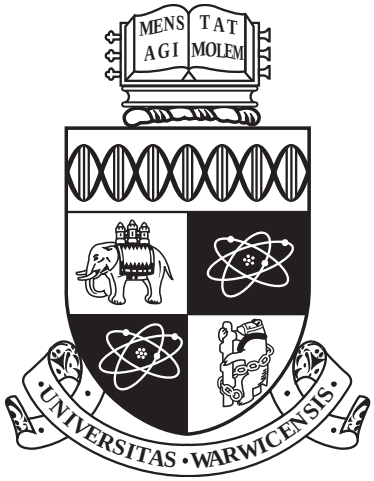

Functional Data Analysis in Phonetics

by

Pantelis-Zenon Hadjipantelis

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Department of Statistics

December 2013

THE UNIVERSITY OF
WARWICK

Contents

List of Tables

List of Figures

Acknowledgments

I was fortunate to come across people who helped me in this PhD, among those were: Alistair Tucker, Anas Rana, Ashish Kumar, Aude Exbrayat, Chris Knight, Chris Oates, Davide Pigoli, David Springate, Deborah Walker, Gui Pedro de Mendonça, John Coleman, John Moriarty, John Paravantis, Jenny Bowskill, Mónica de Lucena, Nick Jones, Nikolai Peremezhney, Phil Richardson, Quentin Caudron, Robin Ball, Sergio Morales, Yan Zhou and a few anonymous referees.

Special thanks goes to Jonathan P. Evans and my family; while far away, each gave me guidance and courage to keep going.

This work would certainly not be possible without the help of two persons who both went over and above what they signed up for initially: John A.D. Aston and Dimitra Zante, saying "thank you" is an understatement. They made me a better person.

Declarations

Parts of this thesis have been published or are in submission in:

- PZ Hadjipantelis, JAD Aston and JP Evans. Characterizing fundamental frequency in Mandarin: A functional principal component approach utilizing mixed effect models, *Journal of the Acoustical Society of America*, Volume 131, Issue 6, pp. 4651-4664 (2012)

- PZ Hadjipantelis, NS Jones, J Moriarty, D Springate and CG Knight. Function-Valued Traits in Evolution, *Journal of the Royal Society Interface*, Volume 10, No 82, 20121032, (2013)

- PZ Hadjipantelis, JAD Aston, H-G Müller and J Moriarty. Analysis of spike train data: A Multivariate Mixed Effects Model for Phase and Amplitude, (2013) (Accepted for publication in the *Electronic Journal of Statistics*)

- PZ Hadjipantelis, JAD Aston, H-G Müller and JP Evans. Unifying Amplitude and Phase Analysis: A Compositional Data Approach to Functional Multivariate Mixed-Effects modelling of Mandarin Chinese, (2013) (`http://arxiv.org/abs/1308.0868`)

The thesis is PZH's own work except where it contains work based on collaborative research, in which case the nature and extent of PZH's contribution is indicated. This thesis has not been submitted for a degree at another university.

Abstract

The study of speech sounds has established itself as a distinct area of research, namely Phonetics. This is because speech production is a complex phenomenon mediated by the interaction of multiple components of a linguistic and non-linguistic nature. To investigate such phenomena, this thesis employs a Functional Data Analysis framework where speech segments are viewed as functions. FDA treats functions as its fundamental unit of analysis; the thesis takes advantage of this, both in conceptual as well as practical terms, achieving theoretical coherence as well as statistical robustness in its insights. The main techniques employed in this work are: Functional principal components analysis, Functional mixed-effects regression models and phylogenetic Gaussian process regression for functional data. As it will be shown, these techniques allow for complementary analyses of linguistic data. The thesis presents a series of novel applications of functional data analysis in Phonetics. Firstly, it investigates the influence linguistic information carries on the speech intonation patterns. It provides these insights through an analysis combining FPCA with a series of mixed effect models, through which meaningful categorical prototypes are built. Secondly, the interplay of phase and amplitude variation in functional phonetic data is investigated. A multivariate mixed effects framework is developed for jointly analysing phase and amplitude information contained in phonetic data. Lastly, the phylogenetic associations between languages within a multi-language phonetic corpus are analysed. Utilizing a small subset of related Romance languages, a phylogenetic investigation of the words' spectrograms (functional objects defined over two continua simultaneously) is conducted to showcase a proof-of-concept experiment allowing the interconnection between FDA and Evolutionary Linguistics.

Abbreviations

AIC	Akaike Information Criterion
AUC	Area Under the Curve
BIC	Bayesian Information Criterion
CDF	Cumulative Distribution Function
COSPRO	Mandarin COntinuous Speech PROsody Corpora
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DTW	Dynamic Time-Warping
DWT	Discrete Wavelet Transform
EEG	Electroencephalography
FDA	Functional Data Analysis
FPCA	Functional Principal Component Analysis
FFT	Fast Fourier Transform
F_0	Fundamental Frequency
GMM	Gaussian Mixture Model
GPR	Gaussian Process Regression
HMM	Hidden Markov Model
ICA	Independent Component Analysis
INTSINT	INternational Transcription System for INTonation
IPA	International Phonetic Alphabet
JND	Just Noticeable Difference
K-L	Kullback-Leibler (divergence)
LMM	Linear Mixed-effects Model
MAP	Maximum A Posterior
MCMC	Markov Chain Monte Carlo
MDL	Minimum Description Length
MLE	Maximum Likelihood Estimate
MOMEL	MOdélisation de MELodie
MVLME	Multi-Variate Linear Mixed Effect (model)
qTA	quantitative Target Approximation
O-U	Ornstein-Uhlenbeck (process)
PACE	Principal Analysis by Conditional Estimation
PDF	Probability Density Function
PCA	Principal Component Analysis
PGPR	Phylogenetic Gaussian Process Regression
REML	Restricted Maximum Likelihood
RSS	Residual Sum of Squares
SCAD	Smoothly Clipped Absolute Deviation
SRV	Square Root Velocity

Chapter 1

Introduction

In a way the current work tries to challenge one of the most famous aphorisms in Natural Language Processing; Fred Jelinek's phrase: "Anytime a linguist leaves the group the recognition rate goes up"[159] [1]. While this phrase was famously associated with Speech Recognition (SR) and Text-to-Speech (TTS) research, it does echo the general feeling that theoretical and applied work are often incompatible. This is where the current work tries to make a contribution; it aims to offer a framework for phonetic analysis that is both linguistically and experimentally coherent based on the general paradigms presented in Quantitative Linguistics for example by Baayen [16] and Johnson [155]. It attempts to present a way to bridge the analysis of low-level phonetic information (eg. speaker phonemes [2]) with higher level linguistic information (eg. vowel types and positioning within a sentence). To achieve this we use techniques that can be broadly classified as being part of Functional Data Analysis (FDA) methods [254]. FDA methods will be examined in detail in the next chapters; for now it is safe to consider these methods as direct generalizations of usual multivariate techniques in the case where the fundamental unit of analysis is a function rather than an arbitrary collection of scalar points collected as a vector. As will be shown, the advantages of employing these techniques are two-fold: they are not only robust and statistically sound but also theoretically coherent in a conceptual manner allowing an "expert-knowledge-first" approach to our problems. We show that FDA techniques are directly applicable in a number of situations. Applications of this generalized FDA phonetic analysis framework are further shown to apply in EEG signal analysis [118] and biological phylogenetic inference [119].

Generally speaking, linguistics research can be broadly described as following two quite distinct branches. On the one side one finds "pure" linguists and on the other "applied" linguists. Pure linguists are scientists that ask predominantly theoretical questions about how a language came to be. (eg. within the subfields of *Language Development* studies and *Historical linguistics*.) What changes it and drives a language's evolution? (eg. studies of *Evolutionary linguistics* and *Sociolinguistics*.) How we perceive it and how different languages perceive each other? (Questions of *Semantics* and *Pragmatics*.) Why are two languages related or unrelated? (eg. *Comparative* and *Contact linguistics*.) In a way pure linguists ask the same questions a theoretical biologist would ask regarding a physical organism. On the other side one finds scientists that strive to reproduce and understand speech in its conjuncture; in its phenotype. In this field the basic questions stem from Speech Synthesis and Automatic Speech Recognition (ASR), ie. how can one associate a sound with a textual entry and vice versa. How can one reproduce a sound and how can we interpret it? This field of Natural Language Processing has seen almost seismic developments in the last decades. In a way it has changed the way we do science to an almost philosophical level. Just a century ago academic research was almost convinced by the idea of universal rules about everything. Research in general took an almost Kantian approach to Science where universal principles should always hold true; the work in Linguistic theory of Saussure and later of Chomsky echoing exactly that with the idea of "Universal Grammars" [58]; the wish for an *Unreasonable Effectiveness of Mathematics* not only *in Physical sciences* (as had Wigner's eponymous article declared [327]) but also in Linguistics had emerged. And then appeared Jeremy Bentham. Exactly like Benthan's utilitarian approach to Ethics and Politics

[1] The current work treats Computational Linguistics and Natural Language Processing as interchangeable terms; fine differences can be pointed out but they are not applicable in the context of this project.

[2] The smallest physical unit of a speech sound used in Phonetics is called a *phone*; a single phone or a sequence of phones that carry specific semantic context are called *phonemes* and serve as the simplest abstract class used in phonological analysis [159].

[316] that contrasted that of Kant, computational linguists in their task to recognize speech realized that ultimately they wanted "the greatest good for the greater number": identify the most pieces of speech in the easiest/simplest possible way. Data-driven approaches, originally presented as *Corpus* and *Quantitative Linguistics* gained such a significant hold that lead to their own dogma of the *Unreasonable Effectiveness of Data* [120] with certain Speech related research areas like ASR (Automatic Speech Recognition) being almost exclusively dominated by such vocabulary-based approach for classification and recognition tasks [277; 282]. Undoubtedly as time passes the distinctions get increasingly slimmer in most sub-fields as practitioners from both ends of the spectrum realize the advantages that hybrid approaches offer. The current work is building on the success of this latter data-driven approach to offer insights that would serve theoretical needs. It tries to exploit large amounts of data, not only to predict features without caring about the physical interpretation of the statistical methods, the extensive use (and success) of Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs) being a prime example of this [249; 265], but by using the findings of FDA in a theory-constructive way. It attempts to present a linguistic reasoning behind this statistical work.

Firstly in chapter 2 we present the theoretical outline of the linguistic, and in particular phonetic, findings that are most relevant to this work. We first introduce the basic aspects of the phonetic phenomena and investigate both their physiological as well as their computational characteristics. In particular we outline the basic notions behind neck physiology based models [178] and what connotations these carry to our subsequent view of the problem. Additionally we introduce the Fast Fourier transformation and how this transformation can be utilized to extract readings for the natural phenomena we aim to model [251]. We then continue into contextualizing these in terms of linguistic relevance and how they relate with the properties of known language types. We present briefly the current state-of-the-art models in terms of intonation analysis, making a critical assessment of their strengths, shortcomings and, most importantly, the modelling insights each of them offers [158; 305; 90] and we wish to carry forward in our analytic approach. While no single framework is "perfect", all of them are constructed by experts who through their understanding and research show what any candidate phonetic analysis framework should account for. We need to note here that the current project does not aim to present a novel intonation framework; it rather shows a series of statistical techniques that could lead to one. Finally we introduce the concept of linguistic phylogenies [13]; how we have progressed in classifying the relations between different languages and what were the necessary assumptions we had to make in order to achieve this classification. For that matter we also first touch on the questions surrounding the actual computational construction of language phylogenies. We close this chapter by introducing the two main datasets used to showcase the methods proposed: The Mandarin Chinese Continuous Speech Prosody Corpora (COSPRO) and the Oxford Romance Language Dataset. The datasets were made available to the author by Dr. Jonathan P. Evans and Prof. John Coleman respectively. Their help in this work proved invaluable, as without their generosity to share their data, the current work would simply have not been possible. COSPRO is a Mandarin Chinese family of Phonetic Corpora; we focus on a particular member of that family, COSPRO-1, as it encapsulates the prosodic [3] phenomena we want to investigate. Complementary to that, the Romance Language Dataset is a multi-language corpus of Romance language; Romance (or Latin) Languages form one of the major linguistic families in the European territory and, given the currently available in-depth knowledge of their association, present a good test-bed for our novel phylogenetic inference approach with an FDA framework.

Chapter 3 presents the theoretical backbone of the statistical techniques utilized throughout this thesis. Using the structure of the eponymous *Functional Data Analysis* book from Ramsay & Silver [254] as the road-map outlining the course of an FDA application framework, we begin with issues related to the *Smoothing & Interpolation* (Sect. 3.1) of our sample and then assess potential *Registration* caveats (Sect. 3.2). We then progress to aspects of *Dimension Reduction* (Sect. 3.3) and how this assists the practitioner's inferential tasks. We close by introducing the necessary background behind the *Phylogenetics* (Sect. 3.5) applications that will be presented finally in chapter 6. In particular after introducing the basic aspects behind *kernel smoothing*, the primary smoothing technique employed in this work, we address the problem of Registration, ie. phase variations. Here, after standard definitions and examples, we offer a critical review of synchronization frameworks used to regulate the issue of phase variations. While there is a relative plethora of candidate frameworks and we do not attempt to make

[3] Prosody and its significance in our modelling question will be presented in detail earlier in chapter 2.

an exhaustive listing, we present four well-documented frameworks [304; 340; 98; 176] that emerge as the obvious candidates for our problems' setting. We showcase basic differences on a real biological dataset kindly provided by Dr. P.A. Carter of Washington State University. As after each registration procedure we effectively generate a second set of functional data, whose members have a one-to-one correspondence with the data of the original dataset and their time-registered instances [4]; we stress the significance of the differences between the final solutions obtained. The differences observed being the byproducts of the different theoretical assumptions employed by each framework. Continuing we focus on the implications that one working with a high dimensional dataset faces. We address these issues by presenting the "work-horse" behind most dimension-reduction approaches, *Principal Component Analysis* [156], under a Functional Data Analysis setting [121]. Notably, as we will comment in chapter 7, other dimension reduction techniques such as Independent Component Analysis (ICA) [145] can also be utilized. Inside this reduced dimension field of applications we then showcase the use of *mixed effects models*. Mixed effects models (and in particular *Linear mixed effects (LME) models*) are the major inferential vehicle utilized in this project. They allow us to account for well-documented patterns of variations in our data by extending the assumptions behind the error structure employed by the standard linear model [300]. We utilize LME models because in contrast with other modelling approaches (eg. GMMs) they remain (usually) directly interpretable and allow the linguistic interpretation of their estimated parameters. For that reason we follow this section with sections on *Model Estimation* and *Model Selection* (Sect. 3.4.2 & 3.4.3 respectively), exploring both computational as well as conceptual aspects of these procedures. This chapter closes with a brief exposition of Phylogenetics. We introduce Phylogenetics within a Gaussian Process framework for functional data [291]. We offer an interpretation of these statistical procedures phylogenetically and linguistically and join these concepts with the ones introduced in the previous chapter in regards with Linguistics Phylogenetics. Concerning the purely phylogenetic aspects of our work, we finish by outlining the basic concepts behind the estimation of a phylogeny which we will base our work on, in chapter 6.

The first chapter presenting the research work behind this PhD is found in chapter 4. Given a comprehensive phonetic speech corpus like COSPRO-1 we employ functional principal component mixed effects regression to built a model for the fundamental frequency (F_0) dynamics in it. COSPRO-1 is utilized for this task as Mandarin Chinese is a language rich in pitch-related phenomena that carry lexical meaning, ie. different intonation patterns relate to different words. From a mathematical standpoint we model the F_0 curve samples as a set of realizations of a stochastic Gaussian process. The original five speaker corpus is preprocessed using a locally weighted least squares smoother to produce F_0 curves; the smoothing kernel's bandwidth was chosen by using leave-one-out cross-validation. During this step the F_0 curves analysed are also interpolated on a common time-grid that was considered to represent "rhyme time", the rhymes in this work being specially annotated instances of the original Mandarin vowels. Contrary to most approaches found in literature, we do not formulate an explicit model on the shape of tonal components. Nevertheless we are successful in recovering functional principal components that appear to have strong similarities with those well-documented tonal shapes in Mandarin phonetics, thus lending linguistic coherence to our dimension reduction approach. Interestingly, aside from the first three FPC's that have an almost direct analogy with the tonal shapes presented by Yuen Ren Chao, we are in position to recognize the importance of a fourth sinusoid tonal FPC that does not appear to correspond to a known tone shape; based on our findings though we are able to theorize about its use as a transitional effect between adjacent tones. To analyse our sample we utilize the data projected in the lower dimensional space where the FPC's serve as the axis system. We then proceed to define a series of penalized linear mixed effect models, through which meaningful categorical prototypes are built. The reason for using LME models is not ad-hoc or simply for statistical convenience. It is widely accepted that speaker and semantic content effects impose significant influence in the realization of a recorded utterance. Therefore grouping our data according to this information is not only beneficial but also reasonable if one wishes to draw conclusions for the out-of-sample patterns for F_0 variations. This work serves as a first stepping stone in the field of Phonetics research within an FDA framework. Strictly speaking it does not "break new ground" in terms of statistical methodology but rather establishes the validity of an FDA modelling framework for phonetic data. It does this though extremely successfully

[4] Through this work the terms *time-registration* and *time-warping* will be used almost interchangeably; if any distinctions are drawn they will be outlined in the immediate text.

allowing deep insights regarding the F_0 dynamics to emerge while coming from an almost totally phonetic-agnostic set of tools. The overall coherence of this phonetic approach has lead to a journal publication [117].

Augmenting the inferential procedure of the previous chapter, Chapter 5 employs not only amplitude, but also phase information during the inference. Once again we work on the COSPRO-1 phonetic dataset. In this project though we are not only interested in how amplitude changes affect each other but also how phase variational patterns affect each other and how these variations propagate in changes over the amplitudinal patterns. In contrast with the previous chapter's work, this time we employ a multivariate linear mixed effects model (instead of a series of univariate ones) to gain insights on the dynamics of F_0. As previously, a kernel smoother is used to preprocess the sample, the kernel bandwidth being determined by leave-one-out cross-validation. Following that we use the pairwise warping framework presented by Yao & Müller to time-register our data on a common normalized "syllable time"; we view this warped curve dataset as our amplitude variation functions. Through this time- registration step though we are also presented with a second set of functional data, the warping functions associated with each original F_0 instance; these are assumed to be our phase variation functions. We additionally showcase that these phase-variation functions can be seen as instances of compositional data; we explore what connotations those might have in our analysis and we propose certain relevant transformations. Having two functional datasets that need to be concurrently analysed is the reason why we use a multivariate linear mixed model. Therefore, after doing FPCA in each set of functional data separately, while the projection within a set can be assumed orthogonal to each other and allow the employment of univariate models (as in chapter 4), the FPC scores of the amplitude variation are not guaranteed to be orthogonal with that phase variation. Thus given our two domains of variation, amplitude and frequency, we define a multivariate model that incorporates a complex pattern of partial dependencies; the FPC scores being *orthogonal* with other variables in the same variation domain but being *non-orthogonal* with scores from the other variation domain. As seen in section 5.3.8 the computational considerations are not trivial. Our final results allow us to investigate the correlations between the two different areas of variation as well as make reasonable estimates about the underlying F_0 curves. As a whole this work aims to present a first attempt to unify the amplitude and phase analysis of a phonetic dataset within a FDA framework. This joint model is easily generalized to higher dimension functional data. Also, despite us relying on pairwise time-synchronization, our approach can be directly applicable to any other choice of warping framework. At the time of writing this thesis, the work presented here regarding the phonetic analysis of COSPRO-1 has been submitted for publication. An EEG analysis application paper using the same approach for a unified amplitude and phase model utilizing LME models has already been accepted for publication [118].

Chapter 6 provides a first and rather ambitious application of the phylogenetic analysis of a linguistic dataset within a Functional Data Analysis framework. In contrast with the previous work where our data were assumed to lay on a single continuum (time), here we work with two-dimensional instances of functional data: spectrograms. Thus we assume that our data lay on a two-dimensional continuum indexed by time and frequency. The ultimate goal of this project is to show how one would construct a protolanguage (the language found at the root of a linguistic phylogeny) and to offer insights on the underlying phylogenetic relations between the languages of a given phylogeny. In a very self-contained project we start with just voice recordings. We then build from the ground up the whole inferential framework for the phylogenetic associations between the languages of our sample. After the preprocessing of our data, following the methodology presented previously (*smoothing & interpolation* followed by *registration*), we theorize on the underlying evolutionary dynamics and draw analogies with current biologically related concepts. Then we reduce the dimensions of our dataset, reformulate the problem of linguistics Phylogenetics in that confined space and proceed in making model estimation. Based on that optimal model we then offer estimates about a Latin protolanguage we aimed to reconstruct, as well as theorize on the linguistic associations that emerged by this project. In particular, because we recognize the large absolute size in computational terms of our unit of analysis (a spectrogram), we do not employ a non-parametric smoothing technique as previously. We use a two-dimensional smoothing framework based on the two-dimensional Discrete Cosine Transform. Following that we reformulate the pairwise warping framework outlined in section 3.3 in the case of spectrograms instead of simple voice signals. Interestingly, exactly because we can take advantage of our knowledge on spectrograms, we warp across

a single dimension instead of two; this insight greatly simplifying our computational load. Following that, we utilize dimension reduction across a two-dimensional object to produce a lower dimensional representation of our dataset. We then turn to the first issue underpinning any phylogenetic study: *tree estimation*. For that we employ an Ornstein-Uhlenbeck (O-U) process-based approach to interpret the evolutionary dynamics of our sample. This Gauss-Markov process allows us to conceptually account for all major evolutionary factors known and thus consolidate the logical connection between our work and evolutionary insights. Using O-U we find our "most probable" tree given our data as well as the most likely protolanguage estimate. Overall, it is important to note that this work showcases the ability of FDA to provide robust tools to answer questions that just a few decades ago, the questions themselves were not formulated in their respective field. While this work is still not finalized, its overall coherence has allowed a first simulation study to be published [119].

Bringing the findings of this PhD thesis to a close, the final chapter (chapter 7), offers of short summary of the work and the major conclusions drawn by it. It then outlines the issues that presented the obvious limitations surrounding this work. It closes by offering a brief outline of future research directions that could follow from the current work.

In conclusion, the work in this thesis on Functional Data Analysis tries to outline a framework that presents not only a convenient way to model syllables, subwords or words (or any other phonetic modelling unit one chooses to work with), but also a theoretically meaningful representation of those measurements. As mentioned earlier, the current work does not aim to present itself as a study of Acoustical Phonetics but rather as one of Applied Statistics. With this in mind, a short summary of Acoustics literature is offered in the following section to familiarize the reader with basic concepts and results that will contextualize the material to follow.

Chapter 2

Linguistics, Acoustics & Phonetics

Linguistics is formally defined as the study of human language; the ultimate question it tries to tackle though is how communication takes place among humans [205]. Indisputably, one of the first forms of communication between humans had to be audible sounds (or growls depending on one's perspective). As a natural consequence the studies of sounds (*Acoustics*), and of voices in particular (*Phonetics*), have established themselves as major parts of linguistic studies.

The earliest speech "studies" have been documented at approximately 500 BC and regarded Sanskrit grammar phenomena. Evidently though what we broadly describe as Linguistics is more contemporary. Picking a single point in history where a certain "methodological cut" in a scientific field was made is often subjective and hard to declare; nevertheless the author believes that this happened with the 1916 posthumous publication of Ferdinand Saussure's *Course in general linguistics* [63] where the distinction between studying particular parts of speech and a language spoken by the members of a society was formalized [205]. There, among other concepts, Saussure established the treatment of language as a system, or a structure [1] that has to be interpreted as part of interacting components. This theoretical stance has served as the basis for many theoretical and practical breakthroughs in *Computational Linguistics* [130] despite finding itself nowadays getting increasingly superseded by recent advancements both in inter- (eg. *Demolinguistics*) as well as intra-population linguistic studies (eg. *Neurolinguistics*).

Within Linguistics the current thesis focuses on how statistical analysis can offer new insights into Acoustical Phonetics. On their part Phonetics focus on the production (articulatory), transfer (acoustic) and perceptual (auditive) aspects of phonation processes (speech) [178]. More specifically though, Acoustical Phonetics are concerned with the physical properties of speech sounds, as well as the linguistically relevant acoustic properties of such sounds [270]. In the study of Acoustical Phonetics one does not try to formulate more abstract models based on syntactic, morphological or grammatical issues of speech as this falls under the field of Phonology. Specifically, for phonetic analysis, a number of sound properties are of phonetic interest, namely: pulse, intensity, pitch, spectrum or duration of the examined speech sound segment; a speech sound segment which itself can be a consonant, vowel or even the successive voiceless gap between words.

Arguably the human speech production mechanism (or more precisely the *human vocal apparatus*) is a system of multiple independent components; it involves complex motor tasks by the speaker's vocal organs to form articulatory motions under the synchronous emission of air from the lungs [139]. The final product, speech sounds are periodic complex waves characterized by their frequency and amplitude [155]. Broadly speaking, frequency relates to how fast the sound wave propagated oscillates, the amplitude quantifying the intensity of that oscillation. What we perceive though as sound is not usually a single frequency but a mixture of components called harmonics or *formants* [2]. The zeroth harmonic; the fundamental frequency (F_0) is of major interest. The fundamental frequency is commonly understood (somewhat inaccurately) as pitch; it is of interest physiologically because it relates very closely to the actual frequency under which a person's *vocal folds* vibrate when speaking. It dictates a number of secondary speech attributes; thus understanding the mechanics of F_0 empowers many aspects of speech-related analysis [69]. Interestingly not a single ubiquitous definition of F_0 exists. Assuming T_0 to be *the elapsed time between two successive laryngeal pulses* where *measurement starts at a well-specified point*

[1] Saussure's general approach to Linguistics gave raise to the *Structural Linguistics* paradigm.

[2] Babies actually produce pure periodic signals for a short period in their lives [155].

within the glottal cycle, preferably at the instant of glottal closure [24], F_0 is simply $F_0 = T_0^{-1}$. Alternative speech production based definitions place the measurements start at other points of the excitation cycle; if an incremental definition is used then the length of the excitation cycle itself is assumed to be T_0 [132]. More mathematical definitions define F_0 directly as *the fundamental frequency of an (approximately) harmonic pattern in the (short term) spectral representation of the signal* [33]. Finally purely perceptual definitions where F_0 is the *frequency of the sinusoid that evokes the same perceived pitch as the complex sound that represents the input speech signal* are also applicable in a general acoustic sense [306]. All definitions are used in the literature, almost interchangeably and sometimes without being formally stated by the authors using them. The current work ultimately relies on F_0's view as the lead harmonic measured in certain (natural) units (usually Hz or Mel) and being closely related with pitch. Pitch on the other hand is assumed to be a perceptual rather than a physical phenomenon and it has to do with acoustics as much as audition. A somewhat unintuitive terminology caveat relating to the phonation process is the distinction between voiced and voiceless sounds. In respect with the neck physiology of the speaker, when a sound is produced while their vocal folds vibrate, that sound is considered voiced; if a sound is produced with the vocal folds being open, it is considered voiceless [178]. An easy way for an English speaker to listen/feel this distinction is comparing "v" and "f"; a voiced and a voiceless consonant respectively. First pronouncing a long constant "v"; `[vvvvvvvvvvv]` and then comparing this with a long constant "f"; `[fffffffffff]` (a voiceless consonant) one can immediately feel the difference between the two; putting your fingers on your larynx as you alternate between the two consonant sounds makes the physiological difference even more obvious.

Evidently F_0 is not easy to directly measure. Such measurements have been taken, but they are usually intrusive, complicated and rather expensive [178]. On the contrary there are multiple ways of extracting F_0 from a given acoustic recording. The main two methodologies are based on autocorrelation [248] and cepstrum analysis [33].

2.1 Basic Computational Aspects of F_0 determination

Human speech F_0 can range from approximately 50 Hz (a very low-pitched male) to almost 800 Hz (a very high-pitch woman or a small child). This means that based on the context of Nyquist frequency [199], the highest frequency component detectable can be at most half of the sampling frequency used; sub-Nyquist sampling methodologies have been presented in the past years [187] but we will not examine them. As a logical consequence to ensure a fundamental frequency of 400 Hz is detectable, at least an 800 Hz sampling rate must be used. This is not usually an issue as even low-quality speech recordings are done at a minimum 8 KHz rate, but questions do arise on how large a pitch track sample should be in order to evaluate a lower frequency reading; for example the minimum time required for a 50 Hz wave to have a full oscillation is ($(50Hz)^{-1} = 0.02s$) 20ms [3]. This brings us to how autocorrelation pitch tracking methodology works: Given a small time-frame (typically 10ms or 20ms long) of pitch track sample and then displacing it slowly over the rest of the speech tract readings, using a sliding window approach, we record the correlation produced for each successive displacement. The inverse of the displacement (or lag) that produces the largest correlation is then reported as fundamental frequency of that pitch segment [32].

The short-term autocorrelation function α_f for a given signal $x(t)$ for a lag d is formally given as [285]:

$$\alpha_f(d) = \int_{-\infty}^{\infty} x(t)x(t+d)dt \tag{2.1}$$

or discretely in the case of a short-term signal as :

$$\alpha_f(d,q) = \sum_{t=q}^{K-d+q} x(t)x(t+d) \tag{2.2}$$

[3] The phenomenon of using an inadequately low sampling frequency for a given signal can lead to *aliasing*; higher sampling rates and/or low-pass filtering prior to sampling can be used as remedies.

where K is the size of the time-window examined and q the starting point. This is though also the problem with auto-correlation pitch tracking methodology: unsuitable window size can lead to severely miscalculated pitch estimates. The first problem is *pitch-doubling*; during pitch-doubling the shortest time displacement in the sliding window is as short as the half of the F_0 period. The autocorrelation method will then most probably find that the two halves are separate oscillations, consecutively finding that the signal period is half of what it really is and finally report an F_0 that is twice its real value. The second problem is *pitch-halving*; during pitch-halving we use too large of a window and in that case two or more periods can be fitted within the same time-frame. Then the expected period of pitch is assumed to be longer than what it actually is and therefore the algorithm underestimates the reported F_0. Interestingly pitch-halving can occur even if one chooses the F_0 window length correctly; if two displaced pitch periods are more similar than two adjacent ones, an auto-correlation algorithm can still find a larger period exactly because the correlation between the two windows will be stronger in that case [155]. It is worth noting that while auto-correlation methods are the norm [248], cross-correlation methods and *average magnitude difference function* [24] [4] have also been developed but work again in a time-step principal. No established methodology for choosing the window-length exists. Most popular implementations (eg. Praat [32] and Wavesurfer [295]) aside from using standard predetermined window sizes, are based on "how low" F_0 is expected to be; effectively estimating a *maximin* window length of a frame.

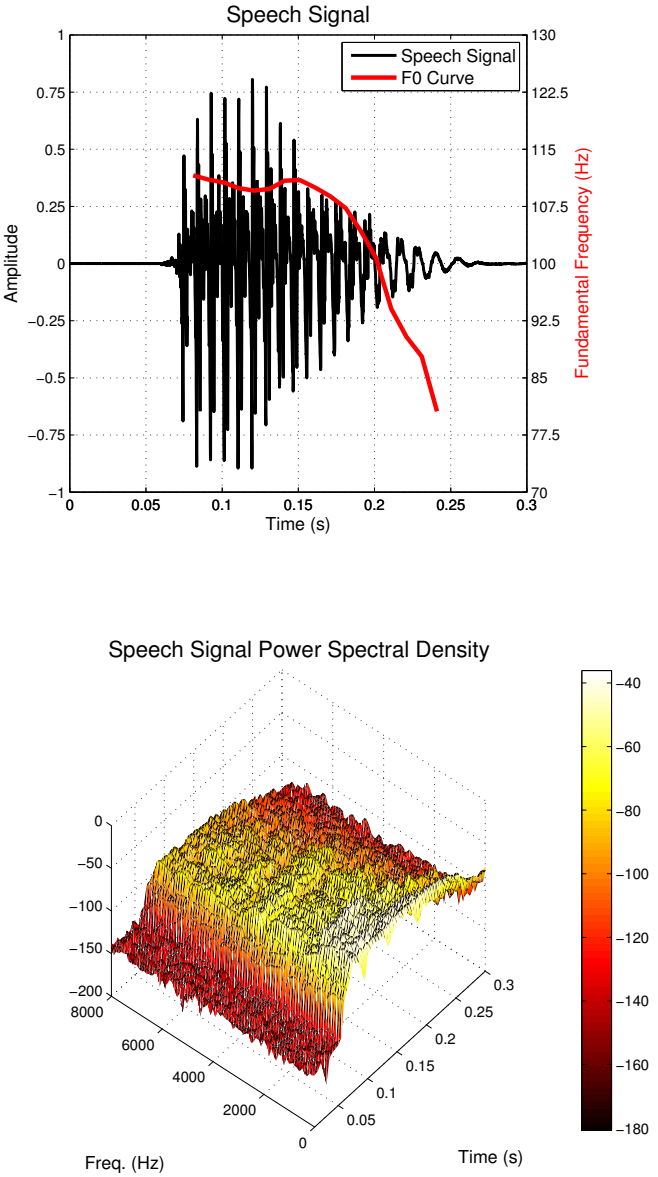

Figure 2.1: The normalized amplitude signal (upper panel) and the related spectrogram (lower panel) of a male French speaker saying the word "un" (œ). The superimposed F_0 never goes above 110Hz (upper panel, red curve) while after the initial excitation (0.075-0.150s) the power density quickly diminishes (lower panel).

Complementary to autocorrelation based methods are the *double-transform* or cepstrum based methods; their back-bone is the Discrete Fourier Transform (DFT). Cepstrum (a word-play of the word spectrum itself) *is the inverse Fourier transform of the natural logarithm of the spectrum. Because it is the inverse transform of a function of frequency, the cepstrum is a function of a time-like variable* [285].

Taking a step back and assuming a time-domain index t, $t \in [0, T]$, T being the total duration of a time sample $x(t)$, the complex spectrum frequency $X(f)$ for any frequency f is the forward Fourier transformation of $x(t)$ [251]:

$$X(\omega) = \int_0^T x(t) e^{-\omega i t} dt, \qquad \omega = 2\pi f \tag{2.3}$$

where it can be immediately seen that the Fourier transformation is a continuous function of frequency

[4] $AMDF(d, q) = \frac{1}{K} \sum_{t=q}^{K+q} |s(t) - s(t+d)|$

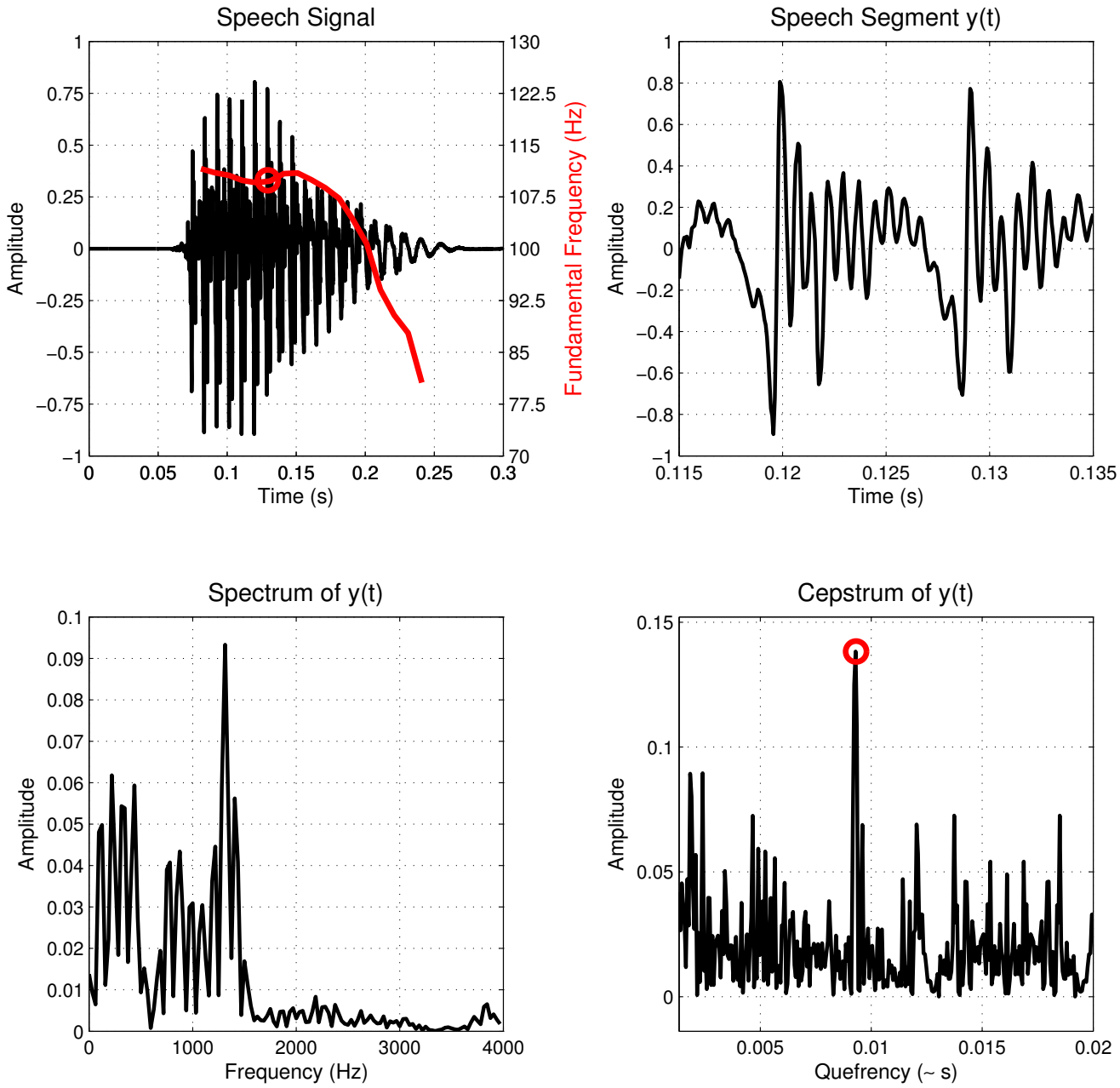

Figure 2.2: Illustration of Cepstral F_0 determination. A male French speaker saying the word "un" (œ)(upper left, black line, left axis) and the corresponding F_0 curve (upper left, red line, right axis, determined using ACF). To estimate the F_0 at as single point (red circle) we use the segment of speech around the estimation point (upper right), we then compute its Power $|dft(y)|$ (lower left) and then the corresponding Cepstrum $idft(log_e(|dft(y)|))$ (lower right). The peak amplitude of the Cepstrum occurs at time $F_0^{-1} = 0.0092$ s $\approx F_0 = 108.7Hz$. A rectangular window was used.

f, the inverse of it expressing the signal as a function of time as:

$$x(t) = \frac{1}{2\pi} \int X(\omega) e^{\omega i t} d\omega. \tag{2.4}$$

Continuing, the instantaneous power $P(t)$ and energy E of a signal $x(t)$ are respectively defined as:

$$P(t) = x^2(t) \quad \text{and} \tag{2.5}$$

$$E = \int P(t)dt \quad = x^2(t)dt \quad \text{where by using Eq. 2.4} \tag{2.6}$$

$$E = \frac{1}{2\pi}\int X(\omega)X(-\omega)d\omega \quad = \frac{1}{2\pi}\int |X(\omega)|^2 d\omega \quad = \int E(\omega)d\omega \tag{2.7}$$

representing that qualitative signal energy is the accumulation of energy spectral density $E(\omega)$. Moving to a discrete signal $x[n]$ now, the discrete-time Fourier transformation of it is defined as:

$$X(e^{i\omega}) = \sum_{t=-\infty}^{\infty} x[t]e^{-i\omega t}dt, \tag{2.8}$$

the inverse of it being defined in respect to its complex logarithm as [24]:

$$\hat{x}[t] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{X}(e^{i\omega})e^{i\omega t}dt \tag{2.9}$$

where the complex logarithm $\hat{X}$ is:

$$\hat{X}(e^{i\omega}) = \log[X(e^{i\omega})]. \tag{2.10}$$

Based on this, one is able to then define the cepstrum of a discrete signal $x[n]$ as:

$$c[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log(|X(e^{i\omega})|)e^{i\omega n}d\omega, \tag{2.11}$$

$$\text{or in a purely discretised form as:} \quad = \sum_{n=0}^{N-1} \log(|\sum_{n=0}^{N-1} x[n]e^{-i\frac{2\pi}{N}kn}|)e^{i\frac{2\pi}{N}kn}, \tag{2.12}$$

and recognize that during cepstral analysis the power spectrum of the original signal is treated as a signal itself. As a result instances of periodicity in that signal will be highlighted in the original signal's cepstrum. Essentially the cepstrum's peaks will coincide with the spacing between the signal's harmonics (Fig. 2.2). As expected the cepstrum is a function of time or *quefrency*, where the quefrency approximates the period of the signal examined (ie. larger quefrencies relate to slower varying components).

One might question the complacency of the two F_0 tracking techniques; the answer though is rather straight-forward and is known as the Wiener-Khinchin theorem [54]. Taking Eq. 2.1 and substituting $x(t)$ using the inverse Fourier transformation, Eq. 2.1 becomes:

$$\alpha_f(d) = \int_{-\infty}^{\infty} x(t)x(t+d)dt \quad = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{\omega it}|X(\omega)|^2 d\omega \quad = \int_{-\infty}^{\infty} e^{\omega it}E(\omega)d\omega \tag{2.13}$$

and it therefore becomes obvious that the Fourier transformation of the energy spectral density $E(\omega)$ (Eq. 2.7) is the autocorrelation function associated with the signal $x(t)$. Other methods for pitch determination based on *Linear Prediction* [250; 289], *Least Squares Estimation* [87; 217] and direct *Harmonic Analysis* [1; 220] have also offered certain practical advantages on application-specific projects but they have failed to established themselves as generic frameworks in comparison with the two methods previously described.

A final point concerning pitch detection, or any general statistical analysis, is that it is based on the assumption of (at least *weak*) stationarity of the dataset examined. In that sense the first and second moments of the sample assumed to be the speech segment examined are "stable" across time. Nevertheless, a speech signal is definitely non-stationary along a speaker's utterance. Even in the short-time analysis framework $(10 - 100ms)$ one might encounter pauses, co-articulation, non-random noise or quantization effects; phenomena that would clearly violate usual stationary assumptions. Excluding more specialized techniques such as non-linear filtering [319], the usual counteraction is to use *windowing* (or *lifting* [33]); an important preliminary step where we *frame* a segment $y(t)$ of our signal $x(t)$ and assume $y(t)$ to be locally stationary. They are many different window types; three of the most common windows of size L are :

- the rectangular $(w(n) = 1)$,

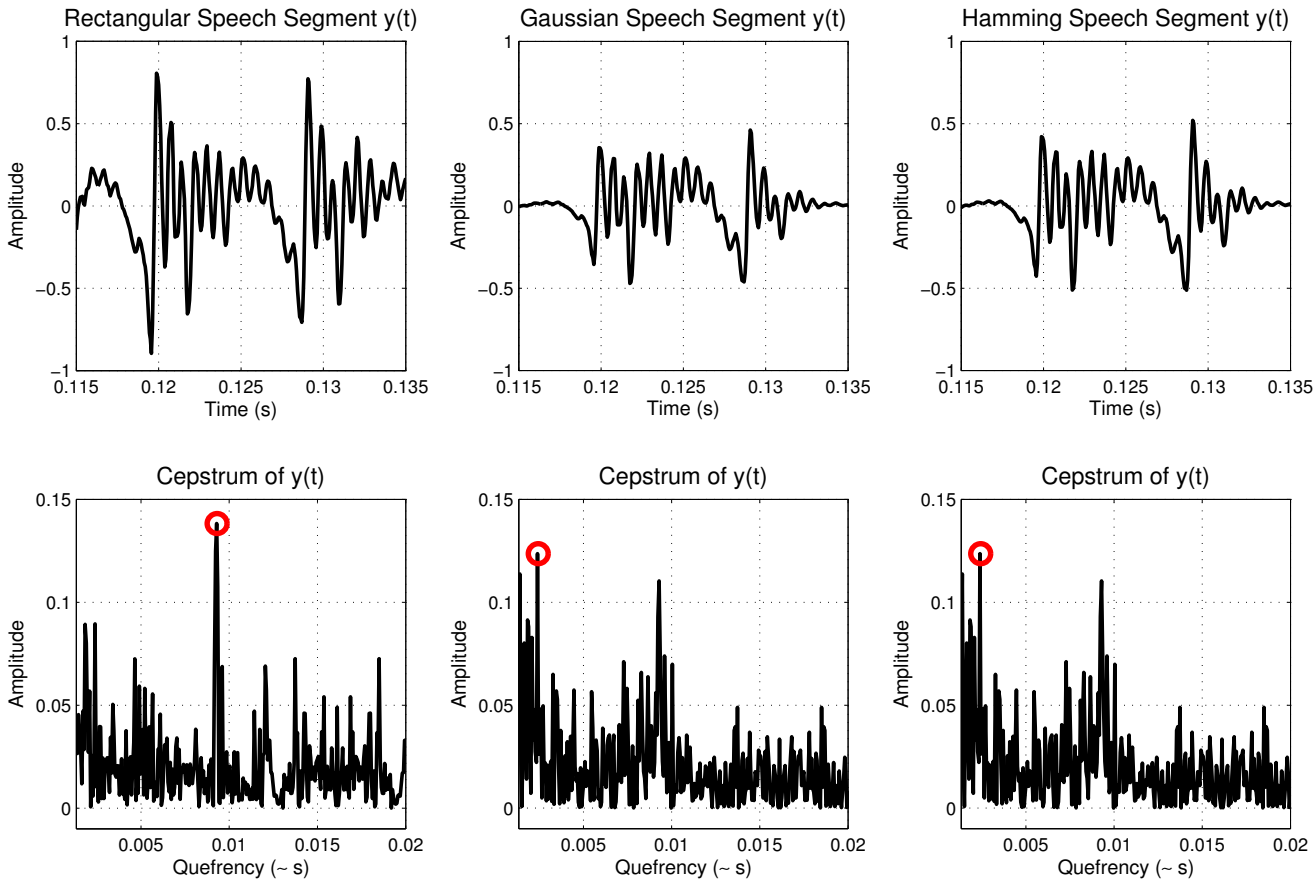

Figure 2.3: Illustration of Cepstral F_0 determination employing different windows. The peak amplitude of the Cepstrum occurs at time $F_0^{-1} = 0.0092$ s $\approx F_0 = 108.7Hz$ in the case of a rectangular window (Left hand-side plots). The peak amplitude of the Cepstrum occurs at time $F_0^{-1} = 0.0024$ s $\approx F_0 = 416.7Hz$ both in the case of a Gaussian and of a Hamming window (Center and Right hand-side plots).

- the Hamming $(w(n) = .54 - .46cos(2\pi\frac{n}{N}), \quad L = N + 1)$,
- the Gaussian window $(w(n) = \exp(-\frac{1}{2}(\frac{n}{\sigma})^2))$,

the latter one being an infinite-duration window. While somewhat trivialized in the context of signal analysis windowing properties such as frame size, type and shift can have very profound effects in the performance of a pitch determination algorithm (Fig. 2.3) and subsequent feature extraction. The relation of windowing and kernel smoothing (Sect. 3.1), as it will be shown later, is all but coincidental and windowing in the sense presented above is a simple reformulation a general "weighting" scheme in mainstream Statistics.

Complementary to this F_0 estimation via the DFT is the task of spectrogram estimation. Without going to unnecessary details, a spectrogram is simply the concatenation of successive Fourier transformations along their frequencies. As in the case of F_0, windowing in terms of segment length and window type significantly affects the resulting two-dimensional function, one of the axes being time (t) and the other frequency (f) (eg. Fig. 2.1 lower panel). We draw attention to a very interesting property of spectrograms: while spectrograms are subject to the same time distortion (See section 3.2) as every other phonetic unit of analysis, time distortion is only meaningful across their time-axis. While *smearing* (or *leakage*) can occur between successive frequency bands [24], this is not time-dependent and is usually associated with recording conditions (eg. room reverberation) and/or windowing. In practice, given we employ conservative choices of windowing, we assume that leakage is minimal. Therefore the frequency axis f is assumed to evolve in "equi-spaced" order with no systematic distortions present; we will reiterate these insights in section 6.2.1.

2.2 Tonal Languages

While pitch in the majority of Indo-European languages is mostly used to convey non-linguistic information such as the emotional state of the speaker, there are many world languages, especially in south-east Asia and sub-Saharan Africa that are tonal [200]. By tonal we mean that the pitch pattern of a pronounced vowel or syllable changes the lexical meaning of a word in a predetermined way [301]. Among tonal languages Mandarin Chinese is by far the most widely-spoken; it is spoken as a first language by approximately 900 million people [53], with considerably more being able to understand it as a second language. It is therefore of interest to try and provide a pitch typology of Mandarin Chinese in a rigorous statistical way incorporating the dynamic nature of the pitch contours into that typology [111; 245]. This interest is not only philological for Indo-European languages' speakers. Discounting the obvious advantage that an accurate typology will offer in automatic speech recognition and speech/prosody production algorithms [69], with the increasing interest in tonal Asian languages as second languages, potential learners could greatly benefit from an empirically derived intonation framework; second language acquisition being an increasingly active research field [140]. Phenomena such as synchronic or diachronic tonal stability, while prominent in tonal languages, are quite difficult to encapsulate in strict formal rules [100] and they are known to manifest in significant learning difficulties.

In particular in Mandarin Chinese there are five tones with five distinct shapes and pragmatic meaning (Fig. 2.4). For example, using the "ma" phoneme with different tones implies:

- Tone1 (mā 媽 "mother") a steady high-level sound,
- Tone2 (má 麻 "hemp") a mid-level ascending sound,
- Tone3 (mǎ 馬 "horse") a low-level descending and then ascending sound,
- Tone4 (mà 罵 "scold") a high-level descending sound and
- Tone5 (ma 嗎 question particle) an unstressed sound.

This means that the rather artificial statement 媽媽罵麻馬嗎 / "Does mother scold the numb horse??" is actually transliterated in Pinyin [5] as: "māma mà má mǎ ma?" where without the pitch patterns shown in the diacritic markings it would be totally incomprehensible [330].

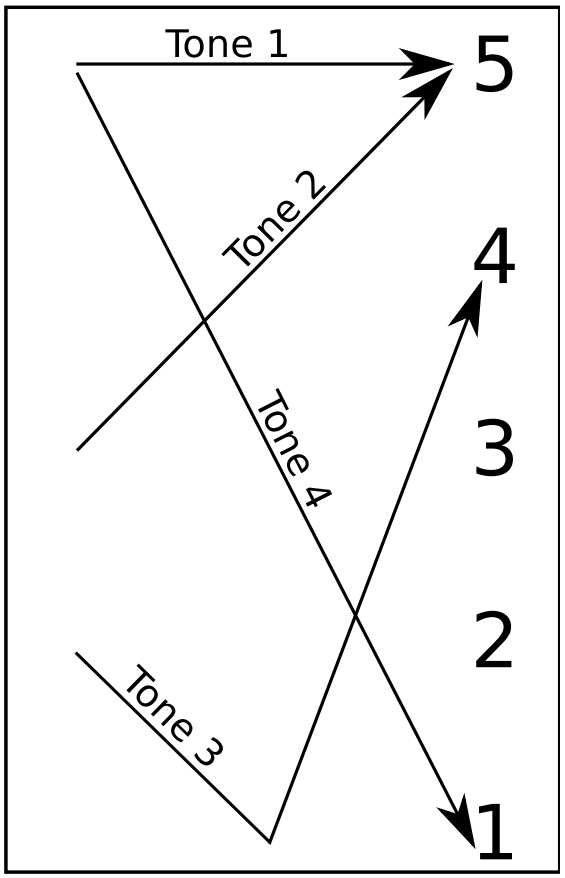

Figure 2.4: Reference tone shapes for Tones 1-4 as presented in the work of Yuen Ren Chao; Tone 5 is not represented as it lacks a general estimate, always being significantly affected by non-standardized down-drift effects. Vertical axis represents impressionistic pitch height.

As Fujisaki recognizes though, while in tonal languages (eg. Mandarin) pitch is modulated by lexical information, para- and non-linguistic effects have significant impact to the speaker utterances [90]. In other words, to assume that the differences observed in intonation pattern is only due to the words' lexical meaning is an oversimplification. Non-lexical effects come into play. In particular by para-linguistic effects we mean contextual and semantic information that affects a user's

[5] Pinyin is the official form of the Latin alphabet transliteration of Mandarin Chinese used by the People's Republic of China [328].

utterance. For example the same speaker might employ different intonation patterns for a formal announcement to that of a private conversation with a friend. Moreover non-linguistic effects related directly to the speaker's physiological characteristics also have prominent influence in the final syllable modulation [333; 117]. As F_0 is associated with the function of the vocal cords, and like most physiological properties of an organism this functionality is influenced by the age, health and general physical characteristics of a speaker, speaker related variation should not be neglected.

2.3 Pitch & Prosody Models

The rhythmic and intonational patterns of a language [159] are known as *prosody*. Prosody and F_0 modelling are intertwined as F_0 is a key component of prosody [6]and this connection is even stronger in tonal languages like Mandarin Chinese. Accurate modelling of the voicing structures enables the accurate modelling of voiced speech segments thus assisting all aspects of speech related studies: synthesis, recognition, and coding [69].

Somewhat generally, the "reference framework" for the analysis and synthesis of intonation patterns is ToBi [158]. As defined by its creators: *"ToBi is a framework for developing community-wide conventions for transcribing the intonation and prosodic structure of spoken utterances in a language variety"*. ToBi defines five pitch accents and four boundary tones based on which it categorizes each respective utterance. ToBi is effectively a complete prosodic system; it has two important caveats. First ToBi is *not* a universal system. There are language-specific ToBi systems that are non-communicative to one another; this is a major shortcoming in its generality. In that sense its rigidity has rendered it too restrictive to model even English varieties (ToBi was specifically developed for the English language originally) leading to the development of IVie [103]. Secondly though it also defines a series of different break types, acting as phrasing boundaries. Break counts are very significant as physiologically a break has a resetting effect on the vocal folds' vibrations; a qualitative description of break counts is provided in Table 2.1. This recognition of the importance of breaks highlights an important physiological characteristic of F_0; while a continuous trajectory is meaningful for temporal modelling, an F_0 trajectory is not a continuously varying parameter along an utterance but rather a series of correlated discrete events that are realized as continuous curves.

Break Type	Meaning
Break 1	Normal syllable boundary. In languages like written Chinese where there is "no alphabet" but the written system corresponds directly to morphemes, this corresponds to a single character. (As syllable segments will often act as our experimental data units, B1 is equivalent to the mean value of the statistical estimates and thus not examined separately as a "dependent variable").
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 2.1: ToBi Break Annotation

Complementary to ToBi is the work of Taylor with the TILT model [305]. *"TILT is a phonetic model of intonation that represents intonation as a sequence of continuously parametrized events."* The interesting thing about TILT is that it effectively places the intonational event not only as its modelling target, but also as its fundamental unit. As such it does not use predetermined labels as ToBi. Instead, each event is characterized by its amplitude, duration and tilt. Tilt (not TILT) is effectively a continuous description of the F_0 curve that is a function of the duration D and the amplitude A of the intonation

[6]Other components being for example *timing* and *loudness*.

pattern examined. In particular:

$$\text{Tilt} = \frac{|A_{rise}| - |A_{fall}|}{2(|A_{rise}| + |A_{fall}|)} + \frac{D_{rise} - D_{fall}}{2(D_{rise} + D_{fall})}. \tag{2.14}$$

The TILT model was in a way influential because it really provided an empirical and continuous representation of F_0. Nevertheless in the end TILT uses three, undoubtedly important numbers to characterize a single curve. This is not "wrong" (the popularity of TILT hinting that these three numbers are highly effective), but ultimately fails to provide a framework that can be directly expanded to account of increasing sample complexity. Additionally it does not account for speaker related information affecting an utterance nor for explicit interaction between successive F_0 curves.

This is one of the main intuitions behind the third and final "reference model" for intonation patterns: the Fujisaki model [90]. The Fujisaki model was introduced by Fujisaki and Ohno in 1997 and was extended mostly by the cooperation of Fujisaki with Mixdorff [7]. Similarly to the TILT model, the Fujisaki model is a quantitative model that does not use explicit labels. The basic modelling assumption behind the Fujisaki model is that the F_0 contour along a sentence is the superposition of both a slowly- and a rapidly-varying component [90]. The slowly-varying component commands the overall curvature of the F_0 contour along the duration of the sentence, the rapidly-varying relates to the lexical tone. This major idea came from the way the F_0 production mechanism is treated: the laryngeal structure being approximated by Fujisaki's earlier work as effectively the step response function of a second-order linear system [133]. Another important theoretical break-through of the Fujisaki model was that it explicitly incorporated speaker related information or better yet uncertainty; for example it assumes that the lower F_0 attainable is a speaker related rather than universal characteristic and that it should be treated as an unobserved random variable. The major shortcoming of the Fujisaki model actually comes from within its design: the idea of a slow-varying down-drift deterministic component is rather restrictive, despite being a reasonable norm. Especially in its original format this assumption fails to account for intonation patterns in Western languages [305]. Also in its original form the Fujisaki model advocated the use of a rigid gradient for each of the rapidly-varying components; a position where the TITL model was definitely more flexible.

A number of other prosodic frameworks have been based on these three basic ones (eg. MOMEL [135], INTSINT [196], qTA [244], etc.) but few have presented a prosodic framework that offers a universal "language-agnostic" approach. The presented work in later chapters of this thesis strives to deliver exactly that.

2.4 Linguistic Phylogenies

Following once again the view of a language as a system that interacts with its environment, the concept of linguistic Phylogenetics is not ungrounded within the general phylogenetic framework [8]. Indeed in the last 15 years there has been a steady increase of papers where language development and biological speciation have been treated as quite similar [320]. Pagel in his eponymous review paper "*Human language as a culturally transmitted replicator*" [230] not only argues on the similarity of genes' and languages' evolutionary behaviour but offers an extensive catalog of analogies between biological and linguistic evolution as well.

Interestingly one might even argue that linguistic phylogenetic studies preceded biological ones at least in the *Western* world. While Aristotle (382-322 BC) was probably one of the first to cluster different animal species in terms of comparative methods [10], Socrates (469-399 BC), among other philosophers, actually realized that language "changed" (or at least "decayed") as time passed [49]. The reason for this observation was relatively simple: Homer's (∼8th century BC) writings while revered as accounts of heroic tradition, they were already *at least* 350 years old at the time of Socrates. People simply realized that Achilles did not speak like them. One of the first to formulate an actual "connection" though between

[7] The author feels that given the amount of work that Hansjörg Mixdorff has published in relation to the Fujisaki model, the Fujisaki-Mixdorff naming scheme would probably be more accurate; eg. see [210; 211; 215; 212; 216; 213; 214] among others.

[8] See section 3.5 for a short introduction in Phylogenetics.

Linguistics and Biology was Gottfried W. Leibniz (1646-1716)[9]. Leibniz advocated the idea of *Natura non facit saltus* ("nature does not make jumps"), gradual change. In addition to that he also advocated the ideas of Monadology: fundamental immaterial units that are eternal were "*the grounds of all corporeal phenomena*" [195]. Those ideas proved fundamental both in Biology and Linguistics. Therefore it is not surprising that the father of modern Biology, Charles R. Darwin (1809-1882) also made similar assertions regarding language in his landmark work *The Origins of Species.* Leaving historical remarks aside, the seminal paper of Cavalli-Sforza et al. [52] changed the way linguistic and genetic information are combined within a single analysis framework in modern times. There the authors focused on the reconstruction of a human phylogeny based on *maximum parsimony* [10] principles but importantly, after pooling genetic data geographically in order to account for heterogeneity, if heterogeneity persisted, they added an "*ethnolinguistic criterion of classification*". That allowed the synchronous derivation of a genetic and a linguistic phylogeny that displayed significant overlap and emphasized that the two fields not only could share methods but also results.

Up until relatively recently *maximum parsimony* trees [321; 107] and *comparative methods* [320] stood as the state-of-the-art in Linguistic Phylogenetics. And while comparative methods were already employed rather broadly within the context of glottochronology [105] the question of computational reconstruction of protolanguages started to emerge [227; 268]. Importantly people began to incorporate the phonetic principals in their tree reconstructions. Research came to the realization that exactly because language was just a human-bounded characteristic, direct analogies with generic Phylogenetics were not only possible, but actually strengthening the theoretical framework used. Language acquisition being associated with children (founder effects), parallel development of characteristics being not as uncommon as originally thought (convergent evolution), insertion-deletion-reversals being usual "units of changes" (the same operations being used in the changes of genetic code) showed that even qualitative linguistic phenomena could be encapsulated within a phylogenetic framework. Evidently the inherent problems of Phylogenetics such as having $(2N-3)!!$[11] rooted-trees for N leaves and being presented with a relatively small amount of data compared to the number of candidate trees [138] did remain, but linguists were nevertheless able to validate a very crucial insight from sociolinguistics: "*while most linguistic structures can be borrowed between closely related dialects, natively acquired sound systems and inflections are resistant to change later in life*" [268]. That meant that essentially a sound system was resistant to change and therefore presented a "good" character for phylogenetic studies. Insights like the positive correlation between rates of change and speaker population size [12] and the coherence of rule-based changes [221; 39] were established. Importantly almost all these techniques rely on binary features [73] or at a best case scenario multi-state ones [231; 106]. While computational linguists recognized the importance of phoneme sequences [39; 38] they do not act on premises of continuous data. As it will be shown in following chapters, the current work is not qualitatively comparable with state-of-the-art multi-state implementations [38] where the number of available training data is significantly larger. In addition, even excluding training sample size issues, the current methodology also acts almost agnostically in relation with semantic information by only using phonetic information. Undoubtedly the choice of discarding semantic information is a strong (yet not uncommon [105; 231]) assumption from a linguistic point of view but it presents itself as a definite progress in the phonetic literature because up until now comparative methods focused on scalar characteristics only.

As a final note we draw attention to the notion of the *molecular clock* of a biological phylogeny [165] and its significance in a linguistic phylogeny. As Gray et al. note, absolute dates in Linguistics are notoriously hard to get [106]. Disregarding the issues that relate to tree uncertainty and lack of concrete evidence, this difficulty is rooted with the quite restrictive assumption of (in this case) *lexical clock* (or a *glottoclock*). Exactly because one asserts that changes "occur in a more-or-less clocklike fashion, so that divergence between sequences should be proportional to the evolutionary time between the two sequences" [165], this assumption is hard to evaluate experimentally [105; 76]. Nevertheless we need this assumption for standard *Maximum Likelihood* methodology to be applicable. It would be therefore interesting to explore a possible application of methodologies that act under the assumption of a non-

[9]The reader will note a significant chronological gap. Aside the obvious need for significant biological and linguistic intellectual capital to be amassed, the Old Testament presented an *issue*; the Biblical story of the Tower of Babel made the question of linguistic phylogenies *somewhat heretic*.

[10]See Sect. 3.5.2 for an overview of tree reconstruction methodologies.

[11]The double or odd factorial where $(2k-1)!! = \frac{(2k)!}{2^k k!}$ or more generally: $(2k-1)!! = \Pi_{i=1}^{k}(2i-1)$

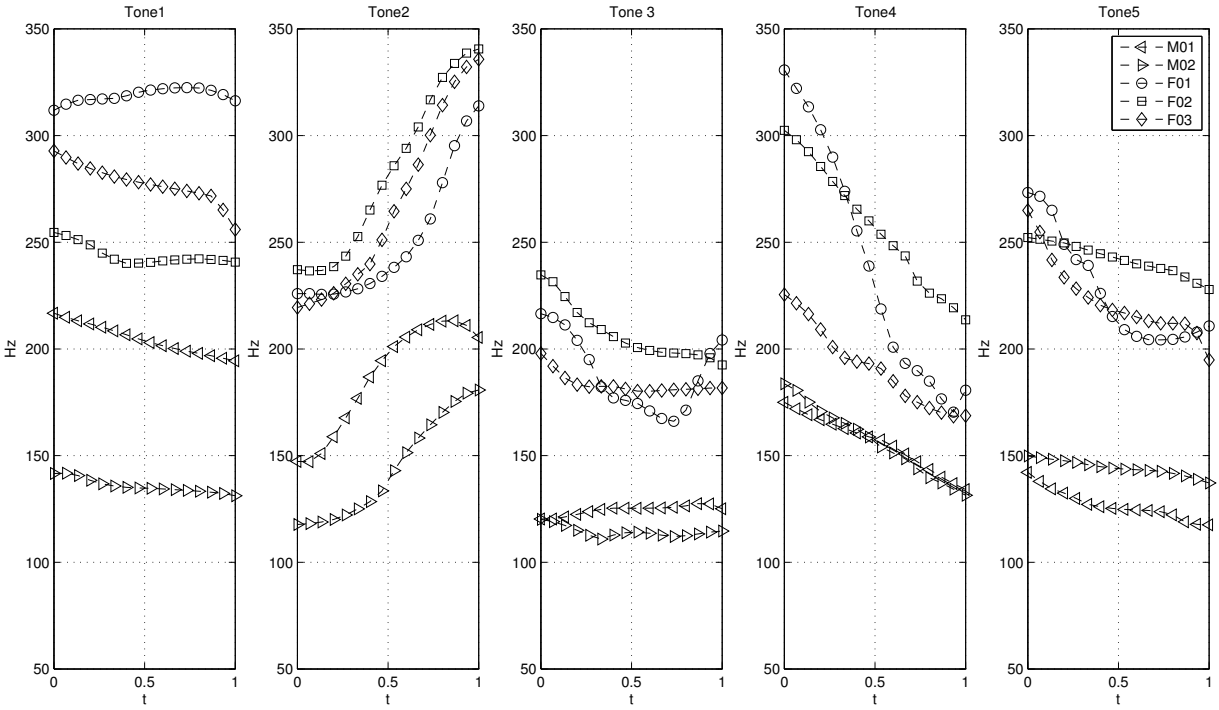

Figure 2.5: Tone realization in 5 speakers from the COSPRO-1 dataset.

universal time- continuum. The current work does explore this in terms of the observational time on the phylogeny's leaves but leaves the question of actual phylogenetic time (and its potential time-distortion) for future work.

2.5 Datasets

The current work employs two functional datasets. Both of them were made available to the author by his respective collaborators.

2.5.1 Sinica Mandarin Chinese Continuous Speech Prosody Corpora (COSPRO)

The Sinica Continuous Speech Prosody Corpora (COSPRO) [311] was collected at the Phonetics Lab of the Institute of Linguistics in Academia Sinica and consists of 9 sets of speech corpora. We focus our attention on the COSPRO-1 corpus; the phonetically balanced speech database consists of recordings of Taiwanese Mandarin read speech. The COSPRO-1 recordings themselves were collected in 1994. COSPRO-1 was designed to specifically include all possible syllable combinations in Mandarin based on the most frequently used 2- to 4-syllable lexical words. Additionally it incorporates all the possible tonal combinations and concatenations. It therefore offers a high quality speech corpus that, in theory at least, encapsulates all the prosodic effects that might be of acoustic interest.

After pre-processing and annotation, the recorded utterances, having a median length of 20 syllables, resulted in a total of 54707 fundamental frequency curves. Each F_0 curve corresponds to the rhyme portion of one syllable. The three female and two male participants were native Taiwanese Mandarin speakers. Using the in-house developed speech processing software package COSPRO toolkit [311; 312], the fundamental frequency (F_0) of each rhyme utterance was extracted at 10ms intervals, a duration under which the speech waveform can be regarded as a stationary signal [131]. Associated with the recordings were characterizations of tone, rhyme, adjacent consonants as well as speech break or pause. Importantly the presented corpus is a real language corpus and not just a series of nonsensical phonation patterns and thus while designed to include all tonal combinations, it still has semantic meaning.

More specifically the syllables are labeled with one of the four lexically specified tones or a sign that are phonologically toneless (tone 5). In addition contextual information is also associated with

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 index break in sentence	$B2$
B3	linear	Position of the B3 index break in sentence	$B3$
B4	linear	Position of the B4 index break in sentence	$B4$
B5	linear	Position of the B5 index break in sentence	$B5$
Sex	0:1	1 for male, 0 for female	Sex
Duration	linear	10s of ms	$Duration$
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 2.2: Covariates examined in relation to F_0 production in Taiwanese Mandarin. Tone variables in a 5-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). Reference tone trajectories are shown in Fig. 2.4.

each curve (see Table 2.2 for a list of covariates included). Fig. 2.5 shows time-normalized example realizations of all 5 tones for all 5 speakers.

2.5.2 Oxford Romance Language Dataset

The Oxford Romance Language Dataset was collected by Prof. John Coleman in the Phonetics Laboratory of University of Oxford between 2012-13. It consists of natural speech recordings of four languages; French, Italian, Portuguese and Spanish. Spanish recordings where classified as American or Iberian Spanish. For the purpose of this study American and Iberian Spanish are treated as distinct languages. The speakers utter the numbers one to ten in their native language and dialect. The dataset is inherently unbalanced; we have seven (7) French speakers, five (5) Italian speakers, five (5) American Spanish speakers, five (5) Iberian Spanish speakers and three (3) Portuguese speakers. We were unable to have records for all 10 digits from all speakers, this finally resulting in a sample of 219 recordings. The sources of the recordings were either collected from freely available recordings from language training websites or standardized recording made by university students.

An important caveat regarding this dataset is it is "real world". This contrasts with the COSPRO dataset that was recorded under phonetic laboratory conditions. The Romance language dataset consisted of recordings people made under non-laboratory settings (eg. classes, offices). It is also heterogeneous in terms of bit-rate sampling, duration and even format. As such before any phonetic or statistical analysis took place, all data were converted in *.wav files of 16Khz. This clearly undermines the quality of the recordings compared to the ones acquired by COSPRO but these conversions were deemed essential

Language	Number of Speakers (F/M)
French	7 (4/3)
Italian	5 (3/2)
American Spanish	5 (3/2)
Iberian Spanish	5 (4/1)
Portuguese	3 (2/1)

Table 2.3: Speaker-related information in the Romance languages sample. Numbers in parentheses show how many female and male speakers are available.

to ensure sample homogeneity. Fig. 2.1 shows a typical waveform reading.

The Romance language dataset is exclusively used for the phylogenetic applications showcased in Chapt. 6 as it provides an obvious "well-examined" [105; 230; 106] sub-sample of the greater Romance languages linguistic family; some "standard members" of the Romance family like Catalan and Romanian were not included. Fig. 2.6 shows an unrooted linguistic phylogenetic tree $\mathbf{T}$ of nominal phylogenetic distances for the languages at hand based on Grey et al on [106].

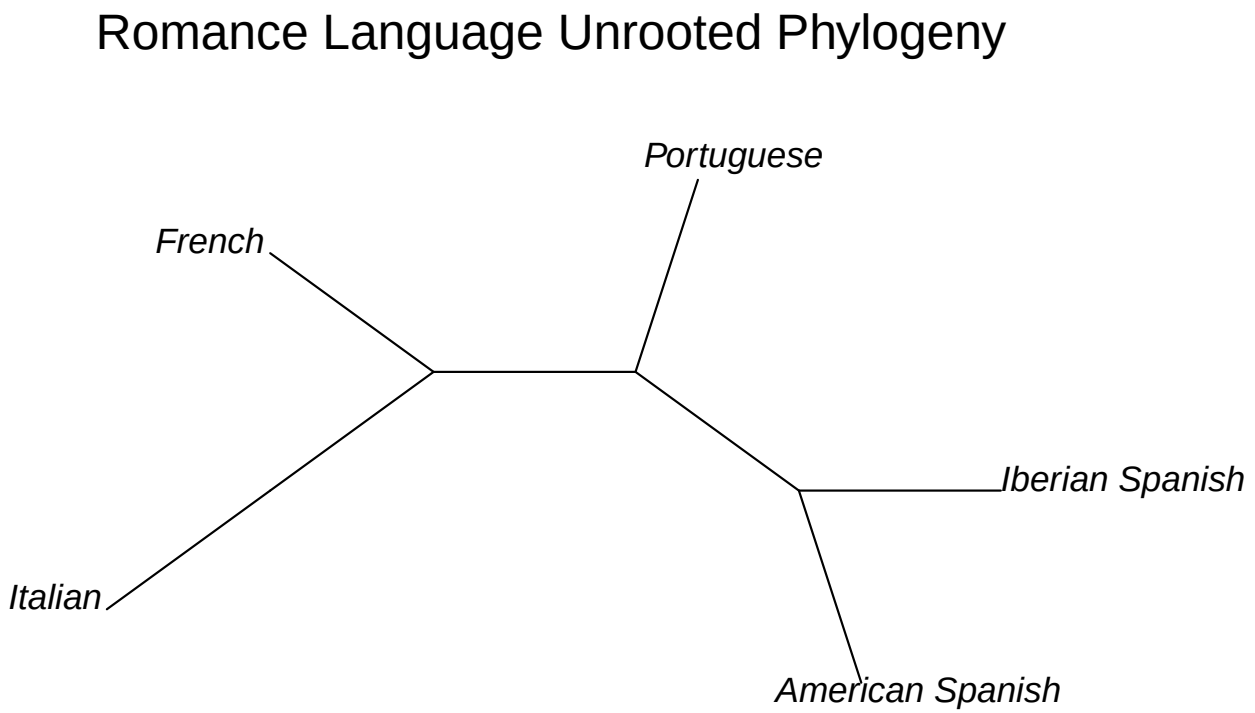

Figure 2.6: Unrooted Romance Language Phylogeny based on [106]. Branch lengths do *not* correspond to lexical clock time.

Chapter 3

Statistical Techniques for Functional Data Analysis

Functional Data Analysis (FDA) defines a framework where the fundamental units of analysis are functions. The dataset are assumed to hold observations from an underlying smoothly varying stochastic process. FDA application works in both a parametric and a non-parametric setting. Using a parametric framework one usually assumes that the underlaying process is a member of a specific class of functions, eg. Gaussian [122], Dirichlet [236], Poisson [147] or some other generalization of point-processes (eg. Cox processes) [40; 26]. Under a non-parametric framework one directly usually utilizes a spline- [254] or a wavelet-based [115] representation of the data. Nevertheless irrespective of the framework used as stated by Valderrama: "*approximations to FDA have been done from two main ways: by extending multivariate techniques from vectors to curves and by descending from stochastic processes to real world*" [317].

In particular if one considers a smooth [1] function $Y(t), t \in T$, if $E\{Y(t)\}^2 < \infty$ and $E\{\int Y^2(t)dt\} < \infty$, Y is said to be squared integrable in the domain T. Additionally if one assumes that the instances of function Y, ie. a functional dataset Y_{ij}, define a vector space in $L^2[0,1]$ that space has a vector space basis spanning it.

A smooth random processes Y can be defined to have a mean function:

$$\mu_Y(t) = E[Y(t)] \tag{3.1}$$

and a symmetric ($C_Y(s,t) = C_Y(t,s)$) auto-covariance function as:

$$\begin{aligned} C_Y(s,t) =& Cov[Y(s), Y(t)] & (3.2)\\ =& E[(Y(t) - \mu_Y(t))(Y(s) - \mu_Y(s))]. & (3.3) \end{aligned}$$

Taking advantage of the symmetric and positive semi-definite nature of the covariance function $C_Y(s,t)$ its spectral decomposition follows by Mercer's theorem [207] as:

$$C_Y(s,t) = \sum_{\nu=1}^{\infty} \lambda_\nu \phi_\nu(s)\phi_\nu(t), \tag{3.4}$$

where $\lambda_1 \geq \lambda_2 \geq ... \geq 0$ are ordered eigenvalues of the operator C_Y with $\sum_{\nu=1}^{\infty} \lambda_\nu < \infty$ (effectively restating its semi-positive definite nature and ensuring that the operator C_Y has a finite trace respectively) and ϕ_ν's are the corresponding and by definition orthogonal, eigenfunctions in $L^2([0,1] \times [0,1])$. Finally given Eq. 3.4, the Karhunen-Loeve expansion of the observations Y [121], registered over a common finite grid indexed by i and j for the curve and time index respectively, is presented as:

$$Y_{ij} = \mu_Y(t_{ij}) + \sum_{\nu=1}^{\infty} \xi_{i\nu}\phi_\nu(t_{ij}) + \epsilon_{ij} \tag{3.5}$$

[1] Contrary to more rigorous definitions of smoothness, here smoothness in relation to a dataset's properties is defined as possessing "one or more derivatives" [254]. This assumption is in place as it allows for a minimum penalization in terms of roughness.

where $\xi_{i\nu}$ is the zero-meaned functional principal component score with variance proportional the corresponding λ_ν; ϕ_ν acting as the basis for the space spanned by Y_{ij}.

Typical functional datasets are collections of curves [255; 11] and shapes [162] but also surfaces [281], general three dimensional objects [174] or even more high-dimensional objects [71; 237] are increasingly considered. It is safe to say that with the increased use of real-time measurement instruments (and data storage resources) dataset of functional forms will become increasingly more common [65]. Intrinsically the main attribute that enables a dataset to be considered functional is that the data themselves are registered on a continuum, that usually being time or space [257]. More "exotic" continua like acidity (pH scale) or molecular mass [170] have already been considered, as micro-array data offer abundance of readings. The current thesis mostly deals with simple one-dimensional instances that are presented in the form of curves, the continua in question being time (in the case of F_0 curves), and pH (in the case of proteome profiles). Nevertheless as shown in certain cases the generalizations from curves (1-D objects) to surfaces (2-D objects) can be relatively straightforward; an example being the speech spectrograms that are shown in section 2.1.

Functional objects have an internal structure which can prove both restrictive and concurrently supportive to the practitioner's insights. On the one hand, arbitrary permutations of a function's values are invalid sample transformations as they distort the continuity of the underlying function model. On the other hand, differentiating a sample can capture the sample's rate of change in a similar manner to differentiating a displacement function results in an object's speed [258]. In general the fact that the space over the dataset is indexed, allows not only for the data to be interpreted over that space but also enables a practitioner to transform the dataset by exploiting that space's physical properties. For example the study of growth curves has often been conducted with respect to the first-difference of the data as this transformation exemplifies changes in growth patterns [254].

Pioneering the concept of FDA was Rao, who first defined a statistical framework where a sample of curves is viewed as the realizations of a smooth stochastic process [261; 262]. The usual smoothness assumptions when working with functional data is that observed longitudinal data are twice differentiable [254]. Approximately at the same time as Rao, Tucker [315] also introduced the idea of a function as the fundamental unit of statistical analysis. From there on most of the functional data analysis literature follows the evolution of spline literature (eg. [6; 60]), only to reach the early 80's when Ramsay published his eponymous article *"When the data are functions"* [257]. At roughly the same time (late 70's) the works of Diggle [67] and Ripley [269] on spatial patterns began formulating the spatial point processes literature that gave raise to subsequent point-processes in FDA. Ramsey's early work was later extended and established by the works of Rice and Silverman [267] and Ramsey and Silverman [254] in an autonomous field of statistical study and has lead to the modern day definition of functional data as given by Ferraty and Vieu: a random variable x being called *functional variable if it takes values in an infinite dimensional space (or functional space)* [83]. Therefore the general form of a functional linear model assuming a function-valued response variable $Y(s)$, $s \in S$, conditional on the functional response variables $X(t)$, $t \in T$ takes the form:

$$y_i(t) = \mu_y(t) + \int_T \beta(t,s)x_i(s)ds + \epsilon_i(t) \tag{3.6}$$

ϵ_i being independent and zero-meaned random errors and $\beta(\cdot,\cdot)$ being a square integrable bivariate regression function [7]. As it will be shown in later sections the current work does not examine cases of functional response variables explicitly. It is based on the orthogonal decomposition of response sample curves as it is presented in Eq. 3.5. This particular format of functional regression was explicitly formulated in this way by Faraway [78], with the extensions to the case of a multi-level design following in early 2000's [252; 72]. Nevertheless the idea behind functional regression was at least partially presented by Massy during 60's and his work in Principal Component regression [326].

In the following sections we outline a standard methodology when working with functional data [254]. We start with smoothing and interpolation, we then conduct data registration (feature alignment) and finally progress into exploring a dataset variability using dimensionality reduction (feature extraction) as well as functional regression techniques (predictive and explanatory analysis).

3.1 Smoothing & Interpolation

When handling functional data one mostly works under the assumptions that he has either a sample of sparse (and possible irregularly sampled) observations [337], or a sample of densely/perfectly observed discretised instances of smooth varying functions [11]. In both cases, one does assume though the existence of a smooth generating function w such that the observed dataset is a collection of function realizations and that the observed "deviations" from smoothness are due to measurement errors and/or simply noise. It is essential therefore as a first step to ensure the sample curves w_i are "smooth". We recognize three different, though not totally unrelated, techniques to achieve this: localized kernel smoothing [57], smoothing splines [115] and wavelet smoothing [8]. The current work employs almost exclusively the first technique.

Examining kernel smoothing one encounters the notion of a kernel, its use is directly analogous with the use of windowing examined earlier in section 2.1. A kernel is a non-negative real-valued integrable function K satisfying the following two requirements:

$$\int_{-\infty}^{+\infty} K(t)dt = 1 \text{ and} \tag{3.7}$$

$$K(-t) = K(t) \text{ for all values of } t. \tag{3.8}$$

A kernel's bandwidth b plays a key role in kernel density estimation; it can be viewed as analogous to the characteristic length scale of a Gaussian process [263] and it informally encodes how far does the correlation between the points of the continuum over the points are measured upon extends. The most basic kernel smoother is the Nadaraya-Watson kernel smoother; it is effectively a weighted mean μ_{NW}, where the weights are given by the kernel function K and bandwidth b used. Therefore evaluating $y(t_i)$ for a given t_i involves the simple calculation of $\mu_{NW}(t_i) = \frac{\sum_{i=1}^{L} K\left(\frac{t_i - t}{b}\right) y(t)}{\sum_{j=1}^{L} K\left(\frac{t_j - t}{b}\right)}$ [62]. Gasser and and Müller [96] having proposed a similar smoother with better properties as: $\mu_{GM}(t_i) = \frac{1}{b}\sum_{i=1}^{L}\int_{s_{i-1}}^{s_i} K(\frac{u-t_i}{b})y(t)$, where $s_i = \frac{x_{i-1}+x_i}{2}$. Generalizing these and in line with Chiou et al. [57] we currently use a locally weighted least squares smoother, S_L, in order to fit local linear lines to the data and produce smooth data-curves interpolated upon a common time-grid of L points on a dimensionless interval $[0, 1]$. We essentially take a weighted average of the points laying within the smoother's bandwidth for a given point t. In particular the form of the (Gaussian) kernel smoother used is the following:

$$S_L\{t; b, (t_i, y(t_i))_{i=1,\ldots,L}\} = \tag{3.9}$$

$$\underset{\alpha_0}{\text{argmin}}\{\min_{\alpha_1}(\sum_{i=1}^{s} K(\frac{t-t_j}{b})[y(t_i) - \{\alpha_0 + \alpha_1(t-t_j)\}]^2))\}$$

the actual kernel used being the Gaussian kernel function $K(x)$

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{3.10}$$

Here, the value of the curve at a point t found at the center of a smoothing window $[t - b, t + b]$ is calculated to equate the intercept of the weighted regression line among the data of the smoothing window. The fixed parameter bandwidth b, which corresponds to a tuning parameter, is estimated using cross-validation [149]. Qualitatively, smaller values of b come to the expense of high-variability where larger values of b to the expense of higher bias as broader smoothing windows incorporate information from possibly "unrelated" distant points.

Cross-validation is done in the following way: For a given bandwidth b and for each curve w_i in our dataset we produce a smoothed estimate for a curve w_i while randomly excluding one of the w_i readings beforehand, $w_i(t_{random})$. We then record the "reconstruction" error associated with this smoothed curve, $w_i(t_{random}) - \hat{w}_i(t_{random})$, and associate a residual sum of squares (RSS) cost with each given value of b we test for. The value resulting in the smaller RSS cost is the one we carry forward in our analysis.

It is important to note that by employing a locally weighted least squares smoother we take account of a second issue that might arise during data recording; irregular sampling. While ideally all sample curves are sampled over the same exact dense grid of points, this is often not the case. In reality missing values (treated as MCAR [2]), non-equidistant sampling points, and/or different sampling rate might result to an irregularly sampled dataset. An irregularly sampled dataset can be problematic when using techniques that rely on the number of readings available (eg. functional principal component analysis).

Therefore by setting the number of estimation points t, L within the context of the smoother S_L, we provide the equidistant grid over which our smoothed reading will lie. As with the case of b, L the number of grid points is found empirically [3]. Arguably the choice of L is less standardized; nevertheless given that one does not use "extreme values" that would result in finite sampling effects, this choice should not affect the analysis in any significant manner [117]. The kernel function K is set to the Gaussian kernel function $K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$. Common alternative kernels are also the Epanechnikov ($K(x) = \frac{3}{4}(1-x^2)\delta_{|x|<1}$) and the triangular kernel ($K(x) = (1-|x|)\delta_{|x|<1}$) [149]. This is in a way analogous with *windowing* a signal to conduct short-time analysis. Fitting a weighted linear model within the kernel's window corresponds roughly to the windowing idea that the signal/sample within that window is stationary and the model's intercept is a good estimate for the overall behaviour of the windows sample.

As an alternative to kernel smoothing, Guo [115] presented a test-case where the smoothing framework employed a generalization of smoothing splines [173]. The basic idea behind the use of spline functions is that one breaks the sample into smaller adjacent sub-samples at some particular break-points (or knots) and fits a different polynomial in each region [4]. In the case of standard spline smoothing, the number of knots, the number of basis functions and the order of the spline employed affect the final result; the computation of smoothed curves laying on the quantification of the roughness of the smoothing. Koopman & Durbin put smoothing splines into state space form where a univariate Kalman filter smoother algorithm provided the fitted smoothing spline [173]. Filtering then constituted recursively estimating the values of the state equation of the associated multivariate Gaussian linear state space model at the predetermined time-points t. Interestingly under this paradigm while originally the curve (or more generally the system at hand) is considered time-invariant, one can directly use time-variant components to encapsulate non-stationarity by requiring certain diffuse priors. This is definitely a welcome functionality but we note that for most phonetics analysis applications syllable phonemics are considered to be stationary along their recorded trajectories.

A third alternative is based on the notion of multi-resolution signal decomposition as this is implemented by using wavelets [202]. Wavelet estimators use projections onto subspaces of $L^2[0,1]$ to represent successive approximations of the dataset but in contrast with a standard orthogonal basis (eg. Fourier polynomials) that might be localized only in a single domain (eg. Fourier polynomials are localized in frequency) wavelets can be localized both in time and in frequency. Therefore as presented by Morris & Carroll one can use wavelets (and in particular the Discrete Wavelet Transformation - DWT) to smooth the original noisy functional data [218]. Under this paradigm, where clearly one must use a continuous mother-wavelet, smoothing is achieved by thresholding; certain wavelet coefficients are "thresholded" in order to exclude parts considered to be noise.

As with the two techniques proposed above, while extremely flexible, wavelet smoothing is in its core a semi-parametric technique. In the case of wavelets smoothing the "parameter" being the choice of the original mother wavelet, in the case of spline smoothing, the type of splines and the order employed, and in the case of kernel smoothing the "parameter" being the choice of kernel and bandwidth. "Purely" parametric techniques have also been employed; a number of other parametric bases have been at times suggested (Fourier polynomials, step-functions, power bases [254]) but they have not enjoyed wide spread use in the functional data literature. Nevertheless a parametric technique [93] is employed in the last section based on the notion of the two dimensional discrete cosine transform (DCT) [299; 272].

Ultimately the use of kernel, splines or wavelets estimators are techniques to a non-parametric

[2] Missing Completely At Random

[3] In practice one sets L equal to the expected number of readings per case for the sample at hand.

[4] An *extreme* case of knot usage is associated with the concept of a smoothing spline where each point of the sample is used as a knot, smoothing being achieved by quantifying the roughness of the resulting function $\int [D''y(t)]^2 dt$ [254].

regression problem; given a signal $y(t)$ they return $y_{smooth}(t)$, such as:

$$y_{smooth} = Hy, \tag{3.11}$$

H being the projection or *hat* matrix. On a conceptual level one might argue that smoothing is a *dimension addition* as it generates readings across a whole continuum that was beforehand unpopulated. Numerically, both spline and wavelet smoothing have been reported to be more computationally efficient than using kernel smoothing [331; 43; 308]. Nevertheless for a phonetic related application the additional computational cost of kernel smoothing does not hinder further analysis. Kernel smoothing gives a semi-parametric way of directly smoothing our sample in a straight-forward manner that is easily adaptable to specific problem mechanics we may wish to adhere to (eg. not smoothing edge readings). An additional advantage of this approach is that we are not making an explicit assumption about the underlying form of the function fitting the data at hand. While this might be problematic if one wishes to have specific mathematic expressions describing the smoothed curve, in applied terms that is not usually an issue.

3.2 Registration

Registration is probably one of the most intrinsic pre-processing procedures; as humans we conduct registration procedures constantly and effortlessly. The need to register stems from the fact that there might be a misalignment between the chronological time we know a signal (or better yet a process) to evolve and the real time we perceive it evolving. Our word recognition mechanism is a prime example: while no new single utterance we ever listen to is completely identical in terms of intensity or tempo with one we have heard before, we nevertheless understand the majority of the spoken words directed to us because we can associate them with a "reference" word we already know. Thus, common patterns between what we hear and what we have already heard allow us to deduce what is communicated (usually). These common patterns are recognized through the registration (or aligning) of the query signal to some reference signal we already know the meaning of. Sakoe and Chiba's seminal work on registration addresses the same problem where it focused on eliminating "*fluctuations in a speech pattern time axis*" [279]. It importantly was one of the first references introducing the concept of a warping function; "*a model of time-axis fluctuations in a speech pattern*" where it could be "*viewed as a mapping from the time axis of pattern A onto that of pattern B*". Registration for functional data tries to achieve the same; align the data by mapping all of them onto a common chronological time-scale facilitating statistical inference. Effectively what we are trying to do is decipher the shape of the function where by shape we mean, as Le & Kendall postulate, "*what is left when the effects associated with translation, scaling and rotation are filtered away*" [185]. As it will be presented later (Chapt. 5), this will allow one to formulate models that account concurrently for variation that can be attributed to registration differences (phase), as well as deterministic variation attributed that is independent of "timing effects" (amplitude).

Given a functional dataset we recognize variation in intensity and tempo to correspond to amplitude and phase variations respectively (Fig. 3.1). We need to therefore register the dataset onto a common time-scale. The obvious way of registering a dataset is through landmark detection. Landmarks are "points of interest" [294]. Assuming a density function D, calculating some property of interest within a neighbourhood Ω, a point $x \,\epsilon\, \Omega$ is consider a landmark L if:

$$L = \{x|\, ||D(x)|| > D_\mu + \lambda D_\sigma \wedge ||D(x)|| \geq \left|\left|D(x')\right|\right| \forall (x')\epsilon\Omega\}$$

where D_μ and D_σ are the average and the standard deviation of the density function D over the entire space X and λ is a user-defined threshold. In practical terms, a point x is a landmark if it exceeds some threshold given by $D_\mu + \lambda D_\sigma$ and is larger or equal of all other points x' in it's neighbourhood Ω (eg. the amplitude peaks in Fig. 2.3) Having found the landmarks, registration transforms the individual time so the landmarks appear synchronized. Formally given a query curve w_i and reference curve w_j where a landmark feature appear at times $t_{\mathbf{i},landmark}$ and $t_{\mathbf{j},landmark}$ respectively, registering w_i onto w_j consists of finding the warping function h_i such that the warped instance of w_i, w_i^*:

$$w_i^*(t) = w_i[h_i(t)] \tag{3.12}$$

allowing (given that the amplitude characteristics of the sample at hand are approximately equal among the realizations examined)

$$w_i^*(t_{\mathbf{j},landmark}) \approx w_j(t_{\mathbf{j},landmark}) \tag{3.13}$$

or as put forward by Ramsay and Silverman the two curves "*have more or less identical argument values for any given landmark*" [254]. Here h_i the warping function, follows the same exact properties as put forward by Sakoe and Chiba [279] but this time under a functional rather than a discrete framework:

- Strict monotonicity : For $t_1 < t_2$ where $t_i \in [0, T]$, $h_i(t_1) \leq h_i(t_2)$
- Boundary conditions: $h_i(0) = 0$ and $h_i(T) = T$
- Continuity conditions: $|t_2 - t_1| < \delta \Rightarrow |h_i(t_2) - h_i(t_1)| < \epsilon$, $\epsilon > 0$ and $\delta > 0$

The basic landmark definition outlines the main "problem" of using landmarks; they are not rigorously defined or guaranteed to be detectable; occasionally they can be truly absent. In such cases registration is either wrong (as we would align incompatible characteristic) or simply impossible (as there would be nothing to align for in the first place). Landmark registration is highly parametrized by the successful detection of landmarks. A more robust measure would be a metric identifying phase variation utilizing the whole shape of the objects aligned; for example least squares criterion [254] utilizing the Procrustes method [256]. In contrast with landmark registration this is an iterative procedure aiming to minimize the registered curves sum of squares errors $REGSEE$ where that is defined as:

$$REGSEE = \sum_{i=1}^{N} \int_T [w_i(\delta_i + t) - \hat{\mu}_w(t)]^2 ds \tag{3.14}$$

$$= \sum_{i=1}^{N} \int_T [w_i^*(t) - \hat{\mu}_w(t)]^2 ds \tag{3.15}$$

where δ_i is the shift related to each point of w_i and $\hat{\mu}_w$ is the empirically estimated sample mean calculated prior to each Procrustes step. Following this rationale, in each iteration we minimize the integrated sum of square differences between the warped objects w^* and the sample mean $\hat{\mu}_w$. This is clearly more robust in the misspecification of landmarks but it does make the restrictive assumption of treating all discrepancies between any two given objects w_i and w_j as products of phase variations. In a way it "warps too much"; it does not incorporate a way of penalizing excessive phase distortion. A number of different

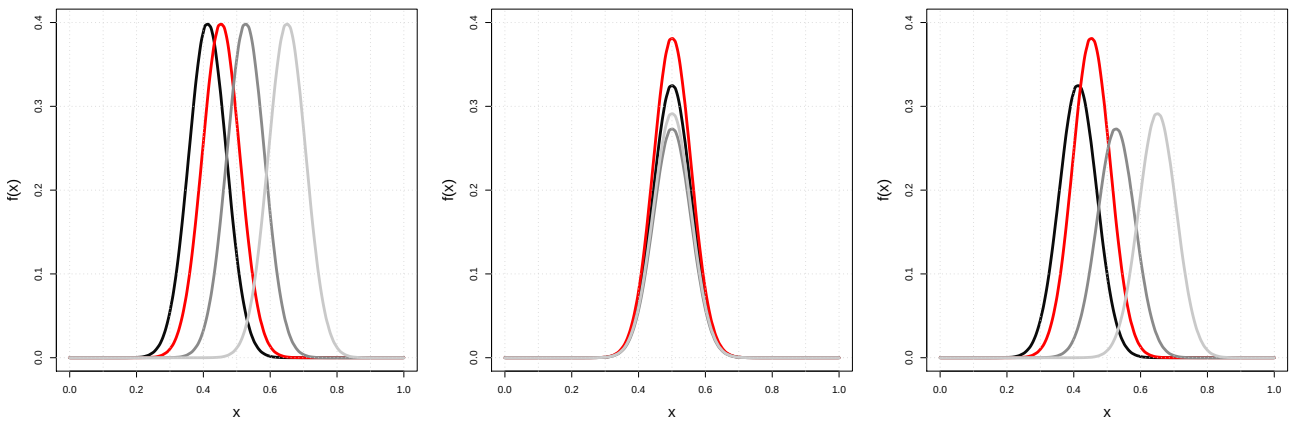

Figure 3.1: Illustration of three different types of variation in functional data. The left subplot shows four data-curves displaying only phase variations. The central subplot shows four data-curves displaying only amplitude variations and the right subplot shows four data-curves displaying concurrent phase and amplitude variations. Here the peak of the "Gaussian bump" serves as an obvious landmark.

approaches to address this; most techniques follow the idea of minimizing a penalizing squared error criterion where the penalization is relative to the squared norm of the warping function h [256] and/or decomposing the sample in terms of splines where the spline coefficients are estimated by combining information from the whole sample [98]. Both time-registration techniques can be utilized not only in terms of the dataset y but also with its derivatives Dy. Dy is not only used to highlight topological characteristics (eg. peaks in the case of landmark registration) but also for optimization purposes (eg. commonly in Procrustes-like methods [254]).

The current work examined two different non-parametric techniques of curve registration: 1. pairwise synchronization [304] and 2. area under the curve synchronization [340]. Both methodologies put forward a number of additional assumptions regarding the nature of the functional dataset W and the corresponding warping function H. In addition to those, two complimentary curve-registration frameworks are show-cased; one based on self-modelling warping functions [98] and one based on the square-root velocity function of the curve samples [209; 176]. We need to stress that despite our critique on the short-comings of landmark registration, for most alignment procedures landmark registration, where usually landmark detection has been done manually by an expert user, serves as benchmark for these non-parametric techniques [168; 304]. If clear and consistent landmark features such as local extrema or inflection points are present in a dataset it is reasonable for them to be exploited; such an approach being forwarded by the concept of "structural average" [168]. Registration studies using functional data analysis in phonetics have already taken advantage of similar characteristics such as the onset and offset of accented vowels in the sentence [113] or the kinematic zero-crossings, time points where the tongue tip is about to move away from the position extrema, during speech [188]. Nevertheless they have also identified certain shortcomings; such having too short trajectories to identify reliable landmarks [114] or being considerably variable to the point of rendering derivative readings uninformative [171]. These giving us further motivation to explore a "landmark-" and "derivative-" free approach.

3.2.1 A non-linguistic dataset

For the next section only we introduce a non-linguistic dataset to showcase differences between the presented time-synchronization frameworks. This is because the complex covariate structure of our linguistic datasets can make difference in registration hard to immediately visualize. The *Flour Beetle Dataset* [148] is a comprehensive dataset of 849 growth curves of flour beetle from larvae to pupae; the average length of the larval period, our T was 17.56 days. We use a sub-sample of 60 random selected growth curves to keep the illustration straightforward. Contrary to Irwin & Carter's approach of using a smoothing spline fit we use a simple Nadaraya-Watson smoother where the bandwidth b was evaluated using cross-validation to give smooth initial curves, as this is the procedure that will be used in our later data analyses.

3.2.2 Pairwise synchronization

Pairwise (curve) synchronization in a broad sense works by employing the Law of Large Numbers [62]; if one averages over the 1-to-1 random mappings of a query curve y_i against a sufficiently large sample of reference curves y_j's, the expected mapping will be the global mapping of the query curve against the reference time of the sample of curves Y (Eq. 3.19). Formally by utilizing the formulation presented by Tang & Müller [303] one can introduce two types of functions, w_i and h_i; w_i and h_i are associated with our observed curve y_i , $i = 1, \ldots, N$ that is the i-th curve in the sample of N curves. As such for a given curve y_i, w_i is the amplitude variation function on the domain $[0, 1]$ while h_i is the monotonically increasing phase variation function on the domain $[0, 1]$, such that $h_i(0) = 0$ and $h_i(1) = 1$, these being the same properties put forward by Sakoe and Chiba [279].

One could distinguish between deterministic and random phase variation. Both can occur within certain experimental settings. Nevertheless the work presented focuses exclusive in the random case.

For generic random phase variation or warping functions h and a sample time domain $[0, T]$, T_i being the duration of the i-th curve, we consider time transformations $u = h^{-1}(\frac{t}{T}))$ from $[0, T]$ to $[0, 1]$ with inverse transformations $t = Th(u)$. Then, the measurement curve y_i over the interval $t \in [0, T_i]$ is

assumed to be of the form:

$$y_i(t) = w_i(h^{-1}(\frac{t}{T})) \Leftrightarrow w_i(u) = y_i(Th_i(u)) \tag{3.16}$$

where $u \in [0, 1]$. As such, a curve y_i is viewed as a realization of the amplitude variation function w_i evaluated over u, with the mapping $h_i^{-1}(\cdot)$ transforming the scaled real time t onto the universal/sample-wide time-scale u as per Eq. 3.12.

Because h_i is a piecewise-linear function, if one views h_i as just the linear interpolation of p predetermined knots $(\zeta_1, \ldots, \zeta_p)$ and given that h_i has to follow specific end-point assumptions the whole function h_i is reduced into estimating p spline coefficients. Complementary to that, w_i, aside the obvious smoothness assumption, is assumed that is of the form:

$$w_i(t) = \mu(t) + \delta V_i(t) \tag{3.17}$$

where μ is a smooth bounded twice differentiable fixed function and importantly V_i is a smooth random trajectory such that:

- $E(V_i(t)) = 0$ and
- V_i, V_i' and V_i'' are all bounded by a constant $C_1 \in (0, \infty)$

Therefore for any t_1 and t_2 in [0,1], if $t_1 < t_2$ then it exists a pair ω_1 and ω_2 in $(0, \infty)$ such that if $\omega_1\omega_2 \leq 0$ any valid pairwise warping function $g_{k,i}(t) = h_k(h_i^{-1}(t))$ has to satisfy $\omega_1 \leq (g(t_1) - g(t_t2))/(t_1 - t_2) \leq \omega_2$. Thus we ensure the statistical identifiability of model (Eq. 3.16) by the exclusion of essentially flat amplitude functions w_i for which time-warping cannot be reasonably identified. Also, it encodes the assumption that the time-variation component reflected by the random variation in h_i asymptotically dominates (but does not account wholly) the total variation. This is a very important modelling assumption as it dictates the interpretation of the samples at hand.

More specifically in computational terms if one defines the pairwise warping function $g_{k,i}(t)$ as the 1-to-1 mapping from the i-th curve time-scale to that of the k-th (Fig. 3.2) and that the inverse of the average $g_{k,i}(\cdot)$ (Eq. 3.21) for a curve i is the curve y_i's corresponding warping function h_i. h_i is therefore a map between individual-specific warped time to absolute time [304]. Because $g_{k,i}(\cdot)$, is a time-scale mapping, it has a number of obvious restrictions on its structure. Firstly, $g_{k,i}(0) = 0$ and $g_{k,i}(1) = 1$.

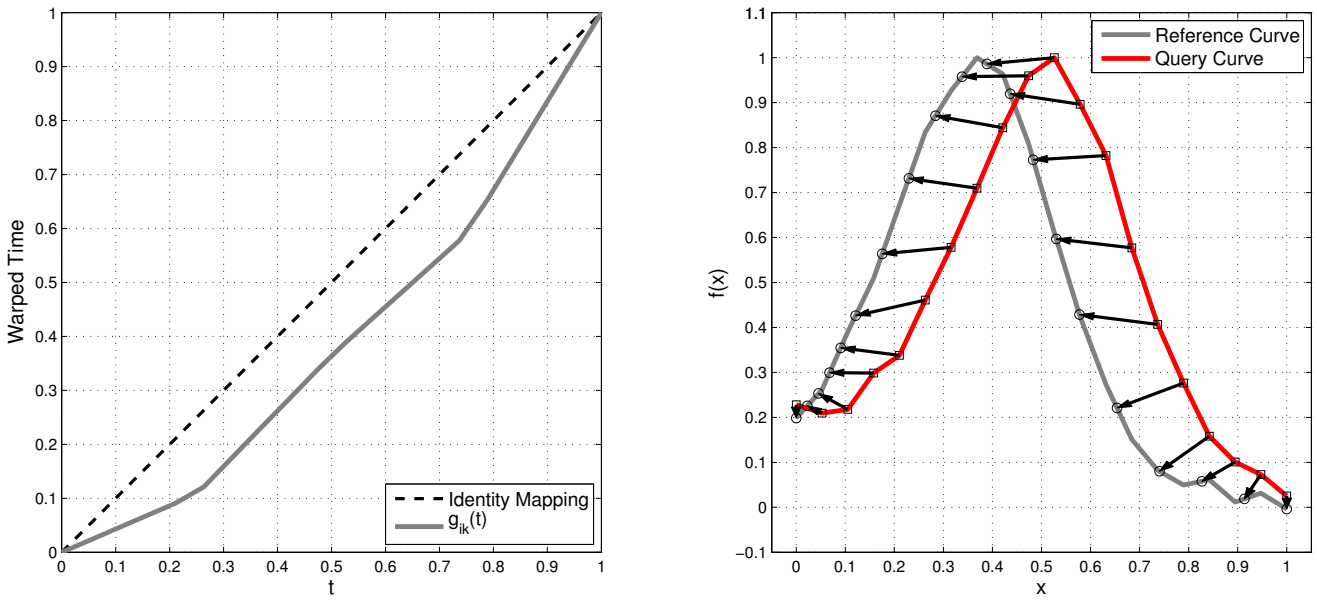

Figure 3.2: Illustration of pairwise warping function. The left subplot shows g_{ik} with respect to an identity warping function (perfect synchronization between curves); the piece-wise linear nature of it is evident. The right subplot shows the query curve's 1-to-1 mapping into the reference curve. The query curve "evolves slower" than the reference curve during its first half.

Secondly, it should be monotonic, i.e. $g_{k,i}(t_j) \leq g_{k,i}(t_{j+1})$, $0 \leqslant t_j < t_{j+1} \leqslant 1$. Finally, $E[g_{k,i}(t)] = t$. This final condition is the one used to ensure we can obtain a final estimate for h_i^{-1} as:

$$h_i^{-1}(t) = E[g_{k,i}(t)|h_i(t)] \tag{3.18}$$

with an equivalent finite sample version being:

$$\hat{h}_i^{-1}(t) = \frac{1}{m}\sum_{k=1}^{m} \hat{g}_{k,i}(t), \quad m \leq N \tag{3.19}$$

These theoretical requirements of the estimation of pairwise warping functions $g_{k,i}$ in practical terms mean that: 1. $g_{k,i}(\cdot)$ needs to span the whole domain, 2. we can not go "back in time" mapping a time-point t_j at a time after the one that a time-point t_{j+1} was mapped at and 3. the time-scale of the sample is the average time-scale followed by the sample curves. With these restrictions in place we can empirically estimate $g_{k,i}(\cdot)$ as $\hat{g}_{k,i}(t) = \text{argmin}_g D(y_k, y_i, g)$ where the "discrepancy" cost function D is defined as:

$$D_\lambda(y_k,y_i, g) = E\{\int_0^1 (y_k(g(t)) - y_i(t))^2 + \lambda(g(t) - t)^2 dt|y_k, y_i\}, \tag{3.20}$$

λ being an empirically evaluated non-negative regularization constant. Intuitively the optimal $g_{k,i}(\cdot)$ minimizes the differences between the reference curve y_i and the "warped" version of f_k subject to the amount of time-scale distortion produced on the original time scale t by $g_{k,i}(\cdot)$. Having a sufficiently large sample of m pairwise warping functions $g_{k,i}(\cdot)$ for a given reference curve y_i, the empirical internal time-scale for y_i is :

$$\hat{h}_i^{-1} = \frac{1}{m}\sum_{k=1}^{m} \hat{g}_{k,i}(t), \tag{3.21}$$

the global warping function h_i being easily obtainable by simple inversion of h_i^{-1}. One can note that this framework is almost directly generalizable in the case of a two-dimensional functional objects that has a single *relevant* axis (See Sect. 6.2.1). Fig. 3.3 shows an example of pairwise warping applied on a real dataset. It is immediately evident that the growth curves appear more aligned when warped than unwarped; the warping functions show small phase variations all across the spectrum of T.

As a final note attention is drawn to the previous fact that estimation of g_{ik} effectively is the estimation of the p predetermined knots for the piecewise linear spline it represents. The standard literature does not appear to provide theoretical results about the convexity of this optimization problem. The default implementation of this method (as implemented in the `MATLAB` package `PACE` [304]) relies on BFGS. A second implementation of the optimization procedure, implemented by the author of this thesis, employs Simulated Annealing. This random search method has given solutions that are qualitative comparable with the ones provided by BFGS but with a significant speed-up ($\sim$ 25x). Preliminary investigation using the linear approximation optimization algorithms (COBYLA [243]) provided results with an even more significant speed-up ($\sim$ 50x +).

3.2.3 Area under the curve synchronization

As the name suggests the area under the curve (AUC) time-registration defines the concept of area under some n-dimensional curve and focuses on normalizing the curve trajectory in a way that the area covered up to a certain point in the trajectories of the sample curves is approximately equal. For that reason this approach is also known as *quantile synchronization* [340]; this also alludes that the main analyses types this approach is intended for, are the analysis of distribution functions [203] as well as any other instance of data where the sample can be considered to represent a density [64]. While the theoretical framework under which an AUC framework is presented treats the data as being density functions, it is important to note that for any (curve) sample where we can assume that the variation observed is dominated by

[5]Warping was implemented in `MATLAB` using `PACE` version 2.16; it contains solver modifications by PZH.

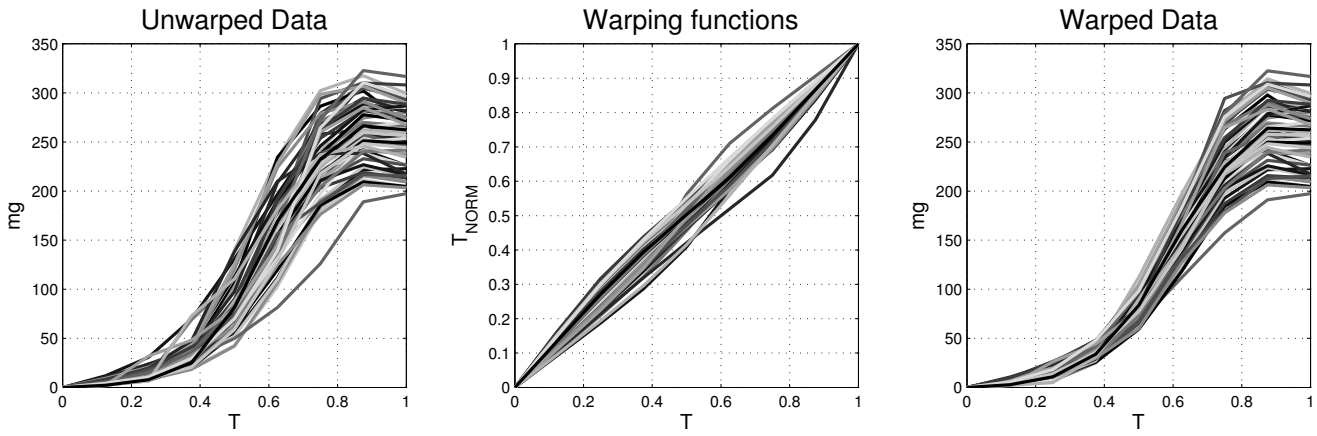

Figure 3.3: Illustration of pairwise warping of 60 beetle growth curves[5]. The left subplot shows the unwarped sample; the middle subplot the warped sample and the right subplot the corresponding warping functions.

the "phase" variation, this AUC time-registration is applicable. An archetypical example of this is the examination of density functions; in that case the phase variation is often considered a nuisance and we want to completely exclude it. Nevertheless, the only question is if it is reasonable to assume that the differentiation observed between the sample instances is due to phase alone or not. For example when comparing microarray data [35], or protein population properties [170]; this is a coherent way of interpreting variation because the observed amplitude differences are due to the internal phase variation of the process observed and at least in theory all external source of amplitudal variation are controlled for. On the contrary for generic phonetic data this approach is almost surely an oversimplification; interestingly though in the case that one has well-defined functional shape-forms (as in the cases of tonal languages) this assumption might hold true.

Similar to the way that pairwise synchronization works by averaging over a sufficient number of 1-to-1 mappings of pairwise synchronization curves g_{ik}, AUC works in the premisses of *quantile averaging* [92]. In particular given a sample of density functions y_i, $i = 1, \ldots, n$, where as before $y_i(t)$, $t \, \epsilon \, [0, T]$ and additionally $y_i(t) \geq 0, \forall \, t \, \epsilon \, [0, T]$ and $\int_0^T y_i(t)dt = 1$, we assume we can work with the smoothed functions. We then calculate the corresponding cumulative distribution functions (CDFs) estimates $\tilde{Y}_i$ and then by simple inversion we get their quantile functions $\tilde{Y}_i^{-1}$. These quantile functions are then treated as synchronization functions and the functional averaging takes place in their domain. In particular the quantile functions can be seen as directly analogous to the inverse warping functions h^{-1} presented in the pairwise synchronization framework above.

Taking this into account the *quantile-synchronized distribution function* $y_{\oplus}$ is estimated as:

$$y_{\oplus}(t) = \phi\{\frac{1}{n}\sum_{k=1}^{n}\tilde{Y}_i^{-1}\}, t'\epsilon[0, T] \tag{3.22}$$

where the density warping map $\phi : Y^{-1} \mapsto y$ maps the synchronized time t' to the natural time t over which the processes are observed. The question that remains to be answered in that of the estimation of the actual time-registration functions h. Taking a step backwards we can define the sample-wide smooth quantile function F_0 as:

$$F_0^{-1}(t) = E\{F^{-1}\} \tag{3.23}$$

where each individual quantile function is modelled as:

$$F^{-1}(t) = F_0^{-1}(t) + \delta(t) \tag{3.24}$$

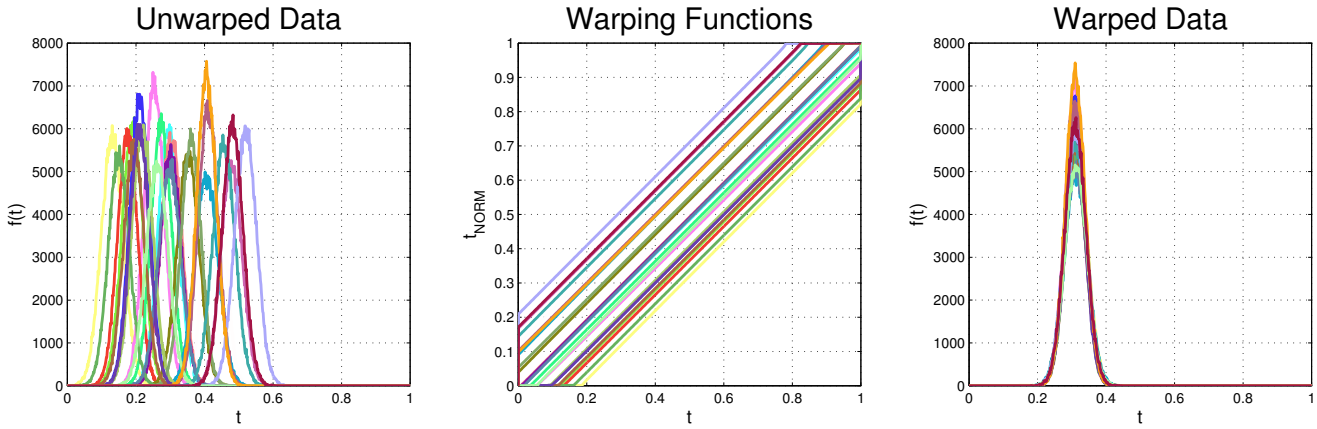

Figure 3.4: Illustration of Area Under the Curve warping. The left subplot shows 25 slightly noisy "Gaussian bumps" that differ mostly due to their initial displacement from the beginning of the recording time. While the signals have different amplitudes, these are treated as independent of their phase variations (and are effectively ignored). The central subplot shows exactly the strongly piece-wise linear nature of warping functions h where they dictate the time that "bump" should occur. The left subplot shows the warped sample curves; all the curves have effectively "collapsed" on the timing followed by F_0^{-1}.

meaning that if random smooth deviation function δ is zero-meaned then we can estimate the sample estimate of $F_0^{-1}(t)$ as:

$$F_0^{-1}(t) = \frac{1}{n}\sum_{k=1}^{n} \tilde{F}_k^{-1}(t). \tag{3.25}$$

and ultimately means that the h^{-1} will be equal to δ plus the inverse of the identity warping function (effectively no-warping) and any difference from absolute normally will be due to F_0 being away from an identity warping function. From there as before simple inversion of h^{-1} will provide the final h estimates. Interestingly AUC can be seen a non-heuristic version of the pairwise warping. In contrast with previously mentioned methodology though, there is no explicit "warping function" formulation, and for that reason it is not based on the assumption of having piecewise linear functions (despite the fact that it will eventually result exactly to that, eg. Fig. 3.4). Nevertheless AUC will suffer from "underwarping" because it is overly restrictive in regards with the sample variational assumptions it assumes; in particular because it assumes phase variation should govern all observed differences and that phase variation is calculated in terms of normalized cumulative density functions, small differences that are unrelated to the actual warping can negate phase effects and underestimate the final phase-variation effects (Fig. 3.5). Despite that though, it presents a conceptually straightforward and computationally cheap way of warping algorithm. As before we examine the performance of the algorithm in a standardized real-dataset (Fig. 3.5); one can easily notice that the growth curves are minimally warped. This expected as the normalized area under the curve corresponding to each is not greatly different from one another. Furthermore initial growth rate differences propagate their influence in later parts of the continuum T and result to higher variability in the warped data Y changes in the start of the curve

3.2.4 Self-modelling warping functions

The idea of self-modelling warping function is rooted in the work for Kneip & Gasser and the concept of structural mean μ [168; 95]. In effect there one assumes that the "structural points", local extrema and inflection points, define a structural mean curve and that the deviations from that curve are what one perceives in a sample's realization. Based on this the generative model of the sample of curves y

[6]Warping was implemented in MATLAB by PZH.

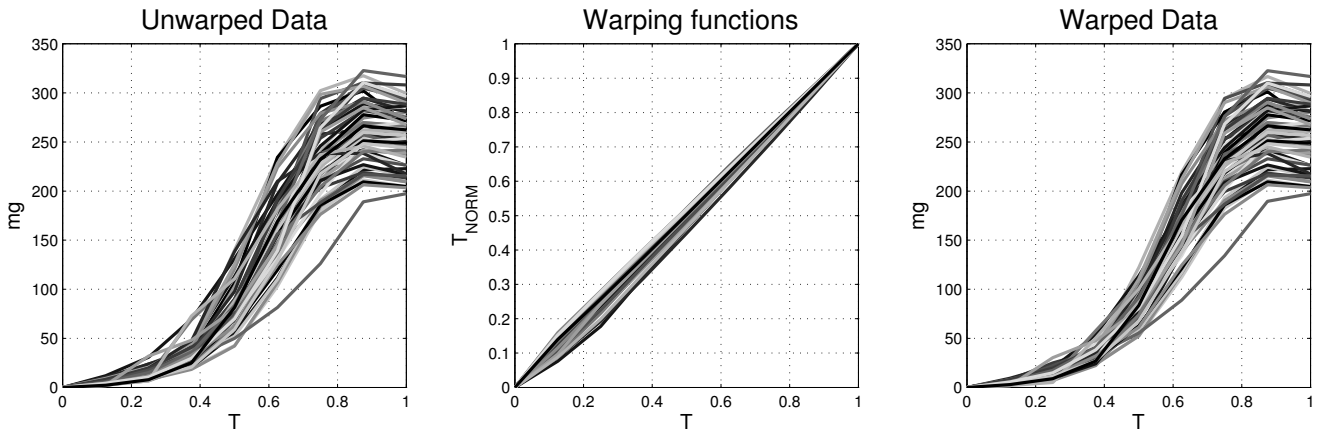

Figure 3.5: Illustration of area under the curve warping of 60 beetle growth curves[6]. The left subplot shows the unwarped sample; the middle subplot the warped sample and the right subplot the corresponding warping functions.

becomes:

$$y_i(t) = a_i \mu\{h_i^{-1}(t)\} + \epsilon_i(t) \tag{3.26}$$

where as before h_i^{-1} is the inverse warping function h_i and ϵ are random errors such that $E\{\epsilon_i(t)\} = 0$. An additional caveat being $a \neq 0$ and $E(a) = 1$. Given this generative model of a sample curves y, the idea of self-modelling warping functions is that one can decompose a warping function h in terms of spectral-like decomposition:

$$h_i(t) = t + \sum_{j=1}^{q} s_{ij}\phi_j(t) \tag{3.27}$$

where as before $i = 1, \ldots, n$, $t \in T$ and this time s_i are the zero-meaned score (or weight vectors) dictating the effect carried from each component $\phi_i(t)$:

$$\phi_i(t) = c_j^T \beta(t). \tag{3.28}$$

However ϕ is not an eigenbasis because $\beta(t)$ is a vector of B-spline basis functions; thus allowing ϕ's to effectively account of variability in different segments of T. This acts as an attempt to "*back-engineer landmark registration*"[98]. This reverse-engineering happens because intuitively one would expect that the landmarks follow roughly at the same points and the variation (phase or amplitudal) between these time points is *irrelevant* to the actual warping. Following that rationale, estimating the spline knot-locations is actually related to estimating the landmark locations. That is why $E(s) = 0$ after all; on average one would assume that the landmark location would be stable, and the deviations from that location would be due to the phase variations. If "nothing happened", ie. $s_{ij} = 0$ for a fixed i, then the warping function associated with that index i should be the identity function, t.

As before, a number of conditions are put forward to ensure identifiability of Eq. 3.26 & 3.27:

- $c_{jk} \geq 0, k = 1, \ldots, p$
- $||c_j|| = 1, j = 1, \ldots, q$
- $C = (c_{jk}) \epsilon R^{q \times p}$ has a block structure such that: $1 \leq K_1 \leq K_2 \leq \cdots \leq p + 1$ where $c_{jk} = 0_{k<K_j, k \geq K_{j=1}}$
- $c_{j1} = c_{jp} = 0 \forall j$

In effect those conditions ensure the sign of the resulting components, its norm, its support and its boundary conditions respectively. With these restrictions in place the actual cost function minimized by the time-registration step is the average integrated square error between the estimated structural mean μ and the warped instance of the function $y_i w_(t)$ defined as:

$$AISE_n = \frac{1}{n}\sum_{i=1}^{n}\int_a^b [y_i\{h_i(t)\} - a_i\mu(t)]^2 h_i'(t)dt \tag{3.29}$$

where the functional structural mean is defined as :

$$\hat{\mu}(t) = \frac{\sum_{i=1}^{n}\hat{a}_i\hat{h}_i' y_i\{\hat{h}_i(t)\}}{\sum_{i=1}^{n}\hat{a}_i^2\hat{h}_i'(t)} \tag{3.30}$$

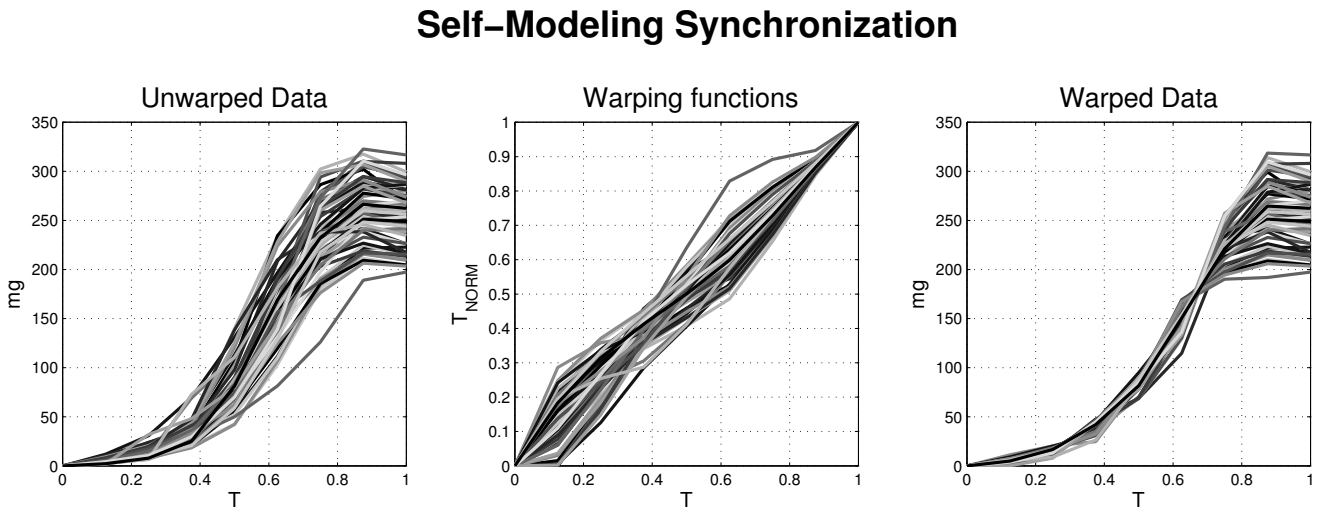

Figure 3.6: Illustration of self-modelling warping of 60 beetle growth curves[7]. The left subplot shows the unwarped sample; the middle subplot the warped sample and the right subplot the corresponding warping functions.

The main critiques regarding the performance of self-modelling warping functions stem from their use of B-splines and their overall idea of a common structural mean. Additionally as AUC synchronization beforehand, this framework follows the concept that "all" variability is due to phase variations and that can prove quite restrictive. In particular looking at Fig. 3.6 where essentially the first 60% of function space is collapsed on the structural mean, all of these issues are exemplified immediately. First one sees that as there is an obvious inflection point in approximately at $.7T$, we have no reason to believe or disbelieve that such an inflection point actually exists in our data; especially given that most the raw data seem not to exhibit such a point. Additionally almost the entire first half of sample Y appears collapsed on a common structural mean, again this might not be wrong but it seems highly unlikely that all growth curves evolve on exactly the same way only to reach an inflection point and then "fan out". One might even argue that this is an artefact of the whole warping process. This happening due to the fact that all amplitude variation is assumed to be "phase related" and the structural mean is modelled by using B-splines: essentially what happens is that up until the inflection point all data are collapsed to the structural mean and then they "fan out" by necessity because of their different values at times $t = T$. Let us stress that this framework critique, and any other warping-related issue can not be assumed a priori to falsify a framework (or justify it for that matter). Each framework has its own different modelling assumptions; given one meets them, any difference is completely justifiable.

[7]Warping was implemented using the `MATLAB` functions provided by D. Gervini : `https://pantherfile.uwm.edu/gervini/www/programs.html`

3.2.5 Square-root velocity functions

Square-root velocity (SRV) based work is a differential geometry based model of elastic curves [209]. Similar to the previous frameworks the curve sample is assumed to be realization of the model:

$$y_i(t) = c_i w(h_i^{-1}(t)) + \epsilon_i \tag{3.31}$$

c being a scaling constant, w being the underlying amplitude variation, ϵ being random noise and h being the corresponding warping function. Importantly this warping function is now viewed not only as a mapping but also as an orientation preserving diffeomorphism on the unit interval where the function y_i is assumed to be observed [176]. Based on that the SRV framework effectively treats each curve y_i as being translation and scaling invariant and defines a continuous mapping $Q : R \to R$ such that:

$$Q \equiv \begin{cases} \frac{\frac{dy_i}{dt}}{\sqrt{\left|\left|\frac{dy_i}{dt}\right|\right|}} & \text{if } \left|\left|\frac{dy_i}{dt}\right|\right| \neq 0 \\ 0 & \text{if otherwise} \end{cases}$$

This is effectively the square-root of the velocity (ie. derivative [8]) function of our original functional y_i.

Going back now to the original space Y of functions y_i and assuming $v_1, v_2 \, \epsilon \, T_y(Y)$, $T_y(Y)$ being the tangent space of Y one can define a corresponding mean in the space. That metric being called the Fisher-Rao Riemannian metric and being defined as the inner product:

$$\langle v_1, v_2 \rangle_y = \frac{1}{4} \int_0^1 \frac{v_1(t)}{dt} \frac{v_2(t)}{dt} \frac{1}{\left|\left|\frac{y(t)}{dt}\right|\right|} dt \tag{3.32}$$

This warping-invariant metric while complicated does have the property that becomes a standard L^2 metric under a SRV framework. As such one can define the distance $d_{FR}(y_1, y_2) = ||q_1 - q_2||$ and based on the isometric property of the warping procedure define $||q_1 - q_2|| = ||(q_1, h) - (q_2, h)||$. Given one can define the Fisher-Rao distance between two warping functions as:

$$d_{FR}(h_1, h_2) = cos^{-1}\left(\int_0^1 \sqrt{\frac{h_1}{dt}} \sqrt{\frac{h_2}{dt}} dt\right) \tag{3.33}$$

as well as the Karcher mean of h to equal:

$$h_{Km} = \underset{h \epsilon H}{\operatorname{argmin}} \sum_{i=1}^{n} d_{FR}(h, h_i)^2 \tag{3.34}$$

and the Karcher mean of the SRV transformed orbits $[\mu]_n$ in the space $S = L^2/H$ to equal:

$$[\mu]_n = \underset{q \epsilon S}{\operatorname{argmin}} \sum_{i=1}^{n} d(q, q_i)^2 \tag{3.35}$$

As Kurtek et al. note the *"Karcher mean $[\mu]_n$ is actually an orbit of functions, rather than a function"*[176]. This means that we are actually looking to find a specific element μ_n of this mean orbit that would ultimately be our "mean" warping function. To do this we perform a two-stage procedure. First we recognize that for each q_i its corresponding h_i should be minimized $||q - (q_i, h)||$ where q is any element of the orbit $[\mu]_n$. Then having found h_i's one can compute the mean h_i, h_{Km} using Eq. 3.34 and then subsequently using the isometry relation mentioned above to find the center of the orbit as $\mu_n = (q, h_{Km}^{-1})$. This μ_n will serve as our template in L^2.

Having found that central orbit computing the actual warping functions h_i are done by solving $h_i = \operatorname{argmin}_{q \epsilon S} ||\mu_n - (q_i, h)||$. We essentially align the individual warping functions to match our template μ_n. Finally given we estimated the warping functions we use them to align our original data y_i.

The important advantage of this approach is its flexibility on multidimensional objects. Applica-

[8] The 2^{nd} derivative informally considered as the *torsion* or *curvature* function.

tions to 3-D objects have already appeared and appear competitive to other custom approaches [174; 177]. Theoretically SRV-framework defines an excellent approach where the metric used is scale-, translation-, rotation- and re-parametrization- invariant. The main critique is that is actually so successful that it "over-warps". By that we mean that it fails to allow for multiple h_i and as such "collapses" the data in a common time-template. As a result if one then tries to analyse these warped data instances the differences among them will be not only due to random amplitude variations; the warped data will still contain phase variations and meaningful statistical inference will be difficult. As a general comment this can work in opposite direction also, if one enforces all the warped data on a single trajectory, this extreme alignment will also enforce spurious phase variations. Interestingly looking at the Fig. 3.7 one immediately spots this tendency to enforce common warping templates in all data. The warping functions plots shows a handful of rather distinct patterns what are assumed to be common among certain instances. This results to "under-synchronized" curves in comparison with all other methodologies examined.

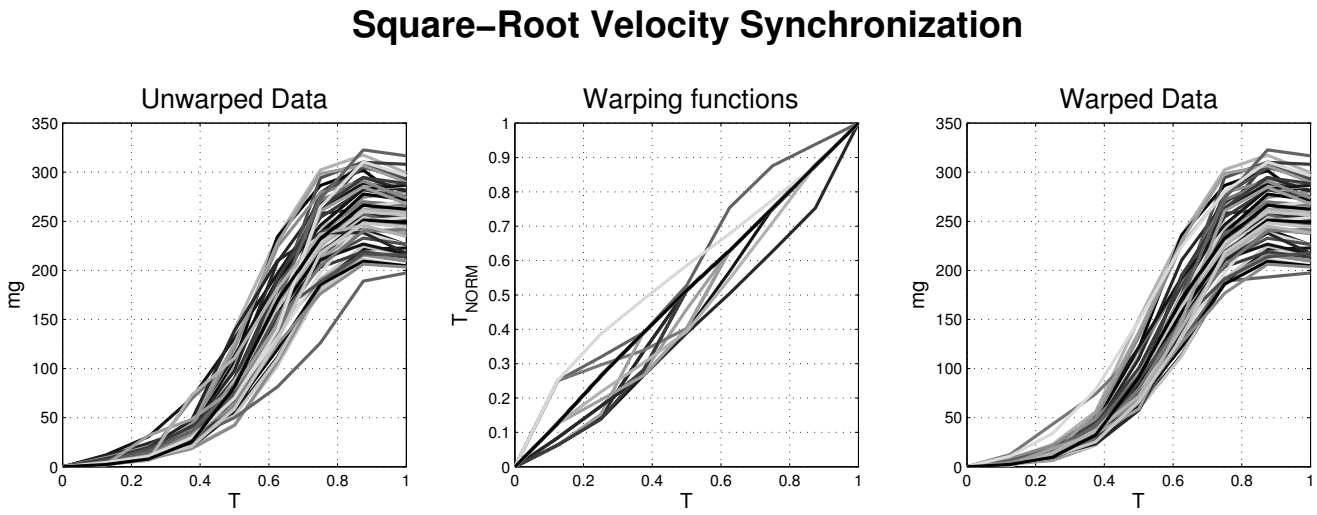

Figure 3.7: Illustration of square-root velocity warping of 60 beetle growth curves[9]. The left subplot shows the unwarped sample; the middle subplot the warped sample and the right subplot the corresponding warping functions.

3.3 Dimension Reduction

As mentioned earlier, functional data can be considered as extending multivariate techniques to a functional domain [64]. Aside the obvious size-constraints, it is possible that we are encoding redundant, unrelated or even misleading information in a high-dimensional dataset. It is therefore to a modeller's benefit to extract features or modes of variations that are informative and less prone to corrupted information. Especially in the cases of two- or three-dimensional data the visualization of a complex dataset is by definition harder than that of a simple dataset and for that reason one would strive to have a more succinct dataset to display; moreover even higher dimensional datasets might have an adequate, in terms of variation explained, representation in two or three dimensions thus allowing their previously impossible visualization. Finally exactly because of the redundancy of information we are expecting, a reduced dimension representation of the dataset could be used as a surrogate dataset for the original high-dimensionality dataset analysed, as not only would it present an obvious "space-complexity" advantage but it could potentially "filter" unstructured information out of original dataset.

Dimension reduction is based on the notion that we can produce a compact low-dimensional encoding of a given high-dimensional dataset. The current work utilizes one main methodology to achieve this task: Functional Principal Components Analysis (FPCA) [121]. FPCA is inherently linear and unsupervised [99]; also it is known to be used in FDA on a number of different application fields. By linear one means that the dataset at hand lies close to a linear subspace and such an accurate approximation of the data can be obtained by using a coordinate system that spans that linear subspace

[9]Warping was implemented using the `R` package `fdasrvf` version 1.4.1.

alone [17]. As such in the case of FPCA the original zero-meaned dataset Y of N observations is assumed to be approximated by the form:

$$\alpha_{\nu,n} = \int_0^T \phi_\nu(t) y_n(t) dt \tag{3.36}$$

where $\phi_\nu(t)$ is the functional principal component of the ν-th order and $\alpha_{\nu,n}$ is the corresponding FPC score where as in Eq. 3.5, $Var(\alpha_{\nu,n}) = \lambda_\nu$. These scores being the projections of the dataset Y into the coordinate systems defined by their respective components or in layman's terms the mixing coefficients dictating "how much" of each components is used to reconstruct the i-th instance of sample Y. In contrast with linear methods, the archetypal non-linear (but still unsupervised) dimension reduction algorithm is that of kernel PCA [284], a number of other popular non-linear algorithms (eg. Locally linear embedding (LLE) [275] and Semi-definite Embedding (SDE) [323]) can also be cast as kernel PCA [99]. In brief in the case of kernel PCA each point Y_i of the original data Y is projected onto a point $\psi(y_i)$ by employing a non-linear transform $\psi(\cdot)$. Then "standard" PCA is performed at that possibly high-dimensional domain; while we will not explore this in any detail, we need to stress that the whole "trick" behind kernel PCA is that one does not need to explicitly compute $\psi(Y_i)$ but rather to compute the $\psi(Y_i)^T\psi(Y_j)$ directly through the use of a valid kernel $K(\cdot,\cdot)$ such that $K = \psi(Y_i)^T\psi(Y_j)$. As mentioned FPCA is unsupervised; by that ones means that there is no prediction variable $\hat{Y}$ (as it would be for instance in the case of linear regression) that it can be used as a "supervisor" indicating the goodness of the solution. On the contrary if for example one had access on some notion of class information in forming the projection, then that information would be beneficial because it would allow informed within-class covariance estimates that would themselves inform the across-class sample covariance. Fisher's Linear Discriminant analysis is a typical example of a supervised linear dimensionality reduction algorithm [17]. As final note, a problem we have re-iterated through the text, is that of the selection of the number of dimensions to retain. This is still an open problem but it is effectively a model selection problem addressed by multiple researchers [141; 208]. The basic solutions stem by reformulating the dimensionality determination task as the optimization of an equivalent information criterion [10]; these Information Criteria materialize even in simple truncation-based heuristics (eg. the *broken stick* model). More formally though the work of Tipping & Bishop in Probabilistic PCA serves as the back-bone framework for this dimension determination tasks in some cases [309].

3.3.1 Functional Principal Component Analysis

Castro et al. [51] work is one of the first to formalize the concept of dimensionality reduction via covariance function eigendecomposition for functional data as it was first presented on Eq. 3.4. This, as with the standard PCA, provides not only a convenient transformation for dimensionality reduction but also as a way to built characterizations of the sample's trajectories around an overall mean trend function [337]. The functional principal components act as the building blocks of our sample. 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:

$$\hat{Y} = M + \sum_{\nu=1}^{m} \alpha_\nu Z_\nu(t), \tag{3.37}$$

is called a m-dimensional model of Y, where M denotes the mean vector of the process ($M = E\{Y\}$), $Z_1, Z_2, ..., Z_m$ are fixed unit length p vectors and $\alpha_1, \alpha_2, ..., \alpha_k$ are scalar variates dependent on Y. Where the mean squared error $S_k^2 = \min E\{||Y - \hat{Y}||\}$ is minimized by the vectors Z_i then $\hat{Y}$ is called the best m-dimensional linear model for Y.

If a process $Y(t)$ is observed at p distinctive times $t_1, t_2, ..., t_p$ it then yields the analogous random vectors $y(t)$, describing the stochastic process $Y = (y(t_1),\ y(t_2),\ ...,\ y(t_p))^T$, fitting perfectly with the theoretical notions of longitudinal data being a variation of repeated measurements. Returning to the

[10] Information Criteria will be discussed in detail in the related *Model selection* section (Sect. 3.4.3).

original notion of a stochastic process $Y(t)$, the m-dimensional linear model for such process is:

$$y_j(t) = \mu(t) + \sum_{\nu=1}^{m} \alpha_{\nu,j}\phi_\nu(t), \tag{3.38}$$

where α_ν are once more the uncorrelated random variables with zero mean and refer to the ν-th principal component score of the j-th subject and ϕ_ν are linear independent basis-functions, of the random trajectories Y_j. This expansion (Eq. 3.38) is referred to as the Karhunen-Loève or FPC expansion of the stochastic process Y [121] where now $\phi_\nu(t)$ refers to continuous pairwise orthogonal real-valued functions in $L^2[0,T]$, as before $\mu(t) = E\{y(t)\}$, $t\epsilon[0,T]$. Similarly the mean squared error is reinstated here as the integrated square error $||y_j(t) - \hat{y}_i(t)||^2 = \int [y(t) - \hat{y}(t)]^2 ds$, with the choice of optimal ϕ's encoding the best m-dimensional model for Y. Empirically finding these unit norm ϕ require first the definition of the sample covariance function $\hat{C}_Y(s,t)$ in a way similar to Eq. 3.3:

$$\hat{C}_Y(s,t) = \frac{1}{N}\sum_{i=1}^{N}(Y_i(s) - \hat{\mu}(s))(Y_i(t) - \hat{\mu}(t)) \tag{3.39}$$

where $\hat{\mu}(t) = \frac{1}{N}\sum_{i=1}^{N} Y_i(t)$ [11] and then the subsequent eigendecomposition of $\hat{C}_Y(s,t)$ for the zero-meaned sample Y as:

$$\hat{C}_Y(s,t) = \sum_{\nu=1}^{N} \hat{\lambda}_\nu \hat{\phi}_\nu(s)\hat{\phi}_\nu(t) \tag{3.40}$$

or equivalently in matrix notation $\hat{C}_Y = \Phi\Lambda\Phi^T$, the later being also known as the *principal axis theorem* [156]. Ultimately, exactly because of the optimality of the FPC's in terms of variance explained, these modes of variations will be the ones explaining the maximum amount of variance in the original sample.

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_i tend to zero rapidly so that the variability is predominantly of large scale [267]. A further important qualitative view of the FPC's is as representing a rotation of the original dataset in order to diagonalize the covariance matrix of the data; thus making the new coordinates of the dataset uncorrelated [28]. This functionality of PCA even allowing it to be reformulated within a phylogenetic framework [266].

In physical terms, smoothness of data is critical so that the discrete sample data can be considered functional [253]. For example as seen in the work on Chen & Müller [55] in the case of two-dimensional functional data, the discretisation and the subsequent interpolation can have significant implications in one's results (the authors advocating a *two-way* FPCA to counter these issues).

As noted in the previous section a number of smoothing techniques have been proposed over the years concerning FPCA; 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 presmoothing, considered to be the optimal choice in the case of local weighting [83], is the one applied in all the cases of this work due to its simplicity and computational ease, yielding smooth sample curves. Finally we draw upon the fact that we do work with a discretised version of a functional sample Y and that a core requirement for FPCA to be applied directly is that the sample Y_i has the same number of equi-spaced readings (see [337] for a case where one can apply FPCA in sparse and irregularly sampled data by employing a conditional expectation procedure). This requirement being easily fulfilled by the smoothing and concurrent interpolation of the sample.

Interestingly a number of regularized or smoothed functional PCA approaches have appeared over the years. In such cases smoothness is imposed in multiple ways. Either by penalizing the roughness of the candidates ϕ based by means of their integrated squared second derivative [254] or by projecting the original sample down to a lower dimensional domain where the data appear smoother, probably by taking advantage of a periodic basis like Fourier polynomials and carry out standard FPCA in that domain. The basic qualitative difference between the two approaches being that in the first case smoothing occurs directly on the FPCA step, while in the later we smooth the data directly. In our primary work with F_0

[11] $\hat{C}_Y(s,t)$ is biased by $\frac{N-1}{N}$ but this is considered asymptotically negligible in the current context.

we smooth and interpolate the data beforehand but we do not impose secondary smoothing techniques as the ones mentioned above; an initial smoothing is adequate and additional smoothing will only draw attention away from our true sample dynamics. On the contrary when working with spectrograms (chapter 6) exactly because we do not smooth the data originally, we do smooth the spectrogram's readings after initial interpolation (section 6.2.1).

3.4 Modelling Functional Principal Component scores

As mentioned in this previous section's introduction the current work does not utilize a functional linear regression approach with a functional response directly as shown in Eq. 3.6. Instead it employs the previously presented dimension reduction methodology to identify Φ and A^{ϕ} as shown in Eq. 3.36. Knowing these we conduct inference related to using A^{ϕ} scores as surrogate data for our sample (as Φ is fixed) and functional regression is formulated as in Eq. 3.38. At this point we recognize that given the structure of our phonetic dataset, simple linear models fail to encapsulate its complexity. We will therefore utilize linear mixed effects models as the appropriate technique to conduct inference on A^{ϕ}.

3.4.1 Mixed Effects Models

Linear mixed effects (LME) models serve as an extension to the standard linear regression model. As Pinheiro and Bates [238] present: "(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". The term "mixed" derives from the fact that LME models combine fixed and random effects. The fixed effects are attributed to an otherwise unknown deterministic component while the group effects are treated as random variables rather than fixed parameters; occasionally these random effects can be treated as *nuisance* [190; 84]. More formally, and using the classical linear mixed effect model notation [325], combined with the distributions notation presented by Faraway [77], a standard univariate fixed effect model with normal errors:

$$a = X\beta + \epsilon \qquad \text{or} \qquad a \sim \mathcal{N}(X\beta, \sigma^2 I) \tag{3.41}$$

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

$$a = X\beta + Z\gamma + \epsilon \quad \text{or} \quad a|\gamma \sim \mathcal{N}(X\beta + Z\gamma, \sigma^2 I) \tag{3.42}$$

where in the presented case a is the vector of length $n \times 1$ readings of the dependent variable a, X is the $n \times k$ model matrix, the vector ϵ of length n encapsulates the random variables representing measurement errors, and β is a vector of length k that contains the linear (fixed) regression coefficients, k being the number of those coefficients. The extension of this model now to account for mixed effects is such that Z is a model matrix $n \times l$ [12] associated with a vector γ of random effects.

In this work, we will assume that the random effects are by definition random variables themselves [15]. As such, the γ vector will follow a multivariate Gaussian distribution $\gamma \sim \mathcal{N}(0, \Sigma_\gamma)$, where Σ_γ represents the covariance matrix of the elements in vector γ and are therefore assumed to vary at random around their mean 0. In a similar manner, the error residual vector ϵ will follow a multivariate Gaussian distribution where $\epsilon \sim \mathcal{N}(0, \Sigma_\epsilon)$ and Σ_ϵ is the covariance matrix for residuals in vector ϵ. In many applications Σ_ϵ is assumed to be diagonal. Under the formulation the variance of a can subsequently be expressed as:

$$Var(a) = Var(Z_\gamma) + Var(\epsilon) = Z\Sigma_\gamma Z^T + \Sigma_\epsilon \tag{3.43}$$

resulting in the unconditional distribution :

$$a \sim \mathcal{N}(X\beta, Z\Sigma_\gamma Z^T + \Sigma_\epsilon) \tag{3.44}$$

[12] $l < n$ in usual cases

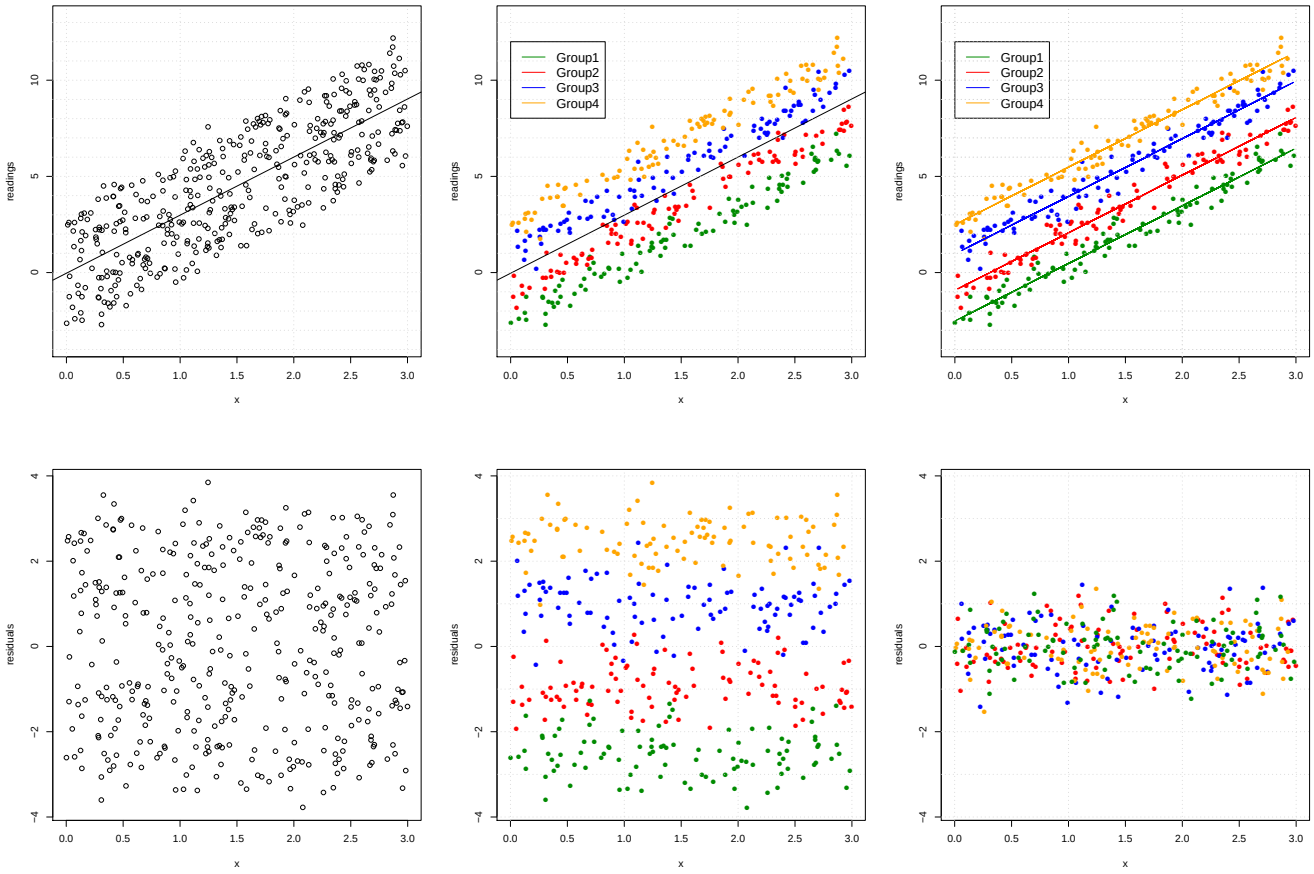

Figure 3.8: Illustration of why LMMs can be helpful: Given the original α data (upper left graph) a standard marginal $\alpha \sim \mathcal{N}(X\beta, \sigma^2 I)$ model would yield seemingly unstructured residuals (lower left graph), however a careful examination of some sample trait Z can reveal a highly structured pattern both in the original data (upper center graph) and the corresponding residuals of the marginal model (lower center graph). Using a model conditional on α, $\alpha|\gamma \sim \mathcal{N}(X\beta + Z\gamma, \sigma^2 I)$ (upper right graph) allows for "truly" unstructured estimates (lower right graph). Importantly "shrinkage" of β might also occur.

Importantly if we are interested in the joint distribution of a and γ that is given as:

$$\begin{bmatrix} a \\ \gamma \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} X\beta \\ 0 \end{bmatrix}, \begin{bmatrix} Var(a) & Z\Sigma_\gamma^T \\ \Sigma_\gamma Z^T & \Sigma_\gamma \end{bmatrix} \right) \tag{3.45}$$

An obvious question is how to differ a fixed from a random effect. This is still an open question [238; 325; 296] but it is significant to identify that this a *design* question. Therefore the important indicator for this design question's answer is whether or not we assume our sample's structure is in itself a random realization that occurred during sampling or it is a fixed structure that arose based on our experimental questions. For example if we wish to characterize the effect that stress has on speaker utterances given some arbitrary qualitative stress-scale, stress is a fixed effect; it is not a question if certain stress states occurred by accident; we wanted them to be recorded. On the contrary, we will inevitably use a number of volunteer speakers who are part of the greater speaker population. While we might try to stratify our sample in a way (eg. by having an equal number of female and male speakers) our sample speakers are not the interest of the study itself, we want to gain insights about the speaker population; if anything, an individual's stress responses are not known beforehand. This also begs another design question regarding the type of random effects used; while in their *vanilla* variant random effects are modelled as *random intercepts* (Fig. 3.8), the inclusion of *random slopes* can also be beneficial. If random slopes are indeed present and are omitted, this may lead to anti-conservative evaluation of fixed-effects [283] and to non-generalizable results [18]. Computationally there are also some less well-documented considerations. Given that we "generally" assume Gaussianity in regard with

the random effects structures and that of identifiability purposes [190] we assume that $\sum_i \hat{\gamma}_i = 0$, a minimum number of random effects level is necessary to have meaningful inference. While no definite reference exist on this matter the norm is "more than 5 or 6 at a minimum" [34].

The next design question is concerned with the structure of the random effects. When a factor (random or fixed) is exclusively measured within a given realization of another random factor (eg. the speakers within a predetermined region (Fig. 3.9), then it is considered to be nested within the levels of the higher levels factor; this is because random effects models are commonly associated with hierarchical models as one might assume there is a certain hierarchy between factors. On the contrary if a factor can be measured within multiple levels of another factor (eg. a sentence being read by multiple users (Fig. 3.10), then it is considered to be crossed with the levels of the other factor [15; 117]. Here we also add a caveat regarding the existence of a balanced or imbalanced design; regardless if we have an equal number of realizations for each random effect level. While important to have some concept of balance, it is not crucial and with the exception of pathological cases it is usually unimportant.

Ultimately the big question about the LME models is the level of inference which we choose to focus on. Is one interested in within-group or across-group variance? ie. Is one interested in conditional or marginal inference? This is once more an open question [190] where the answer is in essence embedded in the question itself: do we aim to produce subject specific (conditional) or population average (marginal) estimates? A marginal model $E(\alpha) = X\beta$ is clearly the distribution of the observed data but it is unable to control for the unobserved random effects; on the other hand the conditional model $E(\alpha) = X\beta + Z\gamma$ does provide that, but in the expense of possible "shrinkage". This means that the inference on the fixed effects might be different from the one made by the marginal model. In short this happens *algebraically* because in comparison with a standard linear / fixed-effects only model, where $\hat{\beta} = (X^T\Sigma_\epsilon^{-1}X)^{-1}X^T\Sigma_\epsilon^{-1}a$, in the case of mixed effects model $\hat{\beta} = (X^T(Z\Sigma_\gamma Z^T + \Sigma_\epsilon)^{-1}X)^{-1}X^T(Z\Sigma_\gamma Z^T + \Sigma_\epsilon)^{-1}a$[13]. As a consequence the resulting diagonal of the mixed-effect error variance will be greater and the resulting $\hat{\beta}$'s will be smaller. In effect we borrow strength from the other sample points to model the variance in a single point therefore the β are smaller (closer to 0), annealing their influence towards a population estimate.

A final note of this introductory section is that while the current work assumes all realizations of γ to be of Gaussian nature, non-Gaussian distributions can be used [189; 240; 274; 136] to model the realization of the random effects offering greater robustness at the cost of a more parametrized model.

[13] This being eventually a GLS estimator.

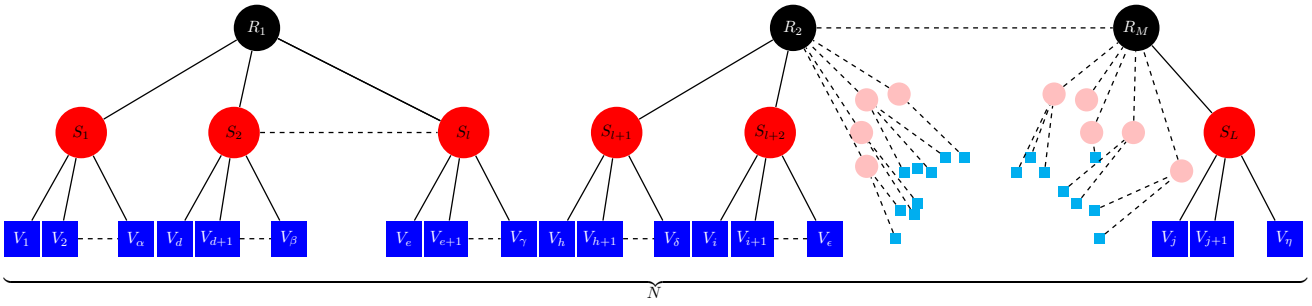

Figure 3.9: Illustration of a simple nested design: Given a linguistic field study of a language, across different regions R (black circles), different speakers S (red circles) are uttering a number of voice recordings V. Clearly a voice recording V_i is only due to a single speaker S_j and a specific speaker (assuming no travelling speakers) is classified as a member of specific population residing at region R_k. Because of the huge combination of lexical instances in a language, a record corpus V has to be treated as a sample rather than an exhaustive list (population). Similarly within a region, one can not realistically hope to record all speakers; the speakers S available are just a random sample from the greater speaker population.

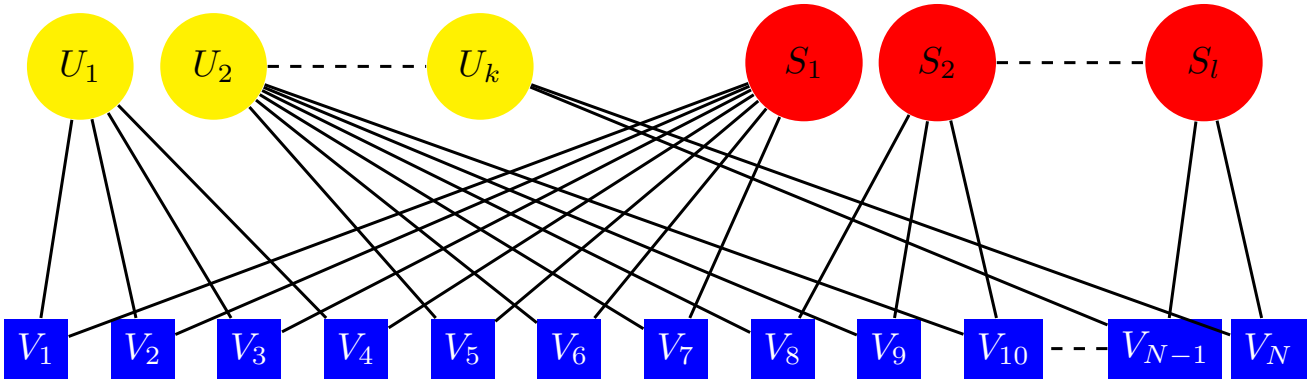

Figure 3.10: Illustration of a simple crossed design: Given a linguistic study where we use a determined set of sample sentence utterances U and a sample of speakers S, because all speakers ultimately say all utterances, the voice recordings V are subjected to random effects both due to the Speaker and Sentence utterance effects.

3.4.2 Model Estimation

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 [238; 300]. Nevertheless, a number of issues in mixed effects models 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 [288]. Moreover the MLEs of β are downwards biased [300] because given that the number of samples in the random vector might be quite small, as in the case of speakers in a linguistic study, 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. In brief, 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.

In particular, when trying to estimate the mixed model via the model's likelihood, we observe that usual maximum likelihood (ML) estimation underestimates the mixed model's variance components [234]; this does not refer to the *shrinkage* effect mentioned previously. Based on Eqs. 3.42 & 3.45 the log-likelihood function $L(a|\gamma)$ can be seen as:

$$L(a|\gamma) = -\frac{N}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_\epsilon| - \frac{1}{2}\epsilon^T\Sigma_\epsilon^{-1}\epsilon, \tag{3.46}$$

$$\epsilon = a - X\beta - Z\gamma \tag{3.47}$$

Additionally exactly because we assume that $\gamma \sim \mathcal{N}(0, \Sigma_\gamma)$, $L(\gamma)$ is of the form:

$$L(\gamma) = -\frac{M}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_\gamma| - \frac{1}{2}\gamma^T\Sigma_\gamma^{-1}\gamma \tag{3.48}$$

and leads to the joint pseudo-log-likelihood estimate (ignoring fixed terms) :

$$L(a,\gamma) = -\frac{1}{2}\gamma^T\Sigma_\gamma^{-1}\gamma - \frac{1}{2}(a - X\beta - Z\gamma)^T\Sigma_\epsilon^{-1}(a - X\beta - Z\gamma). \tag{3.49}$$

Differentiating with respect to γ and β and setting to 0, gives the respective MLE estimates. Then the

mixed model equations of the LMM regression can be written as:

$$\begin{bmatrix} X^T\Sigma_\epsilon^{-1}X & X^T\Sigma_\epsilon^{-1}Z \\ Z^T\Sigma_\epsilon^{-1}X & Z^T\Sigma_\epsilon^{-1}Z + \Sigma_\gamma^{-1} \end{bmatrix} \begin{bmatrix} \beta \\ \gamma \end{bmatrix} = \begin{bmatrix} X^T\Sigma_\epsilon^{-1}a \\ Z^T\Sigma_\epsilon^{-1}a \end{bmatrix} \tag{3.50}$$

From that point onwards if Σ_ϵ and Σ_γ were known we would get estimates for β and γ immediately. The problem is that this is not usually the case, so we have to estimate them. One essentially uses an iterative procedure. First setting β and γ on some arbitrary values ones estimates Σ_ϵ and Σ_γ and then based on the newly estimated Σ_ϵ and Σ_γ β and γ; this iterative procedure is then repeated until convergence. We mention though that these estimates will be biased. Here this bias comes specifically from the fact that we lose degrees of freedom, when we treat our "intermediate" mean estimates of the procedure (sample estimates effectively) as the true population estimates and we subsequently use those to estimate the population variance. Qualitative this is the same phenomenon that is observed when one uses $\frac{1}{N}(y-\bar{y})$ instead of $\frac{1}{N-1}(y-\bar{y})$ to calculate a sample variance. This ambiguity surrounding the LME's degrees of freedom does not only affect estimation; it also significantly complicates the application of Log-likelihood-Ratio Tests [62] making the standard notion of p-value associated with β not straightforward to obtain (eg. "how many" parameters does one account a single random effect to convey? Or, do two random effects with significantly different number of levels encode the same amount of information in terms of degrees of freedom lost?)

To bypass these issues at least partially, we use REML. The restricted Maximum likelihood utilizes a matrix K such that the expression $K^Ta = K^TX\beta + K^TZ\gamma$ no longer contains β, ie. as mentioned above $K^TX\beta = 0 \Leftrightarrow K^TX = 0$. Realistically the matrix K is equivalent to $K = TM$ where:

$$M = I - X(X^TX)^{-1}X \tag{3.51}$$

$$\text{and } T = [I_{q_1}0_{q_2}], \quad q_1 = rank(M), \quad q_2 = N - q_1. \tag{3.52}$$

This together with $K^Ta \sim \mathcal{N}(0, K^T(Z\Sigma_\gamma Z^T + \Sigma_\epsilon)K)$ resulting in the final estimate of the restricted log-likelihood:

$$L_{REML}(\theta) = -\frac{1}{2}[(N-r)\log(2\pi) + \log(|\Psi|) + \overrightarrow{\Omega}^T\Psi^{-1}\overrightarrow{\Omega}] \tag{3.53}$$

where $\Omega = K^Ta$ and $\Psi = K^T(Z\Sigma_\gamma Z^T + \Sigma_\epsilon)K$.

As in the case of MLE (Eq. 3.49) we optimize Eq. 3.53 by employing an iterative procedure based on penalized least squares [21]. The details of the procedure are explained in section 5.3.8 for the case of multivariate A instead of a univariate a but they are directly applicable.

3.4.3 Model Selection

"*All models are wrong but some are useful*" by George E. P. Box is probably one of the most worn statistical quotes. It does highlight though the obvious intuition that a (statistical) model is a simplification of reality that allows the modeller to infer the dynamics behind the model's components. If one is therefore presented with multiple models it is essential he can estimate the performance of different models "*in order to choose the best one*" [129]; this procedure being commonly referred as *model selection*.

Given we have a sample y from an unknown parametric model $m(x;\theta)$, and estimates from an associated predictive model $\hat{y} = \hat{m}(x;\theta)$, an obvious test for the goodness of our estimation is how well our estimate $\hat{y}$ can predict y in terms of mean squared error [62]. For example, assuming a squared loss function given as: $C = (\hat{y}-y)^2$, the expected C equals:

$$E[C] = E[(\hat{y}-y)^2] \tag{3.54}$$

$$= E[(\hat{y} - y - E[\hat{y}] + E[\hat{y}])^2] \tag{3.55}$$

$$= E[\hat{y} - E[\hat{y}]]^2 + E[y - E[\hat{y}]]^2 + 2E[(E[\hat{y}] - y)(\hat{y} - E[\hat{y}])] \tag{3.56}$$

where the final term equates to zero and we get:

$$= E[\hat{y} - E[\hat{y}]]^2 + E[y - E[\hat{y}]]^2 \tag{3.57}$$

$$= var(\hat{y}) + bias^2(\hat{y}) \tag{3.58}$$

We see that the more we decrease the bias of our predictor, the more we increase its variance; the more we overfit our data the closer we get to our actual estimation points. This results in a model m that has poor predictive power for unseen data and poor explanatory power for the population dynamics from which the sample was taken from. This maximum likelihood approach succeeded in giving us the model that maximizes the likelihood function $p(D|m)$, where D are our observed dataset and m a model from our model space M. Unfortunately direct maximization of the likelihood function $p(D|m)$ results in choosing increasingly larger models. To alleviate this limitation of direct maximum likelihood estimation we are using two different approaches: one data-driven, and one based on analytical results. The data-driven approach is based on cross-validation and resampling principals while the analytical approach works by making meaningful approximations between our estimated distribution and the "true" distribution of the data.

As mentioned in section 3.1, *cross-validation* is based on the idea that you exclude a portion of your data as a validation set [28]. In the case of a k-fold cross-validation one randomly partitions his dataset in k (usually of equal size) partitions, uses $k-1$ available partitions to train his model and then the model's fitting is evaluated using the k-th partition excluded. We thus use a $(\frac{k-1}{k})\%$ of our available data each time. This procedure is executed k times, and at the end of it (usually) the performance scores from the k runs are averaged in order to get the final estimate for the model's performance. We then proceed in comparing the different model performances and choose the best one. A similar approach based on resampling the data is *jackknifing*. During jackknifing instead of using a validation and a test set we generate k sub-samples y_{jack} by resampling our original sample y, evaluate our model's performance in that sub-sample y_{jack} and average over the k runs to give the final performance estimate.

A second approach data-driven approach would be to use *bootstrapping* [129]. Focusing on the parametric bootstrap, we first fit the parametric model for which we want to assess the performance of our data. We then resample from that model in order to produce "bootstrapped samples" y_{boot} of size N, N being our original sample size. Repeating this procedure k times we re-fit each time our model using the new y_{boot} produced. Similarly to cross-validation we then average the performance scores from the k runs in order to get the final estimate for the model's performance [14].

Without looking into theoretical problems stemming from resampling, a common problem encountered by all resampling-based approaches is that of computational costs. Both in terms of memory and CPU time, resampling and/or refitting a large number of models is an expensive procedure. Even a simple ordinary least squares model requires usually the Cholesky decomposition of the X^TX matrix or the QR decomposition of the design matrix X; these procedures being of approximate asymptotic order $\frac{1}{3}N^2$ and $\frac{4}{3}N^2$ respectively [101] (the obvious time trade-off between the two being at the computational time of the matrix-matrix multiplication X^TX). Repeating this millions of times can become extremely time-consuming. Finally stating almost the obvious, this inferential procedure is based on random sampling, these results are not strictly deterministic, another sample gives slightly different values.

An optimal solution could be to find a procedure that you can use only *once* and access the "goodness of the model". This is achieved by a series of approximations; the intuition for these approximations comes from two directions. First we want an approximation that tells us how good we do based on "population estimates"; this is why we used resampling after all. Second we recognize that a problem with using a maximum likelihood approach stems from failing to penalize unnecessarily complex models; a problem relating directly to the parsimony principal of Occam's razor [62]; "*it is vain to do with more what can be done with fewer*". Occam's razor is the driving force behind a number of information criteria (IC). The current study relies almost exclusively to Akaike's Information Criterion (AIC) [5], which is established as the "standard" among ICs. A second almost equally popular IC is the Bayesian or Schwarz Information Criterion (BIC or SIC respectively). These information-based model selection criteria aiming to essentially balance model complexity and predictive power, providing a way to rationally penalize each parameter added to the model with the respect to the "explanatory power" it provides.

[14] Resampling can also be formulated in a Bayesian context; there sampling is done from the posterior distribution of the parameters estimated. For each parameter a higher posterior density (HPD) interval over some value $q\%$ can be created from the empirical cumulative distribution function of the sample as the shortest interval for which the difference in the empirical cumulative distribution function values of the endpoints equates with q; ie. they *"minimize the volume among q-credible regions and, therefore, can be envisioned as optimal solutions in a decision setting"*[271].

The theoretical machinery behind AIC is *Kullback-Leibler (K-L) divergence*[15]. K-L divergence is a distance between an unknown distribution $t(x)$ and an approximate distribution $q(x)$ in terms of additional amount of information one needs to use to specify x due to the fact of using $q(x)$ instead of $t(x)$ [28]. Thus K-L divergence is given by:

$$\begin{aligned} KL(t||q) &= -\int t(x)\log q(x)dx - (-\int t(x)\log t(x)dx) \\ &= -\int t(x)\log\frac{q(x)}{t(x)}dx \end{aligned} \tag{3.59}$$

It needs to be stressed that K-L divergence concept is akin to a likelihood ratio statistic. Exactly because AIC reflects "additional" information the smaller it is the better [62]. Clearly for the application of AIC the main issue is that one does not know the $t(x)$ beforehand. Akaike's solution was to estimate it; AIC score is an asymptotically unbiased estimate of the cross-entropy risk. In other words as the sample size $n \to \infty$, the model with the minimum AIC score will possess the smallest Kullback-Leibler divergence. Interestingly despite its rather involved theoretical justification, AIC is computed as :

$$AIC = -2L(\theta) + 2p \tag{3.60}$$

where p is the number of parameters in the model and the $L(\theta)$ is the likelihood of the model used with respect to θ. A general comment is that when the number of parameters in a model, is not significantly smaller than the number of available samples ($40 \geq \frac{n}{p}$), then using a version of AIC correct for smaller samples is desirable [44]: AICc. AICc is defined as:

$$AICc = -2L(\theta) + 2p + \frac{2p(p+1)}{n-p-1} \tag{3.61}$$

where evidently as $\frac{n}{p} \to \infty$ one gets back the original AIC (n being the number of available samples).

AIC (and AICc) approach takes a full frequenist approach regarding model selection and is based on asymptotic behaviour properties of the estimator used (K-L divergence); a Bayesian approach was proposed by Schwarz [287] leading to the formulation of BIC. In brief one assumes that all candidate models are equali-probably (essentially having an un-informative prior) and that the "true" model is among the candidate models; then by finding the model that gives the higher posterior probability[17]:

$$p(m_i|D) = \frac{p(D|m_i)p(m_i)}{p(D)} \tag{3.62}$$

$$\text{where: } p(D) = \sum_{i=1}^{M} p(D|m_i)p(m_i) \text{ and } p(D|m_i) = \int p(D|\theta_i, m_i)p(\theta_i|m_i)d\theta_i \tag{3.63}$$

one finds the "most-probable" model that generated the data (from the subset of examined models of course). Two caveats immediately arise: 1. we are fairly certain that some models are more probable than others and 2. we have no reason to believe that the "true" model is among our candidate models. Nevertheless in a Bayesian approach there is no need to explicitly penalize model complexity as that is incorporated by the integral over the posterior parameter distribution. Given these initial assumptions BIC is calculated as:

$$BIC = -2L(\theta) + p\log(n) \tag{3.64}$$

An important note is that while BIC selection is consistent AIC is not [62]; where one by consistency means that the probability of the "true" model being selected tends to 1 as $n \to \infty$. As Davison [62] shows if a model m' is close to m^{true}, where the respective number of parameters in each model is p and q, if $p - q$ is small (< 10) then it not improbable that one chooses m' instead of m^{true}.

Returning to our original LME model case we already identified that "generic" model estimation

[15]The term *relative entropy* is sometimes used interchangeably with KL-divergence.

should occur within a REML framework. Nevertheless exactly because the matrix K is reformulating the response vector a into $K^T a$ in a model specific way, the residuals associated with two LME models with a different number of fixed effects will not be directly comparable. In practice that can be seen as K changing the "residual" term of the likelihood ($\Omega^T \Psi^{-1} \Omega$). This being even more obvious if we see the alternative formulation of AIC as $\frac{n}{2} \log(RSS) + p$ [123]; $RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$.

Both ICs can be inconclusive; as a general rule of thumb when the absolute difference between two models is less or equal to two (2), there is no obvious reason to select one model over another. In relation to that Burnham and Anderson note: *"A substantial advantage in using information-theoretic criteria is that they are valid for non-nested models. Of course, traditional likelihood ratio tests are defined only for nested models, and this represents another substantial limitation in the use of hypothesis testing in model selection."*[44].

In practical terms constructing and finding the *best* model relating to a process of interest in somewhat heuristic. Two main methodologies are usually employed; forwards and backwards model selection. In the case of forwards model selection one starts with the smallest (or least complex) relevant model for the relationship between independent and dependent variables and through consecutive comparison among the candidate variables the variable that most substantially "betters" the model's fit is added. The process being iterated until *convergence*; ie. no variable can be added that improves the model (based on some information criterion or LR test). Backwards model selection is effectively the opposite. One defines the largest (most complex) relevant model for the dataset at hand and then removes the "least helpful" variable based on the definition of helpfulness. An important point to be made here is that we need to always remember that the interoperability of the model is of interest; forgetting that and employing a stepwise selection technique as forwards or backwards selection process will ultimately result in *data dredging*; essentially discovering causally irrelevant associations between otherwise disassociated physical terms.

It is worth mentioning that a completely different approach to model selection is to conceptually merge the estimation and the selection procedure as this is exhibited in the case of SCAD or LASSO [146]. In these cases one effectively builds in the complexity penalization procedure within the model estimation step. Within a linear mixed effect modelling framework Lan [180] presents the penalization of β while Bondell et al. [36] present an even more generalized approach where β and γ are penalized in an iterative manner. Finally it is notable that following the increasing popularity of ensemble approach in statistical learning [41; 88], ensemble approaches for variable selection have also started to appear [342; 332].

An obvious matter that is often either ignored or left unattended is *data quality*. Bad data will give bad results irrespective of how successful the results might look in explaining the original research question. Missing or corrupted data are an aspect of an analyst's research life and that should never be forgotten. Nevertheless one might advocate, model under-fitting is more damaging than model over-fitting when it comes to model-based inference given a set of fixed quality data [44]. Ultimately the usage of AIC, BIC, MDL or any other model selection method guarantees that under certain assumptions the best candidate model will be chosen from a set of candidate models. If a "good" model is not part of the set of candidate models, it will not be discovered by model selection algorithms.

3.5 Stochastic Processes and Phylogenetics

A major question when examining a dynamical system is *what* leads to it; *how* this system came to be. And while language is definitely a dynamical system, the first questions of this kind were most probably Biology-related. Given that domestication of dogs started prior to 35000 years ago and with certainty the concept of different types of dogs had emerged (as depicted in Ancient Egyptian tombs) 5000 years ago; people must have noticed certain trait regularities were *somehow correlated* [70; 233] and that there was some concept of "ancestry". This giving raise to questions about phyla, ($\phi\grave{\upsilon}\lambda\alpha$, "races") and how did these phyla were generated; namely *Phylogenetics*.

Phylogenetics can be defined as the science of the evolutionary relationships among species [232]; the primary forces behind Evolution being *natural selection*, *random genetic drift* and *founder effects*. This definition though only hints towards their true nature: Phylogenetics are applied stochastic processes in Biology where researchers have employed the concept of random walks (and especially Brownian

Motion (BM) from the late 70's [181; 182]); this concept being finally established by the now seminal work of Felsenstein [80; 81]. As such the major questions behind Phylogenetics: I. reconstructing ancestral states, II. quantifying adaptation, III. dating divergence between species and IV. estimating rates of evolution are simply questions about random walks, their properties and their past states. Mathematically "evolution is viewed as proceeding in two steps: (1) selection, determined by the fitness (i.e., survivability and fecundity) which the trait confers on each individual relative to others, and (2) inheritance, controlled by the mating patterns and genetics of the survivors (breeding adults)" [102]. These two steps convey the basic idea behind evolution and Phylogenetics: Given a population of breeding adults with certain traits, an inherently noisy replication process allows them to propagate those traits based on the fitness potential of those traits. This means that in mathematical terms, for a given trait x we have a Markovian process because :

$$P(X_{child} = x_{child}|X_{parent} = x_{parent}, X_{grandparent} = x_{grandparent}, \ldots, X_0 = x_0) = P(X_{child} = x_{child}|X_{parent} = x_{parent}) \tag{3.65}$$

and if we make the least assumptions possible and assume that the replication process is subjected to Gaussian noise with some intensity of random fluctuations σ_n:

$$X_{child} = X_{parent} + N(0, \sigma_n) \tag{3.66}$$

or by subtracting X_{parent} from both sides, for a given time t:

$$dX(t) = N(0, \sigma_n). \tag{3.67}$$

This being a simplification of the standard (one-dimensional) Brownian motion where a collection of random variables $X(t)$, $t \geq 0$ satisfies [184]:

- $X(t_0) = 0$
- For $t_1 \leq t_2$, $X(t_2) - X(t_1) \sim \mathcal{N}(0, t_2 - t_1)$ and
- $t \to B(t)$ is continuous.

Then if one builds the concept of optimality as X_{opt} when some constant α measures the strength of selection towards that optimum, the final model for evolution becomes:

$$dX(t) = N(0, \sigma_n) + \alpha(X_{opt} - X(t)) \tag{3.68}$$

and we have just stated a simple Ornstein-Uhlenbeck model [125] or the "Hansen model" as it is known in Evolutionary Biology [46], the earlier model of Eq. 3.67 being known as the "method of independent contrasts" [80]. These models thus define a covariance function that we can use in order to define an association between readings. A phylogeny is thereof mathematically "*a rooted binary tree with labeled leaves*"[138].

This inferential framework though immediately raises a number of issues [119] :

- When examining a phylogeny, the empirical information is typically only available for extant taxa, represented by leaves of a phylogenetic tree, whereas evolutionary questions frequently concern unobserved ancestors deeper in the tree.
- The available information for different organisms in a phylogeny is not independent since a phylogeny describes a complex pattern of non-independence; observed variation is a mixture of this inherited variation and taxon-specific variation [56].
- The phylogenetic tree itself is treated as representing the true evolutionary history between its leaves and is not a subject of investigation [144].
- Phylogenetic inference is focused on scalar (or in the best case multivariate) traits. There is already an emerging literature on function-valued traits [167; 307; 298] recognizing that many characteristics

[16] ['dɛɪtə], or ['dætə] depending mostly if you are in England or US respectively.

Why do we understand each other (sometimes)?
Assume one is observing a group of children growing up in a population using a language L with a set dictionary D. Ideally each child is using exactly the same language as any other child as this would ensure perfect intelligibility between them; therefore if X represents the percentage of words having the same semantic meaning with the words in dictionary D but not the same vocalizations, $X_{opt} = 0\%$. Perfect semantic as well as vocal matching is quite improbable though; for instance what constitutes "cold weather" or how does one pronounce "data"[16]is highly dependent on your upbringing as well as surroundings. Additionally even if all children have exactly the same external stimuli, small changes (usually counted in terms of Levensthein distances) due to imperfect language acquisition can be detected between children's and parents' vocalization patterns leading to small fluctuations in the X value of children. Nevertheless these children are mutually intelligible and can directly interact with each other and past generations if they are not "too far away" from X_{opt}. A child with a vocabulary largely different than the one used by all his peers would clearly be unfit in terms of communication and would face extreme evolutionary pressure in social terms; a child with just a very small number of different words though would be still fine. While trivialized, this example does illustrate all three main evolutionary forces: 1. Founder Effects, 2. Random Drift and 3. Natural Selection, each taking the form of parents' states, imperfect language acquisition and communication efficiency respectively.

of living organisms are best represented as a continuous function rather than a single factor or a small number of correlated factors. Such characteristics include growth or mortality curves [242], reaction-norms [166] and distributions [340], where the increasing ease of genome sequencing has greatly expanded the range of species in which distributions of gene [219] or predicted protein [170] properties are available.

We will examine all of the previous questions but we will mostly focus on the last one.

3.5.1 Gaussian Processes & Phylogenetic Regression

A Gaussian Process (GP) is defined as a probability distribution over functions $Y(s)$ such that the set of values of $Y(s)$ evaluated at an arbitrary set of points $s_1, \ldots, s_N$ jointly have a Gaussian distribution. Importantly this means that if one chooses to work with a zero-meaned Gaussian processes, these GPs will be completely specified by their second-order statistics [28]. Drawing analogy with spatial statistics methodology and kriging, these inference would be referred as *simple kriging*. To formulate this function-space view of Gaussian processes we can write a GP as a function $f(x)$ such that:

$$f(x) \sim GP(m(x), k(x, x')) \tag{3.69}$$

where as stated beforehand if $m(x) = 0$, the covariance function $k(x, x')$ can be seen as :

$$K(, x') = E\{(f(x) - m(x))(f(x') - m(x'))\} \tag{3.70}$$

$$= E\{(f(x))(f(x'))\} \quad \text{if } m(x) = 0 \tag{3.71}$$

where $f(x)$ is the value of the function f at point x [263]. Having this very basic formulation in place it is interesting to look more specifically on the covariance functions' level. Covariance functions encode not only the covariance of the sample points among the observed points but also they offer an insight in the dynamics of the whole process.

In addition, the realization of the covariance function K as the covariance matrix K between all the pair of points x and x' specifies a distribution on functions and is known as the Gram matrix. Importantly, because every valid covariance function is a scalar product of vectors, by construction the matrix K is a non-negative definite matrix. Equivalently, the covariance function K is a non-negative definite function in the sense that for every pair x and x' , $K(x, x') \geq 0$, if $K(\cdot, \cdot) \geq 0$ then K is called semi-positive definite. Importantly the non-negative definiteness of K enables its spectral decomposition using the Karhunen-Loeve expansion. Basic aspects that can be defined through the covariance function

are the process' stationarity, isotropy and smoothness [17].

Stationarity refers to the process' behaviour regarding the separation of any two points x and x'. If the process is stationary, it depends on their separation, $x - x'$, while if non-stationary it depends on the actual position of the points x and x'; an example of a stationary process is the Ornstein-Uhlenbeck (O-U) process.

If the process depends only on $|x - x'|$, the Euclidean distance (not the direction) between x and x' then the process is considered isotropic. A process that is concurrently stationary and isotropic is considered to be homogeneous[110]; in practice these properties reflect the differences (or rather the lack of them) in the behaviour of the process given the location of the observer.

Ultimately Gaussian processes translate as taking priors on functions and the smoothness of these priors can be induced by the covariance function [17]. If we expect that for "near-by" input points x and x' their corresponding output points y and y' to be "near-by" also, then the assumption of smoothness is present. If we wish to allow for significant displacement then we might choose a rougher covariance function. Extreme examples of the behaviour, is the Ornstein-Uhlenbeck covariance function and the squared exponential where the former is never differentiable and the latter infinitely differentiable.

Periodicity refers to inducing periodic patterns within the behaviour of the process. Formally, this is achieved by mapping the input x to a two dimensional vector u(x) =(cos(x), sin(x)). As outlined earlier a stochastic process with great biological interest is the O-U process. This is because we recognize that what we ultimately want is a Gaussian-Markov process; a stochastic process that satisfies the requirements of both a Gaussian (in terms of changes) and a Markovian (in terms of finite memory) process. With this in mind using a standard noisy measurement O-U kernel in the context of a phylogenetic GP ($f(L) \sim \mathcal{N}(0, K(L, L, \theta))$) would therefore be resulting in the following covariance structure:

$$K(l_i, l_j) = s_f^2 \exp(-|l_i - l_j|/\lambda) + s_n^2 \delta_{l_i, l_j} \tag{3.72}$$

where for a given trait $f(L)$ on a finite set of co-ordinates "leaf" L, $K(L, L, \theta)$ is the matrix of covariances of pairs (l_i, l_j) with hyperparameters θ; θ being in this case composed by three components:

- s_f^2 : intensity of random fluctuations in evolution due to balance between the restraining forces / amplitude of function variation
- λ : phylogenetic horizon (how many generations back a trait is influence by) / characteristic length scale ("*roughly the distance you have to move in input space before the function value can change significantly*" [263] and
- s_n^2: interspecies variation, changes unaccountable from the relations conveyed by the phylogeny / Gaussian noise.

With this at hand the final estimation due to the predictive distribution is found under a standard maximum likelihood framework where one maximizes the phylogenetic GP's LogLikelihood:

$$\log p(f(L)|\theta) = -\frac{1}{2} f(L)^T K(L, L, \theta) f(L) - \frac{1}{2} log|K(L, L, \theta)| - \frac{|L|}{2} log(2\pi) \tag{3.73}$$

in order to find the optimal values of θ, θ_{opt}.

Through θ_{opt} one is immediately able to answer questions regarding the evolutionary properties of the sample at hand. For example if $s_f^2 << s_n^2$ then is almost obvious that the phylogeny at hand is able to account only for a very small proportion of the observed variance and thus probably the phylogeny is not useful. The same insight being conveyed when $\lambda \to 0$, where effectively this mean that each node is in practice agnostic of all other nodes in the phylogeny and no "information transferral" takes place. In any case the fact is that if one fixes the values of θ the posterior distribution for ancestral states A is immediately available through the posterior of the Gaussian distribution that the (univariate) traits describe. Namely:

$$f(A)|f(L) \sim \mathcal{N}(K(A, L)K(L, L)^{-1} f(L), K(A) - K(A, L)K(L, L)K(A, L)^T). \tag{3.74}$$

While phylogenetic GPR will be revisited in chapter 6, we need to immediately highlight the fact that we not only get a posterior mean estimate for $f(A)$, we are also able to quantify our uncertainty about

that estimate by variance attributed to that point in the phylogeny that is independent of the actual observations value $f(L)$ [157].

Combining this notation with the previously presented concept of an O-U process translates the covariance structure K into a reflection of the perturbations due to selective demands from unconsidered selective factors. These being due, in the case of a language, to semantic correlations of the sounds produced, voicing correlations between the biomechanics of phonation, environmental fluctuations, and obviously random "mutations" that materialize as "corruption" of the initial sounds. Expanding on this, phylogenetic time is the concept that serves as the continuum over which data are observed (in the case of observed leaf nodes) or assumed to exist (unobserved ancestral nodes). To that extend X_{opt} cannot be defined as having a single physical notion but as (under simplifying assumptions) conceptual optimal state where a language conveys using speech perfectly all the information required by its speakers.

3.5.2 Tree Reconstruction

The estimation method presented above makes a critical assumption: The phylogenetic relations among the "leaves" of the phylogenetic tree used are correct. This can't be further than truth; in reality the phylogenetic tree is at best a sensible approximation [61; 154]. Three main approaches have been presented as suitable approximation schemes:

- Distance-based trees
- Maximum Parsimony-based trees
- Maximum Likelihood-based trees

The idea behind distance-based trees is analogous to that used in clustering. One utilizes a metric of similarity between the given extant taxa of the phylogeny and based on that metric a distance matrix is computed. Aside the obvious choice of Euclidean distance other distances metrics eg. Levenshtein distance in Linguistics studies, are popular choices. Using the distance matrix produced, two approaches can be taken. Either a top-bottom or a bottom-up [129]. In the first case one finds a point that best partitions data in two well-distinct partitions and then recursively applies the same splitting among the two resulting partitions. On the bottom-up approach one first merges the two data-points closer together and assigns them in the same cluster. Afterwards the same approach is used but this time the merged cluster is treated as a single point. This approach while rather straightforward has a problematic property: it does not account for the root. This bottom-up approach, known in the Computer Science literature as agglomerative clustering, is the essence of one of the most popular early phylogenetic tree reconstruction algorithms, *neighbour joining* (NJ), the other being *Unweighed Pair Group Method with Arithmetic mean* (UPGMA) [31]. Algorithmically both NJ and UPGMA run in $O(N^3)$ time. While computationally efficient in comparison with other approaches though, both implementations do not guarantee that will result in a tree where no edge lengths are negative; also they are obviously extremely sensitive to the choice of the distant metric used, and as such have been deemed "inappropriate" for most modern day phylogenetic analysis. Often NJ or UPGMA tree serve as an "initial solution" tree for advances methods.

The parsimony based approach views each phylogeny as a model of evolution and tries to fit the most parsimonious model; it is based on the same theoretical principals as model selection: Occam's razor [86]. The parameters of a model in the case of phylogenetic trees though, are evolutionary transitions, roughly speaking speciation events. Unfortunately while this appears reasonably coherent, it often results into over-simplified models. It enforces phylogenetic affinity, by requiring two leaves that exhibit the same trait to be related. While quite plausible, convergent [17] evolution among species have shown this not to be a necessary condition. This manifests in the well-known problem of *long branch attraction*, ie. the clustering together of otherwise unrelated species. The main critique against maximum-parsimony relies in its inconsistency. As with any information measurement used for model selection one would expect the $P(\text{choose the true model}) \to 1$ as the number of sample $N \to \infty$, but maximum parsimony does not guarantee that. Algorithmically maximum parsimony does describe an NP-hard problem and while

[17] As convergent evolution we define the independent evolution of similar features in species of different phylogenies, eg. the presence of wings in bats and birds.

certain well-adapted heuristic algorithms do exist this also tends to make it undesirable. Additionally it is often the case that a number of "equally" good parsimony trees might be produced for a given dataset. In those cases a majority rule is enforced but it is not guaranteed to resolve this collision situation, especially in cases where the data are not highly informative in regards with the phylogeny in question [154]. Ultimately parsimonious reconstructed trees have been generally outperformed by ML-based methods.

The ML-based trees are exactly that; the phylogenies that maximize the likelihood of observing the extant taxa under the evolutionary model assumed [79]. In short, "*likelihood methods produce a number of trees, one of which is usually found to be the most likely tree*"[154]. Under this approach one specifies a model of evolutionary change (eg. the OU model presented beforehand) and then calculates the probability of the data given the evolutionary history presented by the tree. Evidently the quality of this methodology is related to the successful choice of evolutionary model. ML-based methods, in contrast with maximum parsimony based methods do use branch length to calculate the distance between point of the phylogeny; exactly because of that they also enable the practitioner to seamlessly infer ancestral states along the phylogeny in question. The most obvious theoretical limitation of "simple" ML-based methods is the fact they assume a unique rate of change along a phylogeny. Multiple rates can be possibly assumed but especially when one is presented with a smaller dataset, overfitting can be an issue. Obviously standard information criteria (eg. AIC) can be also employed here. In practice one starts with a specified tree derived from the input set (eg. using NJ tree) and then branch lengths are changed in order to produce the "ML-tree" [154]; other methodologies go as far as sampling the whole tree-space, effectively examining 2^N different trees but this is obviously an extremely expensive approach for all but the smallest datasets. Direct generalizations of this approach are presented within a Bayesian setting [143; 273]. While we do not explore this in detail, the presence of priors is used in order to account for prior assumptions regarding certain branch-lengths, evolutionary optima, and other model parameters.

Interestingly none of the proposed methodologies addresses internally the issue of rooting a tree. In general they are two approaches: *mid-point routing* and *outgroup usage* [165]. The first approach assumes that the longest path between two extant taxa denotes the most "archaic" split and therefore the tree is rooted at the mid-point of that path. The second approach is first fitting an unrooted phylogenetic tree $\mathbf{T}$ on the original data and adding an obvious "phylogenetic outlier" to that tree. The new node connecting the original unrooted tree to the outgroup taxon, is considered to be the root of the tree. The rationale behind this technique is straightforward: if for example Greek is added in Romance languages phylogeny the bifurcation between Greek language and all the other Romance languages must be the "oldest" one. Clearly this method can be problematic because one might either pick an outgroup that is actually related to some of the original member of the phylogeny or either the outgroup that is so extreme (for instance a Papuan language in the case of a Romance phylogeny) that the rooting results become "random" as they are no meaningful similarities to start with.

Chapter 4

Amplitude modelling in Mandarin Chinese

4.1 Introduction

As already mentioned, 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 F_0) plays a crucial role in the actual lexical entry of the word; 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 F_0 as a single point by utilizing target values [158; 29] or obtained estimates of the F_0 contour by treating it as a bounded rigid curve through processes of averaging [333]. Such approaches though, by necessity, impose simplifying assumptions which make interpretation difficult when considering a complete corpus of data from a more natural language experiment.

Here, a different approach is adopted as a first attempt to introduce FDA for the phonetic analysis of a language. In the proposed model F_0 curve is characterized as the realization of a stochastic Gaussian process; essentially a generalization of a multivariate Gaussian random variable to an infinite index set [290]. 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.

As a starting point, a smoothing and interpolation procedure is utilized to change the measurement from real-time into that of normalized word time, building partially on the assumption of syllable-synchronization [244]. 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 [77]) and a jackknifing model selection procedure to further enhance and test the robustness of the findings. This use of FPCA and mixed effects modelling offers a generalized semi-parametric approach to the linguistic modelling of Mandarin Chinese F_0.

In particular, a functional principal component analysis (FPCA) is first performed on the dataset's F_0 measurements to extract the principal curves. This "curve-basis" serves as the coordinate system which explains the most variation in the data. Similar approaches might utilize Legendre polynomials [104], quadratic splines [135] 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 [258]. In order to compute the FPC's, the sample mean is subtracted from the data, and then the covariance of the data is calculated, as in section 3.3. 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 one might expect, the two approaches yield very similar results; because this might be of interest (especially regarding the insights provided by FPCA) some intermediate results are shown in the Sect. A.3 of the Appendix.

Building on the FPCA findings, the functional principal component (FPC) scores are used as

the dependent variables in a series of linear mixed effect (LME) models, allowing the scores to act as proxy data for the complete curves. The scores essentially quantify the weight each FPC carries in the final F_0 curve formation, as was discussed in Sect. 3.3.1. LME models allow the inclusion of both fixed and random effects to achieve a flexible modelling of the data (See Sect. 3.4.1 for more details). In the current case, the differences between individual speakers due to genetic, environmental [260] or even chance factors [109] are modelled as a series of random additive effects acting on the F_0 contours [15; 11].

The methodology presented here addresses the issue that, while it has been widely accepted and documented that F_0 undergoes variations due to phonetic processes in speech production attributed to fixed effects (eg. 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 F_0 utterance. This immeasurability 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 [89; 214]. The Fujisaki model implementations have been extended by Mixdorff [212] to account for micro-prosodic effects by taking advantage of the MOMEL algorithm [135]. Other approaches utilize the automatic intonation modelling approach as offered by the INTSINT [196; 134] and/or the TILT [305] algorithmic implementations. Furthermore, the qTA model [244] also builds on Fujisaki's assumption, proposing a description of the physiological mechanisms behind F_0 production, a goal somewhat different from the one here. In the present framework and analogous to the Fujisaki rationale, F_0 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 [75] and Aston et al. have further extended [11], the explanatory power that can be yielded from the application of LME models for F_0 is insightful in cases of tonal languages. In the current study, the F_0 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 analysing 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 final note we need to emphasize that the application of FDA (Functional Data Analysis) in relation with phonetic analysis is not without precedence. Early work of Ramsay et al. [259] 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 [198; 188; 48; 113], as well as with issues of physiological interests [172; 151; 264], have been addressed with FDA. The current work differs from the above mentioned projects by employing an entire corpus as raw

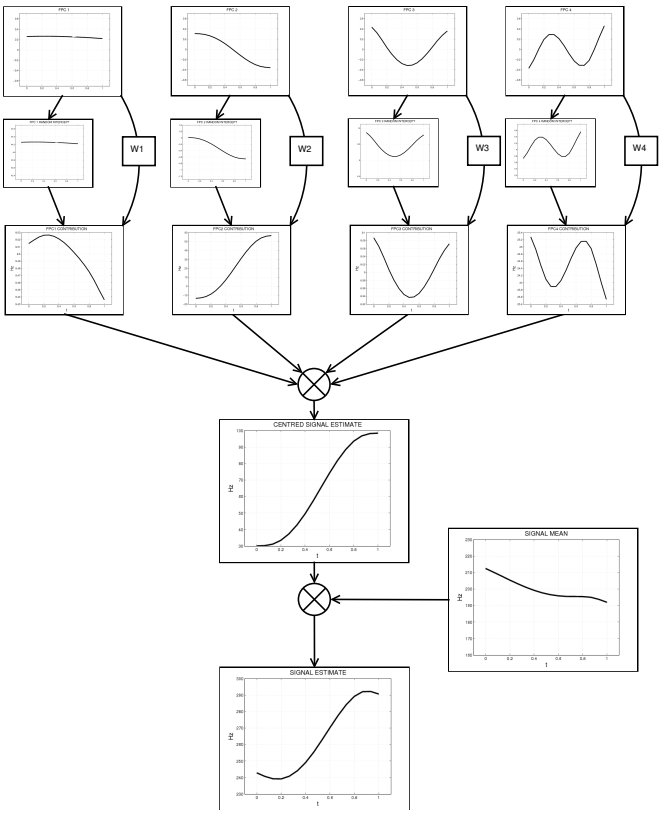

Figure 4.1: The first four eigencomponents (top row) are used to construct the final syllable estimate of F_0 (bottom row). The individual component magnitude (third row) is calculated by using the weight estimates (w_i) obtained as the sum of the relevant utterance covariates from the LME model and the component specific random intercept (2nd 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 F_0 (bottom row).

data. Rather than using a small linguistic sample by a single speaker [214], employing monosyllabic utterances and a small number of sentences [223] and/or frames within the utterances [198; 172; 11; 333; 114] to minimize possible confounds at the data collection level, a large corpus is analysed and the confounds are explicitly modelled. In contrast to existing intonation synthesis algorithms, the current methodology's primary goal is to offer insights into how linguistic and non-linguistic factors are combined in the estimation of F_0 and thus presents an auxiliary approach for existing speech synthesis algorithms in terms of modelling the acoustic shapes of tones.

Overall, the complete procedure to obtain the F_0 estimate is as follows: Once the components FPC_i are fixed, their scores A are used as dependent variables to find the optimal LME models describing those components' dynamics. Their estimates $\hat{a_i} = w_i$ are then used in order to present a reconstructed curve and check the practical alongside the theoretical prowess of our models. A visual summary is offered in Figure 4.1.

4.2 Sample Pre-processing

We focus our attention on modelling fundamental frequency (F_0) curves. The amplitude of F_0, usually measured in Hz, quantifies the rate/frequency of the speaker's vocal folds' vibration. Reiterating on the importance of the F_0 curves analysed to be "smooth", ie. they possess "one or more derivatives" [254], we follow the methodology of Chiou et al. [57] and use a locally weighted least squares smoother, S_L, in order to fit local linear polynomials to the data and produce smooth data-curves interpolated upon a common time-grid of L points on a dimensionless interval $[0, 1]$. As mentioned in section 3.1, different approaches using smoothing splines [115] or wavelets [218] could also be used. The form of the kernel smoother used is shown in Eq. 3.9. Similar to section 3.1 the kernel function K was set to the Gaussian kernel function $K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ providing an infinite support. Smoothing within our context effectively constitutes the estimation of a_0 in Eq. 3.9. The choice for this linear is due to the fact while we do accept that higher order polynomials might reflect that general variation trend across a speaker's utterance, the localized variation patterns are inherently linear. Additionally assuming a higher order polynomial would conceptually reflect a stronger smoothness assumption than the one of double-differentiability we currently examine within the Ramsay & Silverman theoretical framework used.

The curves in the COSPRO sample have an average of 16 readings per case, hence the number of grid points chosen was $L = 16$. 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. In some occasions sign reversals on some of the eigenfunctions of the covariance operators were noted but that carried no impact to our modelling assumptions. The smoother bandwidth b was set to 5% of the relative curve length using leave-one-out cross-validation. As is common in a dataset of this size, occasional missing values have occurred and curves having 5% or more of the F_0 readings missing were excluded from future analysis. These missing values usually occurred at the beginning or the end of a syllable's recording and are most probably due to the delayed start or premature stopping of the recording. We define the newly formed, smoothed and equi-sized version of COSPRO-1's F_0 curves as Y. At this point we need to stress that the use of a common 16-points grid for all Y effectively constitutes a time-normalization (rather than synchronization) step that enforces an "identity" warping function on all Y_i's. While simplistic, this is standard in Linguistic studies; a more sophisticated approach will be explored in chapter 5.

4.3 Data Analysis and Results

Having established the smoothness of the data, the next step in the actual implementation of the m-dimensional linear model for such processes as this is shown in Eq. 3.38. Here it takes the form:

$$y_i(t) = \mu(t) + \sum_{\nu=1}^{m} \alpha_{\nu,i}(t)\phi_\nu(t) \tag{4.1}$$

where $\mu(t)$ is the functional average of the F_0 curves in sample Y and α_ν are the FPC scores associated with the FPC $\phi_\nu(t)$. As such the sample covariance function is calculated using Eq. 3.39. This is shown in Fig. 4.2. Immediately we recognize three significant qualitative findings: 1. The largest variance is exhibited at the initial F_0 curve placement. 2. The covariance appears overall smooth and expectedly confirms the intuition that points located further apart on the F_0 curve exhibit less covariance than points closer to each other. 3. Along the diagonal axis of the covariance matrix a slight "sink" appears. This suggests that probably the edge placements as a whole are more important for the final F_0 realization compared with the middle of the phone.

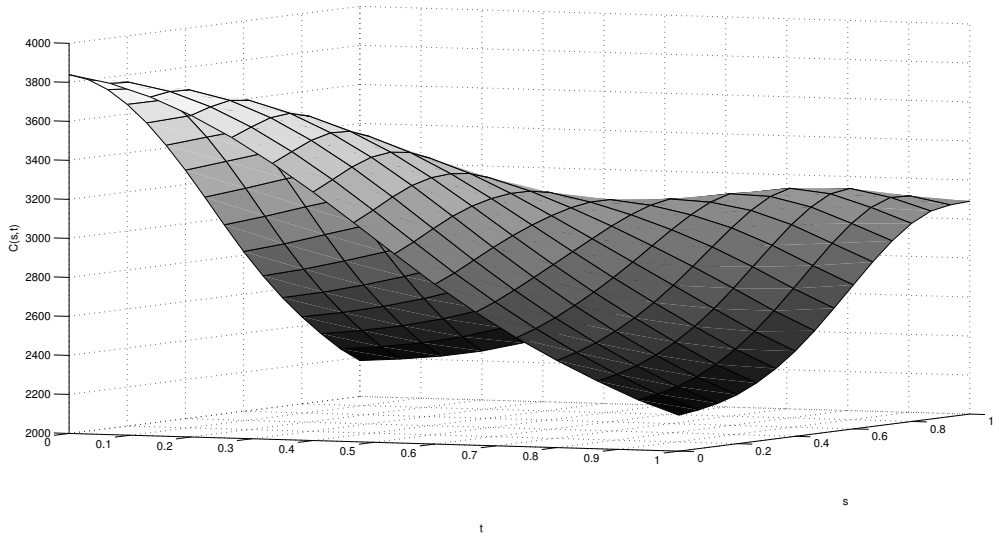

Figure 4.2: Covariance function of the 54707 smoothed F_0 sample curves, exhibiting smooth behaviour. Interpolation on a 16-point grid.

Then, following the same methodology as Aston et al. [11], given that one solves Eq. 3.40 [1], the FPCA $A_{i,\nu}$ scores are estimated as:

$$\hat{A}_{i,\nu} = \sum_{i=1}^{s=16} \{y_i(t_j) - \hat{\mu}(t_j)\}\hat{\phi_\nu}(t_j)\Delta_j, \tag{4.2}$$

where $\Delta_j = t_j - t_{j-1}$ accounts for the discrete nature of our theoretically continuous data. These scores, $A_{i,\nu}$, are the ones finally used for the estimation analysis by the LME models. It must be noted that because by definition the FPC's are orthogonal, their scores are also considered mutually orthogonal. As such we employ a series of univariate LME regressions rather than a single multivariate one. The choice and number of FPC's used is related to the amount of variation that each of these components reflects. 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 F_0 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 10Hz. As advocated by Kochanski [302], in the case of human speech, the JND for pitch motions seems to be rather larger. Black and Hunt [29] show that a 9.9 Hz RMS error is not detrimental to the model's success. This threshold will be used throughout this work; however our approach is flexible enough for other practitioners to utilize with different cut-off thresholds.

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

[1] Numerically this is solved by employing the *Generalized Schur* or *QZ* decomposition.

are relatively sparse within the data, eg. certain tones (eg. Tone 3) are inherently less common than others (eg. Tone 1). 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 modelling 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 FPC's account for the 99.992% of the total variation in the sample (Table 4.1, Columns 2 & 3). Nevertheless, 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).

In addition, 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 five, which were previously deemed as of possible importance (see Table 4.1, Columns 4 & 5). 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 4.1, if a 95% quantile were used, it would suggest that we actually exclude the components that are below 20.7 Hz and ultimately use a significantly narrower F_0 range.

FPC #	Individ. Variation	Cumul. Variation	Hz (99%)	Hz (95%)
FPC_1	88.23	88.23	133.3	101.3
FPC_2	9.78	98.01	55.3	38.3
FPC_3	1.42	99.43	35.8	20.7
FPC_4	0.32	99.75	19.1	9.1
FPC_5	0.11	99.86	8.9	4.2
FPC_6	0.05	99.91	5.7	2.5
FPC_7	0.03	99.94	3.6	1.7
FPC_8	0.02	99.96	2.9	1.2
FPC_9	0.01	99.97	2.4	1.1
FPC_{10}	0.01	99.98	1.8	.85
FPC_{11}	0.01	99.99	1.7	.68
FPC_{12}	0.01	99.99	1.3	.45

Table 4.1: Individual and Cumulative Variation Percentage per FPC. Actual Auditory Variation per FPC (in Hz) (human speech auditory sensitivity threshold $\approx$ 10 Hz).

As mentioned in the previous sections, 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 mean and FPC curves (Figure 4.3) appear smooth through their values. It must be noted though that the 5th and 6th 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 (Figure 4.3). This result concurs with the assumed contour shapes of Grabe et al. [104] where Legendre polynomials $L0$ to $L3$ were utilized for the contour basis of F_0 to examine intonation (See Appendix for actual shapes). In principle, 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 reconstructed by using FPC_1, FPC_2 and FPC_3, as can be seen from the actual shape of those components, with FPC_4 allowing contextual movement between tones.

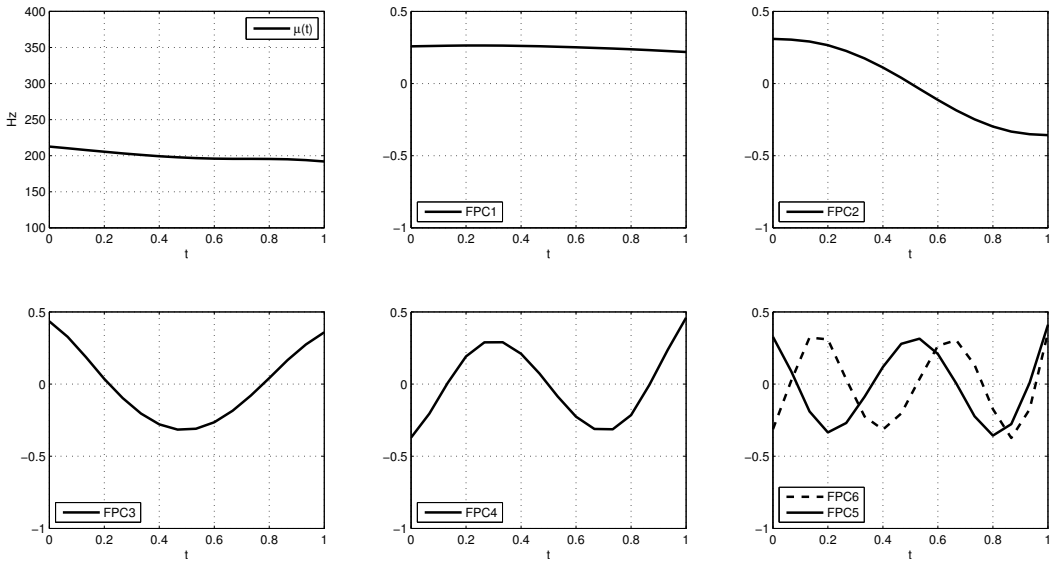

Figure 4.3: Mean Function and 1st, 2nd, 3rd, 4th, 5th and 6th Functional Principal Components. Together these account for 99.994% of the sample variance, but only the first four have linguistic meaning (99.933 % of samples variation) and as such the 5th and 6th were not used in the subsequent analysis.

While the kernel smoothing and interpolation was implemented by a custom built C++ program utilizing the `GSL` package [91], the calculation for the eigenfunction decomposition and the production of the FPC scores were conducted using standard built-in `MATLAB` procedures [204]. The rest of the analysis was carried out in the statistical environment `R` [247]. Except for the obvious standard `R` methods used (`density()`, `lm()`, etc.) the major body of the analysis was done using methods from the statistical package `lme4` [22] for the LME model estimation and prediction. As mentioned in section 3.4.3, we examined the robustness of the selected models also through jackknifing; extensive sub-sampling was implemented using 180 5-sub-sample partitions of our original samples, yielding a total of 900 sub-samples.[2]

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:

$$\begin{aligned} FPC_X = \{&[tn_{previous} * tn_{current} * tn_{next}]+ \\ &[cn_{previous} * tn_{current} * cn_{next}]+ \\ &[(B2) + (B2)^2 + (B2)^3+ \\ &(B3) + (B3)^2 + (B3)^3+ \\ &(B4) + (B4)^2 + (B4)^3+ \\ &(B5) + (B5)^2 + (B5)^3] * Sex + [rhyme_t]\}\beta \\ &+ \{[Sentence] + [SpkrID]\}\gamma + \epsilon. \end{aligned} \tag{4.3}$$

The standard `R` notation is used here for simplicity regarding the interaction effects; [K*L] represents a short-hand notation for [K + L + K:L] where the colon specifies the interaction of the covariates to its left and right [14]. Table 2.2 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 [158]. 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

[2]For a detailed discussion and relevant histograms refer to the Appendix Sect. A.7.

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 (ie. each syllable). Table 2.1 offers a comprehensive list of what each break represents. 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 was 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 modelling of more complex down-drift patterns. This approach allows an analogy to be drawn with the phrase component used in the Fujisaki modelling approach [89]. 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 [313] to encode the 9 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 Appendix, Sect. A.1. As a final note we draw attention to the fact that design-wise we do not incorporate random slopes along with random intercepts for our models. The reason to avoid this is two-fold: first, we assume that especially regarding the speaker random effect, where we have only 5 speakers, over-fitting can be a major problem. Additionally, because we include so many fixed effects interactions between speaker's sex and the break information we do offer "enough flexibility" in our model through its fixed effect so that if random slopes were prominent in our approach, at least we have circumvented partially that issue.

As shown earlier in Table 2.2, 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 3-way interactions and their embedded 2- and 1-way interactions. Three-way interactions have been known to be present in Taiwanese Mandarin and therefore were deemed as significant effects to incorporate [333; 11; 310] 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 [193]). The glottal stop [ʔ] is defined as a 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 [37])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 labelled zero-initial as neither voiced nor voiceless but as 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 (eg. commands have a different F_0 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 4.2).

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 modelling of the second and third FPC's, 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 F_0 curve formation. As mentioned, the first eigenfunction appears to have a shifting effect on the F_0 curve itself, raising or lowering the overall F_0. In contrast, the second, third and fourth eigenfunctions have an average effect on the F_0 curve quite close to 0 over the entire trajectory (as can be easily seen on the plots themselves). Therefore FPC_2, FPC_3 and FPC_4 do not have an overall shifting effect on the curve, but rather only

	FPC_1 Estimate (.025,.975)	FPC_2 Estimate (.025,.975)	FPC_3 Estimate (.025,.975)	FPC_4 Estimate (.025,.975)
Speaker	71.755	4.682	6.976	3.094
	(16.931,133.304)	(1.002,8.669)	(1.644 ,12.970)	(0.722,5.768)
Sentence	30.823	3.497	1.956	0.596
	(28.505,33.124)	(2.893,4.011)	(1.676 ,2.203)	(0.400,0.747)
Residual	118.917	45.089	21.241	12.119
	(118.145,119.601)	(44.799,45.349)	(21.102 ,21.363)	(12.042,12.190)

Table 4.2: Random Effects and parametric bootstrap confidence intervals (2.5%, 97.5%) for the 1st, 2nd, 3rd and 4th FPC scores models as produced by using 10000 samples.

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 F_0 formation does not correspond to any known/formal individual Mandarin tones. Nevertheless, it appears native speakers do indeed exhibit components of sinusoidal-shape in their production of F_0, as FPC_4 accounts for 19 Hz variation, hence represents an audible signal. It is likely that this F_0 curve component is needed to move between different tones in certain tonal configurations, as will be discussed below. Finally we note that the dynamics FPCs used where found to be robust and clearly identifiably when tested in smaller subsets of the COSPRO dataset (Sect. A.4).

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[3]; for each FPC the corresponding model was fit separately from other FPCs. Taking into account the results from AIC and jackknifing, the following model for FPC_1 was chosen:

$$\begin{aligned} FPC_1 = \{&[tn_{previous} * tn_{current}] + [tn_{current} * tn_{next}]+ \\ &[cn_{previous} * tn_{current}[+[tn_{current} * cn_{next}]+ \\ &[cn_{previous} * cn_{next}]+ \\ &[(B2) + (B2)^2 + (B2)^3 + (B3) + (B3)^2 \\ &+ (B3)^3 + (B4) + (B4)^2 + (B4)^3+ \\ &(B5) + (B5)^2 + (B5)^3] * Sex+ \\ &[rhyme_t]\}\beta + \{[Sentence] + [SpkrID]\}\gamma + \epsilon. \end{aligned} \tag{4.4}$$

The first eigenfunction is almost exclusively associated with the speaker's F_0 curve placement. Complex third order tonal interactions were not present. The speaker-identity random effect is significantly high despite the inclusion of speakers' sex as a covariate. Thus, this random effect 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 in F_0 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 F_0 (see Appendix, chapter A.8). 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 neighbouring consonants is of significance for all tone types. Specifically, the voicing of the preceding consonant resulted in a statistically significant lower overall F_0 placement, when compared to the F_0 placement associated with a preceding voiceless consonant. Overall, voiced neighbouring/initial consonants (including epenthetic glottal stop) resulted in lower F_0 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 forma-

[3]For a detailed listing of the relevant covariates and the jackknifing results refer to the Appendix Sections A.6 - A.8.

tion of F_0. 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 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 F_0 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 F_0 estimate to be lower. Furthermore, the break interactions with the 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 F_0 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:

$$\begin{aligned} FPC_2 = \{&[tn_{previous} * tn_{current} * tn_{next}]+ \\ &[cn_{previous} * tn_{current} * cn_{next}]+ \\ &[(B2) + (B2)^2 + (B2)^3 + (B3) + (B3)^2+ \\ &(B3)^3 + (B4) + (B4)^2 + (B4)^3] * Sex+ \\ &[(B5) + (B5)^2 + (B5)^3]+ \\ &[rhyme_t]\}\beta + \{[Sentence] + [SpkrID]\}\gamma + \epsilon. \end{aligned} \tag{4.5}$$

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 kind 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 modelling 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 weight in the eigenfunction's final formation, but their individual impacts are a whole scale of magnitude smaller than the model's residual (See Table 4.2). 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 & 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 neighbouring 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 F_0 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:

$$\begin{aligned} FPC_3 = \{&[tn_{previous} * tn_{current}] + [tn_{current} * tn_{next}]+ \\ &[tn_{previous} * tn_{next}]+ \\ &[cn_{previous} * tn_{current} * cn_{next}]+ \\ &[(B2) + (B2)^2 + (B2)^3 + (B3) + (B3)^2 \\ &+ (B3)^3] * Sex + [rhyme_t]\}\beta+ \\ &\{[Sentence] + [SpkrID]\}\gamma + \epsilon \end{aligned} \tag{4.6}$$

The third eigenfunction possibly plays a dual role. Firstly, it is mostly 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 in transition between syllables. In addition, the speaker's identity random effect appears to play a statistically significant role to the eigencomponent's final weight, 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 FPC_2, 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 modelling 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:

$$\begin{aligned} FPC_4 = \{&[tn_{previous} * tn_{current}] + [tn_{next}] + \\ &[cn_{previous} * tn_{current}] + [tn_{current} * cn_{next}] + \\ &[cn_{previous} * cn_{next}] + \\ &[(B2) + (B2)^2 + (B2)^3 + (B3) + (B3)^2 + \\ &(B3)^3] * Sex + (B4) + (B4)^2 + (B4)^3 + \\ &[rhyme_t]\}\beta + \{[Sentence] + [SpkrID]\}\gamma + \epsilon. \end{aligned} \tag{4.7}$$

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.

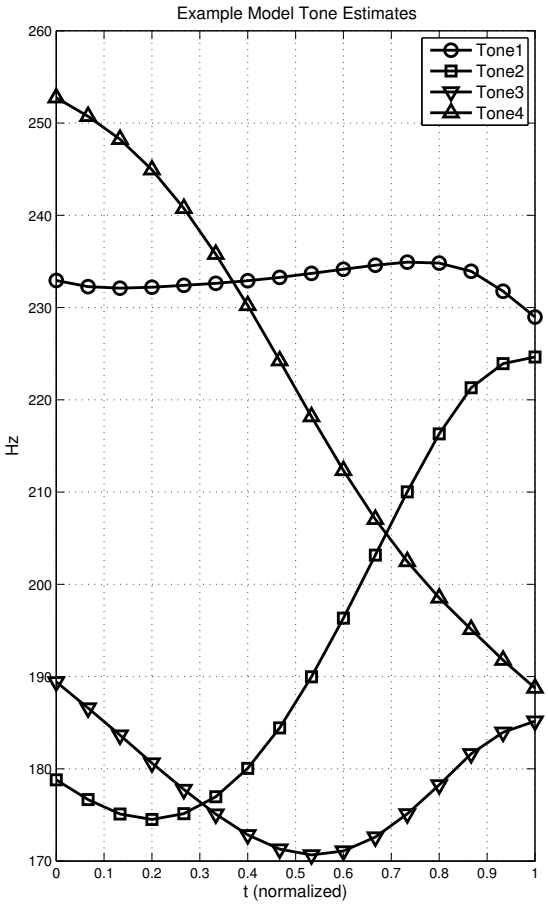

Figure 4.4: Example tone estimates produced by the model utilizing all four FPC's. Tone 5 is not represented as it lacks a general estimate, always being significantly affected by non-standardized downdrift effects. Phonologically, toneless syllables do not specify a pitch target.

Another important issue is that breaks 2 and 3 (prosodic word and phrase) have much influence on the F_0 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 F_0 variance approaches the limit of the explanatory power of our modelling rationale. Therefore, fluctuations smaller than this (small) magnitude are due to articulatory and/or phonetic effects that are beyond the mostly

FPC#	$LM - R_a^2$	$LME - R_a^2$
FPC_1	.6271	.7056
FPC_2	.6109	.6161
FPC_3	.3645	.4136
FPC_4	.1083	.1491

Table 4.3: Adjusted R^2 scores for the selected linear models before and after the inclusion of Speaker and Sentence related random effects.

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 weigh the influence of each respective eigenfunction to the original sample mean, yields the final F_0 estimate (see Figures 4.1 & 4.5). Here the estimates correspond to generic speakers and to estimations of the behaviour 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 (Figure 4.4) generated by the model exhibit qualitatively similar characteristics with those of the YR Chao tone chart (Fig. 2.4).

Table 4.3 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 [77]. 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 F_0 track for rhymes over time. As can be seen in Figure 4.6 [4], the curves estimated from the models are not only fairly good fits to the data on a rhyme by rhyme basis (including the expected estimation error), but the overall

[4]See Appendix, Table A.10, for detailed listing of relevant covariates.

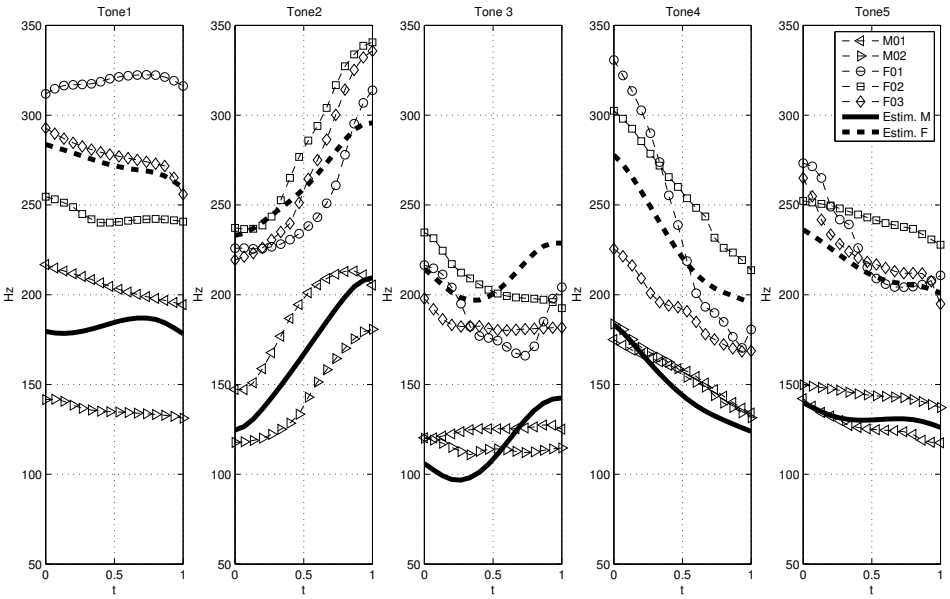

Figure 4.5: One randomly selected syllable for each of the five tones; the functional estimates (bold) for each different tone are shown as well as the corresponding original speaker interpolated data over a dimensionless rhyme time interval t.
(Estimated vowel rhymes: [uei, oŋ, əŋ, uan, ə] for each of the 5 tones respectively. See Appendix Table A.9 for contextual covariate information.)

time normalized track from rhyme to rhyme is captured through the break covariate estimation. 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.

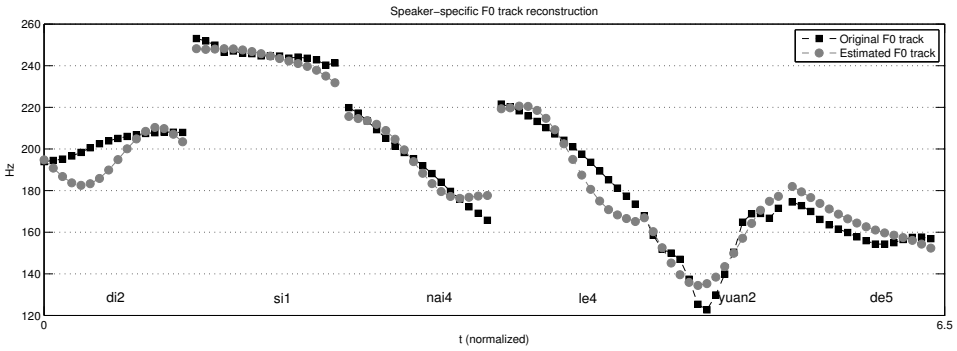

Figure 4.6: Randomly chosen F_0 trajectory over (normalized) time. Here six concurrent F_0 tracks for rhymes are shown for speaker F03. As can be seen, the match is fairly close for most syllables, with the estimates associated with the break information controlling the temporal down drift effects.
(Tonal sequence: 2-1-4-4-2-5 ; Estimated vowel rhymes: [i, ɿ, ai, ə, yɛn, ə]. See Appendix Table A.10 for contextual covariate information.)

4.4 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 dependence of pitch level to the speaker's identity. The influence of triplets in the case of tones 2 and 4 and the subsequent slope-like shape they exhibit is also demonstrated in the case of tone 2 where F_0 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, except FPC_1 (the curve's F_0 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 Prom-on *et al.* in the qTA model [244]. Additionally, an interesting yet not surprising result is that, as the modelling 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 (Figure 4.5[5]) show that the proposed model succeeds in capturing the overall dynamics of the speaker's pronunciation, giving good qualitative and quantitative estimates. This success is obtained despite the fact that the sample exhibits large variance and possible distortion through its measurements even after the initial data were preprocessed. Note that Shih and Kochanski [293] ran into similar issues concerning distorted tone shapes. Collectively, these findings are in line with those of other studies [333], specifically when reviewing the effect of adjacent tones. Durational differences are not taken into account by the current modelling approach. As it will be shown in the next chapter, one can 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. [11] in their study on Luobuzhai Qiang, a tonal Sino-Tibetan language of the Sichuan Province in central-southern China. This fact could reflect a series of shared features among this language family. It could be of interest to review and compare

[5]Figure 4.5 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 Appendix, Table A.9, for detailed listing for relevant covariates.

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 F_0 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 exact 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 phonetic foundations. Therefore, inclusion of such factors as text frequency, 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. The full body of the analysis was re-implemented 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[6] and thus the first four Legendre polynomials did not explain as much of the data as 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 modelled by different FPC's. 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, the notion of JND allows an informal auditory selection procedure to be formed. 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 modelling procedure.

The future steps following from this initial project are fourfold. 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 a framework has already been sketched by Aston et al.[307] and presented with a proof-of-concept biological application by the main author [119] (more details are given in Chapt. 6). Third, by validating this method on a language where many of the effects on F_0 are known, it now becomes possible to investigate numerous effects and their interactions in the production of F_0 in less-studied languages, and to be confident of the results. One can be confident that this will at least for tonal languages be an attainable target as it has already been used for two different applications [11; 117]. Finally the current framework presents a tested methodology that is readily extendible not only towards the investigation of the dynamics within a single set of curves, but also towards the investigation of the interactions between two different sets of curves and the concurrent modelling of their variations. This last point is the focus of the next chapter.

[6] See Appendix Sect. A.2.

Chapter 5

Joint Amplitude and Phase modelling in Mandarin Chinese

5.1 Introduction

As exemplified in the last chapter the modulation of the pitch of the sound is an integral part of the lexical definition of a word. Thus, any statistical approach attempting to provide a pitch typology of the language must incorporate the dynamic nature of the pitch contours into the analysis [111; 245]. Nevertheless pitch contours, and individual human utterances generally, contain variations in both the amplitude and phase of the response, due to effects such as speaker physiology and semantic context. Therefore, to understand the speech synthesis process and analyse the influence that linguistic (eg. context) and non-linguistic effects (eg. speaker) have, we need to account for variations of both types. As seen in the last chapter traditionally, in many phonetic analyses, pitch curves have been linearly time-normalized, removing effects such as speaker speed or vowel length, and these time normalized curves are subsequently analysed as if they were the original data [334; 11]. However, this has a major drawback: potentially interesting information contained in the phase is discarded as pitch patterns are treated as purely amplitude variational phenomena.

In a philosophically similar way to Kneip and Ramsay [169], we model both phase and amplitude information jointly and propose a framework for phonetic analysis based on functional data analysis (FDA) [254] and multivariate linear mixed-effects (LME) models [179]. Using a single multivariate model that concurrently models amplitude, phase and duration, we are able to provide a phonetic typology of the language in terms of a large number of possible linguistic and non-linguistic effects, giving rise to estimates that conform directly to observed data. Following the rationale presented in the previous section, we focus on the dynamics of F_0 [226] in Mandarin Chinese. We utilize two interlinked sets of curves; one set consisting of time normalized F_0 amplitude curves and a second set containing their corresponding time-registration/warping functions registering the original curves to a universal time-scale. Using methodological results from the compositional data literature [2], a principal component analysis of the centred log ratio of the time-registration functions is performed. The principal component scores from the amplitude curves and the time warping functions along with the duration of the syllable are then jointly modelled through a multivariate LME framework.

One notable aspect in our modelling approach is that it is based on a compositional representation of the warping functions. This representation is motivated by viewing the registration functions on normalized time domains as cumulative distribution functions, with derivatives that are density functions, which in turn can be approximated by histograms arbitrarily closely in the L^2 norm. We may then take advantage of the well-known connection between histograms and compositional data [191; 235].

As before our dataset for this work is COSPRO-1. Unfortunately this dataset is prohibitively large to analyse with usual multivariate multilevel computational implementations [22; 116], so a specific computational approach for the analysis of large multivariate LME models is developed. Using the proposed model, we are able to identify a joint model for Mandarin Chinese that serves as a typography for spoken Mandarin. This study thus provides a robust and flexible statistical framework describing intonation properties of Mandarin Chinese with accounting both for amplitude and phase variations.

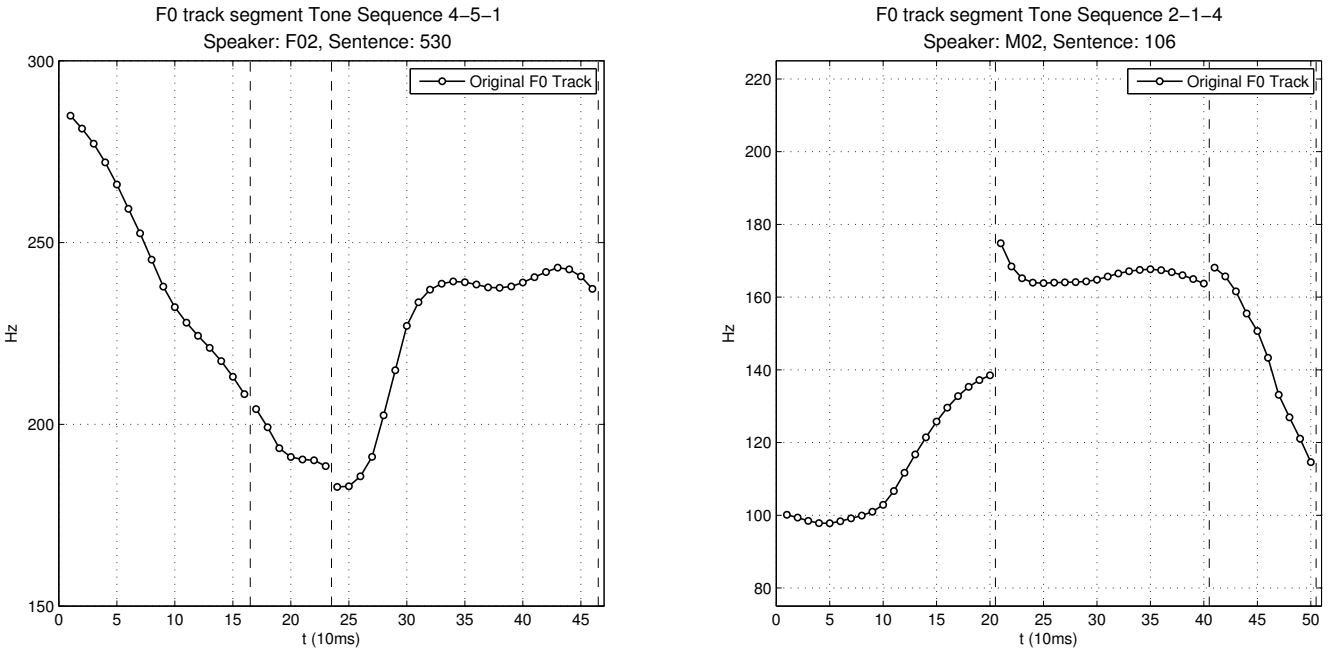

Figure 5.1: An example of triplet trajectories from speakers $F02$ & $M02$ over natural time. F(emale)02 tonal sequence: 4-5-1, M(ale)02 tonal sequence: 2-1-4; Mandarin Chinese rhyme sequences [oŋ-ə-iou] and [ien-in-ʅ] respectively. See Appendix, Sect. A.11 for full contextual covariate information.

5.2 Phonetic Analysis of Mandarin Chinese utilizing Amplitude and Phase

We focus our attention again on modelling fundamental frequency (F_0) curves. The observation units of investigation are brief syllables: F_0 segments that typically span between 120 and 210 milliseconds (Figure 5.1) and are assumed to be smooth and continuous throughout their trajectories. Linguistically our modelling approach of F_0 curves is motivated by the intonation model proposed by Fujisaki [89] where linguistic, para-linguistic and non-linguistic features are assumed to affect speaker F_0 contours. Another motivation for our rationale of combining phase and amplitude variation comes from the successful usage of Hidden Markov Models (HMM) [249; 338] in speech recognition and synthesis modelling. However, unlike the HMM approach, we aim to maintain a linear modelling framework favored by linguists for its explanatory value [15; 75] and suitability for statistical modelling.

Usual approaches segment the analysis of acoustic data as seen in chapter 4. First, one applies a "standard" Dynamic Time Warping (DTW) treatment to the sample using templates [278], registers the data in this new universal time scale and then continues with the analysis of the variational patterns in the synchronized speech utterances [183]. In contrast, we apply Functional Principal Component analysis (FPCA) [51] to the "warped" F_0 curves and also to their corresponding warping functions, the latter being produced during the curve registration step. The warping technique employed is the Pairwise synchronization framework introduced in Sect. 3.2.2. These functional principal component scores then serve as input for using a multivariate LME model.

We use the same set of covariates used in the "amplitude" only investigation presented in the previous chapter. Therefore, in total, aside from Speaker and Sentence information, associated with each F_0 curve are covariates of break index (within word (B2), intermediate (B3), intonational (B4) and utterance (B5) segments), its adjacent consonants, its tone and rhyme type (Table 2.2). In our work all of these variables serve as potential scalar covariates and with the exception of break counts, the fixed covariates are of categorical form [158]. As mentioned previously break counts are very significant as physiologically a break has a resetting effect on the vocal folds' vibrations; the qualitative description of the break counts is provided in the Table 2.1.

5.3 Statistical methodology

5.3.1 A Joint Model

Concurrent phase and amplitude variation is expected in linguistic data and as phonetic datasets feature "dense" measurements with high signal to noise ratios [254], FDA naturally emerges as a statistical framework for F_0 modelling. Nevertheless in all phonetic studies mentioned in chapter 4, the focus of the phonetic analysis has been almost exclusively the amplitude variations (the size of the features on a function's trajectory) rather than the phase variation (the location of the features on a function's trajectory) or the interplay between the two domains. To alleviate this limitation we utilize the formulation presented by Tang & Müller [303] and introduce two types of functions, w_i and h_i. For a given F_0 curve y_i, w_i is the amplitude variation function on the domain $[0,1]$ while h_i is the monotonically increasing phase variation function on the domain $[0,1]$, such that $h_i(0) = 0$ and $h_i(1) = 1$. For generic random phase variation or warping functions h and time domains $[0,T]$, T also being random, we consider time transformations $u = h^{-1}(\frac{t}{T}))$ from $[0,T]$ to $[0,1]$ with inverse transformations $t = Th(u)$. Then, the measurement curve y_i over the interval $t \in [0,T_i]$ is assumed to be of the form:

$$y_i(t) = w_i(h^{-1}(\frac{t}{T_i})) \Leftrightarrow w_i(u) = y_i(T_i h_i(u)) \tag{5.1}$$

where $u \in [0,1]$ and T_i is the duration of the ith curve. A curve y_i is viewed as a realization of the amplitude variation function w_i evaluated over u, with the mapping $h_i(\cdot)$ transforming the scaled real time t onto the universal/sample-wide time-scale u. This being essentially Eq. 3.16. In addition to the generative model presented in Eq. 3.16 though, each curve here can depend on a set of covariates, fixed effects X_i, such as the tone being said, and random effects Z_i, where such random effects correspond to additional speaker and context characteristics. While each individual curve has its own length T_i, the lengths are normalized at the beginning of the analysis of the functional part of the data, and the T_i are included in the modelling as part of the multivariate linear mixed effect framework, allowing not only amplitude and phase, but also duration to be included in the model.

In our application, the curves yi are associated with various covariates, for example, tone, speaker, sentence position. These are incorporated into the model via the principal component scores which can be viewed as taking a common principal component approach [25] to the analysis, where we assume common principal components (across covariates) for the amplitude functions and another common set (across covariates) for phase functions (but these two sets can differ). As will be discussed in section 5.4, this is not a strong assumption in this application. Of the covariates likely present in model, tone is known to affect the shape of the curves (indeed it is in the phonetic textual representation of the syllable), and therefore the identification of warping functions is carried out within tone classes as opposed to across the classes as otherwise very strong (artefactual) warpings will be present in the analysis.

As a direct consequence of our generative model (Eq. 5.1), w_i dictates the size of a given feature and h_i^{-1} dictates the location of that feature for a particular curve i. We assume that w_i and h_i are both elements of $L^2[0,1]$. As a first result w_i can be expressed in terms of a basis expansion:

$$w_i(u) = \mu^w(u) + \sum_{k=1}^{\infty} A^w_{i,k}\phi_k(u), \quad \text{where } \mu^w(u) = E\{w(u)\}. \tag{5.2}$$

The h_i are a sample of random distribution functions which are square integrable but are not naturally representable as a basis expansion in the Hilbert space $L^2[0,1]$, since the space of distribution functions is not closed under linear operations. A common approach to circumvent this difficulty is to observe that $\log(\frac{d}{dt}h_i)$ is not restricted and can be modelled as a basis expansion in $L^2[0,1]$. This observation is done in the following way: first one notices that the integrals of the original functions h_i are by definition continuous and that the exponent of them is by definition positive. As such by reversing this procedure (ie. taking the log of the derivatives of h_i) we ensure that they will define a vector space. In the same manner, a restriction however is that the densities h_i have to integrate to 1, therefore the

random functions $s_i = \log(\frac{d}{dt}h_i)$ are modelled with the unrestricted basis expansion:

$$s_i(u) = \mu^s(u) + \sum_{k=1}^{\infty} A^s_{i,k}\psi_k(u), \quad \text{where } \mu^s(u) = E\{s(u)\}. \tag{5.3}$$

A transformation step is then introduced to satisfy the integration condition, which then yields the representation:

$$h_i(u) = \frac{\int_0^t e^{s_i(u')}du'}{\int_0^1 e^{s_i(u')}du'} \tag{5.4}$$

for the warping functions h_i; the denominator normalizing the final product to ensure the integration condition and allow s_i to be modelled in $L^2[0,1]$. Clearly different choices of bases will give rise to different coefficients A which then can be used for further analysis. A number of different parametric basis functions can be used as basis; for example Grabe et al. advocate the use of Legendre polynomials [104] for the modelling of amplitude. We advocate the use of a principal component basis for both w_i and s_i in Eqs. 5.2 & 5.3, as will be discussed in the next sections, although any basis can be used in the generic framework detailed here. However, a principal components basis does provide the most parsimonious basis in terms of a residual sum of squares like criterion [254].

We note that in order to ensure statistical identifiability of model (Eq. 5.1) several regularity assumptions were introduced in [303], such as the exclusion of essentially flat amplitude functions w_i for which time-warping cannot be reasonably identified, and more importantly, assuming that the time-variation component that is reflected by the random variation in h_i and s_i asymptotically dominates the total variation. In practical terms, we will always obtain well-defined estimates for the component representations in Eqs. 5.2 & 5.3.

For our statistical analysis we explicitly assume that each covariate X_i influences, to different degrees, all of the phone's components/modes as well as influencing the phone's duration T_i. Additionally, as mentioned above in accordance with the Fujisaki model, we assume that each phone component includes Speaker-specific and Sentence-specific variational patterns; we incorporate this information in the covariates Z_i. Then the general form of our model for a given sample curve y_i of duration T_i with two sets of scalar covariates X_i and Z_i is:

$$E\{w_i(u)|X_i, Z_i\} = \mu^w(u) + \sum_{k=1}^{\infty} E\{A^w_{i,k}|X_i, Z_i\}\phi_k(u), \tag{5.5}$$

and

$$E\{s_i(u)|X_i, Z_i\} = \mu^s(u) + \sum_{k=1}^{\infty} E\{A^s_{i,k}|X_i, Z_i\}\psi_k(u). \tag{5.6}$$

Assuming that we have a fixed set of basis functions ϕ and ψ for the amplitude and the phase variation respectively, the scores A_i act as surrogate data for curve y_i. The final joint model for amplitude, phase and phone duration is then formulated as:

$$E\{[A^w_{i,k}, A^s_{i,m}, T_i]|X_i, Z_i\} = X_iB + Z_i\Gamma, \qquad \Gamma \sim \mathcal{N}(0, \Sigma_\Gamma) \tag{5.7}$$

where Γ is assumed to have mean zero and Σ_Γ to be the covariance matrix of the amplitude, phase and duration components with respect to the random effects.

5.3.2 Amplitude modelling

In our study, amplitude analysis is conducted through a functional principal component analysis of the amplitude variation functions in an analogous way to chapter 4. Qualitatively, the w_i are the time-registered versions of the original F_0 samples. Utilizing FPCA, we identify the principal modes of amplitude variation in the sample and use those modes as a basis to project our data to a finite subspace by imposing a finite truncation point on the number of basis terms. Specifically, we define the kernel C^w

of the covariance operator C^w as:

$$C^w(u, u^*) = E\{(w_i(u) - \mu^w(u))\,(w_i(u^*) - \mu^w(u^*))\} \tag{5.8}$$

and by Mercer's theorem[207], the spectral decomposition of the symmetric amplitude covariance function C^w can be written as:

$$C^w(u, u^*) = \sum_{p_w=1}^{\infty} \lambda_{p_w} \phi_{p_w}(u) \phi_{p_w}(u^*), \tag{5.9}$$

where ϕ is treated as the FPCA-generated empirical basis of the amplitude variation functions. Additionally, the eigenvalues λ_{p_w} allow the determination of the total percentage of variation exhibited by the sample along the p-th principal component and show whether the examined component is relevant for further analysis. As shown in chapter 4, the choice of the number of components is based on acoustic criteria [302; 29] with direct interpretation for the data, such that components which are not audible are not considered. Having fixed M_w as the number of ϕ modes / functional principal components (FPC's) to retain, we use ϕ to compute $A^w_{i,p}$, the amplitude projections scores associated with the i-th sample and its p-th corresponding component (Eq. 5.10) as:

$$A^w_{i,p} = \int \{w_i(u) - \mu^w(u)\}\phi_p(u)dt, \qquad \text{where as before } \mu^w(u) = E\{w(u)\} \tag{5.10}$$

where a suitable numerical approximation to the integral is used for practical analysis.

5.3.3 Phase modelling

When examining the warping functions it is important to note that we expect the mean of warping function to correspond to the identity (ie. the case of no warping). Therefore, assuming their domains are all normalized to [0,1], with $t = Th(u)$:

$$u = E\{h(u)\}, \text{where under certain circumstances: } \approx E\{h^{-1}(u)\}, \tag{5.11}$$

and we therefore interpret the deviations from this equality as phase distortions. This clearly also applies conceptually when working with the function $s(u)$. As with the amplitude analysis, phase analysis is carried out using a principal component analysis approach. Utilizing the FPC's of the s_i, we identify the principal modes of variation of the sample and use those modes as a basis to project our data to a finite subspace. Directly analogous to the decomposition of C^w, the spectral decomposition of the phase covariance function C^s is:

$$C^s(u, u^*) = \sum_{p_s=1}^{\infty} \lambda_{p_s} \psi_{p_s}(u) \psi_{p_s}(u^*), \tag{5.12}$$

where $\psi(t)$ is the FPCA-generated empirical basis of the phase variation functions. As in the case of amplitude modelling, the eigenvalues λ_{p_s} allow the determination of the total percentage of variation exhibited by the sample along the p-th mode and help us determine the relevance of the component. As before we will base our selection processes not on an arbitrary threshold based on percentages but on acoustic perceptual criteria [246; 150] for perceivable speed changes. Fixing M_s as the number of ψ modes / functional principal components to retain, we use $\psi(u)$ to compute $A^s_{i,p}$, the phase projections scores associated with the i-th sample and its p-th corresponding component (Eq. 5.13) as:

$$A^s_{i,p} = \int \{h_i(u) - \mu^s(u)\}\psi_p(u)dt, \qquad \text{where as before } \mu^s(u) = E\{s(u)\} \tag{5.13}$$

It is worth stressing that our choice of the number of components to retain will be naturally determined by the phonetic application rather than using a purely statistical criterion. Purely data driven approaches have been developed [208] as well as a number of different heuristics [50] if preferred in another application where no natural choice is available.

5.3.4 Sample Time-registration

The estimation of the phase variation/warping functions is based on the methodology of Tang & Müller, as implemented in the routine WFPCA in PACE [304] as it is presented in section 3.2.2. There, one defines the pairwise warping function $g_{k,i}(t) = h_k(h_i^{-1}(t))$ as the 1-to-1 mapping from the i-th curve time-scale to that of the k-th. The inverse of the average $g_{k,i}(\cdot)$ (Eq. 5.14) for a curve i is then defined as the curve y_i's corresponding warping function h_i. h_i is therefore a map between individual-specific warped time to absolute time [304].

$$\hat{h}_i^{-1}(t) = \frac{1}{m}\sum_{k=1}^{m} \hat{g}_{k,i}(t), \quad m \leq N \tag{5.14}$$

The actual sample registration was conducted using two knots. It also focused on warping together F_0 utterances from the same tonal category (tone 1 curves with other tone 1 curves, tone 2 curves with other tone 2 curves, etc.). Once again we did employ the concept of rhyme time: that is the curves were projected in a common time-grid $T \in [0,1]$. As noted earlier though their durations where recorded so we can reconstruct them in their physical domain.

5.3.5 Compositional representation of warping functions

Two caveats need to be addressed when working with warping functions. First the warping functions themselves do not define a vector space [27]. Second, despite the fact that h_i is treated as a function, in real terms one works with step function approximations to the warping functions h_i, thus being subjected to the limitation stemming from the fact that there is a discrete grid over which the values y_i are recorded. This motivates viewing the warping functions as instances of compositional data. In particular, by compositional data one refers to data where sample space is formed by a positive simplex:

$$\mathbb{S}^d = \{(x_1, \ldots, x_d) : x_j > 0 (j = 1, \ldots, d), x_1 + \cdots + x_d < 1 \tag{5.15}$$

or in a more general form $\sum_{j=1}^{d} x_j = \mathbb{K}$, where $\mathbb{K}$ is some arbitrary constant.

Therefore when examining a warping function h_i the differences in levels between adjacent steps give rise to a histogram that represents the discretised warping function; the function h_i being qualitative similar to a cumulative distribution function [340]. The values of this histogram are then naturally assumed to define a simplex. This is where the connection to the compositional decomposition comes into play as the $\sum \frac{d}{dt} h_i = \mathbb{K}$. Specifically, based on standard compositional data methodology (centred log-ratio transform)[2], the first difference Δ_h of a discretised instance of h_i over an $(m+1)$-dimensional grid is used to evaluate s_i as:

$$s_i = \log \frac{\Delta h_{i,j}}{(\Delta h_{i,1} \cdot \Delta h_{i,2} \cdots \Delta h_{i,m})^{\frac{1}{m}}} \quad j = \{1, \ldots, m\} \tag{5.16}$$

the reverse transformation being:

$$h_i = \sum_j \frac{e^{s_{i,j}}}{\sum_j e^{s_{i,j}}}. \tag{5.17}$$

This ensures that the monotonicity ($h_{i,j} < h_{i,j+1}$) and boundary requirements ($h_{i,1} = 0$, $h_{i,m+1} = 1$) are fulfilled as put in place by the time-registration step. In functional terms it yields the discretised version of Eq. 5.4.

Qualitatively when we employ the centred log-ratio transform for the analysis of the compositional data, we essentially divide the values of Δh_i by their geometric means and then take their logarithms (Eq. 5.16). This allows our sample to have a vector space structure with the standard operations of

perturbation, powering and inner product defined respectively as:

$$z = x \oplus y = C[x_1 \dot{y}_1, \ldots, x_d \dot{y}_d], \tag{5.18}$$

$$z = \lambda \odot x = C[x_1^{\lambda}, \ldots, x_d^{\lambda}], \tag{5.19}$$

$$z = \langle x, y \rangle = \frac{1}{D} \sum_{i>j}^{D} \log \frac{x_i}{x_j} \log \frac{y_i}{y_j} \Leftrightarrow \tag{5.20}$$

$$clr(x) clr^T(y). \tag{5.21}$$

The centred log-ratio transform used here has been the established method of choice for the variational analysis of compositional data; alternative methods such as the additive log-ratio [2] or the isometric log-ratio [74] are also popular choices. In particular, the centred log-ratio, as it sums the transformed components to zero by definition, presents itself as directly interpretable in terms of "time-distortion", negative values reflecting deceleration and positive values acceleration in the relative phase dynamics. Clearly this summation constraint imposes a certain degree of collinearity in our transformed sample [85]; nevertheless it is the most popular choice of compositional data transformation prior to PCA [3; 4] and allows direct interpretation as mentioned above.

5.3.6 Further details on mixed effects modelling

Given an amplitude-variation function w_i, its corresponding phase-variation function s_i and the original F_0 curve duration T_i, each sample curve is mapped on a $M_w + M_s + 1$ vector space of partially dependent measurements. Here, M_w is the number of functional principal components encapsulating amplitude variations, M_s is the number of functional principal components carrying phase information and the 1 refers to the curves' duration. The final discretised form of our generative model for a given sample curve y_i of duration T_i and sets of scalar covariates X_i and Z_i in a mixed effect form is:

$$[A_{i,k}^w, A_{i,m}^s, T_i] = X_i B + Z_i \Gamma + E_i, \qquad \Gamma \sim \mathcal{N}(0, \Sigma_\Gamma), \quad E \sim \mathcal{N}(0, \Sigma_E) \tag{5.22}$$

Σ_E being the diagonal matrix of measurement error variances (Eq. A.3). The covariance structures Σ_Γ and Σ_E are of particular forms; while Σ_E (Eq. A.3) assumes independent measurements errors, the random effects covariance Σ_Γ (Eq. A.2) allows a more complex covariance pattern. In particular, Σ_Γ is assumed to have a highly structured dependency pattern between the amplitude, phase and duration. As a result and in line with previous work [117], we assume independence between FPC projections of the same group (ie. the amplitude FPC's are orthogonal/uncorrelated among themselves due to their PCA construction, the same also being true for the phase FPC's). On the contrary, the amplitude FPC projection scores are not uncorrelated with those of phase and neither projection family is uncorrelated with the duration measurements (Eq. A.2). The choice of an unstructured covariance Σ_Γ for the random effects is necessary; we have found no theoretical or empirical evidence to believe any particular structure such as a compound symmetric covariance structure, for example, is present within the FPC's and/or duration. Nevertheless our framework would still be directly applicable if we chose another restricted covariance (eg. compound symmetry) structure and if anything it will become computationally easier to investigate as the number of parameters would decrease. As mentioned earlier the main purpose of the structure Σ_Γ is account for variation that is not phonetically systematic but a by-product of data generation procedure.

Our sample curves are concurrently included in two nested structures: one based on "speaker" (non-linguistic) and one based on "sentence" (linguistic) (Figure .5.2). We therefore have a crossed design with respect to the random-effects structure of the sample [11; 42], which suggests the inclusion of random effects (Eq. 5.23):

$$A_{n \times p} = X_{n \times k} B_{k \times p} + Z_{n \times l} \Gamma_{l \times p} + E_{n \times p}. \tag{5.23}$$

This generalization allows the formulation of the conditional estimates as:

$$A|\Gamma \sim \mathcal{N}(XB + Z\Gamma, \Sigma_E) \tag{5.24}$$

or unconditionally and in vector form for $\overrightarrow{A}$ as:

$$\overrightarrow{A}_{np\times 1} \sim \mathcal{N}((I_p \otimes X)\overrightarrow{B}_{np\times 1}, \Lambda_{np\times np}), \tag{5.25}$$

$$\Lambda = (I_p \otimes Z)(\Sigma_\Gamma \otimes I_l)(I_p \otimes Z)^T + (\Sigma_E \otimes I_n) \tag{5.26}$$

where X is the matrix of fixed effects covariates, B the matrix of fixed effects coefficients, Z the matrix of random effects covariates, Γ the matrix of random effects coefficients (a sample realization dictated by $N(0, \Sigma_\Gamma)$), $\Sigma_\Gamma = D_\Gamma^{\frac{1}{2}} P_\Gamma D_\Gamma^{\frac{1}{2}^T}$ the random effects covariance matrix, D_Γ the diagonal matrix holding the individual variances of random effects, P_Γ the correlation matrix of the random effects between the series in columns i, j and Σ_E the diagonal measurement errors covariance matrix. Kronecker products ($\otimes$) are utilized to generate the full covariance matrix Λ of $\overrightarrow{A}$ as the sum of the block covariance matrix for the random effects and the measurement errors.

The main advantage of this approach is two-fold: It is both theoretically consistent in the sense that incorporates prior knowledge of phase variation being present in speech, as well as allows insights on how speech amplitude patterns (what practically is "simplistically" interpreted as speech) is affected by random phase variations.

5.3.7 Estimation

Estimation is required in two stages: generating the warping functions and multivariate mixed effects regression estimation. Requirements for the estimation of pairwise warping functions $g_{k,i}$ were discussed in section 5.3.4. In practical terms these requirements mean that: 1. $g_{k,i}(\cdot)$ needs to span the whole domain, 2. we can not go "back in time", i.e. the function must be monotonic and 3. the time-scale of the sample is the average time-scale followed by the sample curves. With these restrictions in place we can empirically estimate $g_{k,i}(\cdot)$ as $\hat{g}_{k,i}(t) = \text{argmin}_g D(y_k, y_i, g)$ where the "discrepancy" cost function D is defined as:

$$D_\lambda(y_k, y_i, g) = E\{\int_0^1 (y_k(g(t); T_k) - y_i(t; T_i))^2 + \lambda(g(t) - t)^2 dt | y_k, y_i, T_k, T_i\}, \tag{5.27}$$

λ being an empirically evaluated non-negative regularization constant, chosen in a similar way to Tang & Müller [304]; see also Ramsay & Li [256]; T_i and T_k being used to normalized the curve lengths. Intuitively the optimal $g_{k,i}(\cdot)$ minimizes the differences between the reference curve y_i and the "warped" version of f_k, subject to the amount of time-scale distortion produced on the original time scale t by $g_{k,i}(\cdot)$. Having a sufficiently large sample of m pairwise warping functions $g_{k,i}(\cdot)$ for a given reference curve y_i, the empirical internal time-scale for y_i is given by Eq. 5.14, the global warping function h_i being easily obtainable by simple inversion of h_i^{-1}. It is worth noting that in Mandarin, each tone has its own distinct shape; their features are not similar and therefore should not be aligned. For this reason, the curves were warped separately per tone, i.e. realizations of tone 1 curves were warped against other

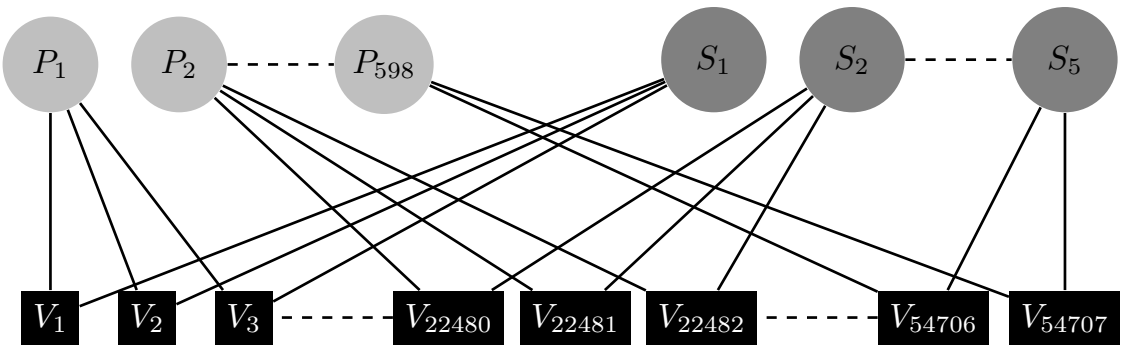

Figure 5.2: The multivariate mixed effects model presented exhibits a crossed (non-balanced) random structure. The vowel-rhyme curves (V) examined are cross-classified by their linguistic (Sentence - P_i) and their non-linguistic characterization (Speaker - S_i).

realizations of tone 1 only, the same being applied to all other four tones.

Finally to estimate the mixed model via the model's likelihood, we observe that usual maximum likelihood (ML) estimation underestimates the model's variance components [234] as mentioned in section 3.4.2. We therefore utilize Restricted Maximum Likelihood (REML); this is essentially equivalent to taking the ML estimates for our mixed model after accounting for the fixed effects X. The restricted maximum (log)likelihood estimates are given by maximizing the following formula:

$$L_{REML}(\theta) = -\frac{1}{2}[p(q-r)\log(2\pi) + \log(|\Psi|) + \overrightarrow{\Omega}^T \Psi^{-1} \overrightarrow{\Omega}] \tag{5.28}$$

where q is the total number of readings $(n * p)$, r the number of fixed effects $(k * p)$, $\Psi = K^T \Lambda K$ and $\Omega = K^T A$; K being the "whitener" matrix such that $0 = K^T(I_p \otimes X)$ [288]. Based on this we are in position to concurrently estimate the random effect covariances while taking into account the possible non-diagonal correlation structure between them. However, because we "remove" the influence of the fixed effects, if we wished to compare models with different fixed effects structures we would need to use ML rather REML estimates. Standard mixed-effects software such as `lme4` [22], `nlme` [239] and `MCMCglmm` [116] either do not allow the kinds of restrictions on the random effects covariance structures that we require, as they are not designed to model multivariate mixed effects models, or computationally are not efficient enough to model a dataset of this size and complexity; we were therefore required to write our own implementation for the evaluation of REML/ML. Exact details about the optimization procedure used to do this are given in the following section.

5.3.8 Multivariate Mixed Effects Models & Computational Considerations

Generalizing from a univariate (Eq. 3.42) to a multivariate mixed effects model is clearly of interest in the case of high dimensional data. There, instead of working with a vector a $(N \times 1)$ one works with matrix A $(N \times p)$ as our dependent variable.

Actual computation of the random effects variances requires a more involved computational approach than maximizing the restricted log-likelihood given in Eq. 5.28 directly; that is because in its straightforward form the estimation of $det(\Psi)$ involves in the case of a multivariate model the Cholesky decomposition of an $np \times np$ full matrix; a very computationally expensive process in terms of computational time and memory. Such an approach does not take advantage of the highly structured nature of K and of the matrices that construct it. To solve this computational issue we use the formulation presented by Bates and DebRoy [21] for evaluating the profiled REML deviance. This means that we optimize not for the variance-covariance and measurement error magnitudes directly but for a ratio between them.

As a result, the vector θ holds $\nu\frac{p(p+1)}{2}$ values, ν being the total number of different random effects structures and p the total number of components in our multivariate MLE. Starting with the model:

$$A = XB + Z\Gamma + E \tag{5.29}$$

where as before A is of dimensionality $(n \times p)$, X is of dimensionality $(n \times k)$, B is of dimensionality $(k \times p)$, Z is of dimensionality $(n \times l)$, Γ is of dimensionality $(l \times p)$ and E is of dimensionality $(n \times p)$. Eq. 5.29 translates in vector notation as:

$$\overrightarrow{A} = (I_p \otimes X)\overrightarrow{B} + (I_p \otimes Z)\overrightarrow{\Gamma} + \overrightarrow{E} \tag{5.30}$$

where we have that:

$$\overrightarrow{E} \sim \mathcal{N}(0, \Sigma_E \otimes I_{n\times n}), \qquad \overrightarrow{\Gamma} \sim \mathcal{N}(0, \Sigma_R \otimes I_l) \tag{5.31}$$

where Σ_E and Σ_R are of dimensions $(p \times p)$; l being the number of levels in the random effects. Significantly Σ_R has a structure that can easily accommodate for sparse patterns of covariance. This structure can be enforced by multiplying the candidate Σ_R by a $0-1$ "boolean matrix" M_{bool} of dimensions $(p \times p)$

that sets to zero all entries not explicitly assumed to be non-zeros by design; effectively updating R as:

$$\Sigma_R^0 = \Sigma_R \circ M_{bool} \tag{5.32}$$

where $\circ$ is the Hadamard product (or Schur product) between two matrices. Imposing in this way can be problematic and lead to non-positive-definite sparse "variance matrices". In such case one might use Tikhonov regularization ($R \leftarrow R + \lambda I,\ \lambda \to 0$); qualitatively this equates with having more noise in the observed values. Following that and given that Σ_R^0 remains a valid covariance matrix, thus being positive definite, it can be expressed as $\Sigma_R^0 = LL^T$ and additionally can be expressed in term of a relative precision factor [238] as:

$$\frac{\Sigma_R^0}{\frac{1}{\sigma^2}} = \Delta\Delta^T. \tag{5.33}$$

We need to draw attention to the fact here that σ^2 can be, and is in our multivariate case of no intrinsic meaning. The variance it expresses is a "sample-wide" variance that does not reflect any single variance of the p dimensions of the model. We can use it nevertheless because of our hypothesis that Σ_E is diagonal, therefore the ratio expressed in Δ can be formulated even if it is only for algorithmic simplicity. Taking this into account, Eq. 5.30 can then be re-written as:

$$\overrightarrow{A} = (I_p \otimes X)\overrightarrow{B} + [(I_p \otimes Z)][\overrightarrow{\Gamma}] + \overrightarrow{E} \tag{5.34}$$

and restate the universal random effects ($pl \times pl$) matrix Σ_{R_U} as:

$$\Sigma_{R_U}^{-1} = SS^T \tag{5.35}$$

where:

$$S = (\Delta \otimes I_l) \tag{5.36}$$

$$\Delta\Delta^T = \frac{\Sigma_R^0}{\frac{1}{\sigma^2}} \tag{5.37}$$

in accordance with the above. We therefore can reformulate our model as the minimization of the following penalized least squares expression:

$$\min_{\Gamma,B} \overrightarrow{A}_{aug} - \Phi(\theta) \begin{bmatrix} \overrightarrow{\Gamma}_{aug} \\ \overrightarrow{B} \end{bmatrix} \tag{5.38}$$

where:

$$A_{aug} = \begin{bmatrix} A \\ 0 \end{bmatrix},\ \overrightarrow{\Gamma}_{aug} = [\overrightarrow{\Gamma}],\ \Phi(\theta) = \begin{bmatrix} Z_{aug} & X_{aug} \\ S(\theta) & 0 \end{bmatrix}, \tag{5.39}$$

$$Z_{aug} = [(I_p \otimes Z)] \text{ and } X_{aug} = (I_p \otimes X) \tag{5.40}$$

A_{aug} being the original $n \times p$ matrix A augmented by a zero $l \times p$ bottom submatrix leading to a final dimensionality of $(n+l) \times p$ and $\Phi(\theta)$ being the augmented model matrix (now of dimensions $p(n+l) \times p(k+l)$). To solve this we form, proceeding analogously to Bates [21], $\Phi_e = [\Phi, \tilde{A}]$ (of dimensionality $p(n+l) \times p(l+k+p)$) and define $R_e^T R_e$ to be the Cholesky decomposition of the $\Phi_e^T \Phi_e$. Thus instead of working with a $(np \times np)$ matrix, we now work with a matrix of dimensions $(p(l+k+p) \times p(l+k+p))$.

In particular, in matrix notion we have the following:

$$\Phi_e^T \Phi_e = \begin{bmatrix} Z_{aug}^T & S(\theta)^T \\ X_{aug}^T & 0 \\ A_{aug}^T & 0 \end{bmatrix} \begin{bmatrix} Z_{aug} & X_{aug} & \overrightarrow{A}_{aug} \\ S(\theta) & 0 & 0 \end{bmatrix} \tag{5.41}$$

$$= \begin{bmatrix} Z_{aug}^T Z_{aug} + \Sigma_{R_U}^{-1} & Z_{aug}^T X_{aug} & Z_{aug}^T \overrightarrow{A}_{aug} \\ X_{aug}^T Z_{aug} & X_{aug}^T X_{aug} & X_{aug}^T \overrightarrow{A}_{aug} \\ \overrightarrow{A}_{aug}^T Z_{aug} & \overrightarrow{A}_{aug}^T X_{aug} & \overrightarrow{A}_{aug}^T \overrightarrow{A}_{aug} \end{bmatrix} \tag{5.42}$$

$$= R_e^T R_e, \text{ where } R_e^T\text{:}$$

$$R_e^T = \begin{bmatrix} R_{ZZ} & R_{ZX} & r_{ZA} \\ 0 & R_{XX} & r_{XA} \\ 0 & 0 & r_{AA} \end{bmatrix} \tag{5.43}$$

where R_{ZZ} and R_{XX} are both upper triangular, non-singular matrices of dimensions $pl \times pl$ and $pk \times pk$ respectively; R_{ZX} is of dimensionality $pl \times pk$. Similarly, r_{ZA}, r_{XA} and r_{AA} are of dimensions $pl \times 1$, $pk \times 1$ and 1×1. As a result the conditional REML estimates for $\overrightarrow{B}$ are given by the solving the following triangular system:

$$R_{XX} \overrightarrow{\hat{B}} = r_{XA}. \tag{5.44}$$

Similarly, we have:

$$\hat{\sigma}^2 = \frac{\underline{r}_{AA}^T \underline{r}_{AA}}{p(n-k)} \tag{5.45}$$

with the profiled log-restricted-likelihood being:

$$-2L_{REML}(\theta) = \log(\frac{|\Phi^T \Phi|}{|\Sigma_{R_U}^{-1}|}) + (p(n-k))[1 + \log(2\pi\hat{\sigma}^2)] \tag{5.46}$$

or the profiled log-likelihood as:

$$-2L_{ML}(\theta) = \log(|\Phi^T \Phi|) + pn[1 + \log(2\pi \frac{\underline{r}_{AA}^T \underline{r}_{AA}}{pn})]. \tag{5.47}$$

Finally, the conditional expected value of Γ is given by the solution of the system:

$$R_{ZZ} \overrightarrow{\hat{\Gamma}}_{aug} = r_{ZA} - R_{ZX} \overrightarrow{\hat{B}} \tag{5.48}$$

and the conditional $\hat{\sigma}_i$, $i = 1, \ldots, p$ for a given component of the original multivariate model equals:

$$\hat{\sigma}_i = \sqrt{\frac{1}{n-k} \Sigma[(\hat{A}_i - A_i)^2 + U_i^2]} \tag{5.49}$$

where U_i is the l-dimensional random vector such that $\hat{A} = X_{aug} \overrightarrow{\hat{B}} + Z_{aug} \Sigma_{R_U} \overrightarrow{U}$ [20]. We briefly note that the current approach of estimation does not examine Bayesian approaches. To that effect after assuming that $\overrightarrow{\hat{B}}$, $\overrightarrow{\hat{\Gamma}}$ and $\overrightarrow{\hat{E}}$ follow a multivariate normal distribution such as :

$$\begin{bmatrix} \overrightarrow{B} \\ \overrightarrow{\Gamma} \\ \overrightarrow{E} \end{bmatrix} \sim N \left(\begin{bmatrix} \overrightarrow{B_0} \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_B & 0 & 0 \\ 0 & \Sigma_R \otimes I_l & 0 \\ 0 & 0 & \Sigma_E \otimes I_n \end{bmatrix} \right) \tag{5.50}$$

where $\overrightarrow{B_0}$ is the prior means for the fixed effects with prior covariance function Σ_B and $\Sigma_R \otimes I_l$ and $\Sigma_E \otimes I_n$ are the expected covariances of the random effects and residuals respectively, the parameters of

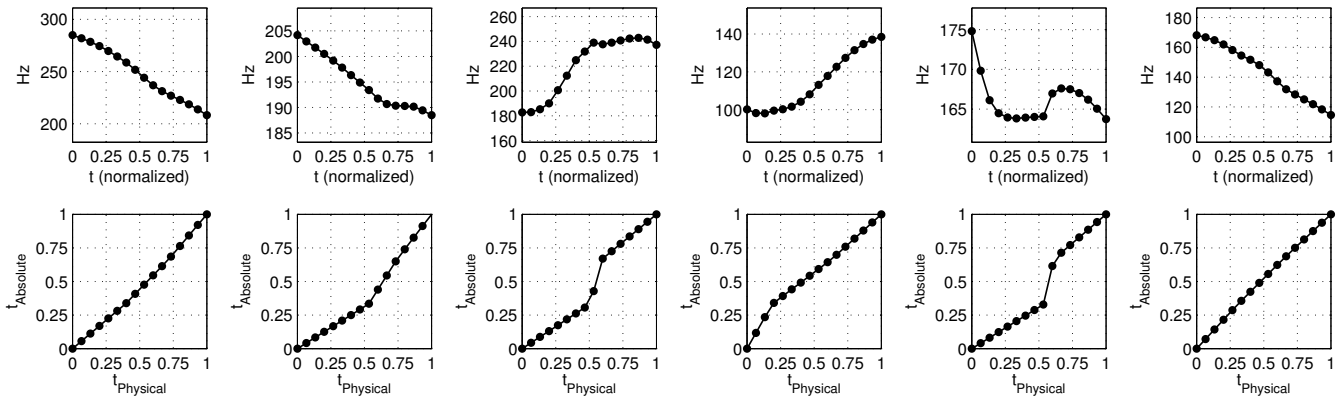

Figure 5.3: Corresponding amplitude variation functions w (top row) and phase variation functions h (bottom row) for the triplets shown in Fig. 5.1.

the mixed model ($\overrightarrow{B}$ and $\overrightarrow{\Gamma}$) can be sampled using Gibbs sampling in an EM methodology to evaluate the respective likelihood. This method being presented by Garcia et al. in [94] and being successfully utilized in the package `R MCMCglmm` by Hadfield[116].

5.4 Data Analysis and Results

5.4.1 Model Presentation & Fitting

We use the smoothing procedure from section 4.2 and note each curve's original time duration (T_i) so that it can be used within the modelling. At this point the F_0 curve sample is not time-registered but has been smoothed and interpolated to lay on a common grid. We register our dataset following the procedure in section 5.3.7. Then using the exposition outlined in Eq. 5.7, the following model is proposed, as it accounts for all the linguistic effects that might be present in a dataset of this form [117]:

$$\begin{aligned} Component_X = \{&[tn_{previous} * tn_{current} * tn_{next}] + [cn_{previous} * tn_{current} * cn_{next}] + \\ &[(B2) + (B2)^2 + (B2)^3 + (B3) + (B3)^2 + (B3)^3 + (B4) + (B4)^2 + \\ &(B4)^3 + (B5) + (B5)^2 + (B5)^3] * Sex + [rhyme_t]\}\beta + \\ &\{[Sentence] + [SpkrID]\}\gamma + \epsilon. \end{aligned} \tag{5.51}$$

The same notation as in chapter 4 is used. First examining the fixed effects structure, we incorporate the presence of tone-triplets and of consonant:tone:consonant interactions, using the same rationale with the previous chapter where both types of three-way interactions are included. We also look at break counts, our only covariate that is not categorical. A break's duration and strength significantly affects the shape of the F_0 contour and not just within a rhyme but also across phrases. Break counts are allowed to exhibit squared and cubic patterns as cubic downdrift has been previously observed in Mandarin studies [11; 117]. We also model breaks as interacting with the speaker's sex since we want to provide the flexibility of having different curvature declination patterns among male and female speakers. This partially alleviates the need to incorporate a random slope as well as a random intercept in our mixed model's random structure. The final fixed effect we examine is the type of rhyme uttered. Each rhyme consists of a vowel and a final -n/ -ŋ if present; rhyme types are the single most linguistically relevant predictors for the shape of F_0's curve as when combined together they form words; words carrying semantic meaning. Examining the random effects structure we incorporate speaker and sentence. The inclusion of speaker as a random effect is justified as factors of age, health, neck physiology and emotional condition affect a speaker's utterance and are mostly immeasurable but still rather "subject-specific". Additionally we incorporate Sentence as a random effect since it is known that pitch variation is associated with the utterance context (eg. commands have a different F_0 trajectory than questions). We need to note that we do not test for the statistical significance of our random effects; we assume they are "given" as any linguistically relevant model has to include them. However if one wished to assess the statistical

	Amplitude/(w)	Phase/(s)
FPC_1	88.67 (88.67)	49.40 (49.40)
FPC_2	10.16 (98.82)	19.25 (68.65)
FPC_3	0.75 (99.57)	9.02 (77.68)
FPC_4	0.22 (99.80)	6.53 (84.19)
FPC_5	0.10 (99.90)	4.34 (88.53)
FPC_6	0.05 (99.94)	2.98 (91.51)
FPC_7	0.02 (99.97)	2.32 (93.83)
FPC_8	0.01 (99.98)	1.96 (95.79)
FPC_9	0.01 (99.99)	1.29 (97.08)

Table 5.1: Percentage of variances reflected from each respective FPC (first 9 shown). Cumulative variance in parenthesis.

	Amplitude/(w)
FPC_1	121.16(121.16)
FPC_2	66.52 (187.68)
FPC_3	31.22 (218.90)
FPC_4	17.50 (236.40)
FPC_5	9.00 (245.39)
FPC_6	4.86 (250.26)
FPC_7	3.64 (253.90)
FPC_8	2.71 (256.61)
FPC_9	1.96 (258.56)

Table 5.2: Actual deviations in Hz from each respective FPC (first 9 shown). Cumulative deviance in parenthesis. (human speech auditory sensitivity threshold ≈ 10 Hz)

relevance of their inclusion, the χ^2 mixtures framework utilized by Lindquist et al. [194] provides an accessible approach to such a high-dimensional problem, as re-sampling approaches (bootstrapping) are computationally too expensive in a dataset of the size considered here; Wei and Zhou having focused on the same problem from an Information Criterion point of view [322]. Fixed effects comparisons are more straightforward; assuming a given random-effects structure, AIC-based methodology can be directly applied [108]. Fitting the models entails maximizing REML of model (Eq. 5.28).

Our findings can be grouped into three main categories, those from the amplitude analysis, those from the phase and those from the joint part of the model. Some examples of the curves produced by the curve registration step are given in Figure 5.3. However, overall, as can be seen in Figure 5.4, there is a good correspondence between the model estimates and the observed data in its original domain when the complete modelling setup is considered.

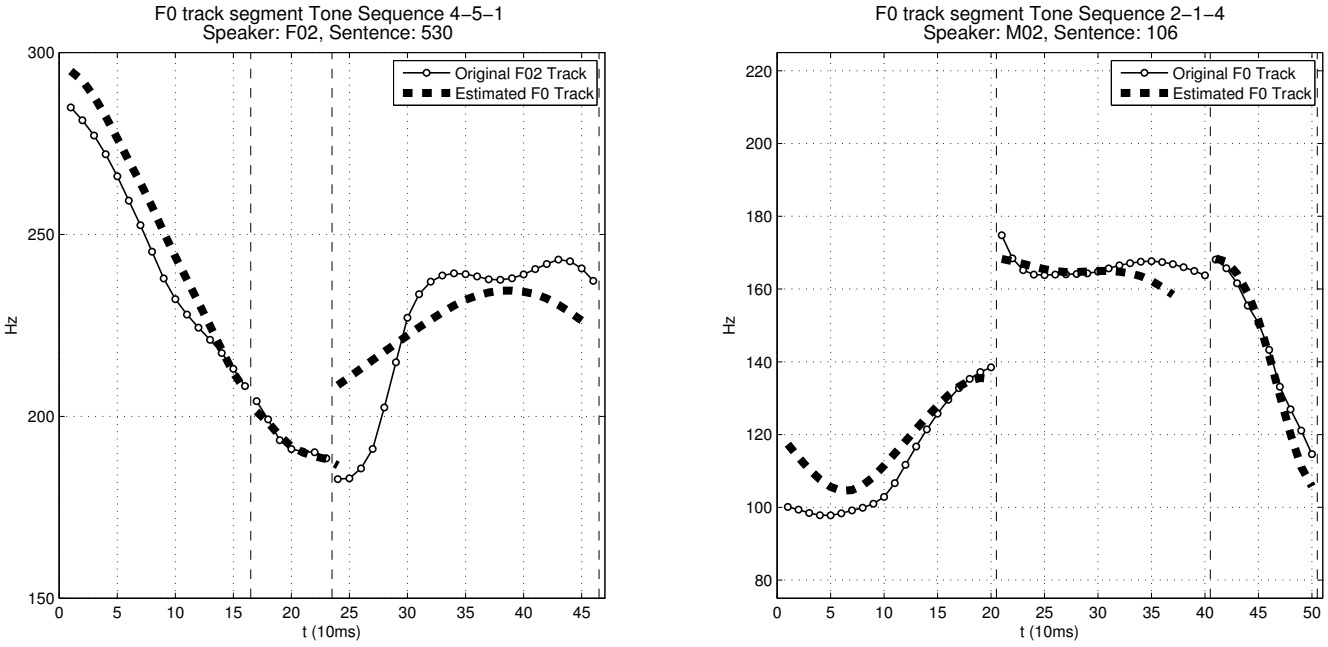

Figure 5.4: Functional estimates (continuous curves) are shown superimposed on the corresponding original discretised speaker data over the physical time domain t.

Empirical findings from the amplitude FPCA: The first question one asks when applying any form of dimensionality reduction is how many dimensions to retain, or more specifically in the case of FPCA how many components to use. We take the same perceptual approach as in chapter 4. Instead of using

an arbitrary percentage of variation, we calculate the minimum variation in Hz each FPC can actually exhibit (Tables 5.1-5.2). Based on the notion of Just Noticeable Differences (JND) [45] we use for further analysis only FPC's that reflect variation that is actually detectable by a standard speaker (F_0 JND: ≈10 Hz; $M_w = 4$). The empirical wFPC's (Figure 5.5) correspond morphologically to known Mandarin tonal structures (Figure 2.4) increasing our confidence in the model. Looking into the analogy between components and reference tones with more detail, $wFPC_1$ corresponds closely to Tone 1, $wFPC_2$ can be easily associated with the shape of Tones 2 and 4 and $wFPC_3$ corresponds to the U-shaped structure shown in Tone 3. $wFPC_4$ appears to exhibit a sinusoid pattern that can be justified as necessary when moving between different tones in certain tonal configurations [117]. The amplitude FPCs derived during the application of FPCA correspond well (if not identically in terms of qualitative characteristics) to the FPCs found during the FPCA step conducted in chapter 4. This being partially expected as ultimately the variational patterns of phonetic sample analysed as mostly due to amplitude variation, therefore one would not expect the lack of registration in Chapt. 4 to lead into significant changes of the main modes of variation in this dataset.

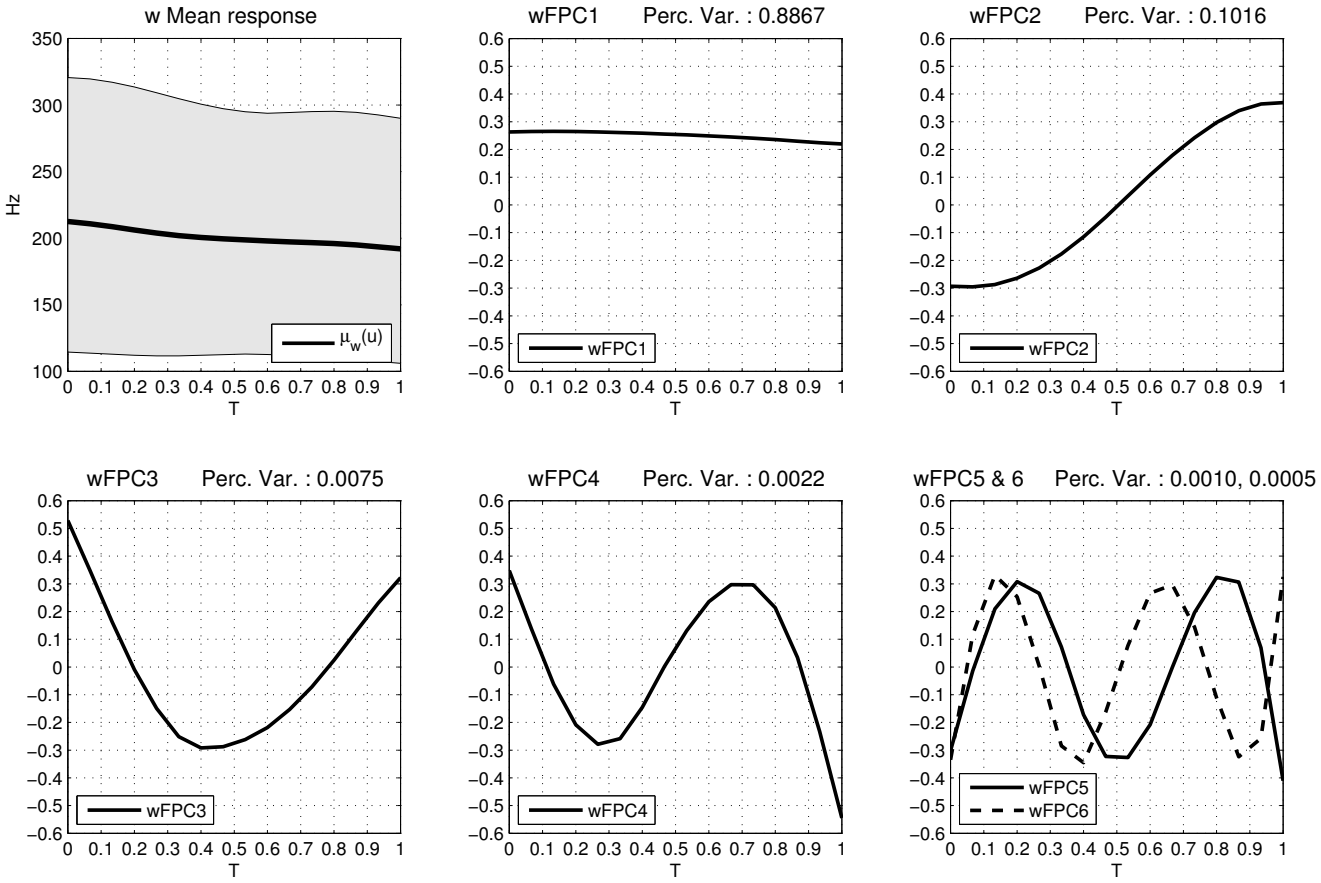

Figure 5.5: W (Amplitude) Functional Principal Components Φ: Mean function ([.05,.95] percentiles shown in grey) and 1st, 2nd, 3rd, 4th, 5th, and 6th functional principal components of amplitude.

Empirical findings from the phase FPCA: Again the first question is how many components to retain. Based on existing Just Noticeable Differences in tempo studies [246], [150], we opt to follow their methodology for choosing the number of "relevant" components (tempo JND: ≈ 5% relative distortion; $M_s = 4$). We focus on percentage changes on the transformed domain over the original phase domain as it is preferable to conduct Principal Component analysis [3]; $sFPC$'s also corresponding to "standard patterns" (Figure 5.6). $sFPC_1$ and $sFPC_2$ exhibit a typical variation one would expect for slow starts and/or trailing phone utterances where a decelerated start leads to an accelerated ending of the word - a catch-up effect- and vice versa. $sFPC_3$ and $sFPC_4$ on the hand show more complex variation patterns that are most probably rhyme specific (eg. ia) or associated with uncommon sequences (eg. silent pause followed by a Tone 3) and do not have an obvious universal interpretation. While the curves in Figure 5.6 are not particularly smooth due to the discretised nature of the modelling, as can be seen in Figure A.7 in the Appendix, the resulting warping functions after transformation are smooth. However it should be

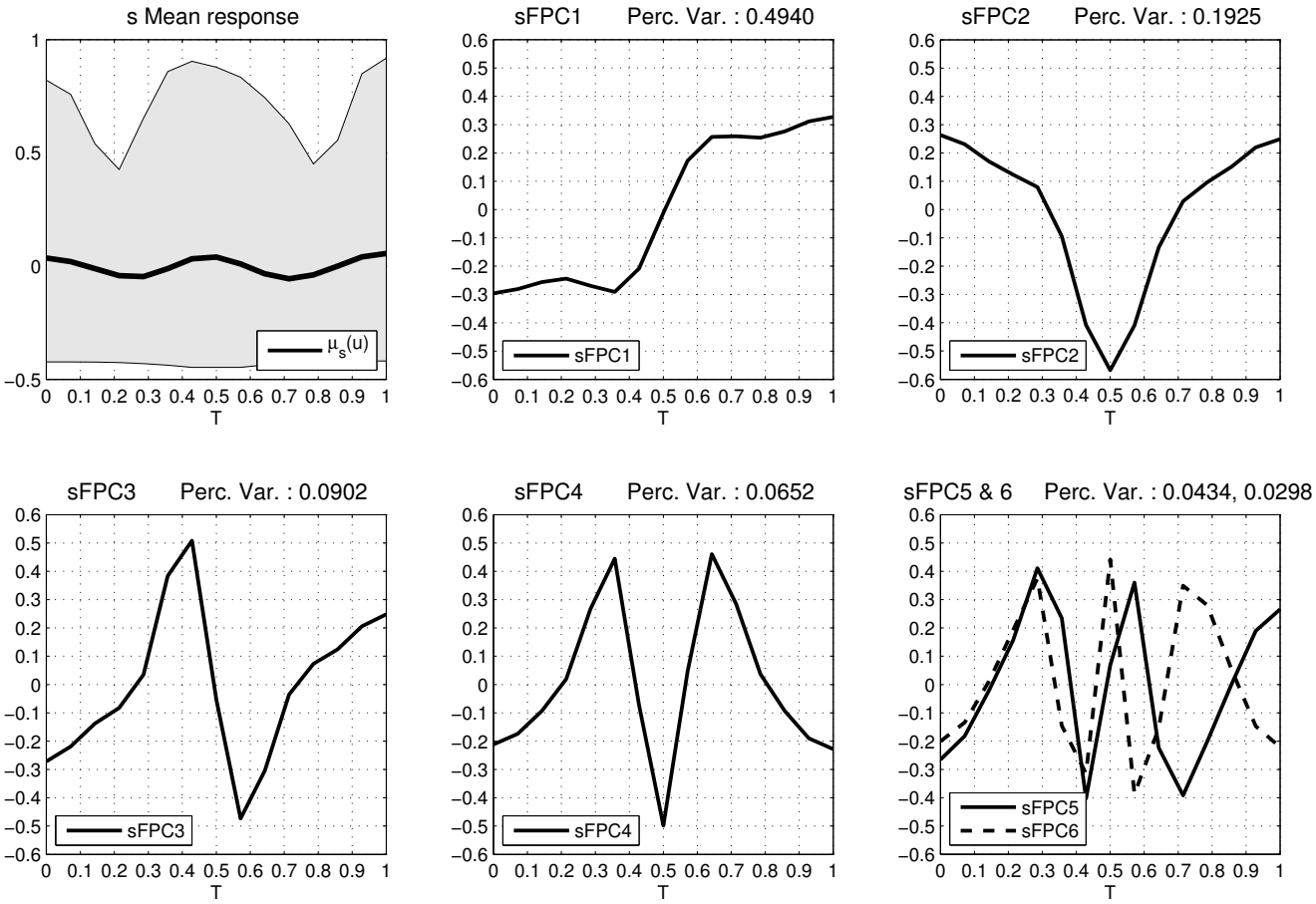

Figure 5.6: H (Phase) Functional Principal Components Ψ: Mean function ([.05,.95] percentiles shown in grey) and 1st, 2nd, 3rd, 4th, 5th, and 6th functional principal components of phase. Roughness is due to differentiation and finite grid; the corresponding warping functions in their original domain are given in Figure A.7 in the Appendix.

noted that the curves in Figure A.7 cannot be combined linearly whereas those from Figure 5.6 can.

Empirical findings from the MVLME analysis: The most important joint findings are the correlation patterns presented in the covariance structures of the random effects as well as their variance amplitudes. A striking phenomenon is the small, in comparison with the residual amplitude, amplitudes of the Sentence effects (Table 5.3). This goes to show that pitch as a whole is much more speaker dependent than context dependent. It also emphasizes why certain pitch modelling algorithms focus on the simulations of "neck physiology"[89; 305; 196]. In addition to that we see some linguistically relevant correlation patterns in Figure 5.7 (see also A.4-A.5 in the Appendix). For example, $wFPC_2$ and duration are highly correlated both in the context of Speaker and Sentence related variation. The shape of the second wFPC is mostly associated with linguistic properties [117] and a phone's duration is a linguistically relevant property itself. As $wFPC_2$ is mostly associated with the slope of phone's F_0 trajectory, it is unsurprising that changes in the slope affect the duration. Moreover, looking at the signs we see that while the Speaker influence is negative, in the case of Sentence, it is positive. That means that there is a balance on how variable the length of an utterance can be in order to remain comprehensible (so for example when a speaker tends to talk more slowly than normal, the effect of the Sentence will be to "accelerate" the pronunciation of the words in this case). In relation to that, in the speaker random effect, $sFPC_1$ is also correlated with duration as well as $wFPC_2$; yielding a triplet of associated variables. Looking specifically to another phase component, $sFPC_2$ indicating mid phone acceleration or deceleration that allow for changes in the overall pitch patterns, is associated with a phone's duration, this being easily interpreted by the face that such changes are modulated by alterations in the duration of the phone itself. Complementary to these phenomena is the relation between the phone duration and $wFPC_1$ sentence related variation. This correlation does not appear in the speaker effects and thus is

[1]See Sect. 5.3.8 for Σ_{R_i}'s definitions.

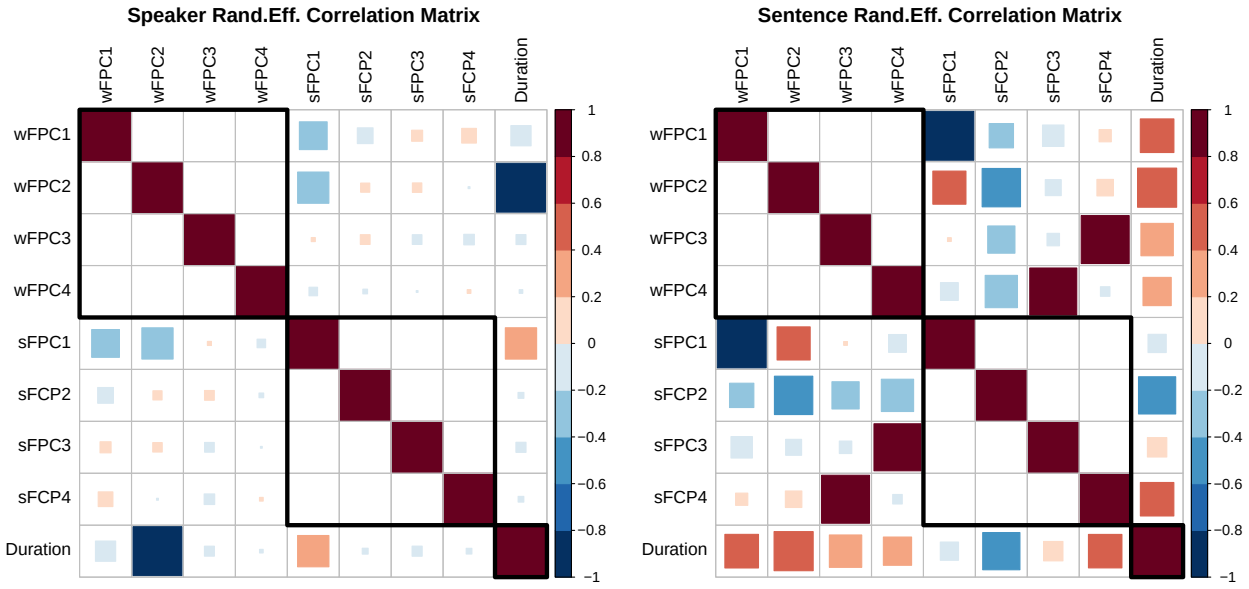

Figure 5.7: Random Effects Correlation Matrices. The estimated correlation between the variables of the original multivariate model (Eq. 5.23) is calculated by rescaling the variance-covariance submatrices Σ_{R_1} and Σ_{R_2}[1]of Σ_Γ to unit variances. Each cell i, j shows the correlation between the variance of component in row i and that of column j; Row/Columns 1-4 : $wFPC_{1-4}$, Row/Columns 5-8 : $sFPC_{1-4}$, Row/Columns 9 : Duration.

likely due to more linguistic rather than physiological changes in the sample. As mentioned previously, $wFPC_1$ can be thought of as dictating pitch-level placement, and the correlation implies that higher-pitched utterances tend to last longer. This is not contrary to the previous finding; higher F_0 placements are necessary for a speaker to utter a more pronounced slope differential and obviously need more time to be manifested.

Interestingly a number of lower magnitude correlation effects appear to associate $wFPC_1$ and sFPC's. This is something that needs careful interpretation. $wFPC_1$ is essentially "flat" (Figure 5.5, upper middle panel), and as such cannot be easily interpreted when combined with registration functions. Nevertheless this shows the value in our joint modelling approach for these data. We concurrently account for all these correlations during model estimation and, as such, our estimates are less influenced by artefacts in individual univariate FPC's.

Estimate	$wFPC_1$	$wFPC_2$	$wFPC_3$	$wFPC_4$	Duration
Speaker	89.245	6.326	3.655	1.330	2.806
Sentence	38.674	4.059	0.045	0.102	0.043
Residual	114.062	44.386	15.399	10.072	4.481

Estimate	$sFPC_1$	$sFPC_2$	$sFPC_3$	$sFPC_4$
Speaker	0.289	0.023	0.022	0.030
Sentence	0.049	0.043	0.042	0.043
Residual	0.959	0.591	0.431	0.370

Table 5.3: Random effects std. deviations.

Examining the influence of fixed effects[2], the presence of adjacent consonants was an important feature for almost every component in the model. Additionally certain "domain-specific" fixed effects emerged too. The syllable's rhyme type appeared to significantly affect duration; the break-point information appeared to influence the amplitude of the F_0 curve and specific consonant-vowel-consonant (C-V-C) triplets to play a major role for phase. Phase also appeared to be related to the rhyme types

[2]Table of $\hat{B}$ and associated standard errors available in `https://tinyurl.com/COSPRO-Betas`

but to a lesser extent.

More specifically regarding duration of the F_0 curve, certain rhyme types (eg. oN, iEn) gave prominent elongation effects while others (eg. u, ɿ) were associated with shorter curves. The same pattern of variability in the duration was associated with the adjacent consonants information; when a vowel was followed by a consonant the F_0 curve was usually longer while when the consonant preceded a vowel the F_0 curve was shorter. Amplitude related components are significantly affected by the utterances' break-type information; particularly B2 and B3 break types. This is not a surprising finding; a pitch trajectory, in order to exhibit the well-established presence of "down-drift" effects [89], needs to be associated with such variables. As in the case of duration, the presence of adjacent consonants affects the amplitude dynamics. Irrespective of its type (voiced or unvoiced), the presence of consonant before or after a rhyme led to an "overall lowering" of the F_0 trajectory. Tone type and the sex of the speaker also influenced the dynamics of amplitude but to a lesser degree. Finally, examining phase it is interesting that most phase variation was mainly due to the adjacent consonants and the rhyme type of the syllable; these also being the covariates affecting duration. This confirms the intuition that as both duration and phase reflect temporal information, they would likely be affected by the same covariates. More specifically, a short or a silent pause at the edge of rhyme caused that edge to appear decelerated, while the presence of a consonant caused that edge to be accelerated. As before, certain rhymes (eg. a, ai) gave more pronounced deceleration-acceleration effects. Tone types, while very important in the case of univariate models for amplitude [117], did not appear significant in this analysis individually; they were usually significant when examined as Vowel-Consonant or Consonant-Vowel pairs. However, this again illustrates the importance of considering joint models versus marginal models, as it allows a more comprehensive understanding of the nature of covariate effects.

5.5 Discussion

Linguistically our work establishes the fact that when trying to make a typology of a language's pitch one needs to take care of amplitude and phase covariance patterns while correcting for linguistic (Sentence) and non-linguistic (Speaker) effects. This need was prominently presented by the strong correlation patterns observed (Figure 5.7). Clearly we do not have independent components in our model and therefore a joint model is appropriate. This has an obvious theoretical advantage in comparison to standard linguistic modelling approaches such as MOMEL [135] or the Fujisaki model [212; 89] where despite the use of splines to model amplitude variation, phase variation is ignored.

Focusing on the interpretation of our results, it is evident that the covariance between phase and amplitude is mostly due to non-linguistic (Speaker-related) rather than linguistic features (Sentence-related). This is also reflected in the dynamics of duration, where the Speaker related influence is also the greatest. Our work as a whole presents a first coherent statistical analysis of pitch incorporating phase, duration and amplitude modelling into a single overall approach.

The obvious technical caveats with this work stem from three main areas: the discretisation procedure, the time-registration procedure and the multivariate mixed effects regression. Focusing on the discretisation, the choice of basis is of fundamental importance. While we used principal components for the reasons mentioned above, there have been questions as to whether a residual sum of squares optimality is most appropriate. It is certainly an open question when it comes to application specific cases [43]. Aside from the case of parametric bases, non parametric basis function generation procedures such as ICA [145] have recently become increasingly more prominent. These bases could be used in the analysis, although the subsequent modelling of the scores would become inherently more complex due to the lack of certain orthogonality assumptions.

Regarding time-registration, there are a number of open questions regarding the choice of the framework to be used. Aside from the pairwise alignment framework we employ [304], as mentioned in section 3.2 we have identified at least two alternative approaches based on different metrics: the square-root velocity function metric [175] and the area under the curve normalization metric [340], that can be used interchangeably, depending on the properties of the warping that are most important. Indeed it has been seen that considering warping and amplitude functions together, based on the square-root velocity metric, can be useful for classification problems [314]. However, we need to stress that each method makes some explicit assumptions to overcome the non-identifiability between the h_i and w_i (Eq.

5.1) and this can lead to significantly different final estimates. Nevertheless, we have reimplemented the main part of the analysis using the AUC methodology of Zhang & Müller [340] (results shown in Appendix, Sect. A.14) and while the registration functions obtained are different, the analysis resulted in almost identical insights for the linguistic roles of w_i and s_i, again emphasising the need to consider a joint model as well as the generality of this approach. The choice of the time-registration framework ultimately relies on the theoretical assumptions one is willing to make and the nature of the sample registered. For this work it is not unreasonable to assume that the pairwise alignment corresponds well to the intuitive belief that intrinsically humans have a "reference" utterance where they "map" what they hear in order to comprehend it [24].

Finally, multivariate mixed effects regression is itself an area with many possibilities. Optimization for such models is not always trivial and as the model and/or the sample size increases, estimation of the model tends to get computationally expensive. In our case we used a hybrid optimization procedure that changes between a simplex algorithm (Nelder-Mead) and a quasi-Newton one (Broyden-Fletcher-Goldfarb-Shanno (BFGS)) [161]; in recent years research regarding the optimization tasks in an LME model has tended to focus on derivative free procedures. In a related issue, the choice of covariance structure is of importance, while we chose a very flexible covariance structure, the choice of covariance can convey important experimental insights. A related modelling approach is that of Zhou et al. concerning paired functional data [341], the dimensionality of their application problem is though smaller and their regression problem more parametrized as they operate in a reduced rank framework. A final note specific to our problem was the presence of only five speakers. Speaker effect is prominent in many components and appears influential despite the small number of speakers available; nevertheless we recognize that including more speakers would be certainly beneficial if they had been available. Given that the Speaker effect was the most important random-effect factor of this study, the inclusion of random slopes might also have been of interest [283; 18]. Nevertheless, the inclusion of generic linear, quadratic and cubic gender-specific down-drift effects presented through the break components allow substantial model flexibility to avoid potential design-driven misspecification of the random effects, and as such random slopes were not included.

In conclusion, a comprehensive modelling framework was proposed for the analysis of phonetic information in its original domain of collection, via the joint analysis of phase, amplitude and duration information. The models are interpretable due to the LME structure, and estimable in a standard Euclidean domain via the compositional transform of the warping functions. The resulting model provides estimates and ultimately a typography of the shape, distortion and duration of tonal patterns and effects in one of the world's major languages.

Chapter 6

Phylogenetic analysis of Romance languages

6.1 Introduction

With the increased availability of computational resources the number and quality of evolutionary trees is increasing rapidly both in Biology [201; 192] and in Linguistics [105; 206]. However, knowing evolutionary relationships through Phylogenetics is only one step in understanding the evolution of their characteristics [336]. Three issues are particularly challenging. The first is limited information: empirical information is typically only available for extant taxa, represented by tips or leaves of a phylogenetic tree, whereas evolutionary questions frequently concern unobserved ancestors deeper in the tree. The second is dependence: the available information for different organisms in a phylogeny is not independent since a phylogeny describes a complex pattern of non-independence; observed variation is a mixture of this inherited and taxon-specific variation [56]. The third is high dimensionality: the emerging literature of biological function-valued traits [167; 307; 298] recognizes that many characteristics of living organisms are best represented as a continuous function rather than a single factor or a small number of correlated factors. This is a slowly emerging trend also in Linguistics [307]. Evolutionary phonetics research has so far focused both on binary and discrete characteristics of a language [224], as well as continuous multivariate ones [320]. Function-valued linguistic traits [307; 114; 112], however, have not been investigated yet in the context of phylogenetic evolution.

In the case of biological characteristics such characteristics include growth or mortality curves [242], reaction-norms [166] and distributions [340], where the increasing ease of genome sequencing has greatly expanded the range of species in which distributions of gene [219] or predicted protein [170] properties are available. On the other side, in the case of Linguistics and more specifically Phonetics and Acoustics, such characteristics can be lip motions, phonation patterns, $F0$ curves or even syllable spectrograms. Using these formulations, a function-valued trait in a phylogeny (irrespective of the phylogeny's type) is defined as a phenotypic trait that can be represented by a continuous mathematical function [166] in one or more dimensions.

Previous work [157] proposed an evolutionary model for function-valued data y related by a phylogeny $\mathbf{T}$. The data are regarded as observations of a phylogenetic Gaussian Process (PGP) at the leaves of $\mathbf{T}$. That work shows that a PGP can be expressed as a stochastic linear operator Q on a fixed set ϕ of basis functions (independent components of variation), so that

$$y = Q_{\mathbf{T}}\phi. \tag{6.1}$$

However, that study does not address the linear inverse problem of obtaining estimates $\hat{\phi}$ and $\hat{Q}$ of ϕ and Q nor how to recover a tree if one is unavailable. Our first contribution in this work is to provide an approach to address these problems in sections 6.2.1 and 6.2.2 via the use of functional principal components analysis (FPCA [121]) and phylogenetic Gaussian process regression (PGPR) respectively.

Hadjipantelis el al. [119] have shown FPCA to work successfully in the case of one-dimensional functional data as curves. We here extend this work in the case of two-dimensional functional data, eg. spectrograms based on two-dimensional techniques that are directly analogous to their one-dimensional

counterparts [19; 164]. As a pre-processing step we also smooth, interpolate and time-warp the sample at hand in order to account for possible noise corruption, uneven signal sizes and phase variation respectively; these steps are described in detail in section 6.2.1. Given this projection framework one refers to Q as the *mixing matrix*, and to the (i, j)th entry of Q as the *mixing coefficient* of the jth basis function at the ith taxon in the tree. It is these mixing coefficients that we model as evolving. For each fixed value of j, the Q_{ij} are correlated (due to phylogeny) as i varies over the taxa. On the contrary the spectrogram basis functions themselves do not evolve in our model and are assumed constant across all stages of a language's evolutionary history.

In section 6.2.2 we focus on the obvious problem inherited in every phylogenetic study: tree reconstruction [82]. While most biological studies focus on either the tree-estimation [186] or the ancestral reconstruction task [127; 119], linguistic phylogenies have far from widely established phylogenies [206; 12]. For that reason given a basic phylogenetic linking relation as this is shown in Fig. 2.6, we optimize using the likelihood of a prespecified evolutionary forward model as a fitting criterion and the data at the leaves as our "evidence". We then construct the "ML-optimal" tree for these leaf-readings and branching relations. It must be noted that our tree-reconstruction and model will be based upon a consensus tree [82] between the ten digits used. The basic methodology behind this step is outlined in section 6.2.2.

In section 6.2.3, we address the problem of estimating the statistical structure of the mixing coefficients by performing phylogenetic Gaussian process regression (PGPR) on the mixing coefficient found in $\hat{Q}$ of each of the jth bases separately; this work essentially applying the work presented in [119] but for a much smaller tree. This corresponds to assuming orthogonality between the rows (i.e. that the coefficients of the different basis functions evolve independently). Given it is commonly argued in the quantitative genetics literature [46] that evolutionary processes can be modelled as Ornstein-Uhlenbeck (O-U) processes, the estimation of the forward operator is reduced to the estimation of a small vector θ of parameters [157]. This model is the O-U model used in section 6.2.2 to estimate the "most likely" trees.

Finally in section 6.3 we clarify the interpretation of these parameters θ in linguistic evolutionary contexts. The estimation of θ is known to be a challenging statistical problem [23]; nevertheless the explicit PGPR posterior likelihood function is used to obtain maximum likelihood (MLE) estimates for θ. In contrast with [119] we can not employ *bagging* [41] due to the small number of taxa (languages) we work with. Recent work from Bouchard et al. [38] also addresses this problem by employing a resampling technique. With lack of a better alternative we report the MLE estimate directly based on multiple initializations of the initial solution θ_0, with θ_0 itself being more constrained (having one less free hyperparameter) in comparison with the θ used in [119]. Clearly, as we utilize "just" 10 words from 5 languages, the robustness of the produced estimates will be sub-optimal but they should still be insightful for the evolutionary dynamics of the languages examined.

The PGPR step also returns a posterior distribution for the mixing coefficient of each basis function at each ancestral taxon in the phylogeny. At any particular ancestor (protolanguage) the estimated basis functions can be therefore combined statistically using the posterior distributions of their respective mixing coefficients, to provide a two-dimensional function-valued posterior distribution. Since the univariate posterior distributions of mixing coefficients are Gaussian, and the mixing is linear, the posterior for the function-valued trait has a closed form representation as a Gaussian process (Eq. 6.25) which provides a major analytical and computational advantage for the approach.

We close this chapter by commenting on the phonetic properties of the ancestral estimates as well as the general insights provided by this phylogenetic analysis. Overall, our methods (sections: 6.2.1, 6.2.2 & 6.2.3), and results (Sect. 6.3) appropriately combine developments in functional data analysis with the evolutionary dynamics of quantitative phenotypic traits.

6.2 Methods & Implementations

6.2.1 Sample preprocessing & dimension reduction

As shown in section 2.1, spectrograms can be assumed to provide a full two-dimensional characterization of a syllable's phonetic properties within the limitation of their physical characteristic (sampling frequency, window length and type). Here we utilize them under the assumption that all possible phonetic characteristics of syllables, starting with the zeroth harmonic F_0 and going all the way up to higher

formants [1] (eg. F_2 or F_3) assumed to be of importance in Indo-european languages [97], are reflected in those syllables' spectrograms.

The original acoustic dataset was first resampled at 16Khz; using that the spectrograms were computed by using a window length of 10ms. This resulted into a window size of 160 readings per frame. Because we used a 16 Khz sampling rate, our maximal effective frequency detected is 8Khz, the Nyquist frequency of our sampling procedure. A Gaussian window was used during windowing of each frame. The original power spectral density is shown after a $10\log_{10}(\cdot)$ transform so it is depicted in decibels (dB).

Despite having an otherwise perfectly balanced grid with no missing values, we can not exclude instances of noise corruption because of the rather heterogeneous sample quality as well as the non-laboratory recording conditions during the sample's generation. For this reason we employ a penalized least squares filtering technique for grid data [93] which is based on the discrete cosine transformation in two dimensions; this is in contrast with our work in the previous sections where we used a kernel smoother. Here because we wanted to keep our implementation fast and efficient we chose a parametric basis for our data. The basic idea behind this parametric assumption stems from the use of Eq. 3.11 as a smoother. We see that effectively the smoothed data are the projections of the original data in another domain. Choosing to penalize the roughness of our data by the use of their second-order difference (their second derivative in the case of functional data), Eq. 3.11 can be re-expressed as a penalized regression system of the form:

$$(I + sB^T B)\hat{y} = y \tag{6.2}$$

where s corresponds to the smoothing parameter used, B to the second order differencing matrix and as always I is the identity matrix. The tridiagonal square matrix B being defined as:

$$B_{i,i-1} = \frac{-2}{r_{i-1}(r_{i-1}+r_i)}, B_{i,i} = \frac{2}{r_{i-1}r_i}, B_{i-1,i} = \frac{-2}{r_i(r_{i-1}+r_i)} \tag{6.3}$$

for $2 \leq i \leq N-1$ where N is the number of elements in $\hat{y}$ and r_i represents the step between $\hat{y}_i$ and $\hat{y}_{i+1}$. Assuming repeating border elements ($y_0 = y_1$ and $y_{N+1} = y_N$) then: $B_{1,1} = -B_{1,2} = r_1^{-2}$ and

[1] F_0 is *not* a formant as it does not refer to acoustic resonance.

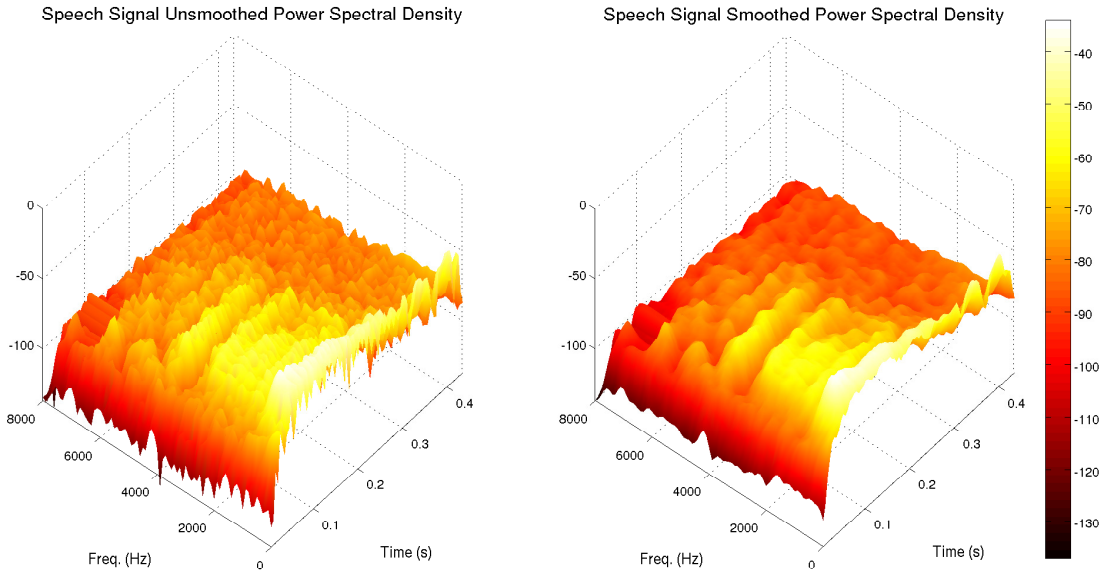

Figure 6.1: Unsmoothed and smoothed spectrogram of a male Portuguese speaker saying "un" (ũ(ŋ)). It is immediately evident that throughout all frequencies there is small-scale unstructured variation that the smoothing algorithm filters out.

$-B_{N,N-1} = B_{N,N} = r_{N-1}^{-2}$. When if $r_i = 1$ for $i = 1, \ldots, N$ matrix B is of the form:

$$B = \begin{bmatrix} 1 & -1 & 0 & \cdots & \cdots & 0 \\ -1 & 2 & -1 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & -1 & 2 & -1 \\ 0 & \cdots & \cdots & 0 & -1 & 1 \end{bmatrix} \tag{6.4}$$

Obviously if $s \to 0$ no smoothing takes places as one retrieves the original signal directly and if $s \to \infty$ one just recovers the second order polynomial fit to the data [324]. Given that B has an eigendecomposition of $B = U\Lambda U^T$, Λ being the diagonal matrix with the eigenvalues of B, Eq. 6.2 can be rewritten as:

$$\hat{y} = U(I + s\Lambda^2)^{-1}U^T y. \tag{6.5}$$

The computational efficiency of this approach comes from the realization that as Garcia presents: "*U^T and U are actually n-by-n type-2 discrete cosine transform (DCT) and inverse DCT matrices, respectively*"[93], the orthogonal form of type-2 DCT kernel matrix being:

$$[C_2]_{i,j} = \sqrt{\frac{2}{N}}\xi(i)cos(\frac{i(j+\frac{1}{2})\pi}{N}), \quad i,j = 0,1, \ldots, (N-1) \tag{6.6}$$

$$\xi(p) = \begin{cases} \sqrt{\frac{1}{2}} & \text{if } p = 0 \text{ or } p = \text{N}, \\ 1 & \text{if } p = 1,2, \ldots, N-1 \end{cases} \tag{6.7}$$

and thus resulting in the equation:

$$\hat{y} = [C_2^{-1}]((I + s\Lambda^2)^{-1}[C_2]y). \tag{6.8}$$

Then taking advantage of the known eigenvalues formulas for tridiagonal matrices like B [339], $(I + s\Lambda^2)$ can also be rewritten as $1 + s(2 - 2cos((i-1)\pi/n))^2$ where i corresponds to the i-th eigenvalue of the original matrix B. We then define $\Gamma = (I + s\Lambda^2)^{-1} = diag([1 + s(2 - 2cos((i-1)\pi/n))^2]^{-1})$ giving the final estimate of y as:

$$\hat{y} = [C_2^{-1}](\Gamma[C_2]y). \tag{6.9}$$

One can immediately see the computational efficiency of Garcia's algorithm compared to standardized smoothing techniques as well as compared to standard matrix decompositions. Especially in regards with this second claim, even the most "efficient" matrix decomposition for the solution of a least squares problem, the Cholesky decomposition is of $\frac{1}{3}n^3$ order complexity [225], while the 2-D DCT [2] (and IDCT) is of the order $n^2 \log(n)$ [297], yielding significant speed-ups even for small datasets. Finally while Garcia advocates the use of generalized cross-validation for a choice of s, the current implementation used $s = 0.5$, this value being determined by qualitatively examining the resulting smoothed spectrograms. The generalization of this technique to the two-dimensional object employs simply the two-dimensional DCT instead of the one-dimensional, the two-dimensional DCT being especially popular as it is the back-bone of the well-known JPEG format [272] for digital pictures. Finally after smoothing is conducted, the sample is interpolated over a common time grid assumed to represent "word time".

Two important caveats need to be mentioned: First, using Garcia's method we enforce a discrete transformation on functional data. Second, this smoothing methodology is based on the theoretical assumption that a function is periodic and extends outside the domain over which it is observed. The

[2]The two-dimensional DCT takes the one-dimensional DCT of each column followed by a one-dimensional DCT of each row of the resulting matrix [30].

first caveat, is an oversimplification that as mentioned is done for the sake of computational efficiency. It cannot hide the fact though that higher order fluctuations might be truncated as only 64 two-dimensional basis functions are used. What can be argued though is that given the relatively small sample from which we want to draw conclusions, the choice of 64, highly informative in the case of two-dimensional patters, basis is not limiting the insights behind our analysis; it does not "meaningfully" exclude information. The second caveat concerns the theoretical foundations of this type-2 DCT smoothing framework and is more ambiguous. In standard periodic signals the assumption of "extending outside the observable domain" might be non-restrictive one; in the current case though and especially when examining a frequency continuum where the concept of negative values is a highly not trivial one conceptually (assuming that one can interpreter "negative time" as going *back in time*), this approach can be questionable. Countering this second caveat is based on dynamics of the physical system we investigate. In the case of frequencies, one has practically no fluctuations below a very low threshold. Frequencies below 20Hz are effectively out of our vocal range. Thus assuming that the border of "zero-th" fluctuations extends "in negative frequencies" does not meaningfully alter the boundary condition we employ. These two caveats were made not cancel the efficiency or the elegance behind Garcia's method of smoothing, they were done because one should not naively move methodologies from a discrete domain to a continuous one; if he chooses to do so, he must be able to offer a meaningful interpretation of the assumptions imposed.

In addition to noise distortions, as mentioned earlier, phase distortions are almost certain to exist in any acoustic signal. Here using spectrograms as our acoustic signal units of analysis, we are presented with two-dimensional instead of one-dimensional objects. While in general in a two-dimensional object phase variation cannot be assumed to influence a single dimension exclusively, under specific circumstances all variation can be assumed to occur along a single "relevant axis". In particular when one focuses on the analysis of spectrograms, an inherently two-dimensional object over a frequency and a time axis, phase variations are relevant only in the context of time; frequency can be assumed to occur in absolute time as the phonation procedure of speaker affects only the timing of the sound excitation and not the amplitude of it (at least directly). One can therefore reformulate the original pairwise warping criterion from simple one-dimensional objects as curves (as in chapter 5 where pairwise curve synchronization was utilized) to slightly more complex two dimensional objects. Assuming $y_i(t,f)$ and $y_k(t,f)$ being two spectrograms with an equal size of frequency index, their "discrepancy" cost function D' is:

$$D'_\lambda(y_k, y_i, g') = \\ E\{\int_{f=0}^{F_{Nyq}} \int_{t=0}^{1} (y_k(g'(t), f; T_k) - y_i(t, f; T_i))^2 + \lambda(g'(t) - t)^2 dt df | y_k, y_i, T_k, T_i\}, \tag{6.10}$$

or in its discretised version:

$$D'_\lambda(y_k, y_i, g') = E\{\sum_{f=0}^{r} \sum_{t=0}^{1} (y_k(g'(t), f; T_k) - y_i(t, f; T_i))^2 + \lambda(g'(t) - t)^2 | y_k, y_i, T_k, T_i\}, \tag{6.11}$$

where as in 3.2.2, λ is an empirically evaluated non-negative regularization constant, T_i and T_k are used to normalize the spectrograms time lengths and $g'_{k,i}(\cdot)$ is the pairwise warping function mapping the time evolution of $y_i(t,f)$ to that of $y_k(t,f)$. Thus we are led to the one-dimensional reformulation of the cost function D' as:

$$D'_\lambda(y_k, y_i, g') = E\{\sum_{t=0}^{1} (\overrightarrow{y_k}(g'_r(t); T_k) - \overrightarrow{y_i}(t; T_i))^2 + \lambda(g'(t) - t)^2 dt | \overrightarrow{y_k}, \overrightarrow{y_i}, T_k, T_i\}, \tag{6.12}$$

where $\overrightarrow{y_k}$ is the concatenated across frequencies vectorized form of the spectrogram y_k and g'_r is the version of the pairwise warping function mapping $g'_{k,i}(\cdot)$ repeated r times, r being the number of discrete points along the frequency axis f. This ultimately being a two-dimensional version of Eq. 5.27. Thus similar to the one-dimensional case of the pairwise warping curves, Eq. 3.21 is used to recover the final warping function by taking advantage of the Law of Large numbers; giving a two-dimensional version of the pairwise synchronization framework presented in Sect. 3.2.2. Fig. 6.2 shows the subtle changes warping induces to a spectrogram's structure in our dataset. With the completion of this step we are presented

with 219 smoothed and warped spectrograms. Importantly the warping itself was done within *digit* and *gender* clusters. That means that the speakers of different genders uttering a specific digit (irrespective of their language) had their utterances time-registered only among themselves. We made this choice for two reasons: first, we know from previous findings that intonation dynamics differ significantly between speakers of opposite sexes [117], second, we also know that registration of completely unrelated data will produce spurious results; for example, the word "un" ([ɛ̃]) and the word "quatro" (['kwatro]) (French for one and Spanish for four respectively) will exhibit different inclination patterns and the time-registration procedure will fail to recognize meaningful similarities to exploit. For modelling purposes the "word time" T was represented by a vector of 100 equi-spaced values between 0 and 1.

Spectrograms are almost by definition objects with a complex internal structure; as instances of functional data they appear as two-dimensional functions of time and frequency. While it is possible to directly work in this function-space for computational efficiency and conceptual conciseness given a dataset y of function-valued traits as shown in Eq. 6.1, we would like to find appropriate estimates $\hat{Q}$ and $\hat{\phi}$ of the mixing matrix Q and the basis set ϕ respectively. The first task is to identify a good linear subspace S of the space of all continuous functions by choosing basis functions appropriately. Evidently these basis functions in the case of spectrogram data will be two dimensional. The purpose of this task is to work, not with the function-valued data directly, but with their projections in S. As formalized in Sect. 3.3 we may say that the chosen subspace S is good if the projected data approximate the original data well while the number of basis functions is not unnecessarily large, so that S has the "effective" dimension of the data. The warped spectrograms W, as in previous sections, are assumed to be adequately expressed as:

$$W_i(u,f) = \mu^W(u,f) + \sum_{k=1}^{\infty} A_{i,k}\phi_k(u,f), \text{ where: } \mu^W(u,f) = E\{W(u,f)\} \tag{6.13}$$

where as before $u \in [0,1]$ is the absolute time-scale the spectrograms are assumed to evolve in and f is the frequency domain (here modelled as the domain between 0 and $8Kz$ in $100Hz$ intervals).

Before applying FPCA to our sample we recognize that the ultimate goal of this work is to provide *language-specific* descriptions; scalar estimates that can be utilized within the context of a phylogenetic tree. Additionally we know that *digit-wide* FPC's would be unrealistic as they would combine non-comparable variation patterns, and that it would be beneficial to incorporate the minimum

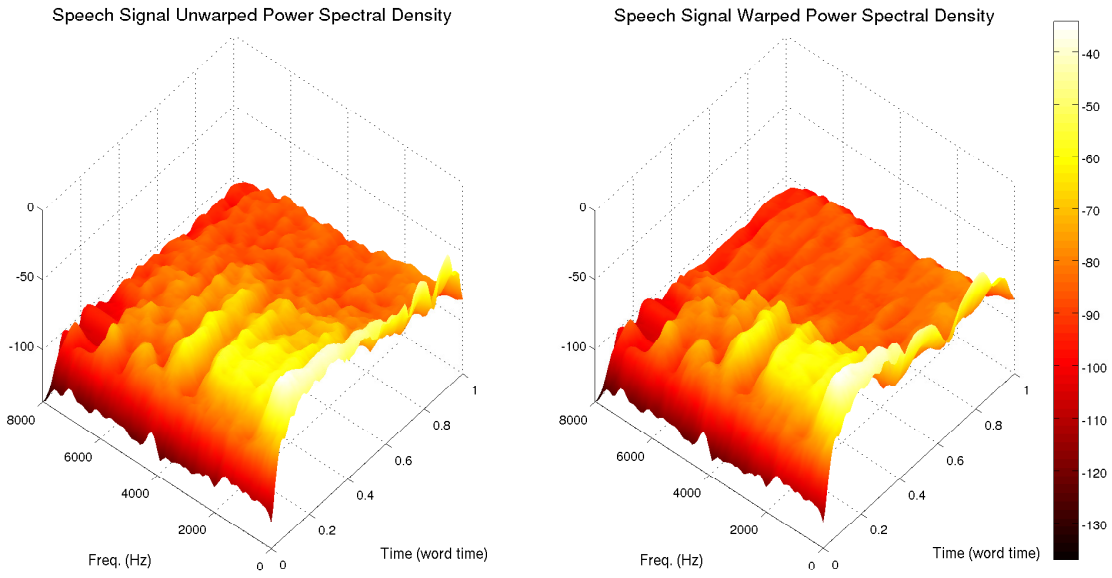

Figure 6.2: Unwarped and warped spectrogram of a male Portuguese speaker saying "un" (ũ(ŋ)). Notice how the warped instance of the spectrogram is registered on a universal "word time" rather than absolute time; ridges among formant frequencies appear more prominently.

prior knowledge that the sex of the speaker has at least "some influence" in the phonetic characteristic encoded by the spectrogram. In a manner similar with section 5.3, given W^d, the spectrograms for a given digit d one formulates:

$$E\{w_i^d(u,f)|X_i^d\} = \mu^{w,d}(u,f) + \sum_{k=1}^{\infty} E\{A_{i,k}^d|X_i^d\}\phi_k^d(u,f). \tag{6.14}$$

Given the structure of our data, we use a fixed effect rather than a mixed effect model to account of speaker variation within a given language l. The reason for this design choice is that we do not have enough speaker realizations to provide meaningful estimates in certain cases. For example, we have a single male speaker in Spanish and in Portuguese; a random effects model could not meaningfully decompose the variation due to the sex of the speaker and the variation due to speaker's unique characteristic. Taking that into account, our final estimates for the *language-specific* FPC scores $\beta_0^{d,l}$ are given by the *language-gender* interaction model:

$$E\{A_{i,k}^{d,l}|X_i^{d,l}\} = X_i^{d,l}\beta^{d,l} \tag{6.15}$$

where $(\beta^{d,l})^T = [\beta_0^{d,l_1}, \beta_0^{d,l_2}, \beta_0^{d,l_3}, \beta_0^{d,l_4}, \beta_0^{d,l_5}, \beta_1^{d,l_1}, \beta_1^{d,l_2}, \beta_1^{d,l_3}, \beta_1^{d,l_4}, \beta_1^{d,l_5}]$ and the design matrix $X_i^{d,l}$ is simply a $n \times m$ indicator matrix, where n equals the number of all speakers uttering digit d and m equals $2 * 5$, such that:

$$X = \begin{bmatrix} \delta_{l_1} & \dots & \delta_{l_5} & \delta_{l_1}^{sex} & \dots & \delta_{l_5}^{sex} \end{bmatrix} \tag{6.16}$$

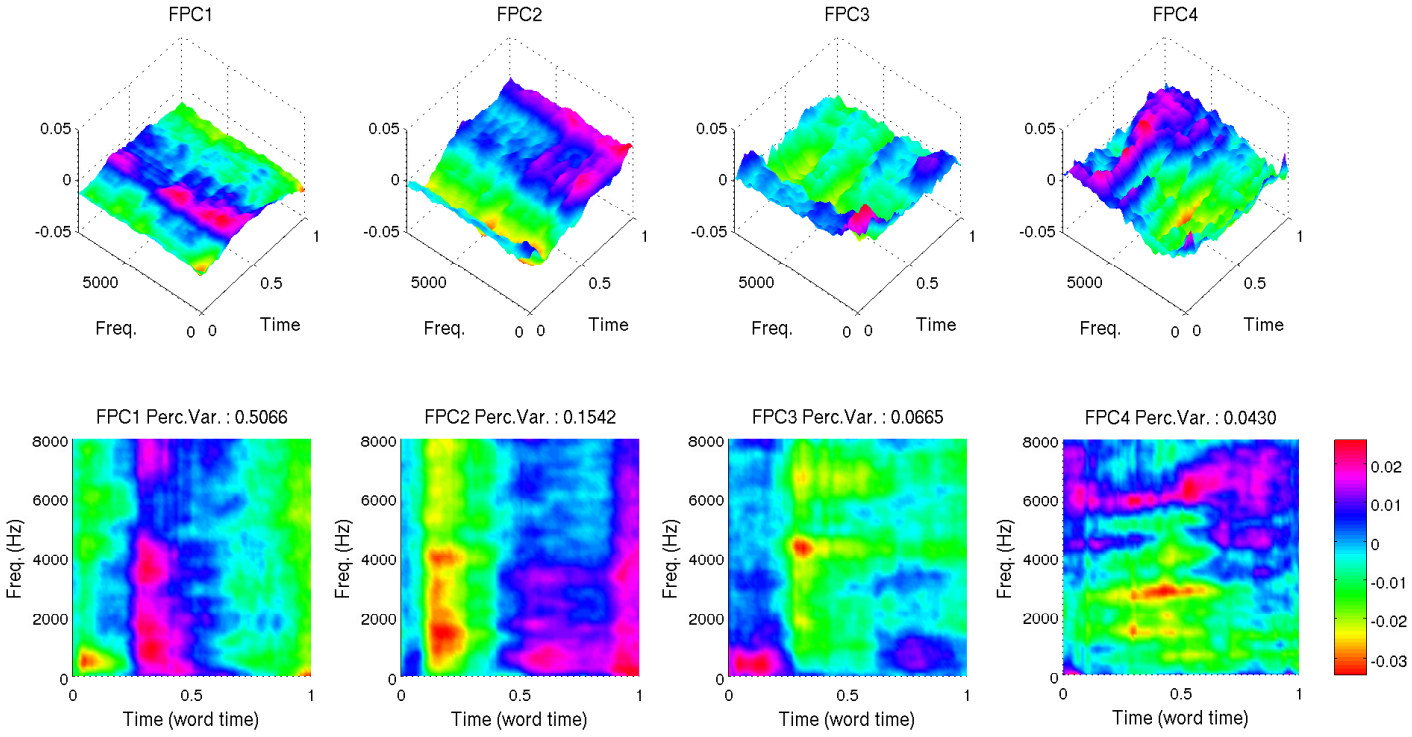

Figure 6.3: Functional Principal Components for the digit *one* spectrograms. Two different views are shown. Top row shows the viewing angle from a $(-50, 50)$ azimuth rotation and vertical elevation; bottom row shows the viewing angle from a $(0, 90)$ viewpoint (completely top to bottom). It is immediately seen that the majority of variation is encapsulated by the first two FPC's. Mean spectrogram shown in Appendix, Fig.A.11.

where the column vectors δ_{l_i} and $\delta_{l_i}^{sex}$ are defined respectively as:

$$\delta_{l_i} \begin{cases} 1 & \text{for language } l_i, \\ 0 & \text{otherwise} \end{cases} \tag{6.17}$$

and

$$\delta_{l_i}^{sex} = \begin{cases} 1 & \text{for language } l_i \text{ iff} \\ & \text{the speaker is male,} \\ 0 & \text{otherwise} \end{cases} \tag{6.18}$$

where $i = \{1, 2, 3, 4, 5\}$ corresponds to each of the five languages represented in the current tree. In that way using these averaged scores [3] we are offered effective representatives of language l for a specific digit d investigated by combining all our digit-specific readings. This allows us to create in a way "language exemplars" scores (β_0^{d,l_i}), these scores being the ones used for the phylogenetic analysis. Clearly one could also construct "language exemplar" spectrograms[4], that could be then utilized to compare the final protolanguage against. Notably, given the design matrix X used, the protolanguage estimates will correspond to female speakers, as the male gender effects should be encapsulated in the $\beta_{1,k}^{d,l}$ term that is not carried forward in the analysis. Examining these artificial spectrograms, interestingly American Spanish appear to downplay the effect of the second vowel in their utterance compared to the other "two-vowel" languages, while on the contrary French (somewhat expectedly given the phonation of "un" in French) have a strong almost singly peak-like spectrogram. The actual interpretation of the first three FPC's is almost obvious; the first FPC captures the variation due to the phonation of the first vowel present in the utterances of digit *one*[5]. Even if another vowel exists (as in the cases of Italian and Spanish), that vowel is not as strongly stressed as the first one; it is therefore expected that the major point of variance will be at the beginning of the word. This finding is in accordance with the finding of the previous chapters where in all cases the beginning of a syllable exhibits greater influence in the syllables dynamics than the other parts. The second and the third FPC's encapsulate the presence of the second vowel. They reflect a phonation event occurring in the second half of the word utterance. It can be also argued upon investigating the second FPC's shape, that it partially compliments the first FPC; it allows the difference between the two vowels to come forward more strongly. In similar but less pronounced manner, the third FPC also compliments the first FPC but in a more localized manner; the highly localized frequency drop in the amplitude of the third FPC, occurring approximately in the center of the word's first half, is counter-balanced by an overall amplitudal increase in the lower frequencies of the word's half. For the fourth FPC it could be argued that the long ridge exhibited approximately at the 6KHz band is a speaker specific construct. One would not expect phylogenetically attributed phonetic variation in that range as it is highly speaker dependent, in the sense that this might be due to a specific speaker's dynamics or (more worryingly) to speaker specific recording equipment. For that reason we do not examine higher order FPC's. As seen in Table 6.1 these components exhibit variation that is rather small in percentage terms and taking into account that we have "just" 22 instances of the digit *one*, it is not reasonable to believe these FPC's generalized to sample-wide variation patterns. In particular individual variation reflected by each FPC quickly falls below a value ($1/22 \approx 4.5\%$) that could be attributed to a variational pattern present to a single spectrogram.

6.2.2 Tree Estimation

As first mentioned in section 2.4 we begin with an unrooted linguistic phylogenetic tree $\mathbf{T}$ which has arbitrary branch-lengths where only the branching events are set. As previous work has commented [119], branch length distributions are surprisingly consistent across organisms [318]; with that in mind, we make the assumption that the same effect is prominent in a linguistic phylogeny. Utilizing the scalar mixing coefficients associated with each FPC in $\hat{\phi}(u, f)$ one treats these coefficients as "the data at the tips". One then constructs a maximum likelihood consensus tree. In particular, based on the work of Hansen [124] that was later popularized by the work of Butler and King [46], the evolutionary

[3] See Appendix, Table A.12 for actual values.

[4] See Appendix, Fig. A.13. $E\{w^{d,l}\} = \mu^{w,d}(u, f) + \sum_{k=1}^{\infty} \beta_{0,k}^{d,l} \phi_k^d(u, f)$

[5] In IPA these are encoded as: [tũ(ŋ)] in Portuguese, ['u:no] in Italian, ['u:no] in Spanish and [ɛ̃] in French.

FPC #	Individ. Variation	Cumul. Variation	FPC #	Individ. Variation	Cumul. Variation
FPC_1	50.66	50.66	FPC_7	2.18	85.67
FPC_2	16.42	66.08	FPC_8	1.82	87.50
FPC_3	6.65	72.74	FPC_9	1.71	89.20
FPC_4	4.30	77.04	FPC_{10}	1.60	90.80
FPC_5	3.58	80.61	FPC_{11}	1.33	92.14
FPC_6	2.89	83.50	FPC_{12}	1.17	93.31

Table 6.1: Individual and cumulative variation percentage per FPC.

model assumed is that of an Ornstein-Uhlenbeck stochastic model. We find the ML tree associated with each coefficient by doing a random search. To generate candidate branch lengths we assumed that the distribution of branch lengths approximated that of a log-normal $log(b) \sim \mathcal{N}(\mu, \sigma^2)$; this assumption is supported by empirical investigation of tree contained in Tree-fam [192]. We do this because as we assume the notions of *glottoclock* in Linguistics and *molecular clock* in Biology to share the same intrinsic meanings in their respective fields, we consider that the observed diffusion patterns will also be similar in a qualitative level. For the sake of generality we do not assume that the tree at hand is ultrametric (in an ultrametric tree all the extant taxa are on the same time-depth in the tree; this being formally expressed as $d(t_i, t_j) \leq d(t_i, t_k) = d(t_j, t_k)$ for every triplet i, j, k of extant taxa nodes). After finding the ML-optimal trees for each of the k projections utilized, we construct the consensus tree for the languages at hand. The consensus tree is constructed by applying the median branch length (MBL) rationale [82]: given that we have k candidate trees with the same branching topology, the consensus tree is constructed by assigning to each edge of the consensus tree, the median branch length from the k candidate "ML-optimal" trees associated with each branch. One in effect computes the "median tree". A number of complementary methodologies have also been proposed with variants of the majority-rule consensus tree being the most popular [186; 137]. We do not advocate a majority-rule consensus tree on the grounds that our sample is quite small and therefore bootstrapping techniques (as those are extensively used for the generation of majority-rule consensus trees) are not reliable. Additionally we also do not examine a possible clustering of correlation effects in the phylogenies and a subsequent clustering that they might induce [82]. Finally we do not explicitly examine the possibility of a multifurcating tree, ie. a trifurcation or higher degree branching events. While such events might have some gravity in the case of small population linguistic phylogenetic studies, where one can assume rapid branching of different groups of people [38], we do not find it plausible for cases of widely spoken languages as the ones found in the Romance language family. We do though allow for arbitrary small edge lengths, so we can in effect facilitate this possibility as one trifurcation would be associated with a zero branch length for an internal edge.

Implementing these assumptions we begin with the unrooted linguistic phylogenetic tree $\mathbf{T}$ with 5 leaves, shown in Fig. 2.6. This tree is based on [106]; American Spanish have been added though as a distinct language. We make the assumption that American Spanish share a common ancestor with Iberian Spanish, with that "Spanish protolanguage bifurcation" occurring more recently than any other linguistic bifurcation event in the examined Romance language phylogeny. Having fixed the branching structure of the tree we assign at its leaves the FPC scores. Each 5-language FPC score grouping is considered independent not only along the scores associated with the same digit but also with the scores from the other digits. As we are using digits *one* to *ten*, having generated 4 FPC surfaces for each digit, we test 40 different sets of "data at the tips". Using the O-U model we tested against 5120 candidate trees and reported as the optimal tree, the tree with the maximum likelihood for that given set of FPC scores. To conduct this testing step the function `fitContinuous()` from the `R` package `geiger` [128] was utilized; for each candidate tree branch sample "fitting", 700 random initializations of the routine were tested. As mentioned, while the branching events are treated as fixed, the branch lengths are not. Candidate branch lengths b were sampled from a log-normal such that $log(b) \sim \mathcal{N}(-2.29, 1.66^2)$; the actual values of μ and σ shown here were estimated by using the trees publicly available in Tree-fam [6].

[6] See Appendix, Fig. A.12.

Tree-fam ver. 8 [192] contains 16604 trees in total; for this task though, trees with less than 5 or more than 20 nodes were excluded from the analysis because we assumed that they do not present plausible exemplars for a Romance languages linguistic phylogeny. The reasons behind this heuristic rule are three: First, smaller trees may often convey domain-specific relations even within a biological setting. Second, larger trees are also less plausible as linguistic exemplars because they often aggregate different families of organisms with well understood distinctions in a way that is irrelevant for linguistics. Third, based on existing literature [221; 224; 106] Romance languages are not assumed to incorporate more than approximately 20 leaves. Based on these points this cut-off resulted in a 3593 tree sub-sample that was ultimately used to estimate $\hat{mu}$ and $\hat{\sigma}$. Having estimated the 40 "ML-optimal" trees, the "median tree" (show in Fig. 6.4) was constructed and assumed to be the tree that most accurately reflects our modelling assumptions as well as the universal linguistic phylogenetic association between the languages examined.

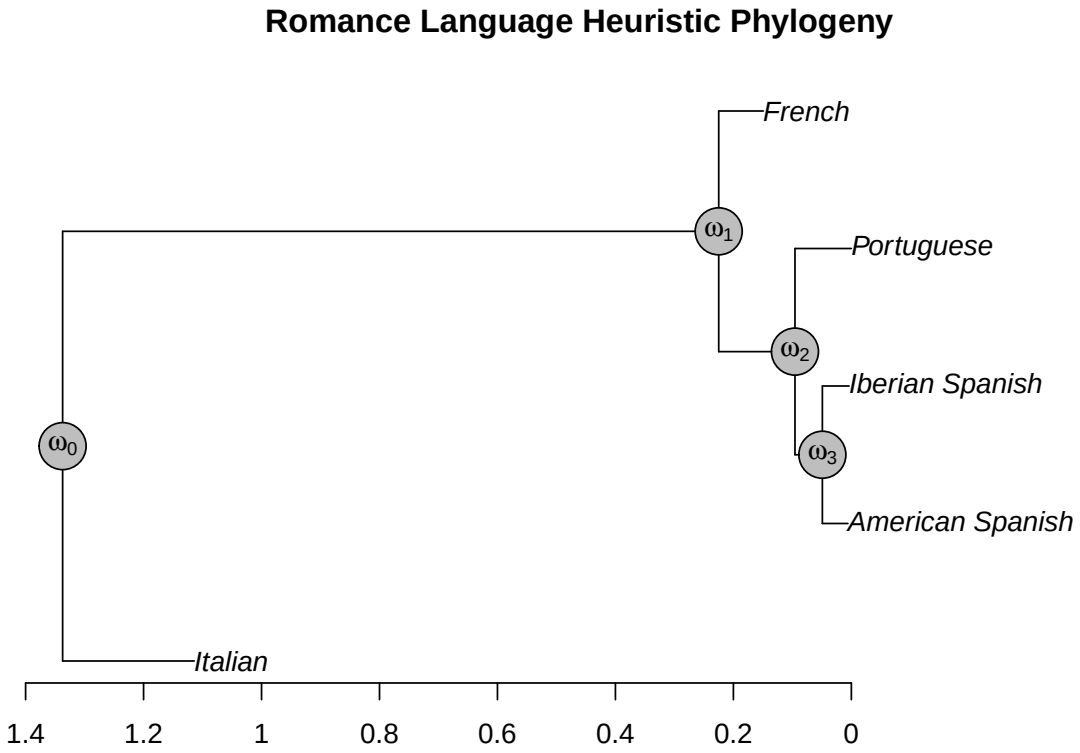

Figure 6.4: Median branch length consensus tree for the Romance language dataset, gray ω_i circles corresponding to protolanguages, Italian emerging as being clearly the modern language closer to a "universal" Romance protolanguage indexed as ω_0. All 10 digits were used to construct this tree.

6.2.3 Phylogenetic Gaussian process regression

As already noted, FPCA returns the estimated mixing coefficients at tip taxa, $\hat{Q}$ as well as the basis $\hat{\Phi}$. The next step in our linguistic phylogenetic study is to perform PGPR [157] separately on each 5-member mixing coefficient set associated with a basis $\phi_j(u, f)$. We assume knowledge of the phylogeny $\mathbf{T}$ (as constructed in the previous step), in order to obtain posterior distributions for all mixing coefficients throughout the tree $\mathbf{T}$. This means that we get estimates for the internal linguistic taxa in $\mathbf{T}$ as well as the leaves themselves.

Gaussian process regression (GPR) [263] is a flexible Bayesian technique in which prior distributions are placed on continuous functions. Its range of priors includes the Brownian motion and Ornstein-Uhlenbeck (O-U) processes, which are by far the most commonly used models of character evolution [125; 46]. (Gaussian and the Matérn kernels enjoying popularity in spatial statistics literature [68].) Its implementation is particularly straightforward since the posterior distributions are also Gaussian processes and have closed forms. Using notation standard in the Machine Learning literature (see, for example, [263]), a Gaussian process may be specified by its mean surface and its covariance function $K(\theta)$, where θ is a vector of parameters. Since the components of θ parameterize the prior distribution,

they are referred to as *hyper*parameters. The Gaussian process prior distribution is denoted:

$$f \sim \mathcal{N}(0, K(\theta)).$$

If x' is a set of unobserved coordinates and x is a set of observed coordinates, the posterior distribution of the vector $f(x')$ given the observations $f(x)$ is:

$$f(x')|f(x) \sim \mathcal{N}(A, B) \tag{6.19}$$

where

$$A = K(x', x, \theta)K(x, x, \theta)^{-1}f(x), \tag{6.20}$$

$$B = K(x', x', \theta) - K(x', x, \theta)K(x, x, \theta)^{-1}K(x', x, \theta)^T \tag{6.21}$$

and $K(x', x, \theta)$ denotes the $|x'| \times |x|$ matrix of the covariance function K evaluated at all pairs $x'_i \in X', x_j \in X$. Equations 6.20 and 6.21 convey that the posterior mean estimate will be a linear combination of the given data and that the posterior variance will be equal to the prior variance minus the amount that can be explained by the data. The interpretation of these in the context of a phylogenetic tree is twofold: first that all ancestral states can be expressed as linear combinations of the observed leaf states and second that the covariance among this data will be only due to phylogenetic associations. If phylogenetic associations are not present a phylogenetic model will return a simple arithmetic mean as its estimate A (Eq. 6.20) and the covariance structure B will be zero-th as the non-phylogenetic fluctuations are considered independent to each other in your generative model (Eq. 6.23). Additionally, the log-likelihood of the sample $f(x)$ is

$$\log p(f(x)|\theta) = -\frac{1}{2}f(x)^T K(x, x, \theta)^{-1}f(x) - \frac{1}{2}\log(det(K(x, x, \theta))) - \frac{|x|}{2}\log 2\pi.^7 \tag{6.22}$$

It can be seen from Eq. 6.22 that the maximum likelihood estimate is subject both to the fit it delivers (the first term) and the model complexity (the second term). Obviously this does not constitute a full model selection procedure as with the cases examined in Sect. 3.4.3. AIC scores are occasionally used but given one fixes the number of parameters of the model a priori (as it will be shown immediately afterwards we fix that number to 2), AIC score changes are exclusively due to changes in the model's likelihood/parameters θ. Thus, Gaussian process regression is non-parametric in the sense that no assumption is made about the structure of the model: the more data gathered, the longer the vector $f(x)$, and the more intricate the posterior model for $f(x')$. This views GPR as non-parametric, stemming from Machine-Learning literature [290]. A counter-argument could be that GPR is an extremely parametric procedure that just "fits k numbers to dataset", k being the number of hyperparameters utilized in θ. Complementary to this approach, where the results of a two-dimensional FPCA are utilized, is the work of Shi et al. [292] on Gaussian Process Functional regression, there the mean function is modelled by functional regression model and the covariance structure by a Gaussian process. To that extend the current methodology is more simplified and does not explicitly model the mean function separately. On the other hand the methodology of Shi et al. by using B-splines in order to model the mean structure assume that underlying structure can be assumed to be piecewise polynomial while the proposed methodology based on $FPCA$ offers an empirical alternative to that assumption.

Phylogenetic Gaussian Process regression (PGPR) extends the applicability of GPR to evolved function-valued traits as spectrograms. A *phylogenetic* Gaussian process is a Gaussian process indexed by a phylogeny $\mathbf{T}$, where the function-valued traits at each pair of taxa are conditionally independent given the function-valued traits of their common ancestors. When the evolutionary process has the same covariance function along any branch of T beginning at its root (called the *marginal covariance function*), these assumptions are sufficient to uniquely specify the covariance function of the PGP, $K_{\mathbf{T}}$. As we assume that $\mathbf{T}$ is known in our inverse problem based on the tree-estimation step presented above,

[7] While we do not focus on computational matter explicitly we draw attention to the fact that one does not need to compute the inverse covariance matrix $K(x, x, \theta)^{-1}$ nor the determinant $det(K(x, x, \theta))$ directly. For the purposes of evaluating $\log p(f(x)|\theta)$, in a manner similar to section 5.3.8 one utilizes the Cholesky decomposition of the matrix.

the only remaining modelling choice is therefore the marginal covariance function. As can be seen from Eq. 6.23, K is a function of patristic distances on the tree rather than Euclidean distances as standard in spatial GPR.

In phylogenetic comparative studies, where one has observations at the leaves of $\mathbf{T}$, the covariance function $K_{\mathbf{T}}$ may be used to construct a Gaussian process prior for the function-valued traits, allowing functional regression. In the model that we use, this is equivalent to specifying a Gaussian prior distribution for the set of mixing coefficients used. This may be done by regarding those coefficients as observations of a univariate PGP. As noted in [157], if we assume that the evolutionary process is Markovian and stationary, then the modelling choice vanishes and the marginal covariance function is specified uniquely: it is the stationary O-U covariance function. If we also add explicit modelling of non-phylogenetically related variation at the tip taxa, the univariate prior covariance function has the unique functional form presented in Eq. 6.23. We do not assume knowledge of the parameters of Eq. 6.23 however. To estimate them we use the consensus tree generated in section 6.2.2, shown in Fig. 6.4, and the two-dimensional basis functions generated in section 6.2.1, shown in Fig. 6.3. This fixes the experimental design for our simulation and inference.

Commenting on the specific parameters chosen for the phylogenetic O-U processes, as in [124] we refer to the *strength of selection parameter* α and the *random genetic drift* σ_n: we add superscripts j to these parameters to distinguish between the four different O-U processes. With this notation, the mixing coefficients for a specific basis have the following covariance function:

$$K^j_{\mathbf{T}}(\mathbf{t}_i, \mathbf{t}_g) = (\sigma^j_f)^2 \exp\left(-2\alpha^j P_T(\mathbf{t}_i, \mathbf{t}_g)\right) + (\sigma^j_n)^2 \delta^e_{\mathbf{t}_i, \mathbf{t}_g} \tag{6.23}$$

where $\sigma^j_f = \sqrt{\frac{(\sigma^j)^2}{2\alpha^j}}$, $P_T(\mathbf{t}_i, \mathbf{t}_g)$ denotes the phylogenetic or patristic distance (that is, the distance in $\mathbf{T}$) between the ith and gth tip taxa, σ_n is defined as above, and

$$\delta^e_{t_i, t_g} = \begin{cases} 1 & \text{iff } t_i = t_g \text{ and } t_i \text{ is a tip taxon,} \\ 0 & \text{otherwise} \end{cases}$$

adds non-phylogenetic variation to extant taxa as discussed above, ie. δ^e evaluates to 1 only for extant taxa, thus σ_n quantifies within-species genetic or environmental effects and measurement error in the i-th mixing coefficient. As a direct consequence the patristic distance which is effectively the sum of the evolutionary time between the ith and gth tip taxa and their common ancestor offers the space upon which evolutionary differences are defined. This is an important modelling assumption: estimates for latent ancestral states will account *only* for phylogenetic variation between the taxa. All non-phylogenetic variation has to be accounted for in the extant taxa level. Therefore, we see from Eq. 6.23 that the proportion of variation in the mixing coefficients attributable to the phylogeny is $\frac{(\sigma^j_f)^2}{(\sigma^j_f)^2 + (\sigma^j_n)^2}$. Clearly if this ratio tends to 0, non-phylogenetic variation dominates our sample and phylogenetic inference is impossible. In the Gaussian process regression literature in Machine Learning, $\frac{1}{2\alpha}$ is equivalent to ℓ, the characteristic length-scale [263] of decay in the correlation function and in the following work we work with the latter.

Aiming to provide the best possible basis in terms of an RSS reconstruction criterion along with the minimal amount of prior assumptions, we use the FPCA-generated basis. In general, there is no reason for our inference procedure to be sensitive to the particular shape of the basis functions; indeed other bases eg. ICA-based [145] could easily be employed. Concerning inference for a specific digit d (eg. *one*) the four simple two-dimensional orthogonal functions shown in Fig. 6.3 were therefore chosen as examples. For computational purposes each basis function was stored numerically as a matrix of dimensions 81 by 100, so that the basis matrix ϕ^d was in this case size 4×8100, each row storing a different basis function. This is in accordance with standard methodology used in spectrogram and face recognition analysis where an image is represented as a concatenated vector [241; 19]. As we will discuss in the final section, given that someone is willing to make certain assumptions about the noise structure applicable, a variety of different models is also available [19; 197; 142].

The mixing coefficients generated by FPCA are stored in Q^d. Our modelling assumption is that the mixing coefficients for distinct basis functions $\phi^d_1, \phi^d_2, \phi^d_3, \phi^d_4$ are statistically independent of each other as they are produced using standard FPCA. It is therefore sufficient to describe the stochastic process

generating the mixing coefficient for each basis independently using the phylogenetic model proposed above (Eq. 6.23). We need to emphasize again at this point that we focus on one digit d, where $d = 1$ in this case. The only instance where all 10 digits were combined, was in the previous subsection for the construction of the *MBL* tree.

The "extant" function-valued trait at tip taxon i is thus $\sum_{j=1}^{4} Q_{i,j}\phi_j$ (a vector of length 8100), while the ancestral function-valued trait at internal taxon g is $\sum_{j=1}^{4} H_{g,j}\phi_j$, H storing the values of the mixing coefficients in the ancestral (historical) states. As commented above, the ancestral function-valued traits exhibit only phylogenetic variation, while the extant function-valued traits exhibit both phylogenetic and non-phylogenetic variation. Of course, it is not possible to reconstruct non-phylogenetic variation using phylogenetic methods. Non-phylogenetic variation is nevertheless a "fact of life" concerning the data at the extant taxa and we need to account for it explicitly. As Hadjipantelis et al. [119] have demonstrated though, this noise does not prevent the reconstruction of the phylogenetic part of variation for ancestral taxa.

Commenting further on the role of parameters in the phylogenetic O-U process described above in Eq. 6.23, exceptionally *small* characteristic length-scales ℓ relative to the tree patristic distances, practically suggest taxa-specific phylogenetic variation, ie. non-phylogenetic variation. This holds also in its reverse: exceptionally *large* characteristic length-scales suggest a stable, non-decaying variation across the examined taxa that is indifferent to their patristic distances, again suggesting the absence of phylogenetic variance among the nodes.

Since the posterior distributions returned by PGPR depend on the hyperparameter vector θ, we must estimate θ in order to reconstruct ancestral function-valued traits; the estimation procedure correcting for the dependence due to the phylogeny. Maximum likelihood estimation (MLE) of the phylogenetic variation, non-phylogenetic variation and characteristic-length-scale hyperparameters σ_f^j, σ_n^j and ℓ^j respectively may be attempted numerically using the explicit prior likelihood function (Eq. 6.22).

Estimating hyperparameters is commonly hindered by problems of non-identifiability [263; 160] and, as a direct consequence, concurrent estimation of all components of $\theta^j = (\sigma_f^j, \sigma_n^j, \ell^j)$ is problematic. As commented by Beaulieu et al. [23], *the influence of sample size on the bias and precision of α is particularly pronounced*, in our setting this problem is even more evident. In particular given that we have only 5 languages estimating 3 hyperparameters we realize that our estimation procedure is going to suffer. Thus we propose fixing the length scale ℓ. This does not mean that we enforce phylogenetic variation but rather that we fix the distance over which the covariance can meaningfully occur. If there is "nothing but non-phylogenetic variation", that will be reflected by the $\frac{(\sigma_f)^2}{(\sigma_n)^2} \to 0$. Hadjipantelis et al. [119] have shown that overall θ estimates may be further improved if one knows a priori the value of the ratio $\frac{(\sigma_f)^2}{(\sigma_n)^2}$, which is closely related to Pagel's λ [228]. We do not examine this possibility here though as we have little prior knowledge over the sample's phylogenetic dynamics in a linguistic application. The final estimated parameters $\hat{\theta}$ are shown in Table 6.2. It is immediately seen that only one FPC, the second FPC, encapsulates plausible phylogenetic associations. This is not surprising, given our small sample; plausible associations might not be provided "enough structure" from the tree itself for them to come forward as significant effects. In particular, seeing the hyperparameter estimates for the first FPC of digit *one*, θ^1, it is striking that all variation is considered to be non-phylogenetic; we expect that because, as mentioned, FPC_1 appears to encapsulate mostly the presence of the initial vowel in each word. Given that all words in the sample start with variants of "u" there are not enough differences to be accounted within a phylogeny. To a lesser extent the opposite can be attributed for FPC_3. It encodes a highly specialized pattern of counter-balanced variation between high-frequency early-timed and low-frequency later-timed vocal excitations within the same word, however being so "specialized" there is not enough structure for it to come across as phylogenetically relevant (if indeed it is). Seeing both these FPC's we see that our decision to fix ℓ does not seem unreasonable, given the gross absence of any phylogenetic signal. On the contrary, examining FPC_2 we witness a noisy but plausible phylogenetic variation. The length-scale ℓ here might not be optimal but it does not preclude the detection of phonetic associations due to a phylogeny. In a way we expected this; as mentioned in the earlier section, FPC_2 is mostly modelling the possible interplay between the second and the first vowel of a word (and if there is no second vowel it comes out close to zero). This is a strong association which is not so specific as

in the case of FPC_3. We can not say that FPC_2 is certainly encapsulating phylogenetic variations, if anything the $\frac{\sigma_f^j}{\sigma_n^j}$ being close to unity signifies the significant presence of non-phylogenetic variation; it nevertheless seems to offer plausible insights.

The final FPC analysed, FPC_4, gives peculiar hyperparameter choices. One could naively even say that it is encapsulating *only* phylogenetic signal. This is clearly not the case as we very well understand that it is practically impossible for a phylogenetic trait (assuming that one is encoded by FPC_4) to have retained absolutely no "non-phylogenetic" variability across a phylogeny. What is more, if we investigate the actual number (700.814), we see it is significantly higher than the standard deviation of the sample coefficients it tried to model originally (279.430). In effect it "amplifies" the phylogenetic variation to such a level so that it acts as "non-phylogenetic" variation with strong practically constant variational amplitude across all nodes. θ^4 are just artefacts of the numerical optimization procedure used. We do not expect any phylogenetic variation in FPC_4 and clearly this choice of $\theta's$ reflects just that. It does draw attention though to the fact that if numerical optimization methods are employed one has to always question the significance of their results not only technically but also conceptually. As mentioned above sample size is an issue that has significant impact on the power of this analysis; we revisit this point in section 6.4.

θ^i #	σ_f^i	σ_n^i	$\frac{\sigma_f^i}{\sigma_n^i}$
1	$2.802{*}10^{-4}$	2095.101	$1.337{*}10^{-7}$
2	363.358	370.084	0.982
3	$1.074{*}10^{-5}$	473.240	$2.270{*}10^{-8}$
4	700.814	$7.988\ {*}10^{-8}$	$1.383{*}10^{9}$

Table 6.2: The MLE estimates for the hyperparameters in Eq. 6.23 for digit *one*. Each row corresponds to a given estimate of the vector θ^i. These estimates provide the maximum likelihood value for Eq. 6.22. When ℓ is denoted as non-applicable, it is because there is no phylogenetic variation in the sample.

6.3 Results: Protolanguage reconstruction & Linguistic Insights

Having been presented with function-valued data, we extracted the functional basis $\hat{\phi}^d(u, f)$ and the associated mixing coefficients $\hat{Q}^d$, (Sect. 6.2.1) and estimated the most relevant tree (Sect. 6.2.2). We then performed PGPR (Sect. 6.2.3) on each mixing coefficient set associated with a specific basis, to obtain the univariate Gaussian posterior distribution for the mixing coefficient at any internal taxon $\mathbf{t}'$. As discussed in Sect. 6.2.3, the Gaussian process prior distribution has covariance function (Eq. 6.23). Because estimating σ_f^j and ℓ^j alone is challenging [23] (although the estimation improves significantly with increased sample size), and we have further increased the challenge by introducing non-phylogenetic variation, we fix the length scale ℓ across all simulations. The actual number of it was set to equate the median branch length within the tree used. Previous works [119; 286] have commented that the median branch length can serve as a good approximate measure for a phylogenetic horizon in the absence of other information; θ^j constitutes thereof of σ_f^j and σ_n^j only.

We substitute $\hat{\theta}^j$ into Eq. 6.23. Taking a simple and direct approach, our estimate $\hat{\phi}^d$ obtained in Sect. 6.2.1 may then be substituted into Eq. 6.14 to obtain the function-valued posterior distribution $f_{\mathbf{t}'}$ for the function-valued trait at taxon $\mathbf{t}'$. Since our estimated basis functions are stored numerically as vectors of length 8100, this gives the same discretisation for the ancestral traits.

Conditioning on our estimated mixing coefficients $\hat{Q}^d$ for the tip taxa for a given basis function j, the posterior distribution of $G_{j\mathbf{t}'}$ is:

$$G_{j\mathbf{t}'} \sim \mathcal{N}(\hat{A}_j, \hat{B}_j) \tag{6.24}$$

where the vector $\hat{A}_j$ and matrix $\hat{B}_j$ are obtained from Eq.'s 6.20 and 6.21, taking $f(x) = \hat{Q}_j^d$, $x' = \mathbf{t}'$ and $\theta = \hat{\theta}^j$ respectively for our observed values at the extant taxa, estimation coordinates and hyperparameter vector. Therefore since our prior assumption is that the mixing coefficients of any two bases are

statistically independent of each other, it follows from Eq. 6.1 that:

$$f_{\mathbf{t}^*} \sim \mathcal{N}(\Sigma_{j=1}^{k} \hat{A}_j \hat{\phi}_j(u,f), \Sigma_{j=1}^{k} \hat{B}_i^T \hat{\phi}_i(u,f) \hat{B}_i). \tag{6.25}$$

The component specific marginal distributions of this representation (mean and standard deviation) are shown in Table 6.3, the reconstructed protolanguage for the root node ω_0 is shown in Fig 6.5.

i	$\hat{A}_i$	$\sqrt{\hat{B}_i}$
1	0.000	2095.10
2	147.679	504.70
3	0.000	473.24
4	(~~6.328~~) 0	618.91

Table 6.3: The posterior estimates for the parameters of $G_{j\mathbf{t}'}$ for digit *one* at the root node ω_0. Each row corresponds to a given estimate of the marginal distribution $G_{j\mathbf{t}'}$. As mentioned FPC_4 does not reflect true phylogenetic variation, therefore $\hat{A}_4$ despite having a non-zero value, for the protolanguage reconstruction purposes a zero value is used.

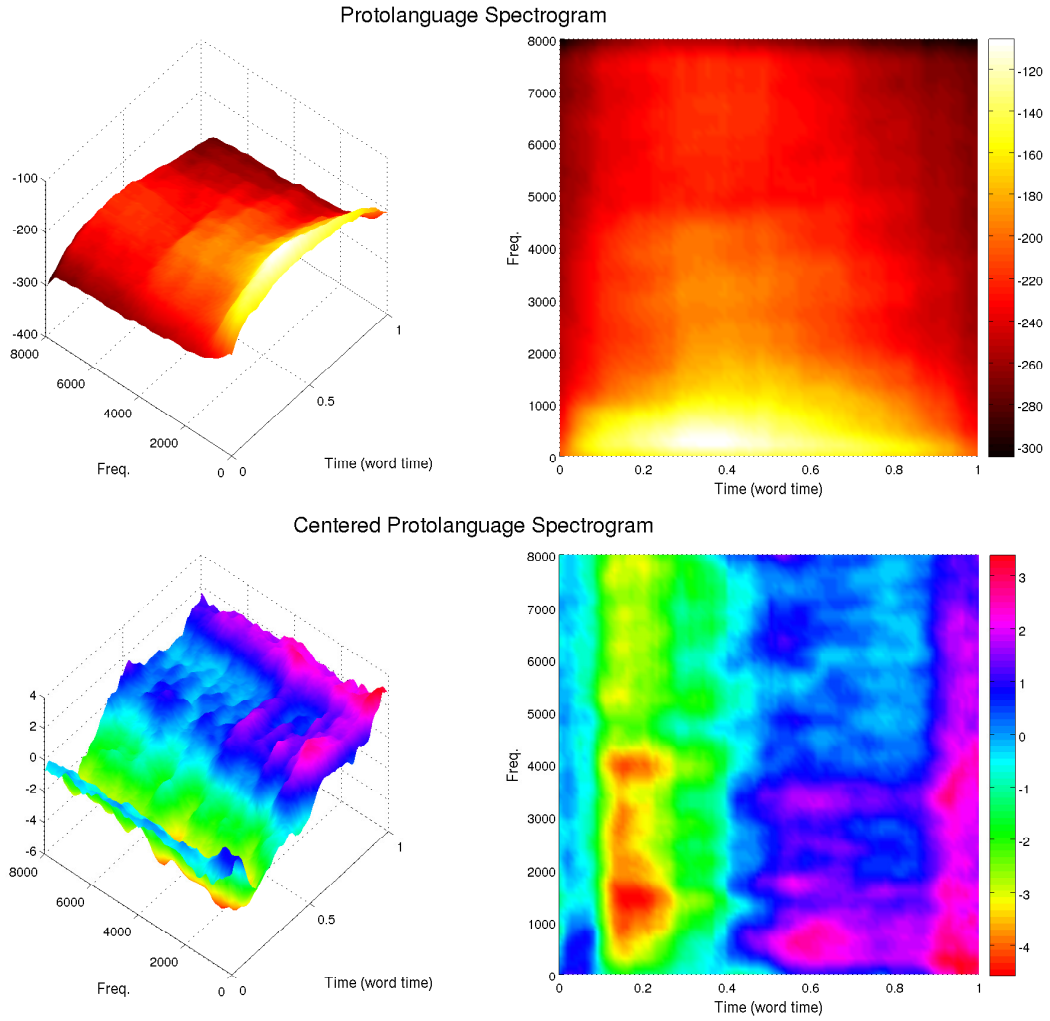

Figure 6.5: Centred (compared to the average of the 22, digit *one* spectrograms) spectrogram for the Romance protolanguage (top row) and the actual spectrogram of the protolanguage (bottom row).

Examining first the table 6.3 one notices two striking but entirely expected facts. First for FPC_1 and FPC_3 our mean estimate is zero. This is expected as no phylogenetic signal is detected. Given that

the ML estimate of zero-meaned Gaussian is zero, that is expectedly returned. This is what effectively also happens in the case of FPC_4; a prediction of 6.328 in the context of a Gaussian with $\sigma = 618.91$ is "the mean". The small deviance observed exists due to the overfitting we already noted during the estimation of the θ^4; exactly because of this extremely strong artificial phylogenetic variance, a small deviation from the mean is observed due to the structure of the phylogeny. On the other hand FPC_2 does present certain degree of phylogenetic variance thus being away from the "naive mean" of zero. The second striking finding is the magnitude of the estimates' standard deviation. In all cases (even in the case of FPC_2), standard deviation is very large. We do expect that though, given that we have in effect "just five numbers", noting also that these are in accordance with the standard deviations observed in the raw data initially (Table A.12); the effects of these large confidence intervals can be visually inspected in Fig. A.14. Continuing with the examination of the actual Romance protolanguage spectrogram we see first the main effect conveyed by the centred spectrogram (Fig. 6.5, top row), this being that it lowers the stress of a first vowel. This means in a way, it tries to downplay the stress of the first vowel found in "uno" and amplify the second half of the utterance, thus leading to a final unimodally stressed word for the digit *one* so that even though it has a second vowel, it is not distinctively stressed; much like the constant power-spectrum pattern observed in American Spanish. We have to note here that this finding is not unexpected given the mean of original data; there the sample mean does appear as a single-stressed word. This is also foreseeable by the fact the only FPC that actually influences this spectrogram is FPC_2, the FPC we do know to have exactly this property controlling the interplay between the second and the first vowel. With these findings we would expect American Spanish (the most mild two-vowel utterance of the three examined), to be the closest to the "Romance protolanguage" in terms of their utterance of the digit *one*.

6.4 Discussion

We restricted ourselves to this phylogenetic problem because, as shown in [157], a wide class of phylogenetic Gaussian process models have exactly the linear decomposition in Eq. 6.1. Given Eq. 6.1 in the setting of mathematical inverse problems where, given data y, a phylogeny $\mathbf{T}$ and function-valued data y at its leaves, we wish to infer the forward operator $Q_{\mathbf{T}}$ and model ϕ such that

$$y = Q_{\mathbf{T}}(\phi). \tag{6.26}$$

Even though the data y are of discrete dimensions and typical of small number of correlated factors per tip taxon, a variety of statistical approaches are available (e.g. see [280], [116]). When the data are functions, phylogenetic Gaussian processes (PGP) [157; 163] have been proposed as the forward operator.

Our dimensionality reduction methodology in Sect. 6.2.1 can be easily varied or extended. For example, any suitable implementation of PCA may be used to perform the initial dimension reduction step: in particular, if the data have an irregular design (as happens frequently with function-valued data), the method of Yao et al. [337] may be applied to account for this; the ICA step then proceeds unchanged. Additionally, assuming someone is prepared to make assumptions about the local features of the noise structure, a number of two dimensional PCA techniques have been proposed especially in regards with the analysis of eigenfaces [197; 335], with the *2DPCA* algorithm appearing as the popular choice. As commented in section 3.3.1, sparse two-dimensional PCA are already available in the case one is presented with irregularly sampled data [55] [8], nevertheless our implementation does not suffer from an irregular or sparse grid; as Cheng and Müller comment *"we (C. & M.) found that in the dense regular case, these two approaches give nearly identical results"*. We also note that while we employed a principal component analysis approach, other dimensionality reduction decompositions could also prove successful. Similar work with one-dimensional functional data has used ICA (and in particular the *CubICA* implementation of ICA rather than the most widely employed *FastICA* [145]). We nevertheless caution the direct application of ICA as, given the small number of variables, even small changes in the ordering of the samples can manifest in differences in the quality of the produced independent components. In addition to that, if the mixing matrix Q is Gaussian, then PCA returns scores which are approximately statistically independent equating the results of ICA; however, no general positive

[8] The authors having shown extremely different results in the case; view Sup. Material of [55].

statement can be made about independence of the entries of $\hat{Q}$.

In order to approximate the solution of the inverse problem in Eq. 6.1, the assumption of orthogonality between the basis functions stored in ϕ was made. This is clearly more restrictive than the maximal independence being incorporated by ICA but on the other hand it is more robust to the smaller samples utilized here. We can rank though these eigensurfaces by their respective eigenvalues and acquire a better understanding of the phylogenetic processes assumed to take place.

Tree estimation followed a standard phylogenetic framework; it was not the main focus of this work. Despite that a phylogeny was generated that corresponds closely to the phylogenies published in the literature on this subject [206; 106]. While we examined the prospect of using a weighted median based on the variance explained from each FPC used during the construction of the consensus tree, we decided against it because we wanted to keep in line with current literature.

Numerically our work on hyperparameter estimation in Sec. 6.2.3 cannot directly circumvent the effects of overfitting due to small sample size [23; 59] by employing bagging in order to bootstrap our sample [119]. Recognizing this, we restrict our approach to a problem of smaller dimensions. Conceptually our work on hyperparameter estimation, when taken together with Sec. 6.2.1, relates to the character process models of [242] and orthogonal polynomial methods of [167], which give estimates for the autocovariance of function-valued traits. Writing out Eq. 6.1 for a single function-valued trait (at the ith tip taxon, say), our model may be viewed as:

$$f(x) = \sum_{j=1}^{4} g_{i,j}\phi_j(x) + \sum_{j=1}^{4} e_{i,j}\phi_j(x) \tag{6.27}$$

where the mixing coefficient $q_{i,j}$ has been expressed as the sum of $g_{i,j}$, the genetic (i.e. phylogenetic) part of variation, plus $e_{i,j}$, the non-phylogenetic (eg. environmental) part of variation, just as in these references. Then the autocovariance of the function-valued trait is:

$$E[f(x_1), f(x_2)] = \sum_{j=1}^{4} \left((\sigma_j^f)^2 + (\sigma_j^n)^2 \right) \phi_j(x_1)\phi_j(x_2). \tag{6.28}$$

The estimates of σ_j^f and σ_j^n obtained in section 6.2.3 may be substituted into Eq. 6.28 to obtain an estimate of the autocovariance of the function-valued traits under study. This estimate has the attractions both of being positive definite (by construction) and taking phylogeny into account. Eq. 6.28 being practically the phylogenetic variant of the GP regression model with a functional mean structure outlined by Shi et al. in [292]. To that extend Shi et al.'s work on batch functional data offers an interesting alternative formulation on the treatment of linguistic corpora where "batching" can be assumed both in terms of speaker-speciic recordings but more importantly language.

Overall, this formulation of the ancestral state reconstruction and evolutionary dynamics investigation tasks allows us to have a full inverse problem view of the phylogenetic regression problem [76; 152]. Clearly there is an obvious issue of time-reversibility and how one treats time as a one-way continuum but we will not expand on the matter. It is noteworthy that this question, especially in relation with the notion of a *molecular clock* that dictates the rate of evolutionary change of biological characters, is one of the central questions in Evolutionary Biology [143; 9]; the notion of *glottoclock* being subjected to the same conceptual issues.

Various frameworks exist which could be used to generalize the method presented in Sec. 6.2.3, to model heterogeneity of evolutionary rates along the branches of a phylogeny [266] or for multiple fixed [46] or randomly evolving [126; 23] local optima of the mixing coefficients. For the stationary O-U process the optimum trait value appears only in the mean, and not in the covariance function, and so does not play a role as a parameter in GPR (see [263]). We have not implemented such extensions here, effectively assuming that a single fixed optimum is adequate for each mixing coefficient. Nonetheless our framework is readily extensible to include such effects, either implicitly through branch-length transformations [229], or explicitly by replacing the O-U model with the more general Hansen model [126]; in that respect Butler & King [46] have already shown an implementation with multiple local optima. Numerically these are implemented by reformulating Eq. 6.23 to have subtree specific covariance terms through the use of further $\delta_{\mathbf{t}_i,\mathbf{t}_g}$-like functions.

On a similar manner we briefly comment on the theoretical as well as numerical implications of the sample size used. Small sample size undermines the statistical power of any study. This is because there might already be a small probability of finding a true effect and in addition even if a true effect is found it might be exaggerated [47]. Moreover, one can not assume that the probability of *Type II* errors is 0. As mentioned the "standard" solution employed in Phylogenetics (and other disciplines) to partially alleviate these concerns is bootstrapping. Unfortunately because of our small sample, resampling approaches are not directly applicable at least in theory. In addition, because ultimately we are solving an optimization problem we cannot ignore the potential presence of local extrema in the log-likelihood function L. For these realistic concerns we can comment three things: first, that we do not find phylogenetic association "everywhere". Indeed, based on biological insights one will always expect environmental (ie. non-phylogenetic) variation to be prominent and this is clearly put forward by our estimates. Second, the mode of variation found to exhibit phylogenetic associations is the most linguistic plausible. Finally, the absolute magnitude of the phylogenetic signal found is well within reasonable estimates both in terms of overall FPC variation as well as intra-FPC variation. For these reasons we believe that the analysis conducted holds at least some weight and it presents the most plausible findings for the sample available.

Commenting exclusively on our linguistic findings, we draw attention to two areas; the theoretical implications from our consensus tree and the protolanguage estimate for digit *one*. Regarding the tree that was generated using all digits, it is undoubted that Italian is closer to all other Romance languages examined to a Romance protolanguage. Even our simplifying procedure and small data confirmed that. What would be interesting would be to add Romanian in the set of examined languages as Romanian is the other "popular choice" of a modern language that closely approximates the original Latin root [222]. In the same manner the addition of Catalan would be helpful as it would allow a finer geographical grid. We must not forget that the carriers of a "language gene" are humans and humans are subjects to spatial constraints. This finding is independent of the fact that specific words may appear closer to the tree's protolanguage as is the case with digit *one* examined here. Second, regarding our protolanguage estimate of *one*, we believe that the "closer" estimate to a protolanguage estimate of digit *one* is probably American Spanish. As American Spanish appears closer to a "Romance phylogenetic mean" we might even argue that exactly because of that deviation as a population effect, the current speakers might retain more characteristics of an archaic version of the languages compared to the "original language's" newer version. This phenomenon is, at least partially, exhibited in the case of Quebec and Metropolitan French [329]; a Latin language case where the Atlantic barrier resulted in the linguistic evolution of two mutually intelligible but distinct descendants of Classical French.

In conclusion, we have proposed a modelling framework for the phylogenetic analysis of phonetic information within the greater Functional Data Analysis framework. We believe that this is probably one of the first applications of Linguistics functional Phylogenetics. We strived to combine established methodologies, while making the least amount of theoretical assumptions possible and always trying to draw direct analogies that our statistical manipulations of the sample have in the actual sample space. Whether these were in the preprocessing steps with time-warping the spectrograms, the dimension reductions procedure conducting FPCA or the actual phylogenetic Gaussian process regression, we feel that the current chapter just scratches the surface of another fruitful and insightful area of research.

Chapter 7

Final Remarks & Future Work

In Phonetics the use of Functional Data Analysis is often under-represented. That despite the fact that FDA appears more theoretically coherent in a number of situations than the current multivariate approaches used. The current thesis showcases the issues and insights that an FDA framework entails when applied in Phonetics. It starts with a simple application, disregarding issues of time-distortion. While somewhat simplified, that approach can offer a deep understanding about the linguistic components involved. We are able to identify linguistically meaningful associations with a minimal amount of prior linguistic knowledge. Our work then moves to formulate a framework where one can conduct concurrent analysis of amplitude and phase information. Through this we are presenting a first quantitative insight on how these two domains interact. We close by showcasing an application of FDA in Linguistic Phylogenetics. While highly experimental, this approach shows a first paradigm of a Functional Data Analysis approach in Phylogenetics as a whole. We are able to draw certain conclusions and conduct a full evolutionary inference procedure in data that would otherwise be either ignored or deemed unsuitable for such an analysis.

We draw three key conclusions, each from the respective chapters of this thesis. The first one being that Functional Data Analysis is an appropriate tool to utilize in Phonetic Corpus Linguistics, allowing complex behaviours to be directly analysed. As shown in chapter 4, a number of vowel sequences and consonant-vowel sequences significantly influence the final F_0 utterance. In most cases the patterns recognized adhere to known grammatical rules. The important thing is that these grammatical rules were not incorporated in the analysis but were recognized as already documented by an expert in the field (Dr. Evans). This is significant evidence for the usefulness of FDA in the analysis of Phonetic data and we feel confident that more languages with less well-documented grammatical rules can also benefit. The second conclusion is that the intertwined nature of amplitude and phase information in a phonetic dataset cannot be ignored a priori. Chapter 5 makes a definite case that there are non-negligible correlations between the two domains. This in a way shows the partial "inability" for the warping algorithms presented to fully separate the two domains of variation. If they were fully successful then one should not observe any strong correlation between the two domains; something that our findings obviously disprove. Clearly there is a question of whether or not one tries to put more weight in this requirement; after all the previous step (that ignored phase information) was successful. We need to emphasize though that one needs to know what is ignored before he decides to exclude it from the analysis. For instance, we see that F_0 slope changes are correlated with changes in the duration patterns of a speaker. So if one aims to model each are of variation independently and then combine them, important effects might be excluded. On the other hand, if one aims to model finer speaker F_0 effects, vowel duration is largely immaterial. Finally we advocate that Functional Data Analysis can be used for Phylogenetics with great potential. Chapter 6, as well as the closely related research paper [119], show that the combination of FDA and known phylogenetic techniques is far from inapplicable or tedious to achieve. Instead, well-established Phylogenetics techniques are almost directly applicable within a FDA framework and those techniques can be used to answer questions that FDA did not recognize as a potential field of applications. In this work we presented a framework under which one is in position of evaluating the evolutionary relations between languages (or any other object of analysis used within a phylogeny), as well as making estimates about their prior states. While our current insights were "trivial" [1], our estimates are not; reconstructing

[1]Finding that Italian is closer to a Latin ancestral language than French, Portuguese and Spanish does not constitute

protolanguages is an open question in Linguistics and our flexible and tested approach is advancing the current literature.

There are a number of future works that can directly stem from the current project:

- Data integration. A logical next step will be an application of the methods shown in chapter 6 on a bigger dataset. "Bigger", referring to the number of languages explored, the number of speakers included as well as the number of words utilized. The number of languages explored is the obvious short-coming of the current work. The small phylogeny it employs does not assist the asymptotic assumptions made nor allows for standard bootstrapping techniques to be employed. We already identified two "easily" sampled languages [2]. It would also be reasonable to include an outgroup language if we want to achieve a deeper understanding of the tree-reconstruction step. Secondly, more speakers per language will allow for better "language exemplar" word estimate. As it stands, our language exemplars are based on a simple *language-gender* interaction model; more sophisticated models can be used if one has more data. Finally, the sample analysed, is a design choice that can have very pronounced effects. The Romance languages recording data are just one dataset analysed; there are a number of specially constructed datasets for Phylogenetics. Moreover even in a bigger dataset, all word instances are used as "independent" with one another; something that is clearly an oversimplification. This is definitely an aspect that any Linguistic study should consider and while documented since early applications of Linguistic Phylogenetics [76], it has not received proper attention. This last issue is in practice a data integration problem, as different avenues provide different datasets that have non-obvious ways of being combined.

- Protolanguage acoustic estimates. Ultimately the question of a protolanguage is "how it sounded". Direct reconstruction of signal from a spectrogram is problematic exactly because one "loses" the phase information of the Short Time Fourier Transform encapsulated in its imaginary part. There are certain methodologies of how to reconstruct a signal from its spectrogram but they are not immediately applicable. Translating our findings to actual acoustic signal is something that would not only benefit current research but also open up a whole new different field of research as currently there is no concept of *Historical Acoustics* exactly because there are no "acoustic fossils" to be analysed.

- Multilevel functional time-registration. As Di et al. have shown multilevel FPCA can be insightful [66]. Nevertheless in a best case scenario a researcher recognizes that certain instances "can not be meaningfully warped together" and uses a cluster specific warping [304]. In particular, Di et al. [66] propose a model for a functional dataset Y such that:

$$y_{ij}(t) = \mu(t) + \eta_j(t) + \sum_{k=1}^{N_1} \xi_{ik}\phi_k^{(1)}(t) + \sum_{l=1}^{N_2} \zeta_{ijl}\phi_l^{(2)}(t) + \epsilon_{ij}(t) \tag{7.1}$$

$$\xi_{ik} \sim N(0, \lambda_k^{(1)}, \quad \zeta_{ijl} \sim N(0, \lambda_l^{(2)}), \quad \epsilon_{ij}(t) = N(0, \sigma^2) \tag{7.2}$$

  where $\mu(t)$ is the overall functional mean, $\eta_j(t)$ is the component specific functional mean, ξ_{ik} and ζ_{ijl} the level 1 and level 2 principal component scores and $\phi_k^{(1)}(t)$ and $\phi_l^{(2)}(t)$ the level 1 and 2 eigenfunctions respectively. As such, one is given the ability to recognize two different domains of variation within unwarped data that the current methodologies would naively merge. Instead, a time-distortion model can be used where instead of the standard:

$$y_{ij}(t) = w_i(h_i^{-1}(t)) + \epsilon_{ij} \tag{7.3}$$

  where as before w_i is the amplitude and the h_i the phase variation function, the time-registration is conducted over two levels. One over the reduced dimension data $y_{ij}^{(1)}(t)$:

$$y_{ij}^{(1)}(t) = \mu(t) + \sum_{k=1}^{N_1} \xi_{ik}\phi_k^{(1)}(t) = w_i^{(1)}(h_i^{-1(1)}(t)) + \epsilon_{ij}^{(1)} \tag{7.4}$$

an advancement in the Evolutionary Linguistics literature.

[2] Catalan and Romanian.

and a second one over the reduced dimension dataset $y_{ij}^{(2)}(t)$ for each j individually:

$$y_{ij}^{(2)}(t) = \eta_j(t) + \sum_{l=1}^{N_2} \zeta_{ijl}\phi_l^{(2)}(t) = w_i^{(2)}(h_i^{-1(2)}(t)) + \epsilon_{ij}^{(2)} \tag{7.5}$$

where clearly:

$$y_{ij}(t) = y_{ij}^{(1)}(t) + y_{ij}^{(2)}(t) + \epsilon_{ij}(t). \tag{7.6}$$

This exposition offers a simple sketch of a broader idea. As multilevel techniques become more prominent, time-warping methodology will have to account for "level-related" design properties instead of plainly segmenting the data based on empirical insights.

- Decomposition of functional data. Having made the choice of working in a lower dimensional space than the one a functional dataset lies originally, the choice of dimension reduction framework is an open one. In the current work we used FPCA because it is the most well-studied and widely used decomposition. There was no intrinsic reason why one would not use another basis, this choice becoming even more prominent when we consider functional objects being defined on more than one continuum. In particular, as shown in [119] ICA [145] can be fruitful if used in conjunction with FPCA for curves. Especially for phylogenetic applications, where one might not be presented with phylogenetic information, this tandem dimension reduction approach can provide highly robust components. Nevertheless, we chose not to use this methodology in chapter 6. We decided that because we believed concatenating a two-dimensional functional object in a one-dimensional vector, while understood and investigated in the context of FPCA, is not in the context of ICA and that could jeopardize the statistical coherence of our approach. There is already a small literature of ICA applications for two-dimensional objects (eg. [276; 153]); nevertheless especially taking into account the small language specific sample that we might be presented with, issues of multiple optima in the negentropy function used by the ICA algorithm cannot be ignored and therefore caution is still needed. Therefore we recognize decompositions complementary to FPCA and ICA for two-dimensional objects in particular, as a potentially fruitful field of investigation.

It is thus evident that there is a series of emerging challenges for functional data analysis, both theoretical and applied. It appears that functional data analysis, generalizing on the findings of multivariate techniques, can offer tools for the further advancement of Phonetics as well as applied research in general.

Bibliography

[1] T. Abe, T. Kobayashi, and S. Imai. Robust pitch estimation with harmonics enhancement in noisy environments based on instantaneous frequency. In *Proceedings of Fourth International Conference on Spoken Language*, volume 2, pages 1277–1280. IEEE, 1996.

[2] J. Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):139–177, 1982.

[3] J. Aitchison. Principal component analysis of compositional data. *Biometrika*, 70(1):57–65, 1983.

[4] J. Aitchison and M. Greenacre. Biplots of compositional data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 51(4):375–392, 2002.

[5] H. Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, 1974.

[6] P. M. Anselone and P. J. Laurent. A general method for the construction of interpolating or smoothing spline-functions. *Numerische Mathematik*, 12:66–82, 1967.

[7] J. Antoch, L. Prchal, M. R. De Rosa, and P. Sarda. Functional linear regression with functional response: application to prediction of electricity consumption. In *Functional and Operatorial Statistics*, pages 23–29. Springer, 2008.

[8] A. Antoniadis, G. Gregoire, and I. W. McKeague. Wavelet methods for curve estimation. *Journal of the American Statistical Association*, 89(428):1340–1353, 1994.

[9] S. Aris-Brosou and Z. Yang. Effects of models of rate evolution on estimation of divergence dates with special reference to the metazoan 18s ribosomal rna phylogeny. *Systematic Biology*, 51(5):703–714, 2002.

[10] Aristotle. History of animals, $\sim$ 350bc. `http://classics.mit.edu/Aristotle/history_anim.html`. [Accessed Oct. 23, 2013.].

[11] J. A. D. Aston, J. Chiou, and J. P. Evans. Linguistic pitch analysis using functional principal component mixed effect models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(2):297–317, 2010.

[12] Q. D. Atkinson. Phonemic diversity supports a serial founder effect model of language expansion from africa. *Science*, 332(6027):346–349, 2011.

[13] Q. D. Atkinson and R. D. Gray. Curious parallels and curious connections-phylogenetic thinking in biology and historical linguistics. *Systematic Biology*, 54(4):513–526, 2005.

[14] R. Baayen. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge University Press, UK, 2008. Chapt.4.

[15] R. Baayen, D. Davidson, and D. Bates. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412, 2008.

[16] R. H. Baayen. *Analyzing linguistic data*. Cambridge University Press Cambridge, UK, 2008.

[17] D. Barber. *Bayesian Reasoning and Machine Learing*. Cambridge University Press, 2012. Chapt. 12, 16, 19 & 21.

[18] D. J. Barr, R. Levy, C. Scheepers, and H. J. Tily. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278, 2013.

[19] M. S. Bartlett, J. R. Movellan, and T. J. Sejnowski. Face recognition by independent component analysis. *Neural Networks, IEEE Transactions on*, 13(6):1450–1464, 2002.

[20] D. Bates. Penalized least squares versus generalized least squares representations of linear mixed models. lme4's vignette, 2012.

[21] D. Bates and S. DebRoy. Linear mixed models and penalized least squares. *Journal of Multivariate Analysis*, 91(1):1–17, 2004.

[22] D. Bates and M. Maechler. *lme4: Linear mixed-effects models using S4 classes*, 2013. R package version 0.99999911-1 (date last viewed : 2/5/13).

[23] J. M. Beaulieu, D.-C. Jhwueng, C. Boettiger, and B. C. OMeara. Modeling stabilizing selection: expanding the ornstein-uhlenbeck model of adaptive evolution. *Evolution*, 8(66):2369–2383, 2012.

[24] J. Benesty, M. M. Sondhi, and Y. Huang. *Springer handbook of speech processing*. Springer, 2008. Chapt. 9, 10, 21 & 46.

[25] M. Benko, W. Härdle, A. Kneip, et al. Common functional principal components. *The Annals of Statistics*, 37(1):1–34, 2009.

[26] G. Biau, B. Cadre, Q. Paris, et al. Cox process learning. Technical report, aboratoire de Statistique Théorique et Appliquée (LSTA), Université Pierre et Marie Curie (UPMC) - Paris VI, 2013.

[27] D. Billheimer, P. Guttorp, and W. F. Fagan. Statistical interpretation of species composition. *Journal of the American Statistical Association*, 96(456):1205–1214, 2001.

[28] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., 2006. Chapt. 1 & 12.

[29] A. Black and A. Hunt. Generating f0 contours from tobi labels using linear regression. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 3, pages 1385–1388 vol.3, 1996.

[30] J. Blinn. What's that deal with the dct? *Computer Graphics and Applications, IEEE*, 13(4):78–83, 1993.

[31] H.-J. Böckenhauer and D. Bongartz. *Algorithmic aspects of bioinformatics*. Springer-Verlag Berlin Heidelberg, 2007. Chapt. 11.

[32] P. Boersma. Accurate short-term analysis of the fundamental frequency and the harmonic-to-noise ratio of a sampled sound. *Proceedings of the Institute of Phonetic Sciences*, 17:97–110, 1993.

[33] B. P. Bogert, M. J. Healy, and J. W. Tukey. The quefrency alanysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe cracking. In *Proceedings of the Symposium on Time Series Analysis*, pages 209–243. Wiley, New York, 1963.

[34] B. Bolker. Draft r-sig-mixed-models faq. `http://glmm.wikidot.com/faq`. [Accessed Sep. 10, 2013.].

[35] B. M. Bolstad, R. A. Irizarry, M. Åstrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.

[36] H. D. Bondell, A. Krishna, and S. K. Ghosh. Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics*, 66(4):1069–1077, 2010.

[37] M. L. Borroff. *A Landmark Underspecification Account of the Patterning of Glottal Stop*. PhD thesis, Department of Linguistics, Stony Brook University, New York, 2007.

[38] A. Bouchard-Côté, D. Hall, T. L. Griffiths, and D. Klein. Automated reconstruction of ancient languages using probabilistic models of sound change. *Proceedings of the National Academy of Sciences*, 110(11):4224–4229, 2013.

[39] A. Bouchard-Côté, P. Liang, D. Klein, and T. L. Griffiths. A probabilistic approach to language change. In *Advances in Neural Information Processing Systems*, pages 169–176, 2007.

[40] P. R. Bouzas, M. J. Valderrama, A. M. Aguilera, and N. Ruiz-Fuentes. Modelling the mean of a doubly stochastic poisson process by functional data analysis. *Computational statistics & data analysis*, 50(10):2655–2667, 2006.

[41] L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

[42] B. A. Brumback and J. A. Rice. Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association*, 93:961–976, 1998.

[43] A. Bruns. Fourier-, hilbert- and wavelet-based signal analysis: are they really different approaches? *Journal of Neuroscience Methods*, 137(2):321 – 332, 2004.

[44] K. P. Burnham and D. R. Anderson. *Model selection and multi-model inference: a practical information-theoretic approach*. Springer, 2002.

[45] P. Buser and M. Imbert. *Audition*. MIT Press, 1st edition, 1992. Chapt. 2.

[46] M. A. Butler and A. A. King. Phylogenetic comparative analysis: A modelling approach for adaptive evolution. *American Naturalist*, 164(6):683–695, 2004.

[47] K. S. Button, J. P. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. Robinson, and M. R. Munafò. Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 2013.

[48] 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(6):4456–65, 2008.

[49] L. Campbell. *Historical Linguistics: An Introduction*. MIT Press, 1998. Chapt. 8 & 12.

[50] R. Cangelosi and A. Goriely. Component retention in principal component analysis with application to cDNA microarray data. *Biology Direct*, 2:2+, 2007.

[51] P. Castro, W. Lawton, and E. Sylvestre. Principal modes of variation for processes with continuous sample curves. *Technometrics*, 28(4):329–337, 1986.

[52] L. L. Cavalli-Sforza, A. Piazza, P. Menozzi, and J. Mountain. Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. *Proceedings of the National Academy of Sciences*, 85(16):6002–6006, 1988.

[53] Central Intelligence Agency. The CIA World Factbook. [Accessed Jul. 27, 2012. World:People and Society: Languages].

[54] C. Chatfield. *The Analysis of Time Series: An Introduction*. CRC press, 2003. Chapt. 6.

[55] K. Chen and H.-G. Müller. Modeling repeated functional observations. *Journal of the American Statistical Association*, 107(500):1599–1609, 2012.

[56] J. M. Cheverud, M. M. Dow, and W. Leutenegger. The quantitative assessment of phylogenetic constraints in comparative analyses: Sexual dimorphism in body weight among primates. *Evolution*, 39(6):pp. 1335–1351, 1985.

[57] 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(2):405–423, 2003.

[58] N. Chomsky. On certain formal properties of grammars. *Information and control*, 2(2):137–167, 1959.

[59] D. C. Collar, B. C. O'Meara, P. C. Wainwright, and T. J. Near. Piscivory limits diversfication of feeding morphology in centrarchid fishes. *Evolution*, 63(6), 2009.

[60] P. Craven and G. . Wahba. Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31:377–403, 1979.

[61] C. W. Cunningham, K. E. Omland, and T. H. Oakley. Reconstructing ancestral character states: a critical reappraisal. *Trends in Ecology & Evolution*, 13(9):361–366, 1998.

[62] A. C. Davison. *Statistical models*. Cambridge University Press, 2003. Chapt. 4.

[63] F. de Saussure and R. Harris. *Course in General Linguistics (Open Court Classics)*. Open Court, 1998.

[64] P. Delicado. Functional k-sample problem when data are density functions. *Computational Statistics*, 22(3):391–410, 2007.

[65] P. Delicado, R. Giraldo, C. Comas, and J. Mateu. Statistics for spatial functional data: some recent contributions. *Environmetrics*, 21(3-4):224–239, 2010.

[66] C. Di, C. Crainiceanu, B. Caffo, and N. Punjabi. Multilevel functional principal component analysis. *Annals of applied statistics*, 3(1):458–488, 2009.

[67] P. J. Diggle, J. Besag, and J. T. Gleaves. Statistical analysis of spatial point patterns by means of distance methods. *Biometrics*, pages 659–667, 1976.

[68] P. J. Diggle, J. Tawn, and R. Moyeed. Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(3):299–350, 1998.

[69] J. Droppo and A. Acero. A fine pitch model for speech. In *INTERSPEECH*, pages 2757–2760, 2007.

[70] A. S. Druzhkova, O. Thalmann, V. A. Trifonov, J. A. Leonard, N. V. Vorobieva, N. D. Ovodov, A. S. Graphodatsky, and R. K. Wayne. Ancient dna analysis affirms the canid from altai as a primitive dog. *PLoS ONE*, 8(3), 03 2013.

[71] I. L. Dryden. Statistical analysis on high-dimensional spheres and shape spaces. *Annals of statistics*, pages 1643–1665, 2005.

[72] T. E. Duncan, S. C. Duncan, F. Li, and L. A. Strycker. Multilevel modeling of longitudinal and functional data. *Modeling intraindividual variability with repeated measures data: Methods and applications*, pages 171–201, 2002.

[73] M. Dunn, A. Terrill, G. Reesink, R. A. Foley, and S. C. Levinson. Structural phylogenetics and the reconstruction of ancient language history. *Science*, 309(5743):2072–2075, 2005.

[74] J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3):279–300, 2003.

[75] J. P. Evans, M. Chu, J. A. D. . Aston, and C. Su. Linguistic and human effects on F0 in a tonal dialect of Qiang. *Phonetica*, 67:82–99, 2010.

[76] S. N. Evans, D. Ringe, and T. Warnow. Inference of divergence times as a statistical inverse problem. *Phylogenetic Methods and the Prehistory of Languages. McDonald Institute Monographs*, pages 119–130, 2004.

[77] 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.

[78] J. J. Faraway. Regression analysis for a functional response. *Technometrics*, 39(3):254–261, 1997.

[79] J. Felsenstein. Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of molecular evolution*, 17(6):368–376, 1981.

[80] J. Felsenstein. Phylogenies and the comparative method. *American Naturalist*, 125(1):1–15, 1985.

[81] J. Felsenstein. Phylogenies And Quantitative Characters. *Annual Review of Ecology and Systematics*, 19:445–471, 1988.

[82] J. Felsenstein. *Inferring phylogenies*, volume 2. Sinauer Associates Sunderland, 2004. Chapt. 20 & 30.

[83] F. Ferraty and P. Vieu. *Nonparametric functional data analysis: theory and practice*. Springer Verlag, New York, 2006. Chapt. 1.

[84] A. Fielding and H. Goldstein. Cross-classified and multiple membership structures in multilevel models: An introduction and review. Technical Report RR791, University of Birmingham, Birmingham, UK, 2006.

[85] P. Filzmoser, K. Hron, and C. Reimann. Principal component analysis for compositional data with outliers. *Environmetrics*, 20(6):621–632, 2009.

[86] W. M. Fitch. Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Biology*, 20(4):406–416, 1971.

[87] D. Friedman. Pseudo-maximum-likelihood speech pitch extraction. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 25(3):213–221, 1977.

[88] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, 28(2):337–407, 2000.

[89] H. Fujisaki. Information, Prosody, and Modeling - with Emphasis on Tonal Features of Speech -. In *Speech Prosody 2004*, pages 1–10, 2004.

[90] H. Fujisaki and S. Ohno. Comparison and assessment of models in the study of fundamental frequency contours of speech. In *Intonation: Theory, Models and Applications*, pages 131–134, 1997.

[91] M. Galassi, J. Theiler, J. Davies, and B. Gough. *GNU Scientific Library Reference manual (3rd Ed.)*. Network Theory Limited, 2011.

[92] S. Gallón, J.-M. Loubes, and E. Maza. Statistical properties of the quantile normalization method for density curve alignment. *Mathematical biosciences*, 242(2):129–142, 2013.

[93] D. Garcia. Robust smoothing of gridded data in one and higher dimensions with missing values. *Computational Statistics & Data Analysis*, 54(4):1167–1178, 2010.

[94] L. A. García-Cortés, D. Sorensen, et al. Alternative implementations of monte carlo em algorithms for likelihood inferences. *Genetics Selection Evolution*, 33(4):443, 2001.

[95] T. Gasser and A. Kneip. Searching for Structure in Curve Samples. *Journal of The American Statistical Association*, 90:1179–1188, 1995.

[96] T. Gasser and H.-G. Müller. Estimating regression functions and their derivatives by the kernel method. *Scandinavian Journal of Statistics*, 11(3):171–185, 1984.

[97] C. Gendrot and M. Adda-Decker. Impact of duration on F1/F2 formant values of oral vowels: an automatic analysis of large broadcast news corpora in french and german. *Variations*, 2(22.5):2–4, 2005.

[98] D. Gervini and T. Gasser. Self-modelling warping functions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 66(4):959–971, 2004.

[99] A. Ghodsi. Dimensionality reduction a short tutorial. Technical report, Department of Statistics and Actuarial Science, Univ. of Waterloo, Ontario, Canada, 2006.

[100] J. Goldsmith. Tone languages. In *Encyclopedia of Language and Linguistics (New York: Elsevier Science)*, pages 4626–4628, 1994.

[101] G. H. Golub and C. F. Van Loan. *Matrix Computations (2nd Ed.)*. Johns Hopkins University Press, 1989. Chapt. 4 & 5.

[102] R. Gomulkiewicz and J. H. Beder. The selection gradient of an infinite-dimensional trait. *SIAM Journal on Applied Mathematics*, 56(2):509–523, 1996.

[103] E. Grabe, G. Kochanski, and J. Coleman. The intonation of native accent varieties in the british islespotential for miscommunication. In *English pronunciation models: a changing scene*, pages 311–338, 2005.

[104] E. Grabe, G. Kochanski, and J. Coleman. Connecting intonation labels to mathematical descriptions of fundamental frequency. *Language and Speech*, 50(3):281–310, 2007.

[105] R. D. Gray and Q. D. Atkinson. Language-tree divergence times support the anatolian theory of indo-european origin. *Nature*, 426(6965):435–439, 2003.

[106] R. D. Gray, Q. D. Atkinson, and S. J. Greenhill. Language evolution and human history: what a difference a date makes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1567):1090–1100, 2011.

[107] R. D. Gray and F. M. Jordan. Language trees support the express-train sequence of austronesian expansion. *Nature*, 405(6790):1052–1055, 2000.

[108] S. Greven and T. Kneib. On the behaviour of marginal and conditional aic in linear mixed models. *Biometrika*, 97(4):773–789, 2010.

[109] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan. Primitives-based evaluation and estimation of emotions in speech. *Speech Communication*, 49(10-11):787–800, 2007.

[110] G. Grimmett and D. Stirzaker. *Probability and random processes*. Oxford University Press, Third edition, 2001. Chapt. 5.

[111] W. Gu, K. Hirose, and H. Fujisaki. Modeling the effects of emphasis and question on fundamental frequency contours of cantonese utterances. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(4):1155–1170, 2006.

[112] M. Gubian, L. Boves, and F. Cangemi. Joint analysis of F_0 and speech rate with functional data analysis. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 4972–4975. IEEE, 2011.

[113] M. Gubian, F. Cangemi, and L. Boves. Automatic and data driven pitch contour manipulation with functional data analysis. In *Proceedings of Speech Prosody*, pages 100231:1–4, 2010.

[114] M. Gubian, F. Torreira, H. Strik, L. Boves, et al. Functional Data Analysis as a Tool for Analyzing Speech Dynamics - A Case Study on the French Word était. In *INTERSPEECH*, pages 2199–2202, 2009.

[115] W. Guo. Functional mixed effects models. *Biometrics*, 58:121–128, 2002.

[116] J. D. Hadfield. Mcmc methods for multi-response generalized linear mixed models: The MCMCglmm R package. *Journal of Statistical Software*, 33(2):1–22, 2010.

[117] P. Z. Hadjipantelis, J. A. D. Aston, and J. P. Evans. Characterizing fundamental frequency in Mandarin: A functional principal component approach utilizing mixed effect models. *Journal of the Acoustical Society of America*, 131(6):4651–64, 2012.

[118] P. Z. Hadjipantelis, J. A. D. Aston, H.-G. Müller, and J. Moriarty. Analysis of spike train data: A multivariate mixed effects model for phase and amplitude. *Electronic Journal of Statistics*, 2014. (Accepted for publication).

[119] P. Z. Hadjipantelis, N. S. Jones, J. Moriarty, D. A. Springate, and C. G. Knight. Function-valued traits in evolution. *Journal of The Royal Society Interface*, 10(82), 2013.

[120] A. Halevy, P. Norvig, and F. Pereira. The unreasonable effectiveness of data. *Intelligent Systems, IEEE*, 24(2):8–12, 2009.

[121] P. Hall, H. Müller, and J. Wang. Properties of principal component methods for functional and longitudinal data analysis. *The Annals of Statistics*, 34(3):1493–1517, 2006.

[122] P. Hall, H.-G. Müller, and F. Yao. Modelling sparse generalized longitudinal observations with latent gaussian processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):703–723, 2008.

[123] M. H. Hansen and B. Yu. Minimum description length model selection criteria for generalized linear models. *Lecture Notes-Monograph Series*, pages 145–163, 2003.

[124] T. Hansen. Stabilizing selection and the comparative analysis of adaptation. *Evolution*, 51(5):1341–1351, 1997.

[125] T. Hansen and E. Martins. Translating between microevolutionary process and macroevolutionary patterns: The correlation structure of interspecific data. *Evolution*, 50(4):1404–1417, 1996.

[126] T. Hansen, J. Pienaar, and S. H. Orzack. A comparative method for studying adaptation to a randomly evolving environment. *Evolution*, 8(62):1965–1977, 2008.

[127] T. F. Hansen and K. Bartoszek. Interpreting the evolutionary regression: the interplay between observational and biological errors in phylogenetic comparative studies. *Systematic biology*, 61(3):413–425, 2012.

[128] L. Harmon, J. Weir, C. Brock, R. Glor, and W. Challenger. GEIGER: investigating evolutionary radiations. *Bioinformatics*, 24:129–131, 2008.

[129] T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning*. Springer, 2009. Chapt. 7.

[130] R. R. Hausser. *Foundations of computational linguistics: Man-machine communication in natural language*. Springer, Berlin, 1999. Chapt. 5 & 6.

[131] X. He and L. Deng. Speech recognition, machine translation, and speech translation; a unified discriminative learning paradigm [lecture notes]. *Signal Processing Magazine, IEEE*, 28(5):126 –133, 2011.

[132] W. Hess and H. Indefrey. Accurate time-domain pitch determination of speech signals by means of a laryngograph. *Speech communication*, 6(1):55–68, 1987.

[133] K. Hirose, H. Fujisaki, and M. Yamaguchi. Synthesis by rule of voice fundamental frequency contours of spoken japanese from linguistic information. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 9, pages 597–600, 1984.

[134] 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*, pages 1233–1236, 2007.

[135] 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.

[136] H. J. Ho and T.-I. Lin. Robust linear mixed models using the skew t distribution with application to schizophrenia data. *Biometrical Journal*, 52(4):449–469, 2010.

[137] M. T. Holder, J. Sukumaran, and P. O. Lewis. A justification for reporting the majority-rule consensus tree in bayesian phylogenetics. *Systematic biology*, 57(5):814–821, 2008.

[138] S. P. Holmes. Phylogenies: an overview. *IMA Volumes in mathematics and its applications*, 112:81–118, 1999.

[139] M. Honda. Human speech production mechanisms. *NTT Technical Review*, 1(2):24–29, 2003.

[140] A. Housen and F. Kuiken. Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, 30(4):461–473, 2009.

[141] D. C. Hoyle. Automatic pca dimension selection for high dimensional data and small sample sizes. *Journal of Machine Learning Research*, 9(12):2733–2759, 2008.

[142] P.-S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson. Singing-voice separation from monaural recordings using robust principal component analysis. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 57–60. IEEE, 2012.

[143] J. P. Huelsenbeck, B. Larget, R. E. Miller, and F. Ronquist. Potential applications and pitfalls of bayesian inference of phylogeny. *Systematic biology*, 51(5):673–688, 2002.

[144] J. P. Huelsenbeck, B. Rannala, and J. P. Masly. Accommodating phylogenetic uncertainty in evolutionary studies. *Science*, 288(5475):2349–2350, 2000.

[145] A. Hyvärinen and E. Oja. Independent Component Analysis: Algorithms and Applications. *Neural Networks*, 13(4-5):411–430, 2000.

[146] J. G. Ibrahim, H. Z. Zhu, R. I. G. Garcia, and R. Guo. Fixed and random effects selection in mixed effects models. *Biometrics*, 67(2):495–503, 2011.

[147] J. Illian, E. Benson, J. Crawford, and H. Staines. Principal component analysis for spatial point processes - assessing the appropriateness of the approach in an ecological context. In *Case studies in spatial point process modeling*, pages 135–150. Springer, 2006.

[148] K. Irwin and P. Carter. Constraints on the evolution of function-valued traits: A study of growth in tribolium castaneum. *Journal of Evolutionary Biology*, 2013.

[149] A. Izenman. *Modern Multivariate Statistical Techniques: Regression, Classification and Manifold Learning*. Springer Verlag, New York, 2008. Chapt.6.

[150] E. Jacewicz, R. A. Fox, and L. Wei. Between-speaker and within-speaker variation in speech tempo of american english. *Journal of the Acoustical Society of America*, 128(2):839–50, 2010.

[151] 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.

[152] E. Jaynes. Prior information and ambiguity in inverse problems. *Inverse Problems*, 14:151–166, 1984.

[153] D. Jeong, M. Lee, and S.-W. Ban. (2d)2pca-ica: A new approach for face representation and recognition. In *Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on*, pages 1792–1797, 2009.

[154] C. Jill Harrison and J. A. Langdale. A step by step guide to phylogeny reconstruction. *The Plant Journal*, 45(4):561–572, 2006.

[155] K. Johnson. *Acoustic and auditory phonetics*. Blackwell Publishing, 2nd edition, 2003. Chapt. 1, 4 & 5.

[156] I. Jolliffe. *Principal component analysis*. Wiley Online Library, 2005. Chapt. 3.

[157] N. S. Jones and J. Moriarty. Evolutionary inference for function-valued traits: Gaussian process regression on phylogenies. *J. R. Soc. Interface*, 10(78), 2013.

[158] S.-A. Jun. *Prosodic typology: the phonology of intonation and phrasing*. OUP Oxford, UK, 2006. Chapt.2 The original ToBI system and the evolution of the ToBI framework by Beckman M.E. et al.

[159] D. Jurafsky and J. H. Martin. *Speech and language processing*. Prentice Hall, International edition, 2009. Chapt. 4 & 7.

[160] C. G. Kaufman and S. R. Sain. Bayesian functional ANOVA modeling using gaussian process prior distributions. *Bayesian Analysis*, 5(1):123–149, 2010.

[161] C. T. Kelley. Iterative Methods for Optimization. *Frontiers in Applied Mathematics*, SIAM, 1999.

[162] K. Kenobi, I. L. Dryden, and H. Le. Shape curves and geodesic modelling. *Biometrika*, 97(3):567–584, 2010.

[163] M. Kerr. Evolutionary inference for functional data: Using Gaussian processes on phylogenies of functional data objects, 2012. MSc Thesis - Univ. of Glasgow.

[164] J. Kim, J. Choi, J. Yi, and M. Turk. Effective representation using ica for face recognition robust to local distortion and partial occlusion. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(12):1977–1981, 2005.

[165] J. Kim and T. Warnow. Tutorial on phylogenetic tree estimation. In *Intelligent Systems for Molecular Biology*. Heidelberg, 1999.

[166] J. G. Kingsolver, R. Gomulkiewicz, and P. A. Carter. Variation, selection and evolution of function-valued traits. *Genetica*, 112-113:87–104, 2001.

[167] M. Kirkpatrick and N. Heckman. A quantitative genetic model for growth, shape, reaction norms, and other infinite-dimensional characters. *Journal of Mathematical Biology*, 27:429–450, 1989.

[168] A. Kneip and T. Gasser. Statistical Tools to Analyze Data Representing a Sample of Curves. *Annals of Statistics*, 20:1266–1305, 1992.

[169] A. Kneip and J. O. Ramsay. Combining Registration and Fitting for Functional Models. *Journal of the American Statistical Association*, 103(483):1155–1165, 2008.

[170] C. G. Knight, R. Kassen, H. Hebestreit, and P. B. Rainey. Global analysis of predicted proteomes: functional adaptation of physical properties. *Proceedings of the National Academy of Sciences of the United States of America*, 101(22):8390–8395, 2004.

[171] L. Koenig, J. Lucero, and A. Löfqvist. Studying articulatory variability using functional data analysis. In *Proceedings of the 15 th International Congress of Phonetic Sciences*, pages 269–272, 2003.

[172] 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(124):3158–3170, 2008.

[173] S. J. Koopman and J. Durbin. Fast filtering and smoothing for multivariate state space models. *Journal of Time Series Analysis*, 21(3):281–296, 2000.

[174] S. Kurtek, E. Klassen, Z. Ding, and A. Srivastava. A novel riemannian framework for shape analysis of 3d objects. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010*, pages 1625–1632. IEEE, 2010.

[175] S. Kurtek, A. Srivastava, E. Klassen, and Z. Ding. Statistical modeling of curves using shapes and related features. *Journal of the American Statistical Association*, 107(499):1152–1165, 2012.

[176] S. A. Kurtek, A. Srivastava, and W. Wu. Signal estimation under random time-warpings and nonlinear signal alignment. In *Advances in Neural Information Processing Systems*, pages 675–683, 2011.

[177] J. Laborde, D. Robinson, A. Srivastava, E. Klassen, and J. Zhang. Rna global alignment in the joint sequence–structure space using elastic shape analysis. *Nucleic acids research*, 41(11):e114–e114, 2013.

[178] P. Ladefoged. *A Course in Phonetics*. Patakis, 5th edition, 2010. Chapt. 1 & 2, Greek Translation by Mary Baltazani.

[179] N. M. Laird and J. H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974, 1982.

[180] L. Lan. *Variable Selection in Linear Mixed Model for Longitudinal Data*. PhD thesis, North Carolina State University, 2006. PhD Thesis.

[181] R. Lande. Natural selection and random genetic drift in phenotypic evolution. *Evolution*, 30:314–334, 1976.

[182] R. Lande. Quantitative genetic analysis of multivariate evolution, applied to brain:body size allometry. *Evolution*, 33:402–416, 1979.

[183] V. Latsch and S. L. Netto. Pitch-synchronous time alignment of speech signals for prosody transplantation. In *International Symposium on Circuits and Systems (ISCAS 2011)*, pages 2405–2408. IEEE, 2011.

[184] G. F. Lawler and V. Limic. *Random walk: a modern introduction*. Cambridge University Press, 2010. Chapt. 3.

[185] H. Le and D. G. Kendall. The Riemannian Structure of Euclidean Shape Spaces: A Novel Environment for Statistics. *The Annals of Statistics*, 21(3):1225–1271, 1993.

[186] A. D. Leaché and B. Rannala. The accuracy of species tree estimation under simulation: A comparison of methods. *Systematic Biology*, 60(2):126–137, 2011.

[187] H. Lee and Z. Bien. Sub-nyquist nonuniform sampling and perfect reconstruction of speech signals. In *TENCON 2005 2005 IEEE Region 10*, pages 1–6, 2005.

[188] S. Lee, D. Byrd, and J. Krivokapic. Functional data analysis of prosodic effects on articulatory timing. *Journal of the Acoustical Society of America*, 119(3):1666–1671, 2006.

[189] Y. Lee and J. A. Nelder. Hierarchical generalized linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 619–678, 1996.

[190] Y. Lee and J. A. Nelder. Conditional and Marginal Models: Another View. *Statistical Science*, 19(2):219–238, 2004.

[191] T. Leonard. A Bayesian method for histograms. *Biometrika*, 60(2):297–308, 1973.

[192] H. Li, A. Coghlan, J. Ruan, L. J. Coin, J.-K. Hériché, L. Osmotherly, R. Li, T. Liu, Z. Zhang, L. Bolund, G. K. Wong, W. Zheng, P. Dehal, J. Wang, and R. Durbin. TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Research*, 34:D572–D580, Jan. 2006.

[193] Y.-H. Lin. *The Sounds of Chinese*. Cambridge University Press, 2007. Chapt. 5.

[194] M. A. Lindquist, J. Spicer, I. Asllani, and T. D. Wager. Estimating and testing variance components in a multi-level glm. *Neuroimage*, 59(1):490–501, 2012.

[195] B. C. Look. Gottfried Wilhelm Leibniz. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Stanford University, 2013.

[196] J. Louw and E. Barnard. Automatic intonation modeling with INTSINT. *Proceedings of the Pattern Recognition Association of South Africa*, pages 107–111, 2004.

[197] H. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos. Mpca: Multilinear principal component analysis of tensor objects. *Neural Networks, IEEE Transactions on*, 19(1):18–39, 2008.

[198] J. Lucero and A. Löfqvist. Measures of articulatory variability in VCV sequences. *Acoustics Research Letters Online*, 6(2):80, 2005.

[199] D. J. MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003. Chapt. 11.

[200] I. Maddieson. *The World Atlas of Language Structures Online. Chapt. 13: Tone*. Max Planck Digital Library, Munich, 2011.

[201] D. R. Maddison and K.-S. Schulz. The tree of life web project. `http://tolweb.org`, 2007.

[202] S. G. Mallat. A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 11(7):674–693, 1989.

[203] C. Manté, J. Durbec, and J. C. Dauvin. A functional data-analytic approach to the classification of species according to their spatial dispersion. application to a marine macrobenthic community from the bay of morlaix (western english channel). *Journal of Applied Statistics*, 32(8):831–840, 2005.

[204] MATLAB. *version 7.10.0 (R2010a)*. The MathWorks Inc., Natick, Massachusetts, 2010.

[205] P. H. Matthews. *Linguistics - A very short introduction*. Oxford University Press, 2003. Chapt. 1, 6 & 7.

[206] A. McMahon and R. McMahon. Finding families: quantitative methods in language classification. *Transactions of the Philological Society*, 101(1):7–55, 2003.

[207] 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.

[208] T. P. Minka. Automatic choice of dimensionality for PCA. *Advances in Neural Information Processing Systems*, 15:598–604, 2001.

[209] W. Mio, A. Srivastava, and S. Joshi. On shape of plane elastic curves. *International Journal of Computer Vision*, 73(3):307–324, 2007.

[210] H. Mixdorff. Foreign accent in intonation patterns-a contrastive study applying a quantitative model of the f0 contour. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 3, pages 1469–1472 vol.3, 1996.

[211] H. Mixdorff. Production of broad and narrow focus in german a study applying a quantitative model. In *Intonation: Theory, Models and Applications*, pages 239–242, 1997.

[212] 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, pages 1281–1284. IEEE, 2000.

[213] H. Mixdorff. Speech technology, tobi, and making sense of prosody. In *Speech Prosody 2002, International Conference*, pages 31–38, 2002.

[214] 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*, pages 873–876. ISCA, 2003.

[215] H. Mixdorff and D. Mehnert. Exploring the naturalness of several german high-quality-text-to-speech systems. In *EUROSPEECH*, pages 1859–1862. ISCA, 1999.

[216] H. Mixdorff and C. Widera. Perceived prominence in terms of a linguistically motivated quantitative intonation model. In *INTERSPEECH*, pages 403–406, 2001.

[217] A. Moreno and J. A. Fonollosa. Pitch determination of noisy speech using higher order statistics. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, pages 133–136. IEEE, 1992.

[218] J. S. Morris and R. J. Carroll. Wavelet-based functional mixed models. *Journal of the Royal Statistical Society, Series B*, 68:179–199, 2006.

[219] S. P. Moss, D. A. Joyce, S. Humphries, K. J. Tindall, and D. H. Lunt. Comparative analysis of teleost genome sequences reveals an ancient intron size expansion in the zebrafish lineage. *Genome biology and evolution*, 3:1187–96, 2011.

[220] T. Nakatani and T. Irino. Robust and accurate fundamental frequency estimation based on dominant harmonic components. *The Journal of the Acoustical Society of America*, 116:3690, 2004.

[221] L. Nakhleh, D. Ringe, and T. Warnow. Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages. *Language*, pages 382–420, 2005.

[222] G. Nandris. The development and structure of rumanian. *The Slavonic and East European Review*, 30(74):7–39, 1951.

[223] 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, pages 1644–1647. IEEE, 2002.

[224] G. K. Nicholls and R. D. Gray. Dated ancestral trees from binary trait data and their application to the diversification of languages. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(3):545–566, 2008.

[225] J. Nocedal and S. J. Wright. *Numerical optimization, Springer Series in Operations Research*. Springer, New York, 2006. Chapt. 6.

[226] F. Nolan. Acoustic Phonetics - International Encyclopedia of Linguistics. e-reference edition, 2003.

[227] M. P. Oakes. Computer estimation of vocabulary in a protolanguage from word lists in four daughter languages. *Journal of Quantitative Linguistics*, 7(3):233–243, 2000.

[228] M. Pagel. Inferring evolutionary processes from phylogenies. *Zoologica Scripta*, 26(4):331–348, 1997.

[229] M. Pagel. Inferring the historical patterns of biological evolution. *Nature*, 401(6756):877–84, 1999.

[230] M. Pagel. Human language as a culturally transmitted replicator. *Nature Reviews Genetics*, 10(6):405–415, 2009.

[231] M. Pagel, Q. D. Atkinson, and A. Meade. Frequency of word-use predicts rates of lexical evolution throughout indo-european history. *Nature*, 449(7163):717–720, 2007.

[232] E. Paradis. *Analysis of Phylogenetics and Evolution with R*. Springer, 2012. Chapt. 1.

[233] H. G. Parker, L. V. Kim, N. B. Sutter, S. Carlson, T. D. Lorentzen, T. B. Malek, G. S. Johnson, H. B. DeFrance, E. A. Ostrander, and L. Kruglyak. Genetic structure of the purebred domestic dog. *Science*, 304(5674):1160–1164, 2004.

[234] H. D. Patterson and R. Thomson. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545–554, 1971.

[235] V. Pawlowsky-Glahn and J. Egozcue. Compositional data and their analysis: an introduction. *Geological Society, London, Special Publications*, 264(1):1–10, 2006.

[236] S. Petrone, M. Guindani, and A. E. Gelfand. Hybrid dirichlet mixture models for functional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(4):755–782, 2009.

[237] D. Pigoli and L. M. Sangalli. Wavelets in functional data analysis: Estimation of multidimensional curves and their derivatives. *Computational Statistics & Data Analysis*, 56(6):1482–1498, 2012.

[238] J. Pinheiro and D. Bates. *Mixed-effects models in S and S-PLUS*. Springer Verlag, New York, 2009. Chapt.2.

[239] J. Pinheiro, D. Bates, S. DebRoy, D. Sarkar, and R Core Team. *nlme: Linear and Nonlinear Mixed Effects Models*, 2013. R package version 3.1-109.

[240] J. C. Pinheiro, C. Liu, and Y. N. Wu. Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. *Journal of Computational and Graphical Statistics*, 10(2):249–276, 2001.

[241] B. Pinkowski. Principal component analysis of speech spectrogram images. *Pattern recognition*, 30(5):777–787, 1997.

[242] S. D. Pletcher and C. J. Geyer. The genetic analysis of age-dependent traits: modeling the character process. *Genetics*, 153(2):825–35, 1999.

[243] M. Powell. A direct search optimization method that models the objective and constraint functions by linear interpolation. In S. Gomez and J.-P. Hennart, editors, *Advances in Optimization and Numerical Analysis*, volume 275 of *Mathematics and Its Applications*, pages 51–67. Springer, 1994.

[244] S. Prom-on, Y. Xu, and B. Thipakorn. Quantitative target approximation model: Simulating underlying mechanisms of tones and intonations. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 749–752, 2006.

[245] S. Prom-On, Y. Xu, and B. Thipakorn. Modeling tone and intonation in mandarin and english as a process of target approximation. *The Journal of the Acoustical Society of America*, 125:405, 2009.

[246] H. Quene. On the just noticeable difference for tempo in speech. *Journal of Phonetics*, 35(3):353–362, 2007.

[247] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[248] L. Rabiner. On the use of autocorrelation analysis for pitch detection. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 25(1):24–33, 1977.

[249] L. Rabiner. A tutorial on HMM and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[250] L. Rabiner, M. Cheng, A. Rosenberg, and C. McGonegal. A comparative performance study of several pitch detection algorithms. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 24(5):399–418, 1976.

[251] L. R. Rabiner and R. W. Schafer. Introduction to digital speech processing. *Foundations and trends in signal processing*, 1(1):1–194, 2007.

[252] J. Ramsay. Multilevel modeling of longitudinal and functional data. *Modeling intraindividual variability with repeated measures data: Methods and applications*, pages 171–201, 2002.

[253] J. Ramsay and B. Silverman. *Applied functional data analysis: methods and case studies*. Springer Verlag, New York, 2002. Chapt.1.

[254] J. Ramsay and B. Silverman. *Functional data analysis*. Springer Verlag, New York, 2005. Chapt. 3, 4 & 7.

[255] J. Ramsay, X. Wang, and R. Flanagan. A functional data analysis of the pinch force of human finger. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 44(1):17–30, 1995.

[256] J. Ramsay and L. Xiaochun. Curve registration. *Journal of the Royal Statistical Society. Series B (Methodological)*, 60(2):351–363, 1998.

[257] J. O. Ramsay. When the data are functions. *Psychometrika*, 47:379–396, 1982.

[258] J. O. Ramsay. Functional components of variation in handwriting. *Journal of the American Statistical Association*, 95:9–15, 2000.

[259] 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(99):3718–3727, 1996.

[260] S. Rangachari and P. Loizou. A noise-estimation algorithm for highly non-stationary environments. *Speech Communication*, 48(2):220–231, 2006.

[261] C. Rao. Some statistical methods for comparison of growth curves. *Biometrics*, 14(1):1–17, 1958.

[262] C. Rao. The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves. *Biometrika*, 52(3/4):447–458, 1965.

[263] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006. Chapt. 2.

[264] 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(1):223–239, 2009.

[265] S. Renals, N. Morgan, H. Bourlard, M. Cohen, and H. Franco. Connectionist probability estimators in hmm speech recognition. *Speech and Audio Processing, IEEE Transactions on*, 2(1):161–174, 1994.

[266] L. J. Revell. Size-correction and principal components for interspecific comparative studies. *Evolution*, 63(12):3258–3268, 2009.

[267] 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(1):233–243, 1991.

[268] D. Ringe, T. Warnow, and A. Taylor. Indo-european and computational cladistics. *Transactions of the philological society*, 100(1):59–129, 2002.

[269] B. D. Ripley. Modelling spatial patterns. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 172–212, 1977.

[270] P. Roach. *English phonetics and phonology : a practical course*. Cambridge University Press, 3rd edition, 2000. Chapt. 20.

[271] C. P. Robert. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation (Springer Texts in Statistics)*. Springer-Verlag New York, 2001. Chapt. 5.

[272] N. Roma and L. Sousa. Review: A tutorial overview on the properties of the discrete cosine transform for encoded image and video processing. *Signal Processing*, 91(11):2443–2464, 2011.

[273] F. Ronquist. Bayesian inference of character evolution. *Trends in Ecology & Evolution*, 19(9):475–481, 2004.

[274] G. Rosa, D. Gianola, and C. Padovani. Bayesian longitudinal data analysis with mixed models and thick-tailed distributions using mcmc. *Journal of Applied Statistics*, 31(7):855–873, 2004.

[275] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[276] H. Safavi, N. Correa, W. Xiong, A. Roy, T. Adali, V. R. Korostyshevskiy, C. C. Whisnant, and F. Seillier-Moiseiwitsch. Independent component analysis of 2-d electrophoresis gels. *Electrophoresis*, 29(19):4017–4026, 2008.

[277] T. Sainath, B. Ramabhadran, D. Nahamoo, D. Kanevsky, D. Van Compernolle, K. Demuynck, J. Gemmeke, J. Bellegarda, and S. Sundaram. Exemplar-based processing for speech recognition: An overview. *Signal Processing Magazine, IEEE*, 29(6):98–113, 2012.

[278] H. Sakoe. Two-level dp-matching–a dynamic programming-based pattern matching algorithm for connected word recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 27(6):588 – 595, 1979.

[279] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 26(1):43–49, 1978.

[280] N. Salamin, R. O. Wuest, S. Lavergne, W. Thuiller, and P. B. Pearman. Assessing rapid evolution in a changing environment. *Trends in Ecology & Evolution*, 25(12):692–8, 2010.

[281] L. Sangalli, J. Ramsay, and T. Ramsay. Spatial spline regression models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 75(4):1–23, 2013.

[282] G. Saon and J.-T. Chien. Large-vocabulary continuous speech recognition systems: A look at some recent advances. *Signal Processing Magazine, IEEE*, 29(6):18–33, 2012.

[283] H. Schielzeth and W. Forstmeier. Conclusions beyond support: overconfident estimates in mixed models. *Behavioral Ecology*, 20(2):416–420, 2009.

[284] B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319, 1998.

[285] M. Schroeder, T. D. Rossing, F. Dunn, W. M. Hartmann, D. M. Campbell, and N. H. Fletcher. *Springer Handbook of Acoustics*. Springer Publishing Company, Incorporated, 1st edition, 2007. Chapt. 13 & 16.

[286] R. S. Schwartz and R. L. Mueller. Branch length estimation and divergence dating: estimates of error in bayesian and maximum likelihood frameworks. *BMC evolutionary biology*, 10(1):5, 2010.

[287] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[288] S. R. Searle, G. Casella, and C. E. McCulloch. *Variance components*. John Wiley & Sons, Inc, second edition, 2006. Chapt. 6.

[289] B. Secrest and G. Doddington. An integrated pitch tracking algorithm for speech systems. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '83.*, volume 8, pages 1352–1355, 1983.

[290] M. Seeger. Gaussian processes for machine learning. *International Journal of Neural Systems*, 14:2004, 2004.

[291] J. Q. Shi and T. Choi. *Gaussian process regression analysis for functional data*. CRC Press, 2011. Chapt. 1 & 5.

[292] J. Q. Shi, B. Wang, R. Murray-Smith, and D. Titterington. Gaussian process functional regression modeling for batch data. *Biometrics*, 63(3):714–723, 2007.

[293] C. Shih and G. Kochanski. Chinese Tone Modeling with Stem-ML. In *INTERSPEECH*, pages 67–70, 2000.

[294] R. Sim. Mobile robot localization from learned landmarks. Technical Report CIM-98-03, Centre for Intelligent Machines, McGill University, Montreal, QC, 1998.

[295] K. Sjőlander and J. Beskow. Wavesurfer [Computer program] (Version 1.8.5), 2009.

[296] T. Speed. Terence's Stuff: And ANOVA thing. *IMS Bulletin*, page 16, 2010.

[297] R. Stasinski and J. Konrad. Reduced-complexity shape-adaptive dct for region-based image coding. In *Image Processing, 1998. ICIP 98. International Conference on*, pages 114–118. IEEE, 1998.

[298] J. R. Stinchcombe, "Function-valued Traits Working Group", and M. Kirkpatrick. Genetics and evolution of function-valued traits: understanding environmentally responsive phenotypes. *Trends in ecology & evolution*, 27(11):637–647, 2012.

[299] G. Strang. The discrete cosine transform. *SIAM review*, 41(1):135–147, 1999.

[300] W. W. Stroup. *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications*. CRC Press, Boca Raton, FL, 2013. Chapt. 4.

[301] Z. Su and Z. Wang. An approach to affective-tone modeling for mandarin. In J. Tao, T. Tan, and R. Picard, editors, *Affective Computing and Intelligent Interaction*, volume 3784 of *Lecture Notes in Computer Science*, pages 390–396. Springer Berlin Heidelberg, 2005.

[302] S. Sudhoff. *Methods in empirical prosody research*. Walter De Gruyter Inc. Berlin, 2006. Chapt. 4, Prosody Beyond Fundamental Frequency by Greg Kochanski.

[303] P. Tang and H. Müller. Time-synchronized clustering of gene expression trajectories. *Biostatistics*, 10(1):32–45, 2009.

[304] R. Tang and H.-G. Müller. Pairwise curve synchronization for functional data. *Biometrika*, 95(4):875–889, 2008.

[305] P. Taylor. Analysis and synthesis of intonation using the TILT model. *Journal of the Acoustical Society of America*, 107(3):1697–1714, 2000.

[306] E. Terhardt, G. Stoll, and M. Seewann. Algorithm for extraction of pitch and pitch salience from complex tonal signals. *Journal of the Acoustical Society of America*, 71:679–688, 1982.

[307] The Functional Phylogenies Group. Phylogenetic inference for function-valued traits: speech sound evolution. *Trends in Ecology & Evolution*, 27(3):160–166, 2012.

[308] S. Theis. *Deriving probabilistic short-range forecasts from a deterministic high-resolution model*. PhD thesis, University of Bonn - Universität Bonn, 2005. PhD Thesis.

[309] M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

[310] 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, December 2005.

[311] C. Tseng, Y. Cheng, and C. Chang. Sinica COSPRO and Toolkit: Corpora and Platform of Mandarin Chinese Fluent Speech. In *Proceedings of Oriental COCOSDA*, pages 6–8, 2005.

[312] C. Tseng, S. Pin, Y. Lee, H. Wang, and Y. Chen. Fluent speech prosody: Framework and modeling. *Speech Communication*, 46(3-4):284–309, 2005.

[313] C. Y. Tseng and F. C. Chou. Machine readable phonetic transcription system for Chinese dialects spoken in Taiwan. *Journal of the Acoustical Society of Japan*, 20:215–223, 1999.

[314] J. D. Tucker, W. Wu, and A. Srivastava. Generative models for functional data using phase and amplitude separation. *Computational Statistics & Data Analysis*, 61:50–66, 2013.

[315] L. R. Tucker. Determination of parameters of a functional relationship by factor analysis. *Psychometrika*, 23:19–23, 1958.

[316] W. L. Twining. *Theories of evidence: Bentham and Wigmore*. Stanford University Press, 1985.

[317] M. Valderrama. An overview to modelling functional data. *Computational Statistics*, 22(3):331–334, 2007.

[318] C. Venditti, A. Meade, and M. Pagel. Phylogenies reveal new interpretation of speciation and the red queen. *Nature*, 463(7279):349–52, 2010.

[319] J. Vermaak and A. Blake. Nonlinear filtering for speaker tracking in noisy and reverberant environments. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, volume 5, pages 3021–3024. IEEE, 2001.

[320] T. Warnow. Mathematical approaches to comparative linguistics. *Proceedings of the National Academy of Sciences*, 94(13):6585–6590, 1997.

[321] T. Warnow, D. Ringe, and A. Taylor. Reconstructing the evolutionary history of natural languages. In *Proceedings of the seventh annual ACM-SIAM symposium on Discrete algorithms*, pages 314–322. Society for Industrial and Applied Mathematics, 1996.

[322] J. Wei and L. Zhou. Model selection using modified aic and bic in joint modeling of paired functional data. *Statistics & probability letters*, 80(23):1918–1924, 2010.

[323] K. Q. Weinberger, F. Sha, and L. K. Saul. Learning a kernel matrix for nonlinear dimensionality reduction. In *Proceedings of the twenty-first international conference on Machine learning*, page 106. ACM, 2004.

[324] H. L. Weinert. Efficient computation for Whittaker–Henderson smoothing. *Computational Statistics & Data Analysis*, 52(2):959–974, 2007.

[325] 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.

[326] W.F. Massy. Principal Components Regression in Exploratory Statistical Research. *Journal of the American Statistical Association*, 60(309):234–256, 1965.

[327] E. P. Wigner. The unreasonable effectiveness of mathematics in the natural sciences. *Communications on pure and applied mathematics*, 13(1):1–14, 1960.

[328] Wikipedia. Pinyin — Wikipedia, the free encyclopedia, 2013. [Online; accessed 12-Sept-2013].

[329] Wikipedia. Quebec french — Wikipedia, the free encyclopedia, 2013. [Online; accessed 19-Nov-2013].

[330] Wikipedia. Tone linguistics — Wikipedia, the free encyclopedia, 2013. [Online; accessed 12-Sept-2013].

[331] A.-M. Wink and J. B. T. M. Roerdink. Denoising Functional MR Images: A Comparison of Wavelet Denoising and Gaussian Smoothing. *Medical Imaging, IEEE Transactions on*, 23(3):374–387, 2004.

[332] L. Xin and M. Zhu. Stochastic stepwise ensembles for variable selection. *Journal of Computational and Graphical Statistics*, 21(2):275–294, 2012.

[333] Y. Xu. Effects of tone and focus on the formation and alignment of f_0 contours. *Journal of Phonetics*, 27(1):55–105, 1999.

[334] Y. Xu and Q. E. Wang. Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech Communication*, 33(4):319–337, 2001.

[335] M.-H. Yang. Kernel eigenfaces vs. kernel fisherfaces: Face recognition using kernel methods. In *Automatic Face and Gesture Recognition, Fifth IEEE International Conference on*, pages 215–220. IEEE, 2002.

[336] Z. Yang and B. Rannala. Molecular phylogenetics: principles and practice. *Nature Reviews Genetics*, 13(5):303–14, 2012.

[337] F. Yao, H. Müller, and J. Wang. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470):577–590, 2005.

[338] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann. Making machines understand us in reverberant rooms: robustness against reverberation for automatic speech recognition. *IEEE Signal Processing Magazine*, 29:114–126, Nov. 2012.

[339] W.-C. Yueh. Eigenvalues of several tridiagonal matrices. *Applied Mathematics E-Notes*, 5(66-74):210–230, 2005.

[340] Z. Zhang and H.-G. Müller. Functional density synchronization. *Computational Statistics and Data Analysis*, 55:2234–2249, 2011.

[341] L. Zhou, J. Z. Huang, and R. J. Carroll. Joint modelling of paired sparse functional data using principal components. *Biometrika*, 95(3):601–619, 2008.

[342] M. Zhu and H. A. Chipman. Darwinian evolution in parallel universes: A parallel genetic algorithm for variable selection. *Technometrics*, 48(4), 2006.

Appendix A

A.1 Voicing of Consonants and IPA representations for Chapt. 4 & 5

The voicing characterization was based on the SAMPA-T (Speech Assessment Methods Phonetic Alphabet - Taiwan), and the correspondence between the SAMPA-T system and the IPA system were based on the material provided by Academia Sinica. The following consonants were present in the dataset.

	Labial	Dental	Retro-flex	(Alveo-) Palatal	Velar
Stop	p p^h	t t^h			k k^h
Affricate		ts ts^h	t s $t s^h$	tɕ $tɕ^h$	
Fricative	f	s	s z	ɕ	x
Nasal	m	n			
Approx.		l		j	

This led to the following voicing characterization:

IPA	t	tɕ	ts	t s	f	k	x	k^h	l	m	n
Voiced	0	0	0	0	0	0	0	0	1	1	1
IPA	p	z	p^h	s	s	-	ɕ	t^h	$tɕ^h$	ts^h	$t s^h$
Voiced	0	1	0	0	0	3	0	0	0	0	0

Mandarin Chinese rhymes are transcribed with symbols corresponding to:
[ə, əɻ, a, ai, an, aŋ, au, ei, i, ia, iaŋ, iau, iɛ, iɛn, in, iŋ, iou, ən, əŋ, o, oŋ, ou, u, ɻ, ɿ, ua, uai, uan, uaŋ, uei, uən, uo, y, yɛ, yɛn, yn, yoŋ].

A.2 Comparison of Legendre Polynomials and FPC's for Amplitude Only model

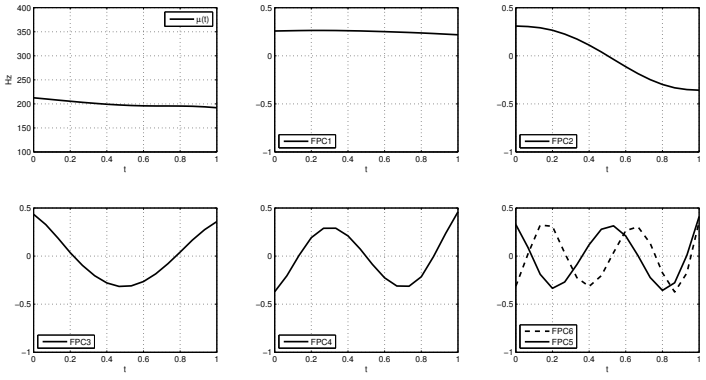

Figure A.1: Sample FPC's normalized on L[0,1]

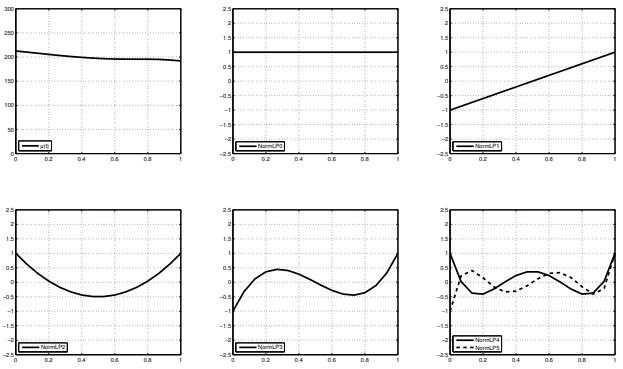

Figure A.2: Legendre polynomials in L[0,1]

The shifted Legendre polynomials are a standard choice for an orthogonal basis expansion in the interval $[0, 1]$ being defined as:

$$\hat{P}_n(x) = \frac{1}{n!}\frac{d^n}{dx^n}(x^2 - x)^n \tag{A.1}$$

While it is clearly seen that the two bases look similar (Fig. A.1 & A.2), Legendre polynomials present a number of limitations compared to Functional Principal Components:
a) Each successive one LP does not reflect diminishing amount of auditory information,
b) given a fixed number of components, the FPC's reflect larger tonal content (eg. FCP1-4 : 243.5 Hz compared to LP0-3 : 170.6Hz),
c) the FPC's are a non-parametric basis so they represent a meaningful basis compared to an arbitrary orthogonal basis and
d) FPC's do not make the implicit assumption that the signal is periodic by nature.

LP#	Hz (99%)	Hz (95%)
LP1	126.6478	95.9651
LP2	28.4172	18.7645
LP3	10.8092	7.1530
LP4	4.7961	2.6287
LP5	9.1977	6.8812
LP6	5.5564	3.5870

LP#	Hz (99%)	Hz (95%)
LP7	10.7919	8.1306
LP8	6.4458	4.2775
LP9	12.9543	9.7760
LP10	7.6454	5.0587
LP11	15.2197	11.4753
LP12	8.7689	5.7868

Table A.1: Auditory variation per LP (in Hz) (human speech auditory sensitivity threshold $\approx$ 10 Hz).

A.3 Speaker Averaged Sample for Amplitude Only model

While the paper rationale follows standard principal component analysis methodology, thus subtracting the sample mean prior to the Karhunen-Loève expansion, the subtraction of speaker-specific means is also possible, yet not desirable from an explanatory perspective. Subtracting the speaker-specific means would result in altering the sample's eigenvectors in a non-obvious way, thus distorting the whole rationale that FPCA is based upon; Miranda et al. 2008 have already presented the optimality of subtracting the sample wide mean. Nevertheless, because of the uniformly scattered nature of our data, the resulting sample covariance surface and functional principal components (Fig. A.3 & A.4 respectively) are almost identical to their non-speaker specific counterparts. Furthermore following the framework outlined in the paper, it can also be seen that the resulting components are not only of same qualitative nature but also carry comparable variational amplitude (Tables A.2). Finally it should be mentioned that subtracting the speaker-specific means would render the speaker related random effect powerless. Thus we would not be able to quantify the effect it carries to each specific component individually (eg. we would miss out the intuition that the FPC_2 is less influenced by the speaker effect than FPC_3.) but give only a very general estimation about the collective influence it carries on the whole reconstructed curve.

FPC#	Hz (99%)	Hz (95%)
FPC_1	112.8356	81.5929
FPC_2	68.1571	43.7576
FPC_3	32.1981	17.584
FPC_4	18.4580	8.8473
FPC_5	8.8254	4.1947
FPC_6	5.4575	2.4836

FPC#	Hz (99%)	Hz (95%)
FPC_7	3.5902	1.7232
FPC_8	2.8679	1.2725
FPC_9	2.3779	1.0709
FPC_{10}	1.8927	0.8478
FPC_{11}	1.6555	0.6787
FPC_{12}	1.0636	0.4531

Table A.2: Auditory variation per FPC in the Speaker-centerd sample (in Hz) (Speaker-Adjusted Sample)

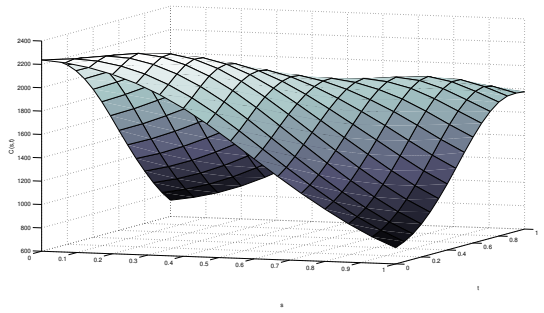

Figure A.3: Covariance function of the 54707 smoothed F_0 sample curves having already subtracted the speaker associated mean from each curve, exhibiting smooth behaviour.

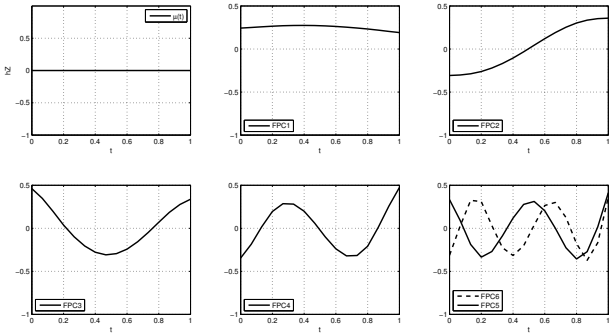

Figure A.4: Mean Function and 1st, 2nd, 3rd, 4th, 5th and 6th Functional Principal Components. Together these account for 99.86% of the sample variance but actually only the first four having linguistic meaning (99.59 % of sample variation); the 5th and 6th were not used in the subsequent analysis.

A.4 Functional Principal Components Robustness check for Amplitude Only model

A.5 Model IDs for Amplitude Only model

Model_ID	DF	prT_cuT_nxT	prT_cuT	prT_nxT	cuT_nxT	prT	cuT	nxT	prC_cuT_nxC	prC_cuT	prC_nxC	nxC_cuT	prC	cuT	nxC	vowel	B2^3	B3^3	B4^3	B5^3	Sex	RandEff
1	245	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
1	247	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	241	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0
2	243	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1
3	155	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
3	157	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
4	151	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0
4	153	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1
5	145	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	0
5	147	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1
6	145	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	1	1	0
6	147	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	1	1	1
7	145	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	1	1	0
7	147	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1
8	145	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	0
8	147	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1
9	127	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	0	0	0	1	0
9	129	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	0	0	0	1	1
10	115	0	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1	1	0
10	117	0	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1
11	150	0	1	1	1	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	0
11	152	0	1	1	1	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1
12	147	0	1	1	1	1	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	0
12	149	0	1	1	1	1	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1
13	147	0	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	0
13	149	0	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1
14	145	0	1	1	1	1	1	1	0	1	0	0	1	1	1	1	1	1	1	1	1	0
14	147	0	1	1	1	1	1	1	0	1	0	0	1	1	1	1	1	1	1	1	1	1
15	145	0	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1	1	0
15	147	0	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1	1	1
16	143	0	1	1	1	1	1	1	0	0	1	0	1	1	1	1	1	1	1	1	1	0
16	145	0	1	1	1	1	1	1	0	0	1	0	1	1	1	1	1	1	1	1	1	1
17	140	0	1	1	1	1	1	1	0	0	0	0	1	1	1	1	1	1	1	1	1	0
17	142	0	1	1	1	1	1	1	0	0	0	0	1	1	1	1	1	1	1	1	1	1
18	127	0	1	0	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0
18	129	0	1	0	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1
19	131	0	1	1	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0
19	133	0	1	1	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1
20	132	0	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0
20	134	0	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1
21	102	0	1	0	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0
21	104	0	1	0	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1
22	103	0	0	0	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0
22	105	0	0	0	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1
23	112	0	0	1	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0
23	114	0	0	1	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1
24	78	0	0	0	0	0	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	0
24	80	0	0	0	0	0	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1
25	54	0	0	0	0	1	1	1	0	0	0	0	1	1	1	1	1	1	1	1	0	0
25	56	0	0	0	0	1	1	1	0	0	0	0	1	1	1	1	1	1	1	1	0	1

Model_ID	DF	prT_cuT_nxT	prT_cuT	prT_nxT	cuT_nxT	prT	cuT	nxT	prC_cuT_nxC	prC_cuT	prC_nxC	nxC_cuT	prC	cuT	nxC	vowel	B2^3	B3^3	B4^3	B5^3	Sex	RandEff
26	79	0	0	0	0	1	1	1	0	0	0	0	1	1	1	1	1	1	1	1	1	0
26	81	0	0	0	0	1	1	1	0	0	0	0	1	1	1	1	1	1	1	1	1	1
27	57	0	0	0	0	1	1	1	0	0	0	0	1	1	1	1	0	0	0	1	0	0
27	59	0	0	0	0	1	1	1	0	0	0	0	1	1	1	1	0	0	0	1	0	1
28	61	0	0	0	0	1	1	1	0	0	0	0	1	1	1	1	0	0	0	1	1	0
28	63	0	0	0	0	1	1	1	0	0	0	0	1	1	1	1	0	0	0	1	1	1
29	57	0	0	0	0	1	1	1	0	0	0	0	1	1	1	1	0	0	1	0	0	0
29	59	0	0	0	0	1	1	1	0	0	0	0	1	1	1	1	0	0	1	0	0	1
30	61	0	0	0	0	1	1	1	0	0	0	0	1	1	1	1	0	0	1	0	1	0
30	63	0	0	0	0	1	1	1	0	0	0	0	1	1	1	1	0	0	1	0	1	1
31	57	0	0	0	0	1	1	1	0	0	0	0	1	1	1	1	0	1	0	0	0	0
31	59	0	0	0	0	1	1	1	0	0	0	0	1	1	1	1	0	1	0	0	0	1
32	61	0	0	0	0	1	1	1	0	0	0	0	1	1	1	1	0	1	0	0	1	0
32	63	0	0	0	0	1	1	1	0	0	0	0	1	1	1	1	0	1	0	0	1	1
33	57	0	0	0	0	1	1	1	0	0	0	0	1	1	1	1	1	0	0	0	0	0
33	59	0	0	0	0	1	1	1	0	0	0	0	1	1	1	1	1	0	0	0	0	1
34	61	0	0	0	0	1	1	1	0	0	0	0	1	1	1	1	1	0	0	0	1	0
34	63	0	0	0	0	1	1	1	0	0	0	0	1	1	1	1	1	0	0	0	1	1
35	54	0	0	0	0	1	1	1	0	0	0	0	1	1	1	1	0	0	0	0	0	0
35	56	0	0	0	0	1	1	1	0	0	0	0	1	1	1	1	0	0	0	0	0	1
36	126	0	1	0	1	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	0
36	128	0	1	0	1	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1
37	120	0	1	0	1	1	1	1	0	1	0	1	1	1	1	1	1	1	1	0	1	0
37	122	0	1	0	1	1	1	1	0	1	0	1	1	1	1	1	1	1	1	0	1	1

A.6 AIC scores (ML-estimated models) for Amplitude Only model

Model_ID	RandEff	DF	AIC1	AIC2	AIC3	AIC4
1	0	301	690545.4	572791.6	493957.2	430858
1	1	303	679331.4	**572518**	490142.8	428528.4
2	0	268	690531.4	572845.2	493988.6	430893.8
2	1	270	679313	572569	490184.2	428566.4
3	0	211	690502.8	572916	494009	430801.2
3	1	213	679299.2	572635.2	**490191**	428480.6
4	0	178	690486.2	572973.8	494014.6	430838.2
4	1	180	679277.8	572690	490210.6	428520.4
5	0	172	690611	572988.4	494016.8	430831.2
5	1	174	679419.2	572701	490210.2	428513.8
6	0	172	692536.2	573028.2	494017	430843.4
6	1	174	680412.6	572734.6	490212.4	428526.6
7	0	172	694361.8	573199	494202.2	430902.8
7	1	174	684558.4	572927.6	490357.4	428556.8
8	0	172	690784.6	573347.4	494367.8	430853.8
8	1	174	679613.6	573073.2	490556	428545.6
9	0	154	698791	573844.2	494728.6	430958
9	1	156	688653	573574	490867.6	428617.8
10	0	142	691728.2	573486.6	497839.6	431481.2
10	1	144	680510.6	573196.8	494299	429154.4
11	0	169	690524.8	572997.2	494083	430917.8
11	1	171	679318	572712.6	490279	428596.2
12	0	166	690599.2	573966.6	494381.6	431543.2
12	1	168	679453	573671.4	490579.2	429233.4
13	0	167	690809.8	574897.4	499304.6	431147.4
13	1	169	679693.8	574584	495824.2	428838.6
14	0	157	690622.4	573990	494478	431636
14	1	159	679473.8	573694.6	490674.8	429322
15	0	158	690846	574920.8	499352.2	431233.2
15	1	160	679732.2	574606.6	495873.4	428920.2
16	0	155	690920.6	575746.6	499676.6	431858.8
16	1	157	679864.4	575423.8	496174.2	429561.6
17	0	146	690945.8	575779.2	499753	431963.2
17	1	148	679888.8	575456.2	496252.4	429661.8
18	0	154	690488.2	573027.6	494026.8	430820
18	1	156	**679297.8**	572744.8	490221	428502.2
19	0	158	691285.6	578417.6	494156.4	430841.2
19	1	160	680233.6	578171.2	490366	428527.2
20	0	159	690739.8	573939.2	494413	431150
20	1	161	679591.2	573644	490663.6	428828.8
21	0	134	691287.4	578465.6	494175.6	430833.8
21	1	136	680254	578219.2	490383.2	**428520.8**
22	0	135	690742.4	573989.2	494423.8	431133.2
22	1	137	679610.8	573694.4	490671.6	428811.2
23	0	139	691548	579330.4	494567	431152
23	1	141	680554	579093.6	490832.4	428834.4
24	0	105	693468	582328	496822.2	432434.2
24	1	107	682954.2	582046	493245	430183.6
25	0	70	728608	582183.8	505304.6	432549.2
25	1	72	682220.8	581871.6	497062.4	430091.4
26	0	83	692059	582129.8	500487	432373.4
26	1	85	681224.8	581850.8	497064	430070.6
27	0	61	733509.4	582753.4	505888	432593
27	1	63	689222.8	582450.4	497615.4	430154.6
28	0	65	699435.6	582726.8	501093.6	432478.6
28	1	67	688974	582456.6	497610.4	430155.8
29	0	61	731863.6	582725	505863.6	432582
29	1	63	688013.6	582435.6	497612.6	430139.6

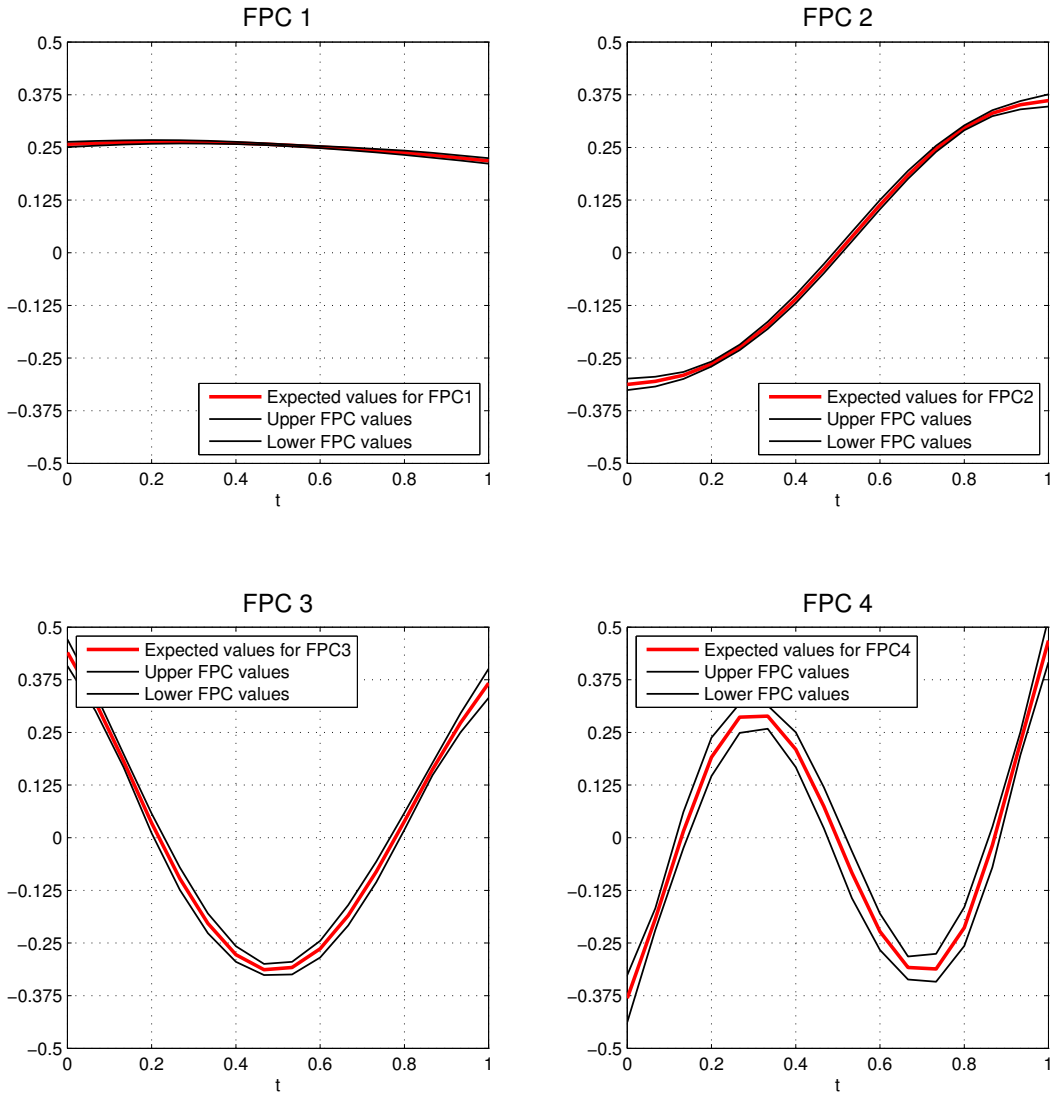

Figure A.5: To assess the robustness of the resulting FPCs, FPCA is conducted in subsamples of the original COSPRO dataset. In particular we segmented the original 54707 curve sample in 10 random subsamples of approximately equal size and computed the FPC of each subsample independently; the procedure was repeated 2048 times resulting in 20480 FPCs curve realizations of each component. The results show strong robustness as the FPC shape characteristic remain universal along all samples.

30	0	65	697040.8	582698.2	501092.4	432461
30	1	67	687555.2	582440.2	497614.2	430140.8
31	0	61	730496.4	582501.2	505606	432566.6
31	1	63	684428.4	582189.6	497350.4	430120
32	0	65	694879.8	582431.8	500790.6	432397
32	1	67	683744.8	582155.4	497352	430105.4
33	0	61	733275.8	582411.6	505483.8	432600.6
33	1	63	689255	582101.8	497187	430147.4
34	0	65	699178.2	582380.2	500643.6	432469
34	1	67	689220.4	582103.2	497180	430131.8
35	0	58	733764.6	582777.2	505884.4	432608.6
35	1	60	690291	582479.6	497613	430168
36	0	145	690526.4	573050.8	494097.2	430899.8
36	1	147	679338.6	572767.2	490291.2	428578.2
37	0	139	690650.4	573065.4	494098.6	430892.8
37	1	141	679481	572778.4	490290.2	428571.8

Table A.4: Best Jackknifing models in bold; sample-wide best AIC models underlined.

A.7 Jackknifing for Amplitude Only model

As mentioned in the main body of the study, in order to account for the robustness of our approach we relied on re-estimating the optimal model under the AIC framework, using random samples of our initial dataset. We implemented 180 runs where in each run we randomly partitioned our data in 5 sub-samples that we then treated as independent and then we subsequently performed FPCA and LME analysis

on each sub-sample. Figure A.6 displays our findings: As we see in all cases the model selected by

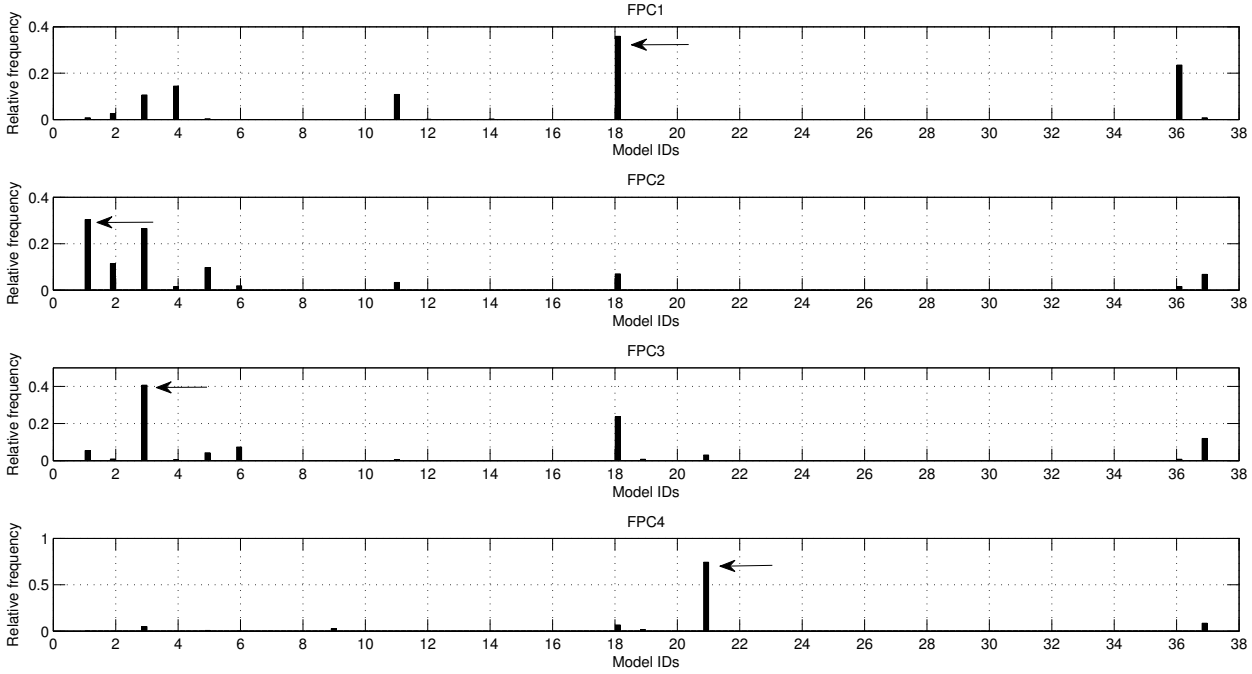

Figure A.6: Bar-chart showing the relative frequencies of model selection under jackknifing; models selected for the thesis' work are pointed with an arrow.

re-sampling does not contradict our finding for the optimal models when utilizing the whole sample. In cases that the jackknifing suggested a different model, we examined if the model proposed by jackknifing was not providing a significantly less optimal fit for the whole dataset. In all 3 cases that fit proposed by jackknifing was qualitatively comparable to that of the whole data. While it is only reasonable that the larger full sample will allow the detection of more rare events, we choose the slightly smaller models proposed by the Jackknifing as to ensure we are not overfitting our sample data.

A.8 Estimate Tables for Amplitude Only model

The estimates were produced using the function lmer() from the `lme4` package. 10000 samples were used to generate the bootstrap estimates. The first two columns show the actual model estimate and the bootstrap mean estimate. Unsurprisingly both estimates are quite close. The next two columns are the lower and higher 95% confidence density intervals. Because these estimates are referring to the final models used, ReML (Restricted Maximum Likelihood) principle was used during their computation.

Table A.5: All fixed effects and 95% confidence intervals for the 1st FPC scores models.

FPC_1 Covariate	Estimate	BoostrapMean	(0.025,	.975)
VoicePRc0	290.6276	288.8879	196.2904	374.8958
VoicePRc1	238.0884	236.3364	144.5801	323.6155
VoicePRc2	239.5133	237.7425	148.8227	326.4128
VoicePRc3	267.4338	265.2797	172.3652	356.5443
tone_2	-158.0311	-157.7912	-188.1032	-127.2074
tone_3	-236.2368	-237.4759	-287.7277	-190.2561
tone_4	-62.6604	-62.9837	-93.7334	-32.6068
tone_5	23.063	23.8788	-33.8707	85.5422
VoiceNXc1	-4.9446	-4.9248	-10.5301	0.8918
VoiceNXc2	-18.3991	-18.5359	-25.11	-11.8492
VoiceNXc3	-22.8719	-22.9102	-28.7349	-16.938
rhyme_ɚ	-3.0514	-2.9682	-22.3141	14.8717
rhyme_a	-46.2323	-46.1669	-54.6929	-38.2717
rhyme_ai	-52.9713	-52.9681	-60.5208	-45.4837
rhyme_an	-13.6657	-13.7609	-21.1268	-6.2696
rhyme_aŋ	-26.0789	-26.1654	-34.4352	-17.9603
rhyme_au	-31.256	-31.2592	-41.3548	-21.4167

Table A.5 – continued from previous page

FPC_1 Covariate	Estimate	BoostrapMean	(0.025,	.975)
rhyme_ei	6.5566	6.6786	-3.1349	15.2371
rhyme_i	9.7728	9.7161	3.3253	16.0293
rhyme_ia	-44.3547	-44.4112	-55.5068	-33.7468
rhyme_iaŋ	-27.5415	-27.8218	-36.8506	-18.3873
rhyme_iau	-29.2383	-29.3334	-38.7381	-19.7165
rhyme_iɛ	-20.1994	-20.0965	-28.3356	-11.14
rhyme_iɛn	-15.4923	-15.5871	-22.7544	-8.5272
rhyme_in	9.5284	9.5247	1.5088	17.3305
rhyme_iŋ	9.5004	9.3795	2.6279	16.2537
rhyme_iou	14.7058	14.8752	6.5932	22.6325
rhyme_ən	2.3474	2.3451	-4.8613	9.2021
rhyme_əŋ	1.5957	1.6013	-5.5467	8.7776
rhyme_o	14.1393	14.4102	-4.6814	34.1881
rhyme_oŋ	19.5683	19.5685	12.5981	25.9998
rhyme_ou	-21.0314	-20.9471	-30.6301	-10.7955
rhyme_u	24.6549	24.6187	18.0004	31.2778
rhyme_ɿ	28.8474	28.8205	19.0472	38.7886
rhyme_ʅ	-16.4383	-16.4949	-23.1142	-10.051
rhyme_ua	-46.813	-46.7781	-69.4845	-23.3788
rhyme_uai	-5.025	-5.0709	-25.3445	13.6476
rhyme_uan	-16.3199	-16.1793	-25.8464	-6.2835
rhyme_uaŋ	-12.967	-12.8079	-25.5231	0.199
rhyme_uei	-8.7169	-8.7723	-15.9956	-0.5403
rhyme_uən	-12.5026	-12.882	-24.7355	-0.1808
rhyme_uo	-8.2602	-8.3119	-14.8086	-1.5086
rhyme_y	6.2325	6.2694	-2.8333	15.4397
rhyme_yɛ	-8.0661	-7.9792	-18.6677	3.0846
rhyme_yɛn	-7.8194	-7.6844	-17.8876	2.7967
rhyme_yn	5.1034	4.7621	-9.1294	19.8606
rhyme_yoŋ	-19.6194	-19.128	-38.6167	0.3638
B2	57.1172	56.6871	29.8611	82.5396
$B2^2$	-37.0934	-36.9394	-49.0051	-24.6728
$B2^3$	5.8102	5.7959	4.1001	7.5416
B3	-54.4534	-54.2685	-57.3363	-51.0481
$B3^2$	4.5697	4.54	4.0036	5.0901
$B3^3$	-0.1073	-0.1061	-0.1312	-0.0808
B4	-5.4305	-5.422	-5.783	-5.0433
$B4^2$	0.0889	0.0888	0.0803	0.0962
$B4^3$	-4e-04	-4e-04	-4e-04	-3e-04
B5	0.0157	0.0112	-0.3048	0.3331
$B5^2$	-0.008	-0.0079	-0.0136	-0.0021
$B5^3$	0	0	0	1e-04
SexM	-376.6931	-374.6379	-514.8648	-230.1063
pr_to11	11.3611	11.3534	-1.0841	24.0511
pr_to12	8.3474	8.5974	-3.7404	21.184
pr_to13	-27.6911	-27.6533	-39.8136	-15.502
pr_to14	-28.098	-27.8887	-39.8373	-15.6739
pr_to15	-8.9191	-8.8892	-22.0691	4.8133
nx_to11	58.0072	57.9872	39.5891	76.1897
nx_to12	95.2594	95.0824	76.6345	113.0094
nx_to13	100.7655	100.6452	81.969	119.7754
nx_to14	91.3527	91.2734	73.0661	109.2332
nx_to15	82.7454	82.5712	62.6679	102.3617
VoicePRc1:tone_2	58.5001	58.6905	47.2799	69.4968
VoicePRc2:tone_2	49.2041	49.2381	37.6237	60.2795
VoicePRc3:tone_2	-1.4369	-0.7953	-20.9772	19.1905
VoicePRc1:tone_3	47.3256	47.3421	34.6181	59.9143
VoicePRc2:tone_3	93.4332	93.3471	80.9414	105.942
VoicePRc3:tone_3	-8.4085	-7.9649	-27.1386	10.557
VoicePRc1:tone_4	61.4711	61.5369	50.1995	72.2303
VoicePRc2:tone_4	39.2997	39.2255	28.1767	49.9861
VoicePRc3:tone_4	54.2863	54.6372	35.6353	73.7137
VoicePRc1:tone_5	69.8676	69.7713	51.959	87.2013
VoicePRc2:tone_5	0.8532	0.5029	-71.0162	72.737
tone_2:VoiceNXc1	-16.4104	-16.4537	-25.0188	-8.3611
tone_3:VoiceNXc1	-0.4894	-0.4076	-11.6462	9.5844
tone_4:VoiceNXc1	26.4155	26.4383	17.756	34.3044
tone_5:VoiceNXc1	6.6469	6.6998	-7.2493	21.0618
tone_2:VoiceNXc2	20.4874	20.5795	10.8909	30.0312
tone_3:VoiceNXc2	24.1558	24.4312	13.1255	35.7025
tone_4:VoiceNXc2	35.579	35.6253	27.0235	44.0081
tone_5:VoiceNXc2	39.2385	39.2833	22.1407	56.9877
tone_2:VoiceNXc3	-17.2567	-17.2356	-25.0453	-8.8607
tone_3:VoiceNXc3	-4.1604	-4.1127	-13.8619	5.6609
tone_4:VoiceNXc3	2.777	2.7105	-4.7611	10.8425
tone_5:VoiceNXc3	12.1734	12.0102	-1.5924	25.376
VoicePRc1:VoiceNXc1	-5.696	-5.8193	-14.1012	2.6284
VoicePRc2:VoiceNXc1	15.7127	15.7708	5.9286	26.3005
VoicePRc3:VoiceNXc1	19.6739	19.5784	1.5294	38.722
VoicePRc1:VoiceNXc2	-23.1075	-23.1455	-31.8518	-14.8163
VoicePRc2:VoiceNXc2	-4.8746	-4.7196	-15.5983	5.6
VoicePRc3:VoiceNXc2	-4.5734	-4.3496	-22.3144	13.9796
VoicePRc1:VoiceNXc3	3.6451	3.5978	-3.1243	11.3571
VoicePRc2:VoiceNXc3	13.9163	14.1232	5.5181	22.573
VoicePRc3:VoiceNXc3	9.1536	9.022	-6.7493	24.7472
B2:SexM	-59.9615	-61.1271	-114.4364	-7.4205
$B2^2$:SexM	22.8947	23.5349	-1.2096	48.5743
$B2^3$:SexM	-2.6511	-2.7509	-6.2431	0.7627
B3:SexM	34.9644	34.8186	29.5378	40.4671
$B3^2$:SexM	-3.2994	-3.2756	-4.2318	-2.3553
$B3^3$:SexM	0.0716	0.0706	0.0273	0.1159
B4:SexM	3.0225	3.0178	2.2673	3.7602
$B4^2$:SexM	-0.0466	-0.0467	-0.0638	-0.0293
$B4^3$:SexM	2e-04	2e-04	1e-04	3e-04
B5:SexM	0.7157	0.7215	0.1842	1.2752
$B5^2$:SexM	-0.0086	-0.0087	-0.0183	0.0011
$B5^3$:SexM	0	0	0	1e-04
tone_2:pr_to11	-60.2238	-60.5252	-76.9758	-42.6763
tone_3:pr_to11	-21.8393	-22.1284	-42.5022	-1.4912
tone_4:pr_to11	-44.4751	-44.4725	-59.981	-28.5539
tone_5:pr_to11	-68.6406	-67.7035	-109.5356	-29.4246
tone_2:pr_to12	-42.6642	-43.4623	-61.2588	-25.4739

Table A.5 – continued from previous page

FPC_1 Covariate	Estimate	BoostrapMean	(0.025,	.975)
tone_3:pr_to12	1.5465	0.943	-19.6645	20.2377
tone_4:pr_to12	-29.4712	-29.5414	-45.2254	-14.0662
tone_5:pr_to12	-75.6004	-74.9938	-114.4153	-37.7349
tone_2:pr_to13	-31.1371	-31.6078	-49.9722	-13.6739
tone_3:pr_to13	25.4096	25.3295	4.4349	44.9896
tone_4:pr_to13	-14.9275	-14.9296	-30.4286	0.8937
tone_5:pr_to13	-112.8046	-112.0038	-152.5997	-74.4657
tone_2:pr_to14	-39.4836	-40.1734	-56.7008	-24.3313
tone_3:pr_to14	-4.0889	-4.6106	-23.7978	13.4256
tone_4:pr_to14	-40.3277	-40.5586	-55.0776	-24.9944
tone_5:pr_to14	-132.2049	-131.8285	-169.4348	-93.2712
tone_2:pr_to15	-59.3155	-59.9608	-79.5509	-40.9645
tone_3:pr_to15	-20.1592	-20.476	-43.5351	1.7043
tone_4:pr_to15	-26.6009	-26.7389	-44.0419	-9.0468
tone_2:nx_to11	30.6341	30.8807	4.041	57.6933
tone_3:nx_to11	-50.2163	-48.9955	-93.5112	-1.4287
tone_4:nx_to11	8.4709	8.9067	-17.9105	35.9684
tone_5:nx_to11	-107.4177	-108.9408	-160.1863	-60.4522
tone_2:nx_to12	13.9168	14.2746	-13.1296	40.0327
tone_3:nx_to12	-80.5023	-78.7329	-124.5513	-30.9264
tone_4:nx_to12	-8.3446	-7.77	-35.6853	18.8548
tone_5:nx_to12	-138.1171	-139.6188	-189.9929	-89.5657
tone_2:nx_to13	35.7075	35.9982	6.5297	63.8812
tone_3:nx_to13	40.4142	42.1176	-4.0413	89.8268
tone_4:nx_to13	3.5714	4.0745	-23.7456	31.9801
tone_5:nx_to13	-103.3212	-104.6503	-156.0392	-53.4573
tone_2:nx_to14	-1.0191	-0.7386	-28.3151	25.9497
tone_3:nx_to14	-78.1722	-76.6589	-120.7483	-29.5356
tone_4:nx_to14	7.8174	8.3478	-18.1823	35.7335
tone_5:nx_to14	-101.4688	-102.9794	-153.7719	-53.769
tone_2:nx_to15	-1.6216	-1.1639	-30.5745	26.7324
tone_3:nx_to15	-68.6251	-67.1657	-112.9532	-18.8865
tone_4:nx_to15	15.9053	16.4964	-12.7233	45.3096
tone_5:nx_to15	-112.194	-113.1722	-169.7683	-58.6162

Table A.6: All fixed effects and 95% confidence intervals for the 2nd FPC scores models.

FPC_2 Covariate	Estimate	BoostrapMean	(0.025,	.975)
VoicePRc0	-55.1365	-54.7171	-78.5651	-31.1631
VoicePRc1	-72.6605	-72.2879	-96.8907	-47.1767
VoicePRc2	-85.1283	-84.7967	-109.3434	-60.047
VoicePRc3	-78.5898	-78.4467	-104.4681	-52.9534
tone_2	-73.7193	-73.9095	-102.9649	-44.96
tone_3	11.2662	9.839	-24.4518	45.2179
tone_4	96.1069	95.4212	68.6346	121.8028
tone_5	44.7093	45.0487	-23.1462	110.7547
VoiceNXc1	-1.2175	-1.222	-3.4634	1.1315
VoiceNXc2	0.9049	0.8273	-1.7245	3.5228
VoiceNXc3	-7.203	-7.2326	-9.7105	-4.8955
rhyme_ɚ	17.0032	17.0258	9.5984	23.8516
rhyme_a	-2.7102	-2.6994	-6.0067	0.2809
rhyme_ai	-3.9288	-3.9343	-6.8222	-1.067
rhyme_an	-5.4668	-5.4977	-8.2447	-2.698
rhyme_aŋ	-8.5036	-8.5368	-11.5707	-5.3685
rhyme_au	-6.9612	-6.9723	-10.6442	-3.4191
rhyme_ei	-3.8492	-3.815	-7.5873	-0.602
rhyme_i	-7.209	-7.2291	-9.6442	-4.7985
rhyme_ia	-0.6193	-0.6377	-4.7092	3.601
rhyme_iaŋ	-1.778	-1.8924	-5.2855	1.7738
rhyme_iau	3.803	3.771	0.1355	7.3157
rhyme_iɛ	8.3833	8.4248	5.3091	11.7694
rhyme_iɛn	3.1438	3.108	0.4882	5.8118
rhyme_in	-9.993	-10.0025	-13.0501	-7.0466
rhyme_iŋ	-4.0968	-4.1457	-6.6685	-1.5281
rhyme_iou	0.1112	0.18	-3.0258	3.1433
rhyme_ən	-8.3946	-8.4058	-11.1002	-5.7295
rhyme_əŋ	-3.2974	-3.2952	-5.8977	-0.6915
rhyme_o	0.0828	0.1913	-6.9279	7.5206
rhyme_oŋ	-11.4959	-11.503	-14.2345	-9.1074
rhyme_ou	1.5709	1.59	-2.0972	5.3308
rhyme_u	-3.2349	-3.2552	-5.801	-0.6302
rhyme_ɿ	-4.0987	-4.1139	-7.7061	-0.3483
rhyme_ʅ	-12.1963	-12.224	-14.6863	-9.9134
rhyme_ua	-11.0644	-11.0337	-19.6088	-2.3057
rhyme_uai	14.4336	14.4252	6.7642	21.4253
rhyme_uan	-6.5375	-6.4904	-10.1722	-2.9514
rhyme_uaŋ	-6.396	-6.3237	-10.898	-1.4799
rhyme_uei	1.2593	1.2383	-1.4841	4.3228
rhyme_uən	-0.473	-0.6101	-5.0683	4.0646
rhyme_uo	-4.9202	-4.933	-7.4728	-2.2334
rhyme_y	-3.2406	-3.218	-6.6975	0.2635
rhyme_yɛ	0.964	0.994	-2.9668	5.191
rhyme_yɛn	0.1873	0.2431	-3.6973	4.1574
rhyme_yn	-0.0506	-0.192	-5.504	5.4832
rhyme_yoŋ	-14.201	-14.0114	-21.4925	-6.9524
B2	49.5581	49.382	39.1802	59.3965
$B2^2$	-17.5691	-17.5039	-22.1279	-12.8292
$B2^3$	1.765	1.7586	1.1173	2.3978
B3	8.5594	8.6278	7.459	9.8386
$B3^2$	-1.2355	-1.2465	-1.4474	-1.0383
$B3^3$	0.047	0.0474	0.0377	0.0571
B4	0.3127	0.3154	0.1851	0.4518
$B4^2$	-0.0037	-0.0038	-0.0068	-9e-04
$B4^3$	0	0	0	0
SexM	12.5843	12.9564	-3.2395	27.8896
B5	-0.0598	-0.0609	-0.1642	0.0443
$B5^2$	9e-04	9e-04	-9e-04	0.0027
$B5^3$	0	0	0	0
VoicePRc1:tone_2	12.1902	12.345	5.7717	18.1931

Table A.6 – continued from previous page

FPC_2 Covariate	Estimate	BoostrapMean	(0.025,	.975)
VoicePRc2:tone_2	52.4587	52.5053	46.6116	58.2919
VoicePRc3:tone_2	-10.2526	-9.7495	-21.1179	2.4176
VoicePRc1:tone_3	19.9557	19.9745	12.3652	26.3248
VoicePRc2:tone_3	60.54	60.6041	53.6109	67.6125
VoicePRc3:tone_3	7.5041	7.723	-2.4652	18.5094
VoicePRc1:tone_4	11.4891	11.5343	5.0592	17.2883
VoicePRc2:tone_4	-8.5501	-8.4677	-14.0362	-3.2256
VoicePRc3:tone_4	10.1887	10.4488	-1.1982	20.967
VoicePRc1:tone_5	7.2615	7.2801	-2.223	16.2358
VoicePRc2:tone_5	18.0003	17.252	-44.0874	81.1793
VoicePRc1:VoiceNXc1	12.1663	12.2032	1.9988	22.0752
VoicePRc2:VoiceNXc1	5.3369	5.492	-4.04	15.021
VoicePRc3:VoiceNXc1	-0.7397	-0.69	-19.6948	18.9607
VoicePRc1:VoiceNXc2	-19.6794	-19.652	-31.8338	-7.026
VoicePRc2:VoiceNXc2	-4.065	-3.9156	-15.6236	7.1063
VoicePRc3:VoiceNXc2	15.923	16.4523	-0.3985	32.5721
VoicePRc1:VoiceNXc3	-5.0862	-5.0067	-13.961	4.0456
VoicePRc2:VoiceNXc3	6.8973	7.1601	-1.6448	15.6579
VoicePRc3:VoiceNXc3	18.9485	18.9442	4.2693	34.1169
tone_2:VoiceNXc1	23.4502	23.4727	19.6506	27.3912
tone_3:VoiceNXc1	13.5009	13.4666	8.2982	18.3188
tone_4:VoiceNXc1	-17.1168	-17.0758	-20.6017	-13.8666
tone_5:VoiceNXc1	-0.265	-0.1967	-6.0675	5.8212
tone_2:VoiceNXc2	14.6606	14.7549	10.6417	18.8869
tone_3:VoiceNXc2	8.836	8.9786	3.9855	13.9276
tone_4:VoiceNXc2	-14.2639	-14.2197	-17.6571	-10.8181
tone_5:VoiceNXc2	4.0263	3.9825	-3.4159	11.4243
tone_2:VoiceNXc3	7.3145	7.3369	3.7046	10.9925
tone_3:VoiceNXc3	18.1506	18.2121	13.2193	22.9061
tone_4:VoiceNXc3	11.0196	11.0255	7.7186	14.1552
tone_5:VoiceNXc3	6.3309	6.2626	0.8324	11.6404
B2:SexM	-3.7263	-4.1555	-24.3888	16.248
$B2^2$:SexM	1.3359	1.5735	-7.7241	11.0957
$B2^3$:SexM	-0.0592	-0.0966	-1.3963	1.1933
B3:SexM	-6.2132	-6.2684	-8.3068	-4.1545
$B3^2$:SexM	0.9787	0.9878	0.6359	1.3348
$B3^3$:SexM	-0.0374	-0.0378	-0.0539	-0.0204
B4:SexM	0.1189	0.1195	-0.0893	0.3365
$B4^2$:SexM	-0.0044	-0.0045	-0.0101	7e-04
$B4^3$:SexM	0	0	0	1e-04
VoicePRc1:tone_2:VoiceNXc1	-21.6617	-21.8552	-33.2943	-10.1615
VoicePRc2:tone_2:VoiceNXc1	-3.0881	-3.0963	-15.0374	9.2013
VoicePRc3:tone_2:VoiceNXc1	-5.3977	-5.8114	-30.5647	19.4289
VoicePRc1:tone_3:VoiceNXc1	-23.649	-23.5841	-36.4229	-11.2344
VoicePRc2:tone_3:VoiceNXc1	-14.2272	-14.2603	-27.46	-2.0278
VoicePRc3:tone_3:VoiceNXc1	-11.7879	-11.5737	-34.0496	10.6165
VoicePRc1:tone_4:VoiceNXc1	-14.3334	-14.3495	-25.69	-2.6545
VoicePRc2:tone_4:VoiceNXc1	-12.3186	-12.6714	-24.0698	-1.2268
VoicePRc3:tone_4:VoiceNXc1	-6.9237	-7.1444	-30.4638	14.7665
VoicePRc1:tone_5:VoiceNXc1	-13.2769	-13.8272	-30.8302	4.1496
VoicePRc2:tone_5:VoiceNXc1	-26.94	-26.6178	-137.743	72.9093
VoicePRc1:tone_2:VoiceNXc2	14.5083	14.3874	1.0476	27.9608
VoicePRc2:tone_2:VoiceNXc2	-0.8828	-1.0245	-14.4198	12.5352
VoicePRc3:tone_2:VoiceNXc2	-29.8622	-30.4682	-51.557	-9.7455
VoicePRc1:tone_3:VoiceNXc2	2.1638	2.167	-12.6299	16.9567
VoicePRc2:tone_3:VoiceNXc2	-16.9368	-17.0992	-30.8034	-3.0288
VoicePRc3:tone_3:VoiceNXc2	-30.1949	-30.8398	-49.928	-11.2423
VoicePRc1:tone_4:VoiceNXc2	19.3024	19.2627	6.2905	32.0349
VoicePRc2:tone_4:VoiceNXc2	13.1436	13.0818	0.6847	26.705
VoicePRc3:tone_4:VoiceNXc2	2.6683	2.3374	-18.6881	24.1459
VoicePRc1:tone_5:VoiceNXc2	20.9372	21.1857	2.3534	39.6359
VoicePRc2:tone_5:VoiceNXc2	-11.8072	-9.3565	-115.4742	95.9099
VoicePRc1:tone_2:VoiceNXc3	0.1367	-0.0074	-10.8838	10.5941
VoicePRc2:tone_2:VoiceNXc3	-7.7753	-7.8538	-18.1848	2.2904
VoicePRc3:tone_2:VoiceNXc3	4.1457	3.7963	-15.8408	22.3894
VoicePRc1:tone_3:VoiceNXc3	3.1164	3.0046	-10.9786	16.4457
VoicePRc2:tone_3:VoiceNXc3	-2.6328	-3.0458	-15.948	9.7904
VoicePRc3:tone_3:VoiceNXc3	-13.7885	-13.683	-32.2288	4.2681
VoicePRc1:tone_4:VoiceNXc3	1.1082	1.0307	-8.9988	11.2283
VoicePRc2:tone_4:VoiceNXc3	-4.3281	-4.5965	-14.3616	5.968
VoicePRc3:tone_4:VoiceNXc3	-25.1533	-25.2413	-44.203	-7.1016
VoicePRc1:tone_5:VoiceNXc3	-4.473	-4.5404	-18.6902	10.3206
VoicePRc2:tone_5:VoiceNXc3	-0.5173	0.063	-72.4659	74.665
tone_3:pr_to10:nx_to10	101.6367	105.0943	11.1516	198.8864
tone_4:pr_to10:nx_to10	115.6575	115.1975	28.8751	199.5071
tone_1:pr_to11:nx_to10	-16.8138	-17.112	-42.498	7.21
tone_2:pr_to11:nx_to10	19.6554	19.3336	-5.8986	43.9835
tone_3:pr_to11:nx_to10	7.9736	7.9675	-66.3041	74.4283
tone_4:pr_to11:nx_to10	15.0227	15.1612	-11.0979	40.9997
tone_5:pr_to11:nx_to10	-11.3525	-11.2083	-79.3497	55.6357
tone_1:pr_to12:nx_to10	-24.8641	-25.4731	-51.5981	2.0528
tone_2:pr_to12:nx_to10	13.3109	12.7534	-10.6155	36.5669
tone_3:pr_to12:nx_to10	35.5947	35.7219	-7.3317	80.7352
tone_4:pr_to12:nx_to10	-7.1849	-7.5416	-28.0149	14.3637
tone_5:pr_to12:nx_to10	-75.606	-75.3686	-149.2602	-1.9087
tone_1:pr_to13:nx_to10	-23.9136	-24.9399	-54.9375	4.5577
tone_2:pr_to13:nx_to10	-26.935	-27.1223	-53.0324	0.6546
tone_3:pr_to13:nx_to10	4.7553	5.2722	-36.0818	49.5299
tone_4:pr_to13:nx_to10	0.1662	0.225	-25.1744	23.1774
tone_5:pr_to13:nx_to10	-105.254	-106.4849	-177.3401	-31.7441
tone_1:pr_to14:nx_to10	-14.0656	-14.3929	-41.2967	11.9205
tone_2:pr_to14:nx_to10	16.6597	16.125	-7.1853	38.9183
tone_3:pr_to14:nx_to10	-52.4262	-51.2365	-96.2843	-11.3077
tone_4:pr_to14:nx_to10	-24.7441	-24.4976	-42.6501	-5.5905
tone_5:pr_to14:nx_to10	-53.9567	-54.3104	-119.0375	12.1411
tone_1:pr_to15:nx_to10	-16.7572	-17.2168	-46.0134	13.4463
tone_2:pr_to15:nx_to10	3.6759	3.2044	-24.1059	28.8141
tone_3:pr_to15:nx_to10	-28.8919	-29.7464	-74.9947	10.7947
tone_4:pr_to15:nx_to10	-1.8321	-1.8609	-24.9081	20.4914
tone_5:pr_to15:nx_to10	-73.2126	-74.0311	-159.7225	4.3332
tone_1:pr_to10:nx_to11	-1.7504	-2.3459	-25.172	22.0864
tone_2:pr_to10:nx_to11	-11.8728	-12.1666	-33.5904	8.9453
tone_3:pr_to10:nx_to11	9.6114	10.4732	-17.7928	39.8643

Table A.6 – continued from previous page

FPC_2 Covariate	Estimate	BoostrapMean	(0.025,	.975)
tone_4:pr_to10:nx_to11	-36.3512	-36.2614	-53.6752	-18.1896
tone_1:pr_to11:nx_to11	-7.4055	-7.9031	-31.2004	14.279
tone_2:pr_to11:nx_to11	32.6939	32.428	13.8937	51.6283
tone_3:pr_to11:nx_to11	22.6017	23.4215	-3.7105	52.2199
tone_4:pr_to11:nx_to11	-2.2867	-2.1628	-17.7332	13.6569
tone_5:pr_to11:nx_to11	-27.9219	-28.8282	-88.2958	34.9256
tone_1:pr_to12:nx_to11	-9.3674	-9.865	-33.1301	13.1215
tone_2:pr_to12:nx_to11	28.1893	27.7952	8.4658	46.3054
tone_3:pr_to12:nx_to11	43.589	44.4484	15.8553	71.8933
tone_4:pr_to12:nx_to11	-13.3802	-13.1781	-28.0788	2.4967
tone_5:pr_to12:nx_to11	-38.2292	-38.9793	-101.0543	25.4858
tone_1:pr_to13:nx_to11	-18.2586	-18.794	-42.0093	3.4051
tone_2:pr_to13:nx_to11	4.6878	4.3521	-15.2653	23.0203
tone_3:pr_to13:nx_to11	34.0424	34.8481	7.7051	62.9437
tone_4:pr_to13:nx_to11	-39.8427	-39.5179	-54.9898	-23.7336
tone_5:pr_to13:nx_to11	-68.6674	-69.5136	-130.6959	-5.8038
tone_1:pr_to14:nx_to11	-7.447	-7.8785	-30.8378	14.6941
tone_2:pr_to14:nx_to11	13.4496	13.0934	-5.9657	31.7754
tone_3:pr_to14:nx_to11	6.0577	6.8493	-19.5702	35.5389
tone_4:pr_to14:nx_to11	-31.0144	-30.8113	-45.3844	-15.3511
tone_5:pr_to14:nx_to11	-51.1698	-52.0104	-113.2777	11.3219
tone_1:pr_to15:nx_to11	-9.1543	-9.4805	-33.2955	14.4519
tone_2:pr_to15:nx_to11	3.6566	3.3315	-16.3827	24.2356
tone_3:pr_to15:nx_to11	23.7225	24.4991	-2.9063	53.9889
tone_4:pr_to15:nx_to11	-16.6219	-16.4375	-32.3829	-0.6164
tone_5:pr_to15:nx_to11	-58.1701	-59.3175	-137.7954	13.6044
tone_1:pr_to10:nx_to12	-13.9319	-14.5387	-37.6165	7.9855
tone_2:pr_to10:nx_to12	-22.959	-23.2272	-44.5176	-2.182
tone_3:pr_to10:nx_to12	17.8095	18.8712	-8.8113	48.5249
tone_4:pr_to10:nx_to12	-29.793	-29.7134	-44.9211	-13.627
tone_1:pr_to11:nx_to12	-7.8789	-8.4937	-31.3398	13.9898
tone_2:pr_to11:nx_to12	20.2484	19.8904	1.0619	38.7764
tone_3:pr_to11:nx_to12	28.5462	29.4817	3.116	55.519
tone_4:pr_to11:nx_to12	8.3141	8.5606	-5.9531	23.386
tone_5:pr_to11:nx_to12	-14.3312	-15.0129	-74.4424	48.1137
tone_1:pr_to12:nx_to12	-12.3109	-12.7462	-35.4923	9.1689
tone_2:pr_to12:nx_to12	11.4785	11.1904	-8.2357	29.5635
tone_3:pr_to12:nx_to12	28.4092	29.2552	3.8419	56.2169
tone_4:pr_to12:nx_to12	1.2757	1.5481	-13.4178	16.8355
tone_5:pr_to12:nx_to12	-30.7632	-31.4806	-91.4351	32.8205
tone_1:pr_to13:nx_to12	-20.104	-20.5774	-43.7844	1.4863
tone_2:pr_to13:nx_to12	-23.0034	-23.4554	-43.1309	-3.9478
tone_3:pr_to13:nx_to12	27.1374	28.0946	1.0162	55.3247
tone_4:pr_to13:nx_to12	-28.744	-28.5364	-43.0742	-13.2919
tone_5:pr_to13:nx_to12	-58.6922	-59.4856	-119.4555	4.4171
tone_1:pr_to14:nx_to12	-13.3235	-13.8781	-37.4411	8.2828
tone_2:pr_to14:nx_to12	-4.6904	-5.0528	-23.8202	13.8281
tone_3:pr_to14:nx_to12	16.395	17.393	-9.1676	44.9638
tone_4:pr_to14:nx_to12	-15.6026	-15.4509	-30.1997	-0.366
tone_5:pr_to14:nx_to12	-37.1466	-38.3692	-98.489	24.1067
tone_1:pr_to15:nx_to12	-12.0483	-12.6333	-35.942	9.578
tone_2:pr_to15:nx_to12	-18.7752	-19.1284	-39.0121	0.2406
tone_3:pr_to15:nx_to12	25.7779	26.6916	-1.1341	55.6438
tone_4:pr_to15:nx_to12	-1.8342	-1.7028	-16.8771	14.3297
tone_5:pr_to15:nx_to12	-46.6468	-48.1566	-118.2517	23.7084
tone_1:pr_to10:nx_to13	-12.0923	-12.6576	-36.9103	11.4675
tone_2:pr_to10:nx_to13	-16.352	-16.4911	-38.4721	4.4972
tone_3:pr_to10:nx_to13	-81.3311	-80.436	-108.1673	-51.3743
tone_4:pr_to10:nx_to13	-23.6362	-23.261	-39.0506	-6.637
tone_1:pr_to11:nx_to13	-7.5688	-8.1399	-31.8408	14.057
tone_2:pr_to11:nx_to13	20.4492	20.0849	0.2266	39.1585
tone_3:pr_to11:nx_to13	-68.1374	-67.1968	-94.0982	-40.2004
tone_4:pr_to11:nx_to13	7.7073	7.8612	-7.4976	23.26
tone_5:pr_to11:nx_to13	-20.9665	-21.8103	-82.9129	41.8535
tone_1:pr_to12:nx_to13	-8.0237	-8.452	-31.8263	13.9852
tone_2:pr_to12:nx_to13	14.2476	13.9358	-5.4354	32.3326
tone_3:pr_to12:nx_to13	-75.0723	-74.1022	-99.8194	-46.515
tone_4:pr_to12:nx_to13	-3.7422	-3.4725	-18.3103	12.1486
tone_5:pr_to12:nx_to13	-40.1174	-40.9195	-102.02	21.6723
tone_1:pr_to13:nx_to13	-25.3501	-25.9055	-49.1369	-3.7995
tone_2:pr_to13:nx_to13	-32.2174	-32.591	-51.1518	-13.3
tone_3:pr_to13:nx_to13	-82.2105	-81.0979	-107.8514	-53.3537
tone_4:pr_to13:nx_to13	-27.7528	-27.585	-42.5845	-12.2096
tone_5:pr_to13:nx_to13	-75.921	-76.7584	-138.5995	-14.762
tone_1:pr_to14:nx_to13	-13.765	-14.1962	-37.9891	7.7865
tone_2:pr_to14:nx_to13	-13.9513	-14.3176	-32.9609	4.0947
tone_3:pr_to14:nx_to13	-86.0466	-85.2191	-111.8112	-57.7592
tone_4:pr_to14:nx_to13	-10.4534	-10.3241	-24.7667	4.4491
tone_5:pr_to14:nx_to13	-41.0237	-41.8405	-102.6446	22.7566
tone_1:pr_to15:nx_to13	-7.2141	-7.9496	-31.368	15.5023
tone_2:pr_to15:nx_to13	-29.5429	-30.0742	-50.3535	-11.0984
tone_3:pr_to15:nx_to13	-120.3475	-119.1951	-146.638	-92.241
tone_4:pr_to15:nx_to13	1.9487	2.2308	-13.4403	18.541
tone_5:pr_to15:nx_to13	-54.075	-55.25	-119.426	10.8672
tone_1:pr_to10:nx_to14	-15.203	-15.7089	-39.6079	8.5117
tone_2:pr_to10:nx_to14	-0.4917	-0.5765	-20.7014	19.5545
tone_3:pr_to10:nx_to14	1.9763	3.0449	-24.4956	31.7442
tone_4:pr_to10:nx_to14	-37.3719	-37.1035	-52.6694	-20.6655
tone_1:pr_to11:nx_to14	-10.3643	-10.9365	-33.8825	11.423
tone_2:pr_to11:nx_to14	25.2936	25.0123	5.7742	43.1298
tone_3:pr_to11:nx_to14	19.9265	20.7707	-6.3479	47.8206
tone_4:pr_to11:nx_to14	-11.8556	-11.6618	-25.8497	2.7874
tone_5:pr_to11:nx_to14	-32.9288	-33.6833	-92.9116	29.632
tone_1:pr_to12:nx_to14	-15.6351	-16.0886	-38.6477	6.1265
tone_2:pr_to12:nx_to14	20.2749	19.8222	0.8886	38.4218
tone_3:pr_to12:nx_to14	25.9886	26.7961	-0.4789	52.8299
tone_4:pr_to12:nx_to14	-14.311	-14.0267	-28.9034	1.1648
tone_5:pr_to12:nx_to14	-53.7245	-54.7109	-115.7413	8.8372
tone_1:pr_to13:nx_to14	-24.222	-24.7703	-47.691	-1.9079
tone_2:pr_to13:nx_to14	-13.2756	-13.5627	-32.5484	4.9443
tone_3:pr_to13:nx_to14	24.4285	25.4143	-0.517	52.2013
tone_4:pr_to13:nx_to14	-37.7049	-37.5202	-51.4598	-22.4737
tone_5:pr_to13:nx_to14	-82.0019	-82.8576	-143.3108	-20.0986

Table A.6 – continued from previous page

FPC_2 Covariate	Estimate	BoostrapMean	(0.025,	.975)
tone_1:pr_to14:nx_to14	-16.2711	-16.7218	-39.237	5.78
tone_2:pr_to14:nx_to14	3.9737	3.6287	-14.9908	22.4613
tone_3:pr_to14:nx_to14	6.1508	7.0322	-19.0234	34.1592
tone_4:pr_to14:nx_to14	-28.0252	-27.8251	-42.3609	-12.6807
tone_5:pr_to14:nx_to14	-49.2009	-50.0778	-110.7767	12.3673
tone_1:pr_to15:nx_to14	-18.618	-19.1626	-42.3024	3.7481
tone_2:pr_to15:nx_to14	-1.6736	-2.1438	-21.6123	16.3799
tone_3:pr_to15:nx_to14	19.8391	20.7008	-6.8748	48.6743
tone_4:pr_to15:nx_to14	-5.7495	-5.6331	-19.9799	9.4225
tone_5:pr_to15:nx_to14	-53.6318	-54.5386	-119.4443	9.1279
tone_1:pr_to10:nx_to15	-15.3183	-15.4899	-48.9972	19.0682
tone_2:pr_to10:nx_to15	11.3191	11.3725	-20.8398	45.4125
tone_3:pr_to10:nx_to15	9.4356	10.3112	-21.5988	43.6474
tone_4:pr_to10:nx_to15	-15.4534	-16.5175	-47.7756	14.7355
tone_1:pr_to11:nx_to15	1.4198	0.8726	-22.0436	23.7766
tone_2:pr_to11:nx_to15	53.572	53.4328	32.255	73.5825
tone_3:pr_to11:nx_to15	26.3094	27.1114	-0.0619	57.1681
tone_4:pr_to11:nx_to15	-1.2357	-1.0688	-17.5302	14.9482
tone_5:pr_to11:nx_to15	-24.6921	-25.4393	-89.5747	40.5193
tone_1:pr_to12:nx_to15	-7.4658	-8.0599	-30.759	14.8021
tone_2:pr_to12:nx_to15	61.7266	61.2917	41.1879	80.8714
tone_3:pr_to12:nx_to15	34.6563	35.4982	8.0884	63.2603
tone_4:pr_to12:nx_to15	-2.5878	-2.3444	-17.2604	13.0575
tone_5:pr_to12:nx_to15	-40.8395	-41.4054	-150.8349	61.799
tone_1:pr_to13:nx_to15	-12.8073	-13.3405	-36.8492	9.3534
tone_2:pr_to13:nx_to15	-13.4868	-13.74	-34.8098	6.6653
tone_3:pr_to13:nx_to15	44.5205	45.2745	18.0732	74.125
tone_4:pr_to13:nx_to15	-40.4052	-40.1977	-55.0641	-24.3558
tone_5:pr_to13:nx_to15	-79.6052	-80.2113	-144.5495	-16.5066
tone_1:pr_to14:nx_to15	-7.6141	-8.1342	-31.5627	14.4889
tone_2:pr_to14:nx_to15	27.7412	27.321	7.6862	45.9698
tone_3:pr_to14:nx_to15	7.4389	8.252	-19.0024	35.0655
tone_4:pr_to14:nx_to15	-29.5352	-29.2832	-44.179	-13.0956

Table A.7: All fixed effects and 95% confidence intervals for the 3rd FPC scores models.

FPC_3 Covariate	Estimate	BoostrapMean	(0.025,	.975)
VoicePRc0	-24.2007	-24.6511	-54.1754	6.5912
VoicePRc1	-23.0395	-23.4703	-53.1449	8.0658
VoicePRc2	-39.6155	-40.0254	-70.7885	-8.882
VoicePRc3	-41.6149	-41.9354	-72.9321	-10.175
tone_2	12.0664	12.027	6.4178	17.5482
tone_3	22.2355	22.4578	13.6794	31.64
tone_4	10.714	10.7838	5.2627	16.3114
tone_5	-2.2683	-2.3923	-13.2311	8.107
VoiceNXc1	5.4562	5.4578	4.3214	6.5622
VoiceNXc2	2.1731	2.2077	0.9238	3.4127
VoiceNXc3	1.6484	1.6615	0.5651	2.8157
rhyme_ɚ	-4.5691	-4.58	-7.8081	-1.0934
rhyme_a	7.4083	7.4035	6.0306	8.9141
rhyme_ai	3.8653	3.8692	2.5233	5.2143
rhyme_an	7.6447	7.6634	6.3592	8.9902
rhyme_aŋ	8.5254	8.5446	7.0654	10.0144
rhyme_au	4.2842	4.2888	2.5714	6.0476
rhyme_ei	3.2304	3.2157	1.6918	4.9915
rhyme_i	-1.7896	-1.7792	-2.9248	-0.654
rhyme_ia	15.0941	15.1045	13.1738	17.0402
rhyme_iaŋ	19.3056	19.3614	17.6562	20.9317
rhyme_iau	14.743	14.7596	13.0929	16.4207
rhyme_iɛ	1.4665	1.4491	-0.1282	2.9047
rhyme_iɛn	15.0395	15.0584	13.7833	16.3103
rhyme_in	6.0707	6.077	4.725	7.5196
rhyme_iŋ	7.3755	7.3988	6.1715	8.5743
rhyme_iou	1.4798	1.4491	0.063	2.9719
rhyme_ɔn	5.4363	5.4442	4.2228	6.696
rhyme_ɔŋ	6.9344	6.9339	5.6872	8.1692
rhyme_o	1.4862	1.4483	-2.0209	4.8291
rhyme_oŋ	6.3025	6.3056	5.1489	7.5854
rhyme_ou	2.6953	2.6842	0.892	4.3809
rhyme_u	-3.462	-3.4517	-4.6697	-2.2819
rhyme_ɻ	-7.2924	-7.2849	-9.0474	-5.5551
rhyme_ɿ	-5.3966	-5.3822	-6.4932	-4.2027
rhyme_ua	9.0999	9.0915	5.0334	13.1846
rhyme_uai	-3.3283	-3.3212	-6.6142	0.2448
rhyme_uan	12.065	12.0414	10.3544	13.7304
rhyme_uaŋ	9.7182	9.6846	7.4229	11.9541
rhyme_uei	1.2739	1.2851	-0.1749	2.5521
rhyme_uən	10.2114	10.2773	8.0492	12.4162
rhyme_uo	1.7555	1.7655	0.5267	2.9456
rhyme_y	-3.7381	-3.747	-5.3617	-2.1157
rhyme_yɛ	3.9225	3.9114	1.9569	5.7962
rhyme_yɛn	18.5084	18.4855	16.6136	20.3134
rhyme_yn	3.2909	3.3556	0.7027	5.8353
rhyme_yoŋ	16.6355	16.5432	13.1849	20.0667
B2	12.6831	12.7634	8.0831	17.5311
$B2^2$	-3.5336	-3.5629	-5.7919	-1.4053
$B2^3$	0.2529	0.2558	-0.048	0.5576
B3	2.5789	2.5462	1.9693	3.108
$B3^2$	-0.3343	-0.3291	-0.4282	-0.2342
$B3^3$	0.012	0.0118	0.0073	0.0164
SexM	-14.9165	-15.1718	-29.7392	-1.2207
pr_to11	8.9777	9.5068	-22.4381	38.122
pr_to12	5.0858	5.7483	-25.323	33.4453
pr_to13	-1.3269	-0.6748	-32.6002	27.8196
pr_to14	2.6105	3.1185	-28.4503	31.8424
pr_to15	3.5195	4.1549	-26.5598	32.4958
nx_to11	1.7628	2.412	-29.4204	31.0962
nx_to12	0.4202	1.0828	-31.1247	28.5619
nx_to13	-1.6538	-1.0591	-33.42	27.6034
nx_to14	5.4889	6.0709	-25.7206	34.5225

Table A.7 – continued from previous page

FPC_3 Covariate	Estimate	BoostrapMean	(0.025,	.975)
nx_to15	0.5241	1.2044	-30.4156	29.4233
VoicePRc1:tone_2	-3.5369	-3.6107	-6.3881	-0.4976
VoicePRc2:tone_2	31.1358	31.1146	28.4621	33.896
VoicePRc3:tone_2	22.4298	22.192	16.6017	27.5998
VoicePRc1:tone_3	-1.8105	-1.8141	-4.7868	1.7504
VoicePRc2:tone_3	19.1386	19.1098	15.9581	22.3832
VoicePRc3:tone_3	6.6268	6.5235	1.5736	11.3142
VoicePRc1:tone_4	-11.1624	-11.1814	-13.7957	-8.1046
VoicePRc2:tone_4	-28.0451	-28.0853	-30.5611	-25.4555
VoicePRc3:tone_4	-15.8456	-15.9633	-20.9926	-10.4787
VoicePRc1:tone_5	-3.6551	-3.6568	-7.9256	0.7715
VoicePRc2:tone_5	22.6037	22.9828	-7.357	51.8202
VoicePRc1:VoiceNXc1	-3.2089	-3.2263	-7.8053	1.6608
VoicePRc2:VoiceNXc1	-6.1985	-6.2728	-10.8227	-1.8854
VoicePRc3:VoiceNXc1	2.0479	2.0268	-7.4384	10.9661
VoicePRc1:VoiceNXc2	-0.2128	-0.22	-6.0502	5.5104
VoicePRc2:VoiceNXc2	-0.8242	-0.8936	-6.1321	4.4928
VoicePRc3:VoiceNXc2	11.2991	11.063	3.4481	18.5324
VoicePRc1:VoiceNXc3	1.6373	1.5979	-2.6063	5.7981
VoicePRc2:VoiceNXc3	-1.1334	-1.2529	-5.2857	2.889
VoicePRc3:VoiceNXc3	6.5301	6.5299	-0.6951	13.3577
tone_2:VoiceNXc1	-0.3429	-0.3512	-2.1593	1.4183
tone_3:VoiceNXc1	-7.7717	-7.7523	-10.0578	-5.3505
tone_4:VoiceNXc1	-10.1751	-10.191	-11.7106	-8.5377
tone_5:VoiceNXc1	-5.7418	-5.7748	-8.6426	-3.0129
tone_2:VoiceNXc2	0.5261	0.4822	-1.4418	2.4477
tone_3:VoiceNXc2	-2.4512	-2.5143	-4.8657	-0.1457
tone_4:VoiceNXc2	-5.0009	-5.018	-6.6158	-3.3183
tone_5:VoiceNXc2	-0.5452	-0.5215	-3.9625	2.9553
tone_2:VoiceNXc3	5.7134	5.701	3.9849	7.3756
tone_3:VoiceNXc3	3.8945	3.8684	1.6643	6.1314
tone_4:VoiceNXc3	0.6201	0.6198	-0.8549	2.195
tone_5:VoiceNXc3	0.942	0.9703	-1.535	3.6009
B2:SexM	-3.8809	-3.6799	-13.3596	6.0878
$B2^2$:SexM	2.042	1.931	-2.5493	6.3209
$B2^3$:SexM	-0.2475	-0.2301	-0.8525	0.3897
B3:SexM	-0.545	-0.5187	-1.5136	0.4405
$B3^2$:SexM	0.0697	0.0655	-0.0989	0.2302
$B3^3$:SexM	-0.0022	-0.002	-0.0099	0.0058
tone_2:pr_to11	11.6652	11.7132	8.5714	14.7508
tone_3:pr_to11	6.4852	6.5255	2.8881	10.2134
tone_4:pr_to11	8.9332	8.921	6.0856	11.622
tone_5:pr_to11	6.4926	6.3181	-0.7318	13.5639
tone_2:pr_to12	7.9949	8.1285	4.9612	11.2857
tone_3:pr_to12	5.2042	5.3038	1.8458	8.9626
tone_4:pr_to12	2.2011	2.201	-0.5528	4.9864
tone_5:pr_to12	2.1578	2.0472	-4.5631	9.007
tone_2:pr_to13	2.411	2.491	-0.5659	5.7141
tone_3:pr_to13	9.6923	9.6998	6.1179	13.4199
tone_4:pr_to13	-2.1128	-2.1245	-4.8033	0.7793
tone_5:pr_to13	-0.0055	-0.1456	-6.6917	6.9196
tone_2:pr_to14	4.5119	4.6308	1.8239	7.4753
tone_3:pr_to14	1.5382	1.6223	-1.6343	5.0004
tone_4:pr_to14	4.4456	4.477	1.6755	7.0474
tone_5:pr_to14	1.5611	1.491	-5.2628	8.1966
tone_2:pr_to15	8.1353	8.2473	4.9865	11.7221
tone_3:pr_to15	4.4769	4.5261	0.472	8.6226
tone_4:pr_to15	3.7708	3.7868	0.6607	6.8479
tone_2:nx_to11	-6.6413	-6.6721	-11.4818	-1.7318
tone_3:nx_to11	-1.231	-1.4497	-9.753	6.4113
tone_4:nx_to11	3.2017	3.1256	-1.6894	8.0713
tone_5:nx_to11	3.2732	3.5261	-5.4756	12.8298
tone_2:nx_to12	-4.7406	-4.7927	-9.4877	0.0691
tone_3:nx_to12	-5.1117	-5.4241	-14.0997	2.7736
tone_4:nx_to12	-0.7796	-0.8803	-5.728	4.0402
tone_5:nx_to12	4.4055	4.6553	-4.1854	14.2764
tone_2:nx_to13	-5.2049	-5.244	-10.2462	-0.2212
tone_3:nx_to13	-3.588	-3.8855	-12.5806	4.1431
tone_4:nx_to13	-0.7865	-0.8699	-5.8747	4.2493
tone_5:nx_to13	4.3033	4.5303	-4.7074	13.8183
tone_2:nx_to14	-4.3714	-4.4101	-9.1412	0.5584
tone_3:nx_to14	-0.8238	-1.0863	-9.4513	6.5894
tone_4:nx_to14	0.068	-0.0215	-4.9312	4.9212
tone_5:nx_to14	2.7575	3.0129	-5.8152	12.2777
tone_2:nx_to15	1.0478	0.978	-4.019	6.1388
tone_3:nx_to15	2.1937	1.9558	-6.8026	10.263
tone_4:nx_to15	-1.6452	-1.7467	-6.7542	3.6445
tone_5:nx_to15	11.1285	11.277	0.9826	21.7781
pr_to11:nx_to11	-6.3663	-6.9343	-35.8879	24.7384
pr_to12:nx_to11	-2.9177	-3.6522	-31.6539	27.8274
pr_to13:nx_to11	-3.0492	-3.7621	-32.3956	28.0283
pr_to14:nx_to11	-3.3115	-3.9073	-33.1018	27.216
pr_to15:nx_to11	-4.617	-5.3461	-34.5598	25.5572
pr_to11:nx_to12	-7.0852	-7.6443	-36.0935	24.051
pr_to12:nx_to12	-5.5488	-6.3061	-33.5511	25.2265
pr_to13:nx_to12	-3.2749	-3.9656	-32.0704	27.2997
pr_to14:nx_to12	-3.6984	-4.2635	-32.9469	27.2943
pr_to15:nx_to12	-5.5957	-6.2678	-34.9123	23.8519
pr_to11:nx_to13	-6.3202	-6.8086	-35.8283	25.3928
pr_to12:nx_to13	-3.6931	-4.3936	-32.5107	27.3731
pr_to13:nx_to13	-3.0741	-3.7061	-32.8057	28.8387
pr_to14:nx_to13	-2.0538	-2.5654	-32.0438	29.1977
pr_to15:nx_to13	-4.3395	-4.9627	-34.2406	26.1625
pr_to11:nx_to14	-9.6354	-10.1275	-38.6825	21.5843
pr_to12:nx_to14	-5.7209	-6.3836	-34.236	24.4036
pr_to13:nx_to14	-4.5489	-5.1748	-33.2572	26.7302
pr_to14:nx_to14	-4.5711	-5.0912	-33.7706	26.3689
pr_to15:nx_to14	-5.8492	-6.4431	-34.9574	24.4991
pr_to11:nx_to15	-12.9451	-13.5376	-42.1258	17.9917
pr_to12:nx_to15	-10.8438	-11.5779	-39.2433	21.0191
pr_to13:nx_to15	-12.7015	-13.4135	-41.6006	17.7332
pr_to14:nx_to15	-7.3909	-7.9891	-36.1214	24.106
pr_to15:nx_to15	-6.8695	-7.5779	-36.0681	24.399

Table A.7 – continued from previous page

FPC_3 Covariate	Estimate	BoostrapMean	(0.025,	.975)
VoicePRc1:tone_2:VoiceNXc1	3.9855	4.0754	-1.4733	9.483
VoicePRc2:tone_2:VoiceNXc1	-7.6385	-7.635	-13.5086	-1.9333
VoicePRc3:tone_2:VoiceNXc1	-0.364	-0.1734	-12.2781	11.2117
VoicePRc1:tone_3:VoiceNXc1	-1.0087	-1.0389	-6.7362	4.9582
VoicePRc2:tone_3:VoiceNXc1	-3.4653	-3.4516	-9.219	2.7144
VoicePRc3:tone_3:VoiceNXc1	-5.707	-5.8182	-15.947	4.8614
VoicePRc1:tone_4:VoiceNXc1	2.7952	2.8033	-2.6344	8.0559
VoicePRc2:tone_4:VoiceNXc1	5.2133	5.3791	-0.2104	10.7107
VoicePRc3:tone_4:VoiceNXc1	-0.521	-0.4149	-10.5925	10.4273
VoicePRc1:tone_5:VoiceNXc1	5.1969	5.4577	-2.9219	13.1671
VoicePRc2:tone_5:VoiceNXc1	2.4031	2.2428	-44.9853	53.0327
VoicePRc1:tone_2:VoiceNXc2	2.2865	2.3413	-4.0576	8.6375
VoicePRc2:tone_2:VoiceNXc2	-7.1758	-7.1079	-13.4436	-0.8515
VoicePRc3:tone_2:VoiceNXc2	-11.171	-10.8947	-20.5228	-1.0919
VoicePRc1:tone_3:VoiceNXc2	-1.6289	-1.6368	-8.6105	5.3565
VoicePRc2:tone_3:VoiceNXc2	-1.5497	-1.4773	-8.1808	5.0167
VoicePRc3:tone_3:VoiceNXc2	0.2703	0.5557	-8.6232	9.5255
VoicePRc1:tone_4:VoiceNXc2	-0.1767	-0.1658	-6.166	5.9312
VoicePRc2:tone_4:VoiceNXc2	-2.5313	-2.5002	-8.9372	3.3139
VoicePRc3:tone_4:VoiceNXc2	-13.4843	-13.3486	-23.4748	-3.2626
VoicePRc1:tone_5:VoiceNXc2	0.154	0.0396	-8.8461	9.0315
VoicePRc2:tone_5:VoiceNXc2	-10.0907	-11.2636	-60.3628	38.1805
VoicePRc1:tone_2:VoiceNXc3	-0.5476	-0.4737	-5.3918	4.5669
VoicePRc2:tone_2:VoiceNXc3	1.3367	1.3723	-3.473	6.2631
VoicePRc3:tone_2:VoiceNXc3	-9.9242	-9.7602	-18.7124	-0.5812
VoicePRc1:tone_3:VoiceNXc3	-7.5122	-7.4661	-13.7963	-0.9006
VoicePRc2:tone_3:VoiceNXc3	-8.2915	-8.0973	-14.0435	-2.029
VoicePRc3:tone_3:VoiceNXc3	-10.311	-10.3553	-18.8454	-1.9403
VoicePRc1:tone_4:VoiceNXc3	-3.2395	-3.2008	-8.0096	1.5005
VoicePRc2:tone_4:VoiceNXc3	-1.3768	-1.2527	-6.2659	3.3877
VoicePRc3:tone_4:VoiceNXc3	-3.6136	-3.5721	-12.0295	5.4762
VoicePRc1:tone_5:VoiceNXc3	-6.3464	-6.3188	-13.3485	0.2029
VoicePRc2:tone_5:VoiceNXc3	-7.1153	-7.3997	-43.351	25.9848

Table A.8: All fixed effects and 95% confidence intervals for the 4th FPC scores models.

FPC_4 Covariate	Estimate	BoostrapMean	(0.025,	.975)
VoicePRc0	1.397	1.3006	-2.7321	5.6755
VoicePRc1	1.4855	1.3873	-2.8497	5.9286
VoicePRc2	10.9811	10.8804	6.5633	15.1911
VoicePRc3	7.9499	7.8076	3.2818	12.4613
tone_2	2.5776	2.6303	0.9794	4.3325
tone_3	3.2176	3.2449	1.3839	5.1827
tone_4	0.849	0.869	-0.5824	2.2754
tone_5	2.3633	2.3066	-1.4231	6.116
VoiceNXc1	5.2725	5.2715	4.7037	5.8579
VoiceNXc2	3.4829	3.4648	2.8004	4.1221
VoiceNXc3	1.6075	1.6045	1.0025	2.2087
rhyme_ɚ	-3.522	-3.5132	-5.4311	-1.7138
rhyme_a	-0.6564	-0.6519	-1.5068	0.1208
rhyme_ai	-0.4601	-0.4613	-1.2343	0.2924
rhyme_an	-1.3065	-1.3167	-2.0403	-0.573
rhyme_aŋ	-2.9934	-3.0035	-3.8436	-2.1912
rhyme_au	-0.556	-0.5588	-1.547	0.3895
rhyme_ei	-0.412	-0.4046	-1.4153	0.4714
rhyme_i	-1.7725	-1.7781	-2.4224	-1.1321
rhyme_ia	-2.4031	-2.4094	-3.5139	-1.321
rhyme_iaŋ	-5.098	-5.1282	-6.018	-4.1644
rhyme_iau	-1.9505	-1.9599	-2.9101	-1.0069
rhyme_iɛ	0.4021	0.4156	-0.417	1.2981
rhyme_iɛn	0.1865	0.1772	-0.5083	0.8856
rhyme_in	-1.388	-1.3903	-2.1917	-0.6207
rhyme_iŋ	-2.0048	-2.018	-2.6869	-1.342
rhyme_iou	0.3379	0.3546	-0.5072	1.154
rhyme_ən	-1.884	-1.8874	-2.6155	-1.1919
rhyme_əŋ	-3.0794	-3.0801	-3.7923	-2.3732
rhyme_o	-0.2095	-0.1857	-2.066	1.7053
rhyme_oŋ	-1.3406	-1.3425	-2.0618	-0.6809
rhyme_ou	-1.5987	-1.5931	-2.5636	-0.5848
rhyme_u	-1.2897	-1.2946	-1.9483	-0.5863
rhyme_ɿ	0.8912	0.8873	-0.0546	1.9033
rhyme_ʅ	1.8975	1.89	1.2202	2.5306
rhyme_ua	-1.7652	-1.7587	-4.0698	0.5496
rhyme_uai	-1.1721	-1.1803	-3.2236	0.6667
rhyme_uan	-0.4153	-0.4013	-1.3571	0.5781
rhyme_uaŋ	-0.0347	-0.0184	-1.2726	1.2356
rhyme_uei	-0.7378	-0.7451	-1.4681	0.0733
rhyme_uən	0.563	0.5248	-0.682	1.7667
rhyme_uo	1.1763	1.1714	0.5236	1.8733
rhyme_y	-1.933	-1.9264	-2.8442	-1.0126
rhyme_yɛ	0.7127	0.7201	-0.3407	1.8151
rhyme_yɛn	0.1926	0.2062	-0.8273	1.2316
rhyme_yn	3.3897	3.354	1.9361	4.8304
rhyme_yoŋ	4.0722	4.1226	2.1635	6.0763
B2	-3.2315	-3.2789	-5.9851	-0.6106
$B2^2$	1.3945	1.4122	0.1755	2.6585
$B2^3$	-0.1673	-0.1691	-0.3427	0.0076
B3	-0.5276	-0.5098	-0.8282	-0.1847
$B3^2$	0.1181	0.1152	0.0602	0.172
$B3^3$	-0.0049	-0.0048	-0.0074	-0.0022
SexM	1.2175	1.3424	-5.1356	7.9374
pr_to11	-0.8875	-0.89	-2.0903	0.3921
pr_to12	-0.3507	-0.3279	-1.5453	0.9354
pr_to13	1.5634	1.5681	0.3748	2.8549
pr_to14	0.228	0.2485	-0.9009	1.4698
pr_to15	0.2866	0.2903	-1.0529	1.668
nx_to11	-1.2466	-1.2283	-2.2694	-0.186
nx_to12	-2.8751	-2.8598	-3.8525	-1.8743
nx_to13	-3.1308	-3.1129	-4.1768	-2.0745
nx_to14	0.1721	0.1898	-0.8887	1.2

Table A.8 – continued from previous page

FPC_4 Covariate	Estimate	BoostrapMean	(0.025,	.975)
nx_to15	-3.1912	-3.1754	-4.2869	-2.1049
B4	0.0483	0.049	0.025	0.0736
$B4^2$	-7e-04	-7e-04	-0.0013	-2e-04
$B4^3$	0	0	0	0
VoicePRc1:tone_2	2.7985	2.8215	1.7065	3.9122
VoicePRc2:tone_2	-3.2316	-3.2284	-4.3921	-2.127
VoicePRc3:tone_2	4.0455	4.1113	2.0885	6.1303
VoicePRc1:tone_3	3.4622	3.4627	2.185	4.7022
VoicePRc2:tone_3	-2.8914	-2.9065	-4.1668	-1.6543
VoicePRc3:tone_3	5.015	5.0563	3.0931	6.9238
VoicePRc1:tone_4	4.3405	4.3468	3.2386	5.4039
VoicePRc2:tone_4	2.7484	2.7391	1.6489	3.8042
VoicePRc3:tone_4	0.3115	0.3463	-1.5828	2.3165
VoicePRc1:tone_5	1.1211	1.1232	-0.5422	2.8286
VoicePRc2:tone_5	-7.8781	-7.8952	-15.205	-0.6614
tone_2:VoiceNXc1	1.5289	1.527	0.6905	2.3966
tone_3:VoiceNXc1	-4.4757	-4.4597	-5.5567	-3.4442
tone_4:VoiceNXc1	-6.4255	-6.4192	-7.2434	-5.6516
tone_5:VoiceNXc1	-4.1648	-4.16	-5.5898	-2.7312
tone_2:VoiceNXc2	1.2838	1.2949	0.3563	2.2374
tone_3:VoiceNXc2	-0.3477	-0.3109	-1.4454	0.8001
tone_4:VoiceNXc2	-4.9084	-4.8994	-5.7567	-4.059
tone_5:VoiceNXc2	-3.5388	-3.5373	-5.2426	-1.7487
tone_2:VoiceNXc3	-3.5439	-3.5457	-4.3464	-2.7316
tone_3:VoiceNXc3	-4.7781	-4.7714	-5.7783	-3.7753
tone_4:VoiceNXc3	-3.2292	-3.2366	-4.001	-2.4348
tone_5:VoiceNXc3	-2.1557	-2.1716	-3.5532	-0.7833
VoicePRc1:VoiceNXc1	-0.8628	-0.8754	-1.6808	-0.0113
VoicePRc2:VoiceNXc1	-2.9106	-2.9045	-3.928	-1.8561
VoicePRc3:VoiceNXc1	-4.6218	-4.6307	-6.4314	-2.6753
VoicePRc1:VoiceNXc2	0.7657	0.76	-0.1128	1.6023
VoicePRc2:VoiceNXc2	-2	-1.9827	-3.0642	-0.9458
VoicePRc3:VoiceNXc2	-2.873	-2.844	-4.713	-1.0069
VoicePRc1:VoiceNXc3	0.4644	0.4599	-0.251	1.2388
VoicePRc2:VoiceNXc3	0.3999	0.4206	-0.4802	1.259
VoicePRc3:VoiceNXc3	-4.0995	-4.1164	-5.6858	-2.5615
B2:SexM	-2.5962	-2.7057	-8.2485	2.7293
$B2^2$:SexM	0.7136	0.7743	-1.7569	3.3184
$B2^3$:SexM	-0.0699	-0.0794	-0.4302	0.2698
B3:SexM	0.1028	0.0889	-0.4479	0.6554
$B3^2$:SexM	-0.0464	-0.0442	-0.1406	0.0503
$B3^3$:SexM	0.0019	0.0018	-0.0026	0.0064
tone_2:pr_to11	-6.3244	-6.3531	-8.0101	-4.5725
tone_3:pr_to11	-2.7697	-2.7952	-4.8525	-0.7216
tone_4:pr_to11	1.3626	1.3618	-0.1779	2.9751
tone_5:pr_to11	-0.6703	-0.5825	-4.7947	3.3122
tone_2:pr_to12	-3.3807	-3.4576	-5.2266	-1.6652
tone_3:pr_to12	-2.7326	-2.7879	-4.8068	-0.7793
tone_4:pr_to12	1.8683	1.861	0.3167	3.3705
tone_5:pr_to12	1.5398	1.5914	-2.3477	5.445
tone_2:pr_to13	-0.0663	-0.1164	-1.9359	1.6222
tone_3:pr_to13	-3.0023	-3.0109	-5.1146	-1.0273
tone_4:pr_to13	2.1397	2.1374	0.5317	3.6741
tone_5:pr_to13	3.7198	3.7891	-0.1484	7.4674
tone_2:pr_to14	-1.4564	-1.5257	-3.1219	0.0544
tone_3:pr_to14	-1.7293	-1.778	-3.7354	0.0653
tone_4:pr_to14	1.0424	1.0162	-0.4234	2.5411
tone_5:pr_to14	1.0445	1.0732	-2.6103	4.9265
tone_2:pr_to15	-1.6132	-1.6769	-3.6302	0.2285
tone_3:pr_to15	-1.312	-1.3421	-3.5382	0.966
tone_4:pr_to15	2.145	2.1269	0.404	3.847

A.9 Covariates for Figures 4.5 and 4.6

SpkrID	SentIdx	RhymeIdx	Tone	PrevTone	NextTone	B2	B3	B4	B5	PrevCons	NextCons	VowelRhyme
*	564	2	1	2	2	2	2	2	2	k	l	uei
*	124	1	2	0	3	1	1	1	1	n	NA	oŋ
*	336	1	3	0	4	1	1	1	1	t	t	əŋ
*	444	4	4	1	5	2	4	4	4	t	t	uan
*	529	3	5	1	1	3	3	3	3	t	ɕ	ə

Table A.9: Specific covariate information for the F_0 curves in Fig. 4.5.

SpkrID	SentIdx	RhymeIdx	Tone	PrevTone	NextTone	B2	B3	B4	B5	PrevCons	NextCons	VowelRhyme
F03	537	33	2.0	1.0	1.0	1	1	33	33	t	s	i
F03	537	34	1.0	2.0	4.0	2	2	34	34	s	n	ɯ[ɻ]
F03	537	35	4.0	1.0	4.0	3	3	35	35	n	l	ai
F03	537	36	4.0	4.0	2.0	1	4	36	36	l	NA	ə
F03	537	37	2.0	4.0	5.0	2	5	37	37	NA	t	yɛn
F03	537	38	5.0	2.0	2.0	3	6	38	38	t	sil	ə

Table A.10: Specific covariate information for the F_0 track in Fig. 4.6.

A.10 Covariance Structures for Amplitude & Phase model

$$\Sigma_\Gamma = \begin{bmatrix} \sigma^2_{\Gamma/w_1} & 0 & \cdots & 0 & \sigma^2_{\Gamma/w_1,s_1} & \cdots & \cdots & \sigma^2_{\Gamma/w_1,s_{p_s}} & \sigma^2_{\Gamma/w_1,T} \\ 0 & \ddots & \ddots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & \sigma^2_{\Gamma/w_{p_w}} & \sigma^2_{\Gamma/w_{p_w},s_1} & \cdots & \cdots & \sigma^2_{\Gamma/w_{p_w},s_{p_s}} & \sigma^2_{\Gamma/w_{p_w},T} \\ \sigma^2_{\Gamma/s_1,w_1} & \cdots & \cdots & \sigma^2_{\Gamma/s_1,w_{p_w}} & \sigma^2_{\Gamma/s_1} & 0 & \cdots & 0 & \sigma^2_{\Gamma/s_1,T} \\ \vdots & \ddots & \ddots & \vdots & 0 & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & \vdots & \vdots & \ddots & \ddots & 0 & \vdots \\ \sigma^2_{\Gamma/s_{p_s},w_1} & \cdots & \cdots & \sigma^2_{\Gamma/s_{p_s},w_{p_w}} & 0 & \cdots & 0 & \sigma^2_{\Gamma/s_{p_s}} & \sigma^2_{\Gamma/s_{p_s},T} \\ \sigma^2_{\Gamma/T,w_1} & \cdots & \cdots & \sigma^2_{\Gamma/T,w_{p_w}} & \sigma^2_{\Gamma/T,s_1} & \cdots & \cdots & \sigma^2_{\Gamma/T,s_{p_s}} & \sigma^2_{\Gamma/T} \end{bmatrix} \tag{A.2}$$

Random Effects covariance structure. The zeros represent the orthogonality constraints arising from principal components.

$$\Sigma_E = \begin{bmatrix} \sigma^2_{E/w_1} & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \sigma^2_{E/w_{p_w}} & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \sigma^2_{E/s_1} & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \sigma^2_{E/s_{p_s}} & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & \sigma^2_{E/T} \end{bmatrix} \tag{A.3}$$

Measurement error / Residual covariance structure; full independence among errors in different components shown.

A.11 Linguistic Covariate Information for Fig. 5.4

SpkrID	SentIdx	RhymeIdx	Tone	PrevTone	NextTone	B2	B3	B4	B5	PrevCons	NextCons	VowelRhyme
F02	530	7	4	2	5	2	4	7	7	dz	d	oŋ
F02	530	8	5	4	1	3	5	8	8	d	NA	ə
F02	530	9	1	5	1	4	6	9	9	NA	sj	iou
M02	106	70	2	2	1	1	3	13	70	n	dj	ien
M02	106	71	1	2	4	2	4	14	71	dj	sp	in
M02	106	72	4	1	4	1	5	15	72	dz‘	d	ɭ

Table A.11: Specific covariate information for the estimated F_0 track in Fig. 5.4.

A.12 Numerical values of random effects correlation matrices for Amplitude & Phase model

$$\hat{P}_{Spkr\_ID} = \begin{bmatrix} 1.00 & 0.00 & 0.00 & 0.00 & -0.29 & -0.09 & 0.05 & 0.08 & -0.15 \\ 0.00 & 1.00 & 0.00 & 0.00 & -0.36 & 0.03 & 0.03 & 0.00 & -0.89 \\ 0.00 & 0.00 & 1.00 & 0.00 & 0.01 & 0.04 & -0.03 & -0.04 & -0.04 \\ 0.00 & 0.00 & 0.00 & 1.00 & -0.03 & -0.01 & 0.00 & 0.01 & 0.00 \\ -0.29 & -0.36 & 0.01 & -0.03 & 1.00 & 0.00 & 0.00 & 0.00 & 0.36 \\ -0.09 & 0.03 & 0.04 & -0.01 & 0.00 & 1.00 & 0.00 & 0.00 & -0.01 \\ 0.05 & 0.03 & -0.03 & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 & -0.04 \\ 0.08 & 0.00 & -0.04 & 0.01 & 0.00 & 0.00 & 0.00 & 1.00 & -0.01 \\ -0.15 & -0.89 & -0.04 & 0.00 & 0.36 & -0.01 & -0.04 & -0.01 & 1.00 \end{bmatrix} \tag{A.4}$$

$$\hat{P}_{Sentence} = \begin{bmatrix} 1.00 & 0.00 & 0.00 & 0.00 & -0.89 & -0.22 & -0.17 & 0.06 & 0.42 \\ 0.00 & 1.00 & 0.00 & 0.00 & 0.41 & -0.55 & -0.09 & 0.10 & 0.57 \\ 0.00 & 0.00 & 1.00 & 0.00 & 0.01 & -0.27 & -0.06 & 0.85 & 0.39 \\ 0.00 & 0.00 & 0.00 & 1.00 & -0.12 & -0.39 & 0.82 & -0.03 & 0.30 \\ -0.89 & 0.41 & 0.01 & -0.12 & 1.00 & 0.00 & 0.00 & 0.00 & -0.12 \\ -0.22 & -0.55 & -0.27 & -0.39 & 0.00 & 1.00 & 0.00 & 0.00 & -0.51 \\ -0.17 & -0.09 & -0.06 & 0.82 & 0.00 & 0.00 & 1.00 & 0.00 & 0.14 \\ 0.06 & 0.10 & 0.85 & -0.03 & 0.00 & 0.00 & 0.00 & 1.00 & 0.42 \\ 0.42 & 0.57 & 0.39 & 0.30 & -0.12 & -0.51 & 0.14 & 0.42 & 1.00 \end{bmatrix} \tag{A.5}$$

A.13 Warping Functions in Original Domain

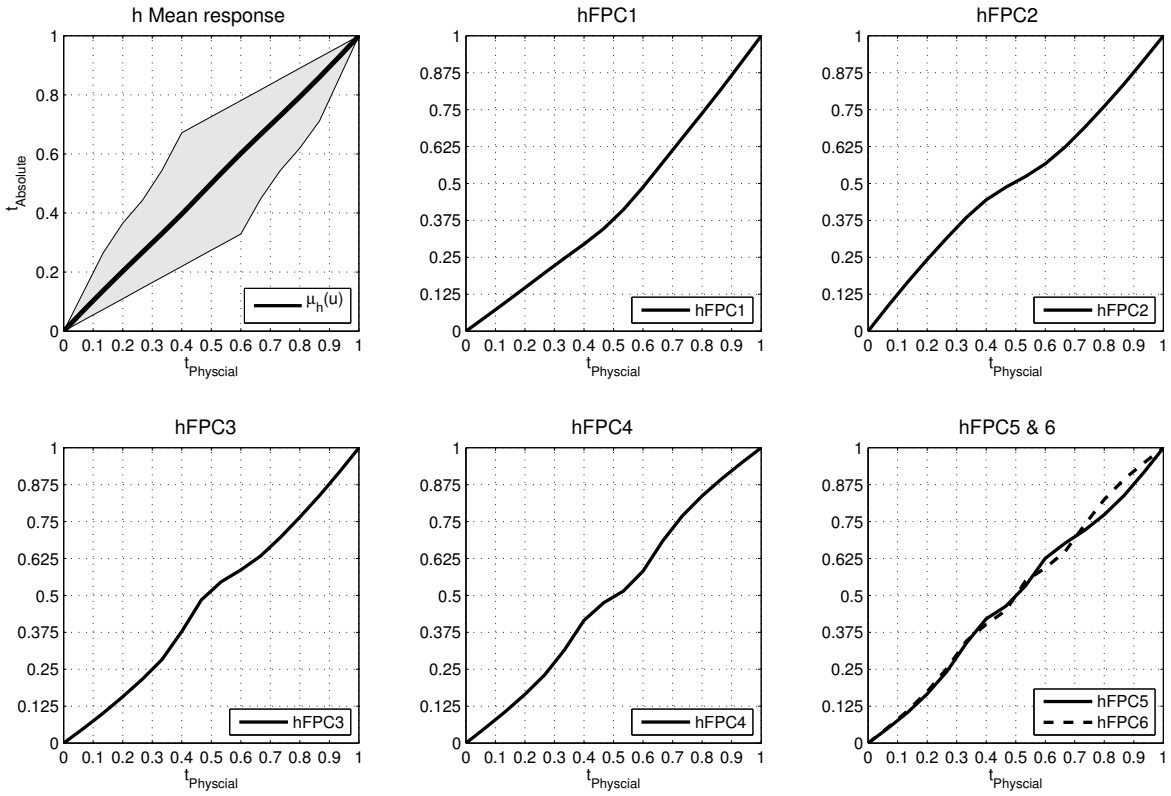

Figure A.7: Modes of variation in the original warping function space due to the components of the transformed domain; produced by applying the inverse transformation on the functional principal components Ψ; gray band around the mean function shows [.05, .95] percentile of sample variation.

A.14 Area Under the Curve - FPCA / MVLME analysis

To verify the generality of the presented framework the core of the analysis in Sect. 5.3 was re-implemented utilizing the Area Under the Curve framework presented in Sect. 3.2.3. The results confirm our assertion that the choice of time-registration framework while crucial does not render the findings from an joint analysis as the one described in the main body of this work, specific to a single framework. The insights offered by the application of FPCA in the new amplitude and phase variation functions H_{AUC} and S_{AUC} (Figures A.8 and A.9 respectively) as well as the insights from the subsequent MVLME analysis of AUC based projections scores (Fig. A.10) communicate very similar insights as the ones offered by PACE in Chapt. 5.

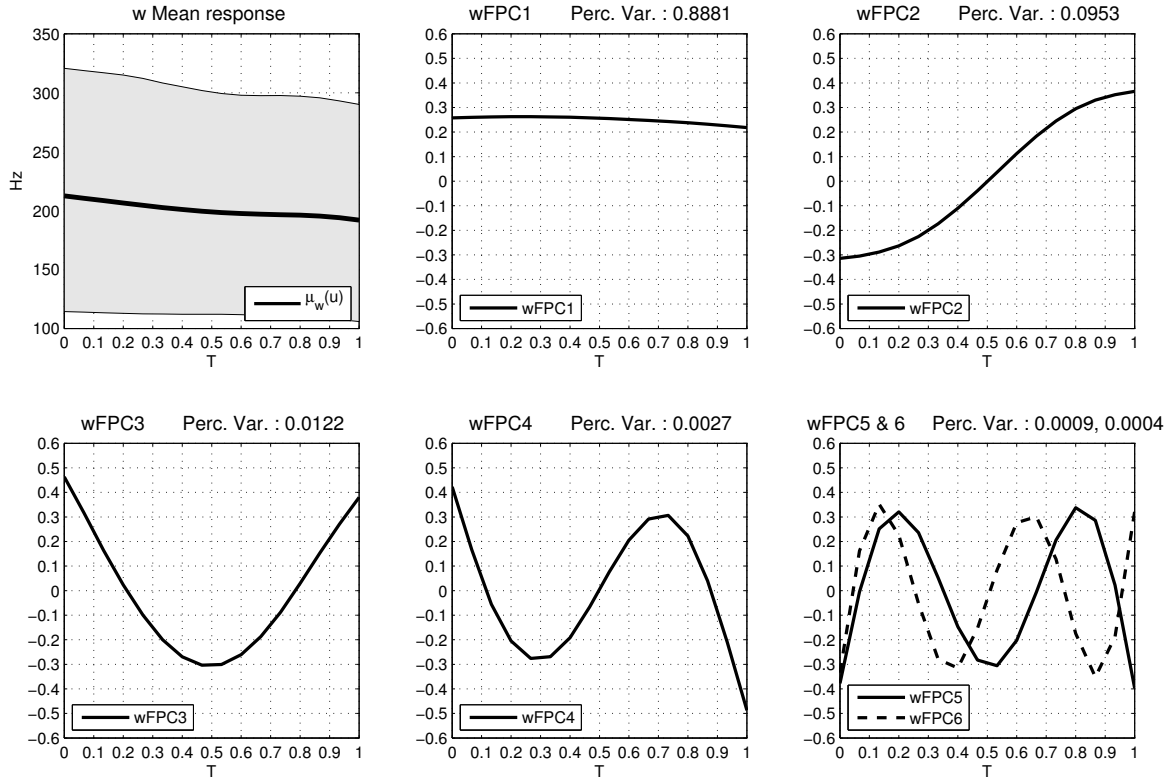

Figure A.8: W_{AUC} (Amplitude) Functional Principal Components Φ_{AUC} computed when using an AUC time-registration framework: Mean function ([.05,.95] percentiles shown in grey) and 1st, 2nd, 3rd, 4th, 5th, and 6th functional principal components of amplitude.

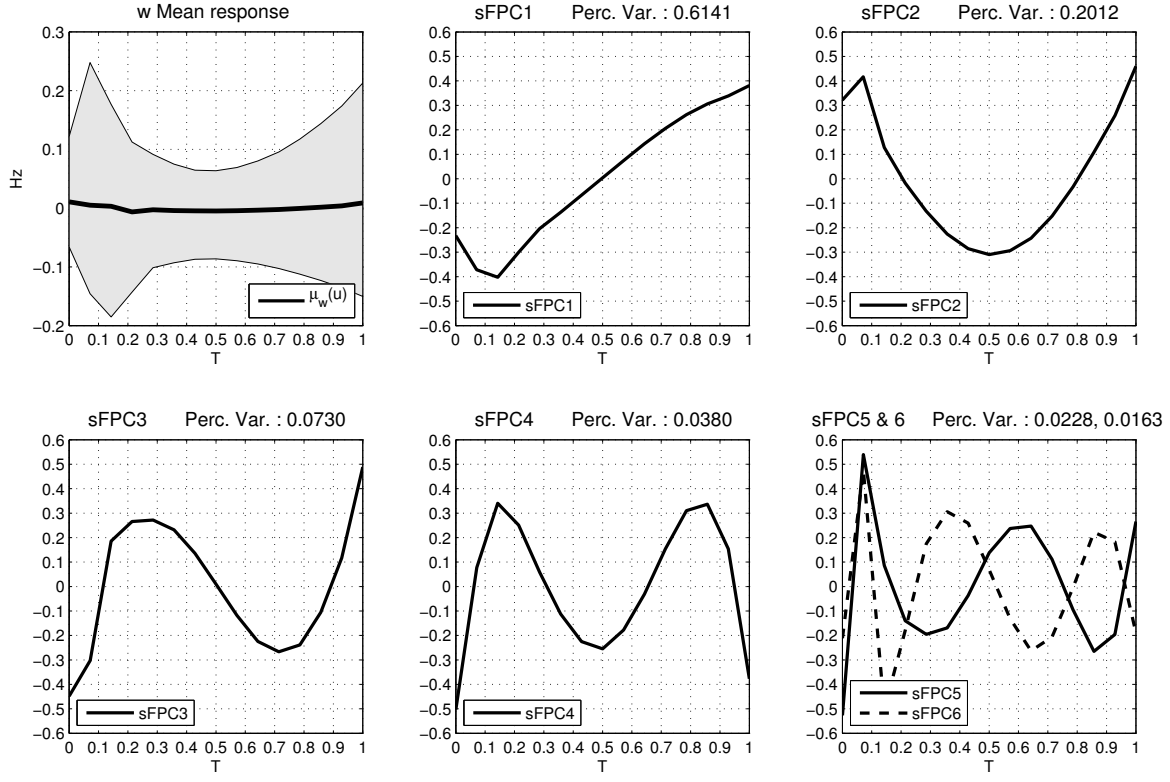

Figure A.9: S_{AUC} (Phase) Functional Principal Components Φ_{AUC} computed when using an AUC time-registration framework: Mean function ([.05,.95] percentiles shown in grey) and 1st, 2nd, 3rd, 4th, 5th, and 6th functional principal components of phase.

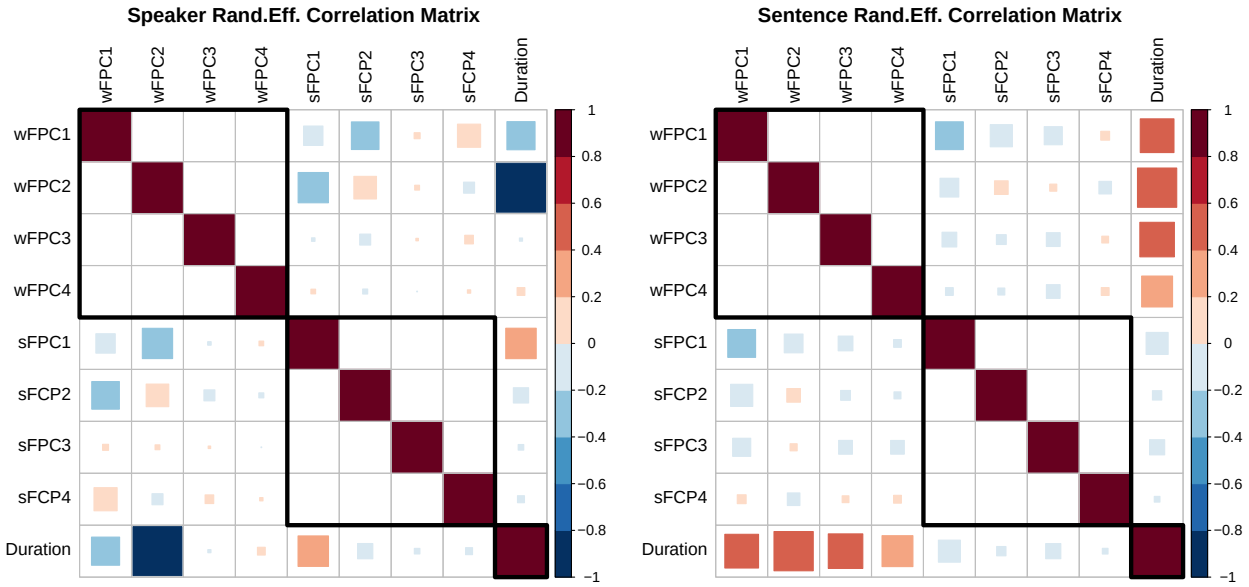

Figure A.10: Random Effects Correlation Matrices using AUC time-registration. The estimated correlation between the variables of the original multivariate model (Eq. 5.23) is calculated by rescaling the variance-covariance submatrices Σ_{R_1} and Σ_{R_2} of Σ_Γ to unit variances. Each cell i, j shows the correlation between the variance of component in row i and that of column j; Row/Columns 1-4 : $wFPC_{1-4}$, Row/Columns 5-8 : $sFPC_{1-4}$, Row/Columns 9 : Duration.

A.15 FPC scores for digit *one*

FPC #	Ital.	Am.Sp.	Ib.Sp.	Port.	Fr.	Std.Dev
1	-1212.400	971.46	-3112.900	1622.70	2685.20	2332.66
2	631.330	310.64	273.530	862.54	-140.20	381.6019
3	342.530	454.02	-142.920	-126.97	-871.65	523.445
4	11.737	368.56	86.611	-407.40	83.43	279.4301

Table A.12: The averaged FPC scores and their sample standard deviation across FPC. Briefly commenting on them: the clear distinction due to the qualitative characteristics of FPC_1 is apparent as the three lower scores are detected in the two-vowel words of Italian and Spanish whereas the higher order FPC_4 appears to almost focus on a distinction between American Spanish and Portuguese, a rather specialized assumption given to the languages in the sample.

A.16 Auxiliary Figures for Chapt. 6

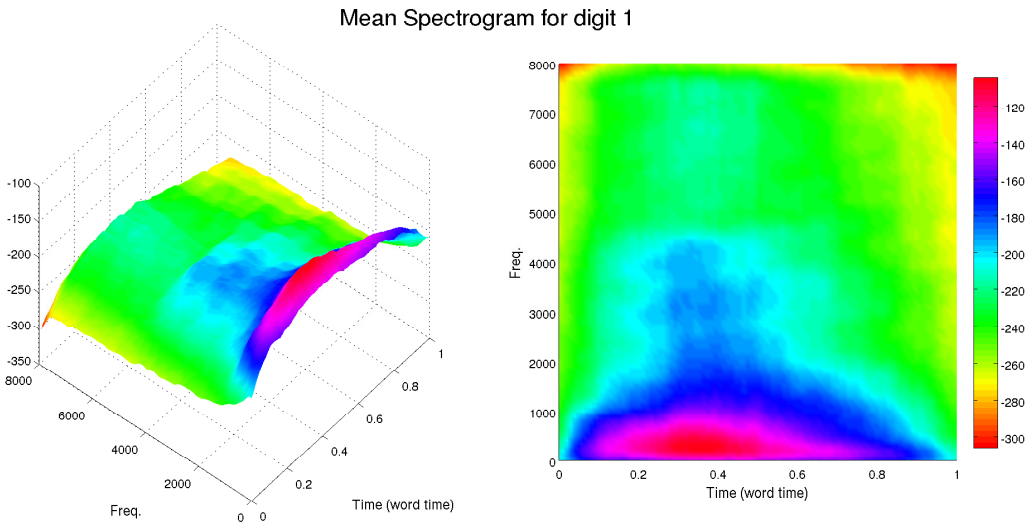

Figure A.11: Mean spectrogram for the instances of digit *one*. It shows a clear variation pattern due to a strong excitation effect in the beginning of a word.

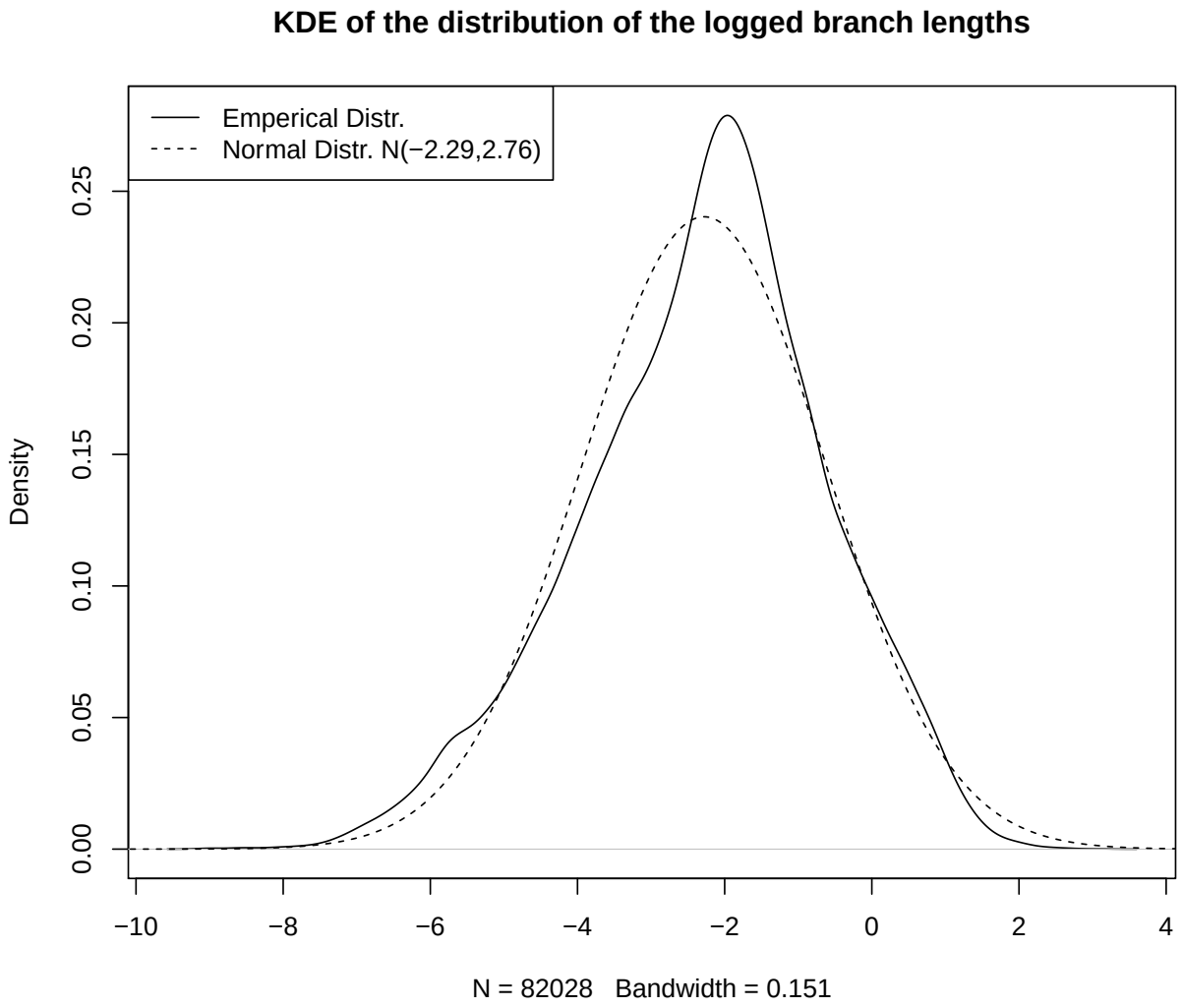

Figure A.12: Empirical distribution of logged branch lengths retrieved from Tree-fam ver.8.0; Skewness = 0.312, Kurtosis = 2.998.

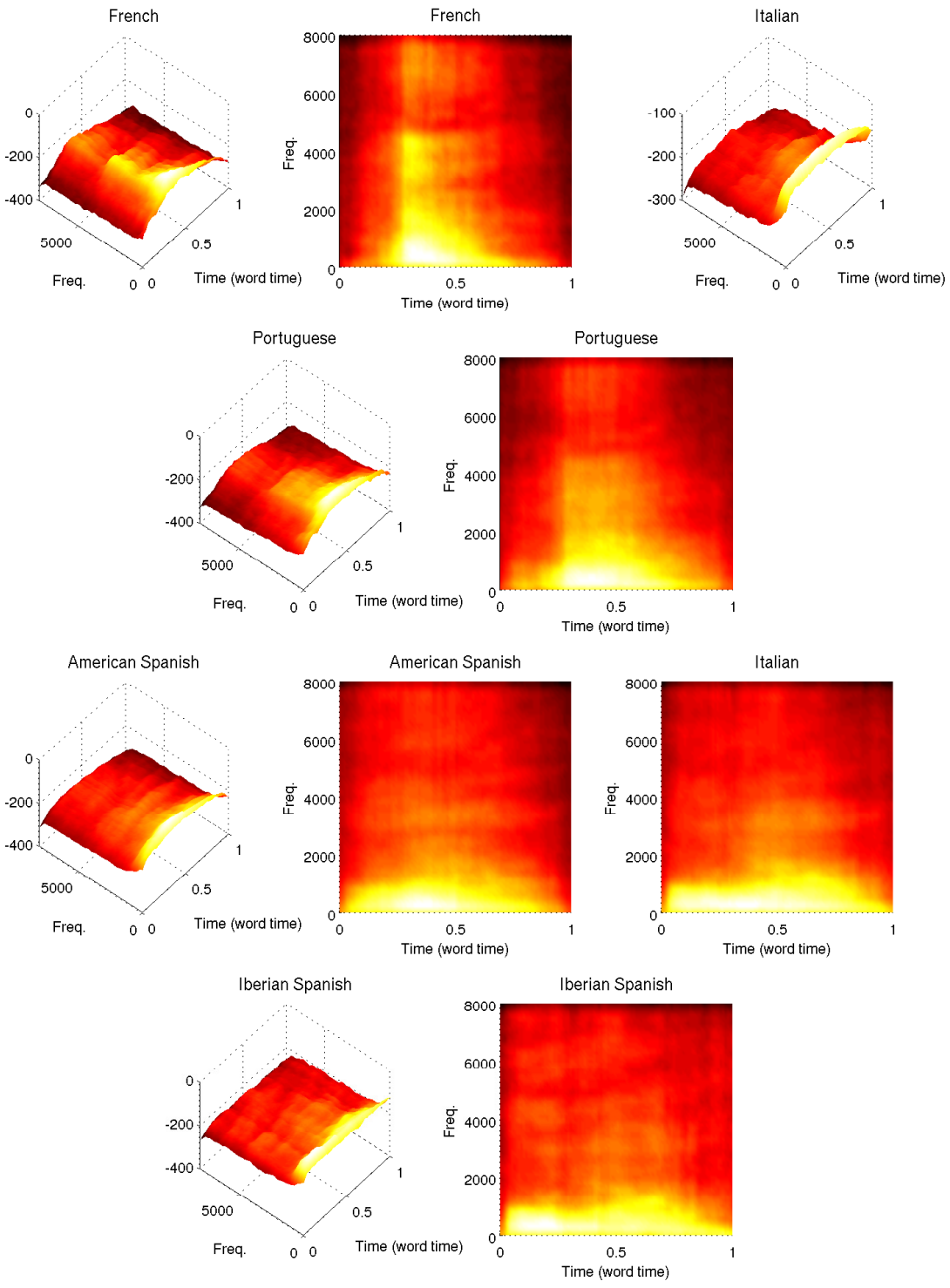

Figure A.13: The language specific spectrograms, as before two different view points are utilized. It is clear that: 1. in French, where the "u" is uttered like "a", more energy is carried on higher frequencies; 2. Italian and the two varieties of Spanish have the most "elongated" spectrograms in terms of power spectral densities as they encapsulate two instead of one vowel; and 3. there is an obvious cut-off in Portuguese, approximately at 5Khz, that is most possibly an artefact of the data rather than a real phenomenon.

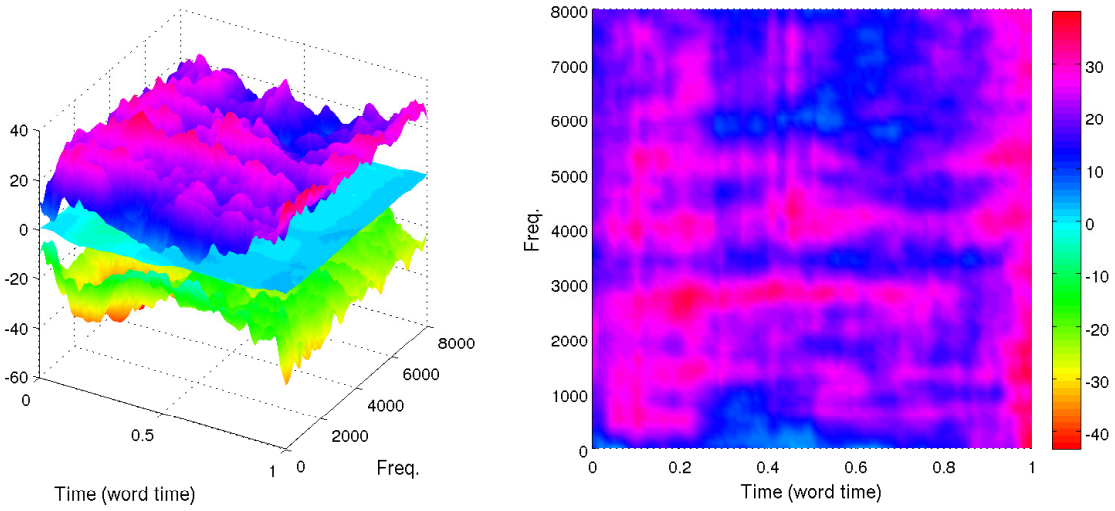

Figure A.14: The protolanguage spectrogram along with its 95% confidence interval (two standard deviations). It is easily seen that the uncertainty regarding our estimate is very strong; in effect the confident intervals give a very wide area over which our estimate may lie. It can be noted that certain frequency bands appear to contain significant variation. It also draws attention to the significant edge effects presented thus emphasizing the important role laboratory conditions play in the quality of the sample; edge effects definitely related (among other things) with background noise and discretisation effects. These confidence intervals are not unexpected though; as seen in Table A.12 the original input coefficients were highly variant and it is only natural for this to propagate along the PGR estimates.