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, UK

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

Jonathan P. Evans
Institute of Linguistics, Academia Sinica, 128 Academia Road Sec 2 Taipei 115 Taiwan R.O.C.

(Dated: September 17, 2011)

A model for fundamental frequency (F0, or commonly *pitch*) employing a functional principal component analysis (FPCA) framework is presented. The language in the presented study is Taiwanese Mandarin; this Sino-Tibetan language is rich in pitch-related information as the relative pitch curve is specified in the syllable of each word's lexical entry. The original 5 speaker corpus is preprocessed using a locally weighted least squares smoother. These smoothed curves are then utilized as input for the computation of the final FPC scores and their corresponding eigenfunctions. These scores are finally utilized in a series of penalized mixed effect models to build meaningful categorical prototypes. These prototypes appeared to confirm known tonal characteristics of the language, as well as suggest the presence of a sinusoid tonal component that is previously undocumented.

PACS numbers: 43.60.Cg, 43.60.Uv, 43.66.Hg
Keywords: Phonetic Analysis; Functional Data Analysis; Linear models

I. INTRODUCTION

A. Theoretical Background

This paper takes a computational statistics approach to acoustic phonetic analysis. Phonetic sound properties of research interest include the pulse, intensity, pitch, spectrum or duration of the examined sound segment, with pitch being the focus of this paper. Speech sounds in particular, consist of periodic waves characterized by their frequency and amplitude. F0 as a speech phenomenon is the major component of what a human listener identifies as pitch. The fundamental frequency (F0) of a speech sound relates (but does not equate) to how *fast* the vocal cords of the speaker vibrate and amplitude quantifies the intensity of that vibration[1]. Because direct measurement from the actual vocal cords are quite difficult to obtain, F0 also has to be defined in acoustic terms, in contrast with pitch which as a notion reflects a perceptual effect within the listener.

While in most Indo-European 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, despite the fact that F0 is obviously derived from a soundwave and thus a curve, linguistic studies treated it as a single point by utilizing target values[2] or by treating the F0 contour as a bounded rigid curve through processes of averaging[3]. Nevertheless, such approaches by necessity impose simplifying assumptions making interpretation difficult when considering a complete corpus of data. Furthermore, it is clear that they offer limited linguistic insights in cases of tonal languages where the pitch contour carries semantic content. To counter this we propose a model where F0 is characterized as a *curve*; acting as the realization of a *stochastic Gaussian process*.

Functional data analysis offers tools for analysing data that *"consist of functions -often but not always, smooth curves"*[4]. In the current study a functional principal component analysis (FPCA) is first performed on the dataset's F0 measurements with the ultimate aim of extracting the component's principal curves. Through this process we are immediately in the position of identifying characteristics regarding the fundamental frequency properties of the corpus analysed. Different approaches might utilize Legendre polynomials[5], quadratic splines[6] or go as far as utilizing Fourier analysis to derive lower and higher ranking polynomials that would correspond to

[a)]Electronic address: J.A.D.Aston@warwick.ac.uk

slower and faster varying components of the utterance. Nevertheless, these basis functions are not derived from the data themselves and are not guaranteed to be optimal as in the case of principal component curves[7]. In addition to that, the functional principal component scores (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. LME models allow the inclusion of several fixed, but also random effects regarding the nature of the data. In the current case, the difference between individual speakers due to genetic, environmental[8] or even chance factors[9] can be modelled as a series of random additive effects acting on the actual F0 contours[10,11]. By combining FPCA with LME models this project is able to propose a possible linguistic description, and subsequently offer explanatory insights, in the case study of Taiwanese Mandarin.

The F0 analysis performed, in the form of a categorical pitch analysis of a tonal language, enables not only the successful phonetic description of a language, but extends the result from a purely phonetic scope to a primitive syntactic scope regarding the language's phonemes by taking account of the syllable's position within its utterance. Moreover, the methodology presented here, addresses the issue that, while it has been widely accepted and documented that tones undergo variations due to phonetic processes in speech production that are attributed to fixed effects (eg. the sex of the speaker), immeasurable variables such as the length of the speaker's vocal cords or the state of her health, also affect the final F0 utterance. This "immeasurability" problem is countered by considering such covariates as random effects. This theoretical perception is not ad-hoc; it lays in direct analogy with the linguistic, para-linguistic and non-linguistic parameters information presented in the work of Fujisaki[12,13], which though in its original approach does not account for microprosodic effects. Mixdorff has extended the Fujisaki model implementations[14] to account for such effects by taking advantage of the MOMEL algorithm[6,15]. Other approaches might utilize the automatic intonation modelling approach as offered by the INTSINT[15,16] and/or the TILT[17] algorithmic implementations. In the present framework, assuming F0 as our dependant variable of interest, standard fixed effects such as the rhyme of a vowel correspond to linguistic effects, sentence variations and break points within the utterance to para-linguistic effects and speaker variations to non-linguistic effects.

As Evans et al. have already presented[18] and Aston et al. have further extended[11], the explanatory power that can be yielded from the application of LME models, is crucial in cases of tonal languages. In the current study the sum of the pitch contours is used; as such, while the two previously mentioned works focused on one position in a frame sentence, in the current project all the words in a read text are investigated, adding new dimensions of complexity and further enhancing the generality of our approach analyzing complete corpus data.

As a starting point, a smoothing and interpolation procedure is utilized to change the measurement from real-time into that of normalized "syllable" time. Next, regression effects are introduced to account for the effects

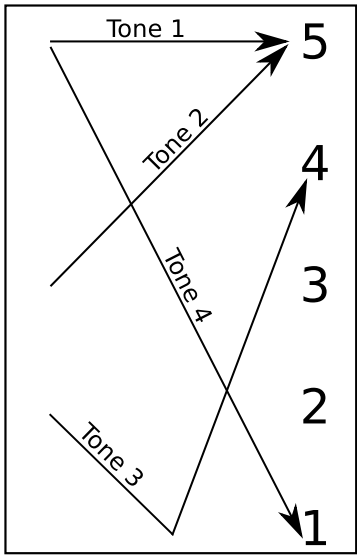

FIG. 1. Chart showing the relative changes in pitch for the four tones of Taiwanese Mandarin

of sentence breaks and help identify significant covariates in the function 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 as such a penalty on the number of regressors in the model seemed appropriate by utilizing an AIC approach (as outlined by Faraway[19]). This use of FPCA and mixed effects modelling offers a generalized semi-parametric approach to the linguistic modelling of Mandarin Chinese F0.

The seminal work for the application of FDA (Functional Data Analysis) in relation to Linguistics was that of Ramsey et al.[20], where X, Y and Z coordinates of lip motion where used in order to infer basic principals of lip coordination. Since then a number of speech production related questions in relation with articulatory issues[21–23] as well as with issues of physiological interests[24–26] have been addressed with FDA. The current work diversifies itself from the above mentioned projects by being the first employing an entire corpus as raw data. It does not limit itself in a small linguistic sample by a single speaker[13], employing monosyllabic utterances and a small number of sentences[27] and/or frames within the utterances[11,21,24]. As a direct consequence it extends the previous works done in F0 realization in Chinese[13,27] in a generalized and robust framework. In contract with existing intonation synthesis algorithms the current methodology's primary target is offering linguistic insights on the tonal mechanism of the language at hand and is presented as a auxiliary approach for existing speech synthesis algorithms.

B. Dataset Presentation

The Sinica Continuous Speech Prosody Corpora 1 (COSPRO-1) is a large-scale comprehensive data-set consisting of Taiwanese Mandarin recordings[28]. A sample

of 5 participants uttered a total of 599 predetermined sentence utterances each, that after phonetic processing resulted in a sum of 54707 frequency curves based on vowel instances only where each curve corresponds to a spoken syllable. All participants were native Taiwanese Mandarin speakers; 3 of them being women and 2 men. The recordings themselves were conducted by the Institute of Linguistics, Academia Sinica in 1994. Using the in-house developed speech processing software package Prosody[29], the fundamental frequency (F0) of the rhyme utterances was extracted and subsequently recorded at 10ms intervals. This interval was judged as the minimum for which actual pitch changes can be realized by a human listener. The recordings also included tone, rhyme, consonant as well as speech break/pause characterizations. The readers will also note that despite the fact that Taiwanese Mandarin is known to specify only four tones (see Figure 1), the analysis presented recognises a fifth; a dummy, short duration tone occurring in unstressed syllables, usually found at the end of utterances or phrases.

II. STATISTICAL METHODOLOGY

A. Functional Data Analysis

Ferraty and Vieu provide the following definition: *"A random variable x is called functional variable if it takes values in an infinite dimensional space (or functional space)"*[30]. Based on this 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*"[31]. As a logical result; given a stochastic process Y(t), $t\epsilon[0,1]$, the sample curves can be thought as having a mean $E[Y(t)] = \mu(t)$ and a covariance $Cov[Y(s), Y(t)] = C(s,t)$. Taking advantage of the symmetric nature of C ($C(s,t) = C(t,s)$) the following spectral decomposition follows by Mercer's theorem[32] for $C(s,t)$:

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

where $\lambda_1 \geq \lambda_2 \geq ... \geq 0$ 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 characterization markers of the sample's trajectories around an overall mean trend function[33], ie each PC gives a representation of the F0 contour components. As Castro et al. briefly summarized on their seminal work on Continuous Sample Curves[34], given a vector process $Y = (y_1, y_2, .., y_p)^T$, where $y_1, y_2, ..., y_p$ are scalar vectors, an expression of the form:

$$Z = M + \sum_{i=1}^{k} \alpha_i Z_i(t), \tag{2}$$

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

$$Y_p(t) = \mu(t) + \sum_{\nu=1}^{k} \alpha_{\nu,p}\phi_\nu(t), \tag{3}$$

where α_ν are once more the uncorrelated random variables with zero mean and are referring to the actual ν-th principal component score of the p_{th} subject and ϕ_ν are linear independent basis-functions, whose random trajectories are Y_p. This expansion (eq.3) is referred to as the Karhunen-Loève or functional principal component expansion of the stochastic process Y[35].

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"*[36]. In physical terms, smoothness of data is critical so that the discrete sample data is considered functional[7]. 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. The latter, kernel smoothing, being advocated as being an *optimal* choice in the case of local weighting[30], 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.[31] 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 centre of a smoothing window which acts into an interval $[t-b, t+b]$ where b is the fixed parameter commonly known as *bandwidth*. The formal definition of the smoother itself being :

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

where K is the kernel function selected, t is the argument of which the smoother S_L is used, b is the smoothing parameter meaning how *big* the window of the smoother will be in relation to actual available data-points and $(t_i, y(t_i))_{i=1,...,s}$ is the actual data scatter-plot consisting of s points. It must also be mentioned that for the data to be a suitable input for the FPCA to take place, besides the smoothing, time normalization is of importance. 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 vowel 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 per case, and based on this estimate, the basis of 16 points is chosen. The analysis was also conducted using 12- & 20-point interpolation so that the impact of the smoothing is more easily identified (but this yielded negligible differences). Furthermore, to ensure that 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 were not smoothed.

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 our 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:

$$\hat{\mu}(t_j) = \frac{1}{s}\sum_{i=1}^{s} y_i(t_j) \tag{5}$$

where s is the number of samples available and j is the number of points in each curve (in this case 16).

The final step in order to calculate the FPCA scores is actually the most straightforward. Following the same methodology as Aston et al.[11] the eigenfunctions are calculated by the spectral analysis of the estimated covariance matrix:

$$\hat{C}(t_k, t_l) = \frac{1}{s}\sum_{i=1}^{s}\{y_i(t_k) - \hat{\mu}(t_k)\}\{y_i(t_l) - \hat{\mu}(t_l)\}, \quad k, l\ \epsilon\{1, .., p\} \tag{6}$$

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 $\lambda_1, \lambda_2, ..., \lambda_\mu$ are the ordered eigenvalues of the system. Finally the FPCA $A_{i,\nu}$ scores are estimated as:

$$\hat{A}_{i,\nu} = \sum_{k=1}^{m}\{Y_i(t_{i,k}) - \hat{\mu}(t_{i,k})\}\hat{\phi_\nu}(t_{i,k})\Delta_{i,k}, \tag{8}$$

where $\Delta_{i,k} = t_{i,k} - t_{i,k-1}$. These scores, $A_{i,\nu}$, are the ones finally used for the estimation analysis by the LME. The choice and number of the FPCs used is related with the amount of variation that each of these components reflect. Because of the large number of available sample utterances, a relatively high number of FPCs is required in order to account for phonetic effects that might occur in just a relatively small number of sample instances. In addition to that, 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 FPCs reflecting variation that has actual *acoustic meaning* are selected. In reality, only pitch fluctuations above 10 Hz threshold can be registered by the human auditory system (just noticeable difference)[37]and this is utilized in the model selection.

B. Linear Mixed Effects Models

Having determined the sample's eigenfunctions and corresponding FPCs, the next step involves the LME model construction and selection. Linear mixed effects models are models in which both random and fixed effects occur linearly in the model's implementation. As Pinheiro and Bates[38] 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 as proposed by West et al.[39] combined with the distributions notion as presented by Faraway[19], a standard fixed effect model with normal errors:

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

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

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

where in the presented case: A_ν is the vector of length p of FPC scores associated with the ν-th FPC, X_ν is the $p \times n$ model matrix, the vector ϵ_ν of length p encapsulates the random variables representing the error in the relation, and β is a vector of length n that contains the linear (fixed) regression coefficients, where n 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 $p \times m$[40] associated with a vector γ of m random effects. We need to stress here that by definition random effects are random variables themselves[10]. As such, the γ vector will follow a multivariate Gaussian distribution such as $\gamma \sim N(0, D)$, where D represents the covariance matrix of the elements in vector γ. In similar manner, the error residual vector ϵ follows also a multivariate Gaussian where $\epsilon \sim N(0, R)$ and R is the covariance matrix for residuals in vector ϵ.

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

$$Var(a) = Var(Z\gamma) + Var(\epsilon) = ZDZ^T + \sigma^2 I \tag{11}$$

resulting in the unconditional distribution :

$$a \sim N(X\beta, \sigma^2 I + ZDZ^T) \tag{12}$$

Model construction requires the use of definition of goodness of fit to the data. 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 our 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 our 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 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. On the opposite side, ML is used for the model selection procedure when we are concerned with model comparisons. That is because, ultimately, the comparison of models involves the comparison of their fixed effects. As ReML will try to transform the fixed effect response in a way as described above, this would lead to a series of different transformations for each model setting, making them not comparable. 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. AIC for each model being defined as:

$$AIC = 2(-l(\hat{\theta}) + n) \tag{13}$$

where n is the number of covariates in the model examined and l the value of the log-likelihood function of model $\hat{\theta}$[41].

Summarising the issues raised in model selection, as these steps involve estimation and comparison of two otherwise nested models, the models were implemented by using ML rather than ReML. As such, the fixed effects are unchanged and the two models are comparable. Using standard statistical methodology AIC values above a |2| threshold were deemed significant enough to reject or accept one model over another. Nevertheless, in the case that values were encountered *close* to or below this threshold, a subsequent simulation approach should be taken. It must be noted though that for models such as the present where large number of covariates is present, this can be extremely cumbersome. Once the model has been selected with AIC, the covariate estimates are then found using the ReML approach. Finally in order to assess significance and give confidence intervals for the resulting estimates, highest posterior density intervals were found[10].

III. RESULTS

We must emphasize that while the statistical robustness of the methods employed is crucial, the phonetic significance and interpretation of the results are the actual targets of this project. Before any further analysis is conducted, raw phonetic data underwent smoothing and interpolating, yielding less noisy sample curves, with equally and "densely" distributed sample times. Because of the high-specificity that the analysis requires, 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 that effects that might only systematically alter a small number of sample curves are not missed in the analysis. Thus, the first 12 FPCs were selected as necessary to incorporate in the modelling procedure. This unusually large number of FPCs was also dictated by the fact that significant regression-related effects might actually appear in a small percentage of the actual sample variates. These 12 FPCs provide projection for the 99.992% of the true variation in the sample (Table I). Nevertheless, in a worst case scenario, even by accounting for such high variation, actual characteristics that may occur in 5 syllables or less within the corpus could be filtered away (based on the residual variation of the discounted FPCs).

Moreover, given the large number of samples, by taking the upper model percentile (99%) of the actual FPC scores and multiplying it by the maximum absolute value of each eigenfunction, we can effectively derive how much of the actual variation is attributed to each component in Hz, the unit that was originally used for measurement. This is of interest because in reality the human auditory system even in optimal cases fails to register frequency differences below a 10Hz threshold and thus any actual variation below that minimum would almost certainly remain unnoticed. This *relativity* cut-off threshold actually excludes all FPCs with rank equal or higher than 5, that were previously deemed as of possible importance (see Table II). As such we take account of concerns such as lopsided sparsity[42] by making sure that the excluded effects are truly inaudible.

FPC #	Individ. Variation	Cummul. 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

TABLE I. Individual and Cumulative Variation Percentage per FPC

As such, 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, the smoothness of the covariance function of this transformation is essential as well as the smoothness of

FPC#	Hz	FPC#	Hz
FPC1	133.3	FPC7	3.6
FPC2	64.0	FPC8	2.9
FPC3	35.8	FPC9	2.4
FPC4	19.1	FPC10	1.8
FPC5	8.9	FPC11	1.7
FPC6	5.7	FPC12	1.3

TABLE II. Actual Auditory Variation per FPC (in Hz) (human auditory sensitivity threshold is $\approx$ 10 Hz)

the eigenfunctions themselves. A visual inspection of our results confirm that the kernel smoothing undertaken was successful with the data being *smooth enough* for the notions of FDA to be applicable. Common assumptions regarding smoothness dictate a twice differentiable functional. The covariance function appears smooth throughout its values (Fig.2) and as well as the mean and FPC curves (Fig.3). It must be commented that the 5-th and 6-th FPCs seem somewhat less smooth in appearance, further signifying that the transformation starts to reach an explanatory threshold and these covariates start to exhibit *noisy* characteristics. 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 eigenfunctions curvature (Fig.3). This result concurs with the assumed contour shapes of Grabe et al.[5] where Legendre polynomials $L0$ to $L3$ were utilized for the contour basis of F0 to examine intonation.

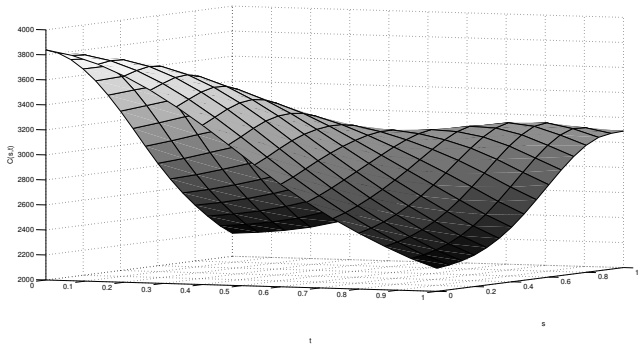

FIG. 2. Covariance function of the 54707 smoothed F0 sample curves, exhibiting smooth behaviour

While the kernel smoothing and interpolation was implemented by a custom built C++ program written by the first author, the calculation for the eigenfunction decomposition and the production of the FPC scores was conducted using standard built-in MATLAB procedures. The rest of the analysis was based on the usage of the statistical environment R[43]. Except for the obvious standard R methods used (qqplot(), lm(), etc.) the major body of the analysis was mostly done by utilizing methods from the statistical package lme4[44](for the LME model estimation and prediction) and languageR[45] (for the MCMC sampling required for the construction of confidence intervals relating to the model's estimators).

The model selection procedure was initiated by selecting the *largest* possible linguistically relevant model and then *de-constructing* it using AIC; excluding covariates that were viewed as insignificant. The following equation presents the original basis equation:

$$\begin{aligned} FPC_X = \{&[tn_{previous} * tn_{current} * tn_{next}]+ \\ &[cn_{previous} * tn_{current} * cn_{next}]+ \\ &[(B2) + (B2)^2 + (B2)^3+ \\ &(B3) + (B3)^2 + (B3)^3+ \\ &(B4) + (B4)^2 + (B4)^3+ \\ &(B5) + (B5)^2 + (B5)^3] * Sex+ \\ &[rhyme_t] + [Sentence] + [SpkrID]\}\beta + \epsilon \end{aligned} \tag{14}$$

Here the standard R notation is used for simplicity regarding the interaction effects; [K *L] representing a short-hand notation for [K + L + K:L] where the colon specifies the interaction of the covariates to its left and right[46]. Table III offers a comprehensive list for what each covariate stands for. It must be pointed out that from the set of fixed effects only *break counts* are of numerical value as all the other fixed covariates are in factor form. Break (or pause) counts 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 disjuncture between two words*", as defined in the ToBi annotations[2]. 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 sentence; essentially signifying a sentence boundary pause. Breaks B3 & B4 represent intermediate or intonational phrase stops respectively. Break annotation is of great importance because physiologically a "break" has a "resetting" effect on the vocal cords' vibrations and thus it's duration and strength significantly affects the shape of the F0 contour. Allowing the break indexes to be form interactions with the speaker's sex; the model can associate different rates of curvature declination among male and female speakers. Finally, it must also be stressed here that the term *rhyme* is used for *convenience*. While Taiwanese Mandarin has 15 vowels, the annotated data used 37 actual rhymes, allowing differentiation within vowels.

A total of 13 possible variables was deemed of possible research interest. As shown in Table III, 11 of them account for fixed effects and 2 for random. The initial model incorporates 3-way interactions and their embedded 2- and 1-way interactions. 3-way interactions have been known to be present in Taiwanese Mandarin and therefore were deemed as a significant effect to incorporate[3,11,47] both in the form previous_tone : current_tone : next_tone interaction as well as a previous_consonant : current_tone : next_consonant interaction. 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. Other than the inclusion of speaker as a random effect, for reasons regarding age, sex, health and emotional condition among

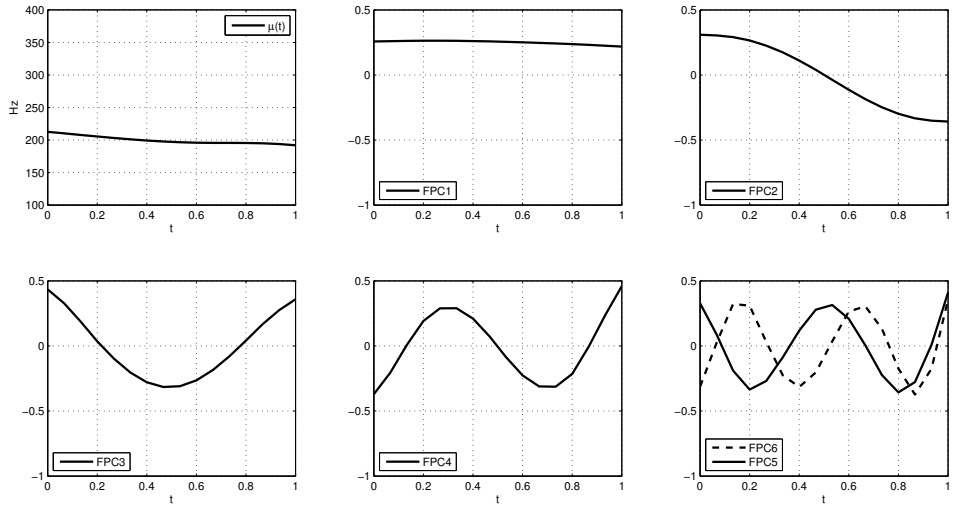

FIG. 3. Mean Function and 1st, 2nd, 3rd, 4th, 5th and 6th Functional Principal Components. Together these account for 99.994% of the sample variance but actually only the first four having linguistic meaning (99.933 % of samples variation) and as such the 5th and 6th were not used in the subsequent analysis.

Effects	Values	Meaning	Notation-mark
Fixed effects			
previous tone	0:5	Tone of previous syllable	$tn_{previous}$
current tone	1:5	Tone of syllable	$tn_{current}$
following tone	0:5	Tone of following syllable	tn_{next}
previous consonant	0:1	0 is voiceless, 1 is voiced	$cn_{previous}$
next consonant	0:1	0 is voiceless, 1 is voiced	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

TABLE III. Covariates examined in relation to the F0-production in Taiwanese Mandarin. Tone variables in a 5-point scale representing tonal characterization, 0 indicating no tone present

others, sentence instance was incorporated as a random effect, as it is known, that pitch variation is associated with the sentence context (eg. Commands have a different F0 trajectory than questions). As noted in the introduction, the analogy between the current approach and the Fujisaki model is further emphasised by the role of the breaks and the random effects in our modelling approach.

The initial analysis presents 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 IV).

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 actual curve placement, third order interactions are present on the modelling of the second FPC, which actually starts incorporating lexical rather than physiological features. In addition, the second eigenfunction appears to reflect a rather substantial 9.2% of the total sample variation; thus significantly affecting the beginning and the end of the curve, providing evidence for the presence of a down-drift effect in the speaker pronunciation. Before continuing it should be also noted that the current findings are in direct analogy with that of Aston et al.[11] on their study on Luobuzhai Qiang, a tonal Sino-Tibetan language of Sichuan Province in central-southern China. This fact could reflect a series of universal features among this language family: it could be of interest

to review and compare these findings with that of other languages, possibly of Indo-European origin, to highlight any differences found.

It will be reasonable before proceeding with the analysis to outline the *role* that each individual eigenfunction plays in the actual F0 curve formation. As mentioned the first eigenfunction appears to have a shifting on the F0 curve itself; contrary to that the second, third and fourth eigenfunctions have a *bending* effect on the F0 curvature allowing it to exhibit content related characteristics. FPC-2, -3 and -4 have an average effect on the F0 curve quite close to 0 over the entire trajectory (as can easily be seen by the plots themselves) and therefore do not have an overall shifting effect in the curve, but only dictate properties of the curve's shape.

Finally it should be pointed out that FPC-4 findings were rather interesting linguistically in the sense that the sinusoid-like F0 formation that is suggested does not correspond to any known/formal individual Mandarin tones. Nevertheless, it appears native speakers do indeed exhibit components of sinusoidal-shape in their final speech utterance, as FPC-4 accounts for 19Hz variation, hence represents audible signal. It is likely that this F0 curve component is needed to move between different tones in certain tonal configurations.

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 modelling 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[48].

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

The first eigenfunction is almost exclusively associated with the actual speaker pitch. As a result, complex third order interactions were not present. On the contrary, the Speaker-Identify random effect is significantly high despite the inclusion of Speaker-Sex as a covariate. Thus, the model captures speaker related variance that can not be accounted for by indexing the sex of the speaker alone. Tones-2, -3 and -4 register lower than Tone-1 ; also, a number of vowel types appear to have a significant associations with the first eigenfunction indicating that a number of vowels have a characteristic influence or shift on F0 (see Supplementary Material). The voicing of the previous consonant of the rhyme is of significance for all tone types; on the contrary, only in the case of Tone-4 does the voicing status of the next initial consonant appear to be significant.

Break types B2, B3 and B4 associated both with males and females appear statistically significant emphasising the importance that the *resetting* effect carries on the actual pitch formation. While B5 breaks, in effect syllable index within the utterance, did not appear significant individually in terms of p-values, AIC deemed them worthy of incorporating as a group yielding a cubic curve, thus noting that while one covariate value might exhibit insignificant effects, the group might be quite important. A more detailed examination of the break term coefficients yields more information about the downdrift effects in the samples. These suggest that, while F0 might exhibit short jumps, reflecting the generally additive effect of B2, as the speaker progresses, and thus the negative effects of B3 and B4 start to carry more weight, the down-drift becomes more prominent forcing the F0 estimate to be lower. Furthermore, the interaction with sex suggest that, male speakers do not exhibit this B2-related effects to such an extent and therefore they drift to lower frequencies more smoothly but also with less intensity as their B3 and B4-related down-drift effects are less prominent.

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

The second eigenfunction scores are actually the only ones that exhibit third order interactions incorporating both *triplet* types tested, previous_tone : current_tone : next_tone and previous_consonant : current_tone : next_consonant. These kind of interactions are of importance as they reflect not only physiological but also linguistic relations in the actual language corpus. At a first glance, only a few *unusual* triples (such as the tone triple 1-4-3 or the tones -2 and -3 being in-between voiced consonants) appear statistically significant. Nevertheless the effects that both third order interactions groups have in the final modelling outcome was found to enhance the whole model in a statistically significant way. As expected from the shape of FPC-2, tones -2 and -4 appear significantly affected by the second eigenfunction, as Tone-2 is the *phonological mirror image* of Tone-4 and vice versa. As such, the two have actual parameter values of opposite signs. Drawing an analogy with the known Mandarin tones; on the one hand, the negative parameter effect in Tone-2 will cause Tone-2 curves to have an upward curvature. On the other hand, the positive parameter effect in Tone-4 will cause an additional downwards bending to the syllable's curvature. Less *rhymes* appear to be associated with this FPC and thus with this shaped contour. Also breaks came through as a significant covariates, despite not having an interaction with

	FPC1 Estimate (95lower,95upper)	FPC2 Estimate (95lower,95upper)	FPC3 Estimate (95lower,95upper)	FPC4 Estimate (95lower,95upper)
Speaker	121.5021	4.7430	7.0617	3.0607
	(72.2604,169.9225)	(2.2167,34.6370)	(4.1190,20.5400)	(1.6901,8.8638)
Sentence	30.4051	4.0693	2.3340	0.8221
	(26.5889,30.9202)	(3.3102,4.5884)	(1.9725,2.5854)	(0.6318,0.9682)
Residual	119.6659	46.1604	22.8243	12.4354
	(119.0102,120.4179)	(45.8731,46.4320)	(22.6990,22.9706)	(12.3624,12.5119)

TABLE IV. Random Effects and 95% highest posterior density confidence intervals for the 1st, 2nd, 3rd and 4th FPC scores models as produced by using 10000 samples.

Sex. Nevertheless, B3, intermediate phrase stops, appeared to significantly influence the eigenfunction's behaviour in association with Sex, presenting a notable exception. Moreover, the voicing nature of the neighbouring consonants proved of importance both individually and in association with the syllable's tone.

$$\begin{aligned} FPC_3 = \{&[tn_{previous} * tn_{current}] + [tn_{current} * tn_{next}] + \\ &[tn_{previous} * tn_{next}] + \\ &[cn_{previous} * tn_{current}] + [tn_{current} * cn_{next}] + \\ &[(B2) + (B2)^2 + (B2)^3 + (B3) + (B3)^2 + \\ &(B3)^3] * Sex + (B4) + (B4)^2 + \\ &(B4)^3 + (B5) + (B5)^2 + (B5)^3 + \\ &[rhyme_t] + [Sentence] + [SpkrID]\}\beta + \epsilon \end{aligned} \tag{17}$$

The third eigenfunction shares a number of characteristics with the first one in its tone and consonant interactions. The similarities though stop here, as the most important covariates in this eigenfunction appears to be *vowel_rhymes*. FPC-3 appears to carry statistically significant associations with a number of different *rhymes*. Also and furthermore emphasising the linguistic 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 in interaction with the current syllable tone appeared extremely relevant to the final curvature realization. Unsurprisingly tones -2 and -3 are the ones having strong association with this FPC; in the case of tone-2 allows for a small initial drop so the that "raise" can be emphasized (or assisted) and in the case of tone-3 it does in essence "shape" the curvature of the tone itself.

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

The "unusual" fourth eigenfunction is the only one showing strong association with the voicing of the next initial consonant. While as expected known tones do not exhibit correlation with this eigenfunction, the interaction between current_tone and next_consonant appears statistically significant in all cases. This eigenfunction appears to reflect strongly localized effects. It must be noted that while only a handful of rhymes appeared to have statistical significant 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 and it is not entirely a purely random artefact due to speaker variation. Another notable issue is that only some break types appear to strongly influence the F0 contour through this eigenfunction. For example, B5 was deemed *not* statistically significant to incorporate in this model indicating that this eigenfunction reflects local shape characteristics, but is not characteristic of the shape of the entire utterance contour trajectory.

Having the covariates proposed for each eigenfunction constructing the actual F0 estimate is straightforward. Choosing the relevant covariates from each FPC for the syllable of interest, summing them up and using that number as a factor to weight the influence of each respective eigenfunction to the original sample mean, yields the final F0 estimate (see Figure 4). Here the estimates are corresponding to generic speakers. These estimates correspond to estimations of the behaviour of the underlying Gaussian process and they do not correspond to specific speakers individually; more specifically, the random effects are set to 0 across all FPCs as 0's are their expected values.

IV. DISCUSSION

Overall, the qualitative analysis of the eigenfunctions suggests the strong dependence of pitch level to the Speaker_ID. The influence of triplets in the case of tones -2 and -4 and the subsequent "drifting" shape they exhibit is also prominently presented. The model puts forward the fact that a number of rhymes have specific shaping attributes which are concurrently speaker and content independent. Finally it proposes that the presence of voiced consonants adjunct to a rhyme alter its original curvature to an audible level that is reflected in the pronunciation of that rhyme.[49]. As it can be seen from the model estimates (Figure 4), the proposed model does succeed in capturing the overall dynamics of the speaker's

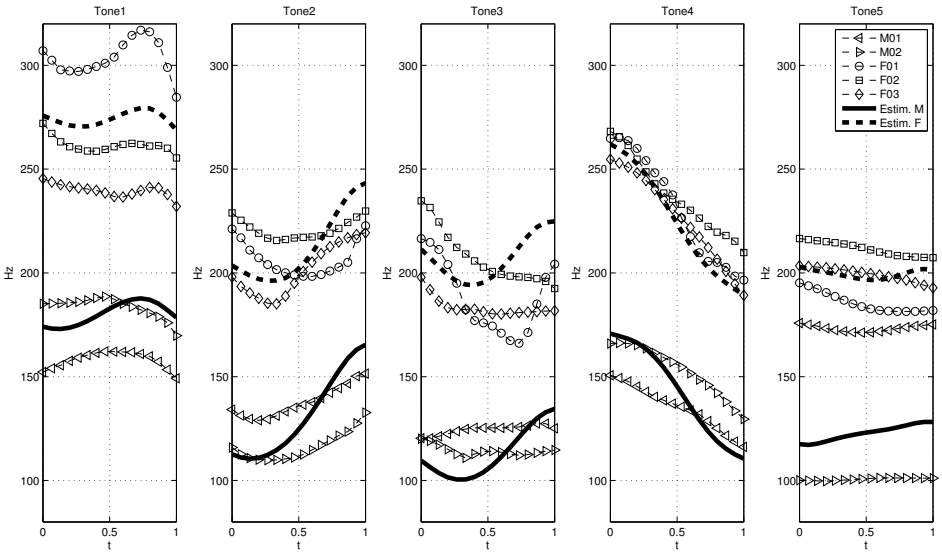

FIG. 4. Randomly selected 5 utterances and subsequently selecting 5 syllables so that all 5 tones are represented; the estimates for each different tone are shown as well as the corresponding original speaker interpolated curves over a dimensionless vowel time interval t .

pronunciation giving good qualitative and quantitative estimates given that the sample was quite noisy.

Each of the *FPCx* models constructed are unit but not scale invariant; alternative models could be postulated for semitones or bark scale following the same exact 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 was not incorporated as a model covariate. While it can be argued that this would upgrade the overall performance of the model, this would nevertheless steer the model away from each sound's phonetic foundations. As such it remains as point of future research. Moreover, because of the time-normalization, observed curvature fluctuations are "per syllable" rather than on an absolute time scale. The full body of the analysis was re-implemented using Legendre polynomials, shifted and normalized in $L_2(0,1)$, as basis instead of FPCs, presenting similar explanatory results, because of their similar shaping with the FPCs. Nevertheless, it must be noted that Legendre polynomials do not represent an optimal basis, under the usage of the RSS of reconstructed curve as a universal metric for the suitability of the curve reconstruction, and overall reflect smaller percentages of sample variation. FPCs represent an empirical basis that outperforms any parametric basis selected.

As outlined in the introduction; the model's novelty is that the curve syllable was assumed to be part of the whole utterance and itself being a continuous random process. In addition to that micro-prosodic phenomena also known to be present are not systematically excluded by the current framework. In that sense, *Statistics* is the mechanism excluding irrelevant or immeasurable components of the sample. As the FPC's are orthogonal to each other, FPC's score account for non-overlapping variations. As such, higher degrees of FPC's might reflect micro-prosodic variation, but as the total amount of information in those considered is below an auditory threshold, this rendering them unnecessary to the actual modelling procedure.

The study at hand provides to our best of knowledge, one of the first robust FDA approach in corpora analysis, as FDA is outlined by Ramsey & Dalzell[50]: a)The data were chosen to lay in the $L_2(0,1)$ space that is being associated with a dimensionless vowel-time, b) the analysis conducted was specifically selected to utilize FPCA, the most insightful tool for our research question, c) we determined a finite dimensional observation vector for our work, namely the FPC score corresponding to each curve, and d) we interpreted out data (Curves) and the meta-data (FPCs) in a manner reasonable and meaningful phonetically as well as statistically.The future goals of this project are two-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 (FPCs, Covariance Surfaces etc.), possibilities arise to infer associations between languages under a functional phylogenetic framework.

[1] 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).
[2] M. Beckman and J. Hirschberg, "The ToBI annotation conventions", Ohio State University (1994).
[3] Y. Xu, "Effects of tone and focus on the formation and alignment of f0 contours", Journal of Phonetics **27**, 55–

105 (1999).
[4] J. Ramsay and B. Silverman, *Applied functional data analysis: methods and case studies* (Springer Verlag, New York) (2002), chapt.1.
[5] E. Grabe, G. Kochanski, and J. Coleman, "Connecting intonation labels to mathematical descriptions of fundamental frequency", Language and Speech **50**, 281–310 (2007).
[6] D. Hirst and R. Espesser, "Automatic modelling of fundamental frequency using a quadratic sline function", Travaux de l'Institut de phonétique d'Aix **15**, 71–85 (1993).
[7] J. Ramsay and B. Silverman, *Functional data analysis* (Springer Verlag, New York) (1997), chapt.6.
[8] S. Rangachari and P. Loizou, "A noise-estimation algorithm for highly non-stationary environments", Speech Communication **48**, 220–231 (2006).
[9] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech", Speech Communication **49**, 787–800 (2007).
[10] R. Baayen, D. Davidson, and D. Bates, "Mixed-effects modeling with crossed random effects for subjects and items", Journal of Memory and Language **59**, 390–412 (2008).
[11] J. Aston, J. Chiou, and J. Evans, "Linguistic pitch analysis using functional principal component mixed effect models", Journal of the Royal Statistical Society: Series C (Applied Statistics) **59**, 297–317 (2010).
[12] H. Fujisaki, "Information, prosody, and modeling-with emphasis on tonal features of speech", in *Speech Prosody 2004, International Conference* (ISCA) (2004).
[13] 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).
[14] 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 on*, volume 3, 1281–1284 (IEEE) (2000).
[15] 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*, 1233–1236 (2007).
[16] J. Louw and E. Barnard, "Automatic intonation modeling with INTSINT", Proceedings of the Pattern Recognition Association of South Africa 107–111 (2004).
[17] P. Taylor, "Analysis and synthesis of intonation using the tilt model", The Journal of the acoustical society of America **107**, 1697 (2000).
[18] 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).
[19] J. Faraway, *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models* (CRC Press, Boca Raton, FL) (2006), chapt.1,8,10.
[20] J. O. Ramsay, K. G. Munhall, V. L. Gracco, and D. J. Ostry, "Functional data analyses of lip motion", Journal of the Acoustical Society of America **6**, 3718–3727 (1996).
[21] J. Lucero and A. Löfqvist, "Measures of articulatory variability in VCV sequences", Acoustics Research Letters Online **6**, 80 (2005).
[22] S. Lee, D. Byrd, and J. Krivokapic, "Functional data analysis of prosodic effects on articulatory timing", Journal of the Acoustical Society of America **119**, 1666–1671 (2006).
[23] D. Byrd, S. Lee, and R. Campos-Astorkiza, "Phrase boundary effects on the temporal kinematics of sequential tongue tip consonants", Journal of the Acoustical Society of America **123**, 4456–65 (2008).
[24] L. L. Koening, J. C. Lucero, and E. Perlman, "Speech production variability in fricatives of children and adults: results of functional data analysis.", Journal of the Acoustical Society of America **5**, 3158–3170 (2008).
[25] M. T. Jackson and R. S. McGowan, "Predicting midsagittal pharyngeal dimensions from measures of anterior tongue position in swedish vowels: statistical considerations", Journal of the Acoustical Society of America **123**, 336–46 (2008).
[26] K. Reilly and C. Moore, "Respiratory movement patterns during vocalizations at 7 and 11 months of age", Journal of Speech, Language, and Hearing Research **52**, 223–239 (2009).
[27] 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 on*, volume 3, 1644–1647 (IEEE) (2002).
[28] C. Tseng, Y. Cheng, and C. Chang, "Sinica COSPRO and Toolkit: Corpora and Platform of Mandarin Chinese Fluent Speech", in *Proceedings of Oriental COCOSDA*, 6–8 (2005).
[29] C. Tseng, S. Pin, Y. Lee, H. Wang, and Y. Chen, "Fluent speech prosody: Framework and modeling", Speech Communication **46**, 284–309 (2005).
[30] F. Ferraty and P. Vieu, *Nonparametric functional data analysis: theory and practice* (Springer Verlag, New York) (2006), chapt.1.
[31] J. Chiou, H. Müller, and J. Wang, "Functional quasi-likelihood regression models with smooth random effects", Journal of the Royal Statistical Society: Series B (Statistical Methodology) **65**, 405–423 (2003).
[32] J. Mercer, "Functions of positive and negative type, and their connection with the theory of integral equations", Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character **209**, 415–446 (1909).
[33] F. Yao, H. Müller, and J. Wang, "Functional data analysis for sparse longitudinal data", Journal of the American Statistical Association **100**, 577–590 (2005).
[34] P. Castro, W. Lawton, and E. Sylvestre, "Principal modes of variation for processes with continuous sample curves", Technometrics **28**, 329–337 (1986).
[35] P. Hall, H. Müller, and J. Wang, "Properties of principal component methods for functional and longitudinal data analysis", The Annals of Statistics **34**, 1493–1517 (2006).
[36] J. Rice and B. Silverman, "Estimating the mean and covariance structure nonparametrically when the data are curves", Journal of the Royal Statistical Society. Series B (Methodological) **53**, 233–243 (1991).
[37] B. West, K. Welch, and A. Galecki, *Linear mixed models: a practical guide using statistical software* (CRC Press, , Boca Raton, FL) (2007), chapt.2,6.
[38] J. Pinheiro and D. Bates, *Mixed-effects models in S and S-PLUS* (Springer Verlag, New York) (2009), chapt.2.
[39] S. Sudhoff, *Methods in empirical prosody research* (Walter De Gruyter Inc. Berlin) (2006), chapt. 4.
[40] $m < n$ in usual cases.
[41] A. Davison, *Statistical Models* (Cambridge University Press) (2003), chapt. 4.
[42] J. van Santen, "Quantitative modeling of segmental duration", in *Proceedings of the workshop on Human Language Technology*, 323–328 (Association for Computational Linguistics) (1993).

[43] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria (2009), URL `http://www.R-project.org`, ISBN 3-900051-07-0 (date last viewed : 2/2/11).
[44] D. Bates and M. Maechler, *lme4: Linear mixed-effects models using S4 classes* (2010), URL `http://CRAN.R-project.org/package=lme4`, r package version 0.999375-35 (date last viewed : 2/2/11).
[45] R. H. Baayen, *languageR: Data sets and functions with "Analyzing Linguistic Data: A practical introduction to statistics".* (2010), URL `http://CRAN.R-project.org/package=languageR`, r package version 1.0 (date last viewed : 2/2/11).
[46] R. Baayen, *Analyzing linguistic data: A practical introduction to statistics using R* (Cambridge University Press, UK) (2008), chapt.4.
[47] 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).
[48] For a detailed listing of the relevant covariates please refer to the Supplementary Material.
[49] Figure 4 Tone1 : Sentence 119, Word 96; Tone2 : Sentence 19, Word 2; Tone3 : Sentence 85, Word 3; Tone4 : Sentence 265, Word 10; Tone5 : Sentence 445, Word 3.
[50] J. Ramsay and C. Dalzell, "Some tools for functional data analysis", Journal of the Royal Statistical Society. Series B (Methodological) **53**, 539–572 (1991).