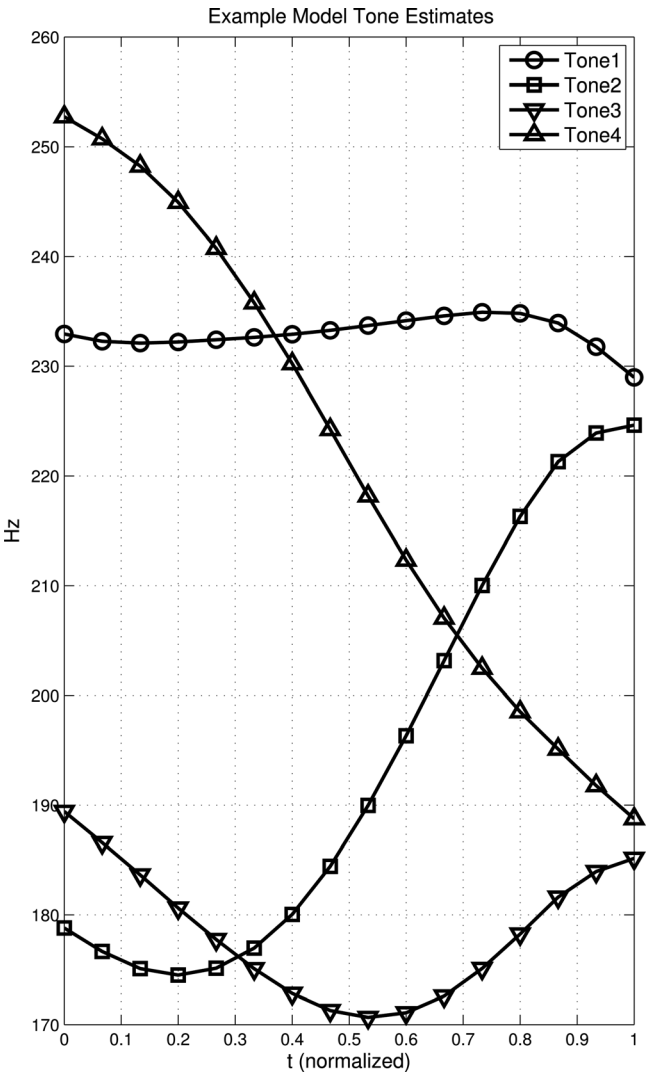

FIG. 3. Example tone estimates produced by the model utilizing all four FPC's. Tone 5 is not represented as it lacks a general estimate, always being significantly affected by non-standardized down-drift effects. Phonologically, toneless syllables do not specify a pitch target.

IV. DISCUSSION

Overall, the presented methodology allows for an analysis of the linguistic corpus at hand. Specifically, the qualitative analysis of the eigenfunctions suggests the strong

TABLE VI. Adjusted R^2 scores for the selected linear models before and after the inclusion of speaker and sentence related random effects.

FPC No.	$LM - R_a^2$	$LME - R_a^2$
FPC1	0.6271	0.7056
FPC2	0.6109	0.6161
FPC3	0.3645	0.4136
FPC4	0.1083	0.1491

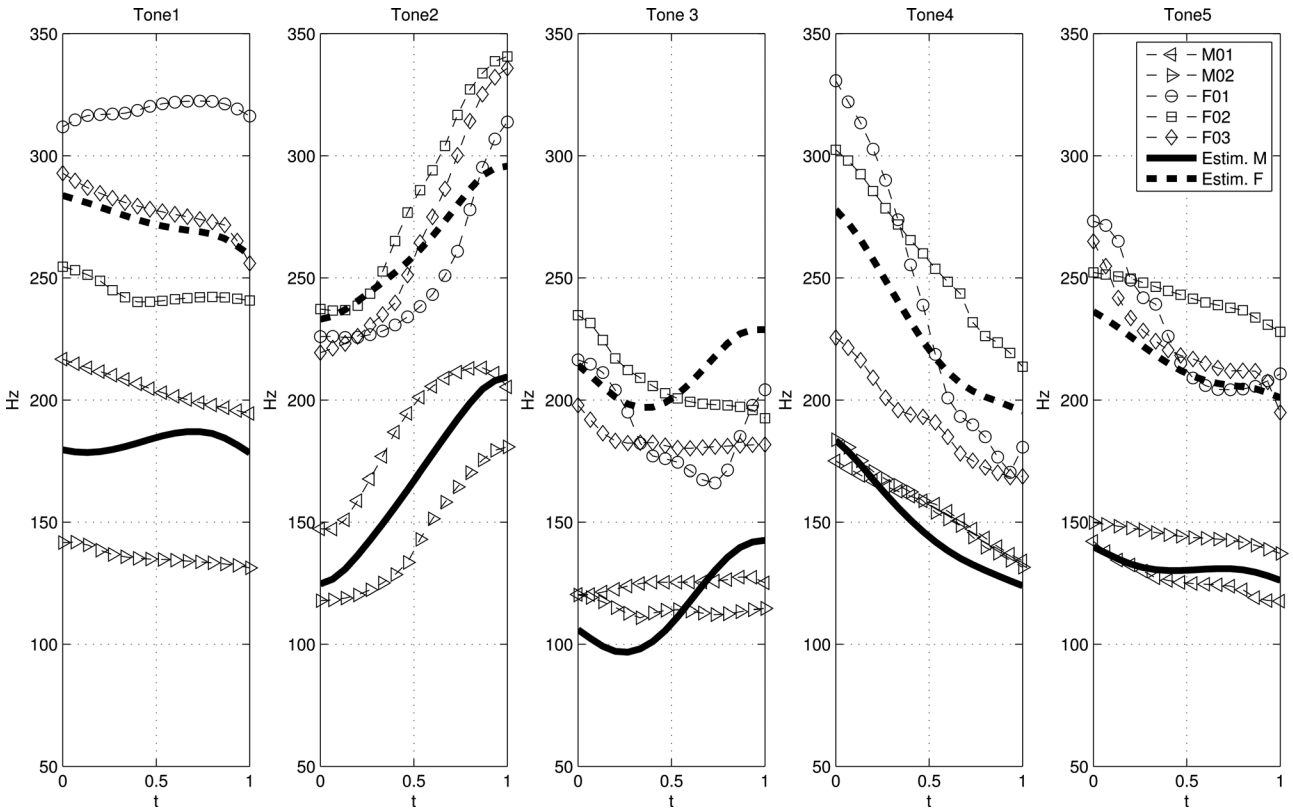

FIG. 4. One randomly selected syllable for each of the five tones; the functional estimates (bold) for each different tone are shown as well as the corresponding original speaker interpolated data over a dimensionless rhyme time interval *t*. [Estimated vowel rhymes: [uei, oŋ, əŋ, uan, ə] for each of the five tones, respectively. See supplementary material (Ref. 16) for contextual covariate information.]

dependence of pitch level to the speaker's identity. The influence of triplets in the case of tones 2 and 4 and the subsequent slopelike shape they exhibit is also demonstrated in the case of tone 2 where F0 initially drops before the rise, the effect being most prominent when tone 2 is spoken after either a tone 1 or tone 2. The model also suggests that statistically significant differences are present on the down-drift effect between speakers of different gender. Nevertheless, excepting FPC1 (the curve's F0 placement component), all the other FPC's did not show significant associations with the speakers' sex, suggesting that males and females have the same generic tone shapes; the actual shaping is statistically gender-independent. Furthermore, the fact that a number of rhymes have specific shaping attributes that are concurrently speaker and sentence independent is also put forward. The model proposes that the presence of voiced consonants adjunct to a rhyme alters its curvature to a noteworthy level; thus it is essentially validating empirically the sequential target approximation assumption used by Promon *et al*. in the qTA model.[23] Additionally, an interesting, yet not surprising, result is that as the modeling procedure focuses on higher order FPC's, higher order breaks (namely, B4 and B5) seem to carry decreasing importance to the final model. This result is in line with the fact that higher order FPC's reflect more localized effects influenced by changes in B2 and B3 indexing. The model estimates (Fig. 4) show that the proposed model succeeds in capturing the overall dynamics of the speaker's pronunciation, giving good qualitative and quantitative estimates. (Tone 1: sentence 564, word 2; tone 2: sentence 124, word 1; tone 3: sentence 336, word 1; tone 4: sentence 444, word 4; tone 5: sentence 529, word 3. See supplementary material,[16] Table IX, for a detailed listing of relevant covariates.) This success is obtained despite the fact that the sample exhibits large

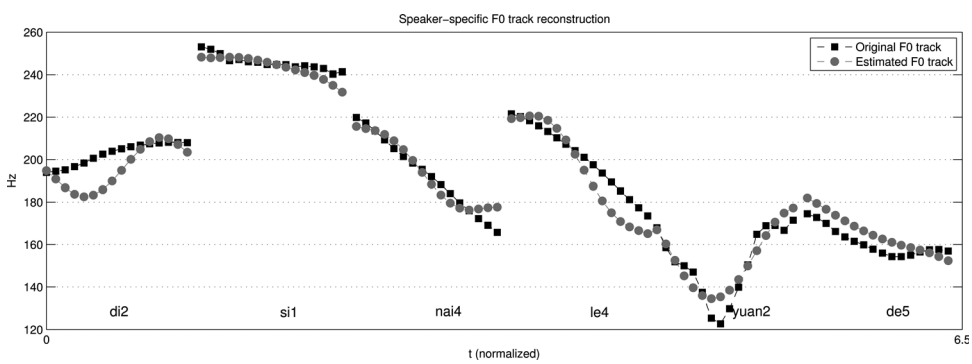

FIG. 5. Randomly chosen F0 trajectory over (normalized) time. Here six concurrent F0 tracks for rhymes are shown for speaker F03. As can be seen, the match is fairly close for most syllables, with the estimates associated with the break information controlling the temporal down drift effects [Tonal sequence: 2-1-4-4-2-5; estimated vowel rhymes: [i, ɿ, ai, ə, yɛn, ə]. See supplementary material (Ref. 16) for contextual covariate information.]

variance and possible distortion through its measurements even after the initial data were preprocessed. Note that Shih and Kochanski[63] ran into similar issues concerning distorted tone shapes. Collectively, these findings are in line with those of other studies,[5] specifically when reviewing the effect of adjacent tones. Durational differences are not taken into account by the current modeling approach. Possible future work would benefit from incorporating time-warping normalization on the rhyme time in order to ensure that possible discrepancies due to durational differences are excluded.

The current findings are also analogous to those of Aston *et al.*[15] in their study on Luobuzhai Qiang, a tonal Sino-Tibetan language of Sichuan Province in central-southern China. It could be of interest to review and compare these findings with those of other languages, especially those that are genealogically and geographically distant, to highlight any differences found in the components recovered from the F0 trajectory.

Each of the FPC_x models constructed are unit but not scale invariant; alternative models could be postulated for semitones or bark scale following the same methodology.

Indeed the analysis was repeated using a semitone scale but the contours recovered were almost identical. Other effects, such as the text frequency of the syllable were not incorporated as model covariates. While it could be argued that this would upgrade the overall performance of the model, this would nevertheless steer the model away from its phonological foundations. Therefore, inclusion of such factors as text frequency, duration, intonation pattern, etc., remains for future research. Moreover, because of the time-normalization, observed curvature fluctuations are per syllable rather than on an absolute time scale. To test our methods against a parametric family of basis functions which has previously been suggested (Ref. 9), the full body of the analysis was reimplemented using Legendre polynomials, shifted and normalized in L_2 $[0, 1]$ as a set of basis functions for the data instead of FPC's. This representation gave very similar explanatory results, because of Legendre polynomials having similar shape to the FPC's. However, as discussed in the Introduction, Legendre polynomials do not represent an optimal basis in terms of most variation of the data explained (see supplementary material,[16] Sec. IV) and thus the first four Legendre polynomials explain a smaller amount of the data's variation than the amount of variation explained by the first four eigenfunctions.

The model's novelty is that while the syllable curve was assumed to be part of the whole utterance as in the Fujisaki approach, the syllable curve itself was treated as a continuous random process modeled by different FPCs. In addition, micro-prosodic phenomena also known to be present are not systematically excluded by the current framework. In that sense, statistical methodology is the mechanism excluding irrelevant or immeasurable components of the sample. As the FPC's are orthogonal to each other, FPC scores account for non-overlapping variations. Higher degrees of FPC's might reflect further micro-prosodic variations than the ones recognized by this study, but as the total amount of information in these FPC's is considered below an auditory threshold, these FPC's are rendered unnecessary to the actual modeling procedure.

The future goals of this project are three-fold. First, by using the model, it may be possible to make meaningful inference from other corpora allowing more realistic speech recognition and speech processing. Secondly, by taking advantage of the surrogate variables generated (FPC's, covariance surfaces, etc.), possibilities arise to infer associations between languages that share common phonological characteristics, under a functional phylogenetic framework. Such framework has already been sketched by Aston *et al.*[64] Third, by validating this method on a language where many of the effects on F0 are known, it now becomes possible to investigate numerous effects and their interactions in the production of F0 in less-studied languages, and to be confident of the results.

ACKNOWLEDGMENTS

J.A.D.A. gratefully acknowledges EPSRC (UK) Grant No. EP/H046224/1 as well as the EPSRC/HEFCE CRiSM grant. J.P.E. gratefully acknowledges NSC Grant No. 97-2410-H-001-067-MY3 from the National Science Council (Taiwan).

[1]M. Schroeder, T. D. Rossing, F. Dunn, W. M. Hartmann, D. M. Campbell, and N. H. Fletcher, *Springer Handbook of Acoustics*, 1st ed. (Springer, Berlin, 2007), Chaps. 13 and 16.
[2]F. Nolan, "Acoustic phonetics—International encyclopedia of linguistics, William J. Frawley," (2003), URL http://www.oxford-linguistics.com/entry?entry=t202.e0008 (e-reference edition, date last viewed 2/2/11).
[3]S.-A. Jun, *Prosodic Typology: The Phonology of Intonation and Phrasing* (Oxford University Press, Oxford, UK, 2006), Chap. 2.
[4]A. W. Black and A. Hunt, "Generating f0 contours from tobi labels using linear regression," in *ICSLP* (1996).
[5]Y. Xu, "Effects of tone and focus on the formation and alignment of f0 contours," J. Phonetics **27**, 55–105 (1999).
[6]C. E. Rasmussen, *Gaussian Processes for Machine Learning* (MIT Press, Cambridge, 2006), Chaps. 1 and 2.
[7]M. Seeger, "Gaussian processes for machine learning," Int. J. Neural Syst. **14**, 2004 (2004).
[8]J. Ramsay and B. Silverman, *Applied Functional Data Analysis: Methods and Case Studies* (Springer Verlag, New York, 2002), Chap. 1.
[9]E. Grabe, G. Kochanski, and J. Coleman, "Connecting intonation labels to mathematical descriptions of fundamental frequency," Lang. Speech **50**, 281–310 (2007).
[10]D. Hirst and R. Espesser, "Automatic modelling of fundamental frequency using a quadratic sline function," Trav. Inst. Phonét. Aix **15**, 71–85 (1993).
[11]J. Ramsay and B. Silverman, *Functional Data Analysis* (Springer Verlag, New York, 1997), Chap. 6.
[12]S. Rangachari and P. Loizou, "A noise-estimation algorithm for highly non-stationary environments," Speech Commun. **48**, 220–231 (2006).
[13]M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," Speech Commun. **49**, 787–800 (2007).
[14]R. Baayen, D. Davidson, and D. Bates, "Mixed-effects modeling with crossed random effects for subjects and items," J. Memory Lang. **59**, 390–412 (2008).
[15]J. Aston, J. Chiou, and J. Evans, "Linguistic pitch analysis using functional principal component mixed effect models," J. R. Stat. Soc., Ser. C, Appl. Stat. **59**, 297–317 (2010).
[16]See supplementary material at http://dx.doi.org/10.1121/1.4714345 for additional information concerning the analysis in this paper.
[17]H. Fujisaki, "Information, prosody, and modeling-with emphasis on tonal features of speech," in *Speech Prosody 2004, International Conference (ISCA)* (2004).
[18]H. Mixdorff, H. Fujisaki, G. P. Chen, and Y. Hu, "Towards the automatic extraction of Fujisaki model parameters for Mandarin," in *Eighth*

European Conference on Speech Communication and Technology (ISCA) (2003).
[19]H. Mixdorff, "A novel approach to the fully automatic extraction of Fujisaki model parameters," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference* (2000), Vol. 3, pp. 1281–1284.
[20]J. Louw and E. Barnard, "Automatic intonation modeling with INTSINT," *Proceedings of the Pattern Recognition Association of South Africa* (2004), pp. 107–111.
[21]D. Hirst, "A Praat plugin for Momel and INTSINT with improved algorithms for modelling and coding intonation," in *Proceedings of the XVIth International Conference of Phonetic Sciences* (2007), pp. 1233–1236.
[22]P. Taylor, "Analysis and synthesis of intonation using the tilt model," J. Acoust. Soc. Am. **107**, 1697–1714 (2000).
[23]S. Prom-on, Y. Xu, and B. Thipakorn, "Quantitative target approximation model: Simulating underlying mechanisms of tones and intonations," in *Acoustics, Speech and Signal Processing, 2006, ICASSP 2006 Proceedings, 2006 IEEE International Conference* (2006), Vol. 1.
[24]J. Evans, M. Chu, J. Aston, and C. Su, "Linguistic and human effects on F0 in a tonal dialect of Qiang," Phonetica **67**, 82–99 (2010).
[25]J. Faraway, *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models* (CRC Press, Boca Raton, FL, 2006), Chaps. 1, 8, 10.
[26]J. O. Ramsay, K. G. Munhall, V. L. Gracco, and D. J. Ostry, "Functional data analyses of lip motion," J. Acoust. Soc. Am. **6**, 3718–3727 (1996).
[27]J. Lucero and A. Löfqvist, "Measures of articulatory variability in VCV sequences," Acoust. Res. Lett. Online **6**, 80 (2005).
[28]S. Lee, D. Byrd, and J. Krivokapic, "Functional data analysis of prosodic effects on articulatory timing," J. Acoust. Soc. Am. **119**, 1666–1671 (2006).
[29]D. Byrd, S. Lee, and R. Campos-Astorkiza, "Phrase boundary effects on the temporal kinematics of sequential tongue tip consonants," J. Acoust. Soc. Am. **123**, 4456–4465 (2008).
[30]L. L. Koening, J. C. Lucero, and E. Perlman, "Speech production variability in fricatives of children and adults: results of functional data analysis," J. Acoust. Soc. Am. **5**, 3158–3170 (2008).
[31]M. T. Jackson and R. S. McGowan, "Predicting midsagittal pharyngeal dimensions from measures of anterior tongue position in swedish vowels: Statistical considerations," J. Acoust. Soc. Am. **123**, 336–346 (2008).
[32]K. Reilly and C. Moore, "Respiratory movement patterns during vocalizations at 7 and 11 months of age," J. Speech Lang. Hear. Res. **52**, 223–239 (2009).
[33]J. Ni, R. Wang, and D. Xia, "A functional model for generation of local components of F0 contours in Chinese," in *Spoken Language, 1996. ICSLP 96. Proceedings, Fourth International Conference (IEEE)* (2002), Vol. 3, 1644–1647.
[34]C. Tseng, Y. Cheng, and C. Chang, "Sinica COSPRO and toolkit: Corpora and platform of Mandarin Chinese fluent speech," in *Proceedings of Oriental COCOSDA* (2005), pp. 6–8.
[35]C. Tseng, S. Pin, Y. Lee, H. Wang, and Y. Chen, "Fluent speech prosody: Framework and modeling," Speech Commun. **46**, 284–309 (2005).
[36]X. He and L. Deng, "Speech recognition, machine translation, and speech translation; a unified discriminative learning paradigm [lecture notes]," Sign. Process. Mag., IEEE **28**, 126–133 (2011).
[37]C. Tseng, Y. Cheng, and C. Chang, "Sinica COSPRO and toolkit—Corpora and platform of Mandarin Chinese fluent speech," in *Oriental COCOSDA* 2005, Jakarata, Indonesia (2005).
[38]F. Ferraty and P. Vieu, *Nonparametric Functional Data Analysis: Theory and Practice* (Springer Verlag, New York, 2006), Chaps. 1 and 4.
[39]J. Chiou, H. Müller, and J. Wang, "Functional quasi-likelihood regression models with smooth random effects," J. R. Stat. Soc.: Ser. B (Stat. Methodol.) **65**, 405–423 (2003).
[40]J. Mercer, "Functions of positive and negative type, and their connection with the theory of integral equations," Philos. Trans. R. Soc. London, Ser. A **209**, 415–446 (1909).
[41]F. Yao, H. Müller, and J. Wang, "Functional data analysis for sparse longitudinal data," J. Am. Stat. Assoc. **100**, 577–590 (2005).
[42]P. Castro, W. Lawton, and E. Sylvestre, "Principal modes of variation for processes with continuous sample curves," Technometrics **28**, 329–337 (1986).
[43]P. Hall, H. Müller, and J. Wang, "Properties of principal component methods for functional and longitudinal data analysis," Ann. Stat. **34**, 1493–1517 (2006).
[44]J. Rice and B. Silverman, "Estimating the mean and covariance structure nonparametrically when the data are curves," J. R. Stat. Soc., Ser. B (Methodol.) **53**, 233–243 (1991).
[45]P. Buser and M. Imbert, *Audition*, 1st ed. (MIT Press, Cambridge, 1992), Chap. 2.
[46]S. Sudhoff, *Methods in Empirical Prosody Research* (Walter De Gruyter, Berlin, 2006), Chap. 4.
[47]J. Pinheiro and D. Bates, *Mixed-Effects Models in S and S-PLUS* (Springer Verlag, New York, 2009), Chap. 2.
[48]B. West, K. Welch, and A. Galecki, *Linear Mixed Models: A Practical Guide using Statistical Software* (CRC Press, Boca Raton, FL, 2007), Chaps. 2 and 6.
[49]$r < n$ in usual cases.
[50]A. Davison, *Statistical Models* (Cambridge University Press, Cambridge, 2003), Chap. 4.
[51]G. Fitzmaurice, N. Laird, and J. Ware, *Applied Longitudinal Analysis* (Wiley-Interscience, New York, 2004), Chap. 7.
[52]M. Galassi, J. Theiler, J. Davies, and B. Gough, *GNU Scientific Library Reference Manual*, 3rd ed. (Network Theory Limited, Bristol, UK, 2009).
[53]MATLAB, version 7.10.0 (R2010a) (MathWorks, Natick, MA, 2010).
[54]R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria (2011), http://www. R-project.org (date last viewed 2/11/11).
[55]D. Bates and M. Maechler, *lme4: Linear mixed-effects models using S4 classes* (2011), http://CRAN.R-project.org/package=lme4, r package version 0.999375-41 (date last viewed 2/11/11).
[56]R. H. Baayen, *Language r: Data sets and functions with "Analyzing Linguistic Data: A practical introduction to statistics*" (2011), http://CRAN.R-project.org/ package=languageR, r package version 1.2 (date last viewed 2/11/11).
[57]R. Baayen, *Analyzing Linguistic Data: A Practical Introduction to Statistics using R* (Cambridge University Press, Cambridge, UK, 2008), Chap. 4.
[58]C. yu Tseng and F. Chiang Chou, "Machine readable phonetic transcription system for Chinese dialects spoken in Taiwan," J. Acoust. Soc. Jpn. **20**, 215–223 (1999).
[59]R. C. Torgerson, "A comparison of Beijing and Taiwan Mandarin tone register: An acoustic analysis of three native speech styles," Master's thesis, Brigham Young University, 2005.
[60]Y.-H. Lin, *The Sounds of Chinese* (Cambridge University Press, Cambridge, 2007).
[61]M. L. Borroff, "A landmark underspecification account of the patterning of glottal stop," Ph.D. thesis, Department of Linguistics, Stony Brook University, New York, 2007.
[62]Y. R. Chao, *A Grammar of Spoken Chinese* (University of California Press, Berkeley, 1968).
[63]C. Shih and G. Kochanski, "Chinese tone modeling with stem-ml," in *ICSLP* (2000), pp. 67–70.
[64]The Functional Phylogenetic Group, J. A. D. Aston, D. Buck, J. Coleman, C. J. Cotter, N. S. Jones, V. Macaulay, N. MacLeod, J. M. Moriarty, and A. Nevins, "Phylogenetic inference for function-valued traits: Speech sound evolution," Trends Ecol. Evol. **27**, 160–166 (2012).