The analysis of Acoustic Phonetic Data: exploring differences in the spoken Romance languages.

Davide Pigoli* [1], Pantelis Z. Hadjipantelis[2], John S. Coleman[3], and John A.D. Aston[4]

[1]Statistical Laboratory, DPMMS, University of Cambridge
[2]Department of Statistics, UC Davis
[3]Phonetics Laboratory, University of Oxford
[4]Statistical Laboratory, DPMMS, University of Cambridge

Abstract

The process of change, particularly understanding the historical and geographical spread, from older to modern languages has long been studied from the point of view of textual changes and phonetic transcriptions. However, it is somewhat more difficult to analyze these from an acoustic point of view, although this is likely to be the dominant method of transmission rather than through written records. Here, we propose a novel approach to the analysis of acoustic phonetic data, where the aim will be to model statistically speech sounds. In particular, we explore phonetic variation and change using a time-frequency representation, namely the log-spectrograms of speech recordings. After preprocessing the data to remove inherent individual differences, we identify time and frequency covariance functions as a feature of the language; in contrast, the mean depends mostly on the particular word that has been uttered. We build models for the mean and covariances (taking into account the restrictions placed on the statistical analysis of such objects) and use this to define a phonetic transformation that allows us to model how an individual speaker would sound in a different language, allowing the exploration of phonetic differences between languages. Finally, we map back these transformations to the domain of sound recordings, allowing us to listen to statistical analysis. The proposed approach is demonstrated using the recordings of the words corresponding to the numbers from "one" to "ten" as pronounced by speakers from five different Romance languages.

1 Introduction

Historical and comparative linguistics is the branch of linguistics which studies languages' evolution and relationships. The idea that languages develop historically by a process

*Address for correspondence: Davide Pigoli, Statistical laboratory, Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, Wilberforce Road, Cambridge, CB3 0WB, United Kingdom. Email: dp497@cam.ac.uk

roughly similar to biological evolution is now generally accepted; see, e.g., Nakhleh et al. (2005). Pagel (2009) claims not only the similarity of genes' and languages' evolutionary behaviour but also offers an extensive catalog of analogies between biological and linguistic evolution.

Interest in language kinships is not by any means restricted to linguistics. For example, the understanding of this evolutionary process is helpful for anthropologists and geneticists, while distances between languages are proxies for cultural differences and communication difficulties and can be used as such in sociology and economic models (Ginsburgh and Weber, 2011). Moreover, the nature of the relationship between languages, and especially the way they are spoken, is a topic of widespread interest for its cultural relevance. We all have our own experience with approaching different languages (and different varieties within each language) and the effort to produce quantitative measurements about speech can shed some light on the subject.

The first step to explore the language ecosystem is to choose how to analyse and measure the differences between languages. A language is indeed a complex entity and its evolution can be be considered from many different points of view. The processes of change from one language to another have been studied for a long time by considering textual and phonetic representation of the words (see, e.g., Morpurgo Davies, 1998, and references therein). This reflects a general normative approach towards languages: for cultural and historical reasons, the way we teach them and the way we think about them are focused on the written expression of the words and their "appropriate" pronunciations. However, this is more a social artifact than a reality of the population, as there is great variation within each language depending on socio-economical and biological attributes, geography and other factors.

The focus of this work is on a more recent development in quantitative linguistics: the study of acoustic phonetic variation, i.e. change on the sounds associated to the pronunciations of words. On one hand, these provide a complementary way to consider the difference between two languages which can be juxtaposed with the differences measured using textual and phonetic representation. On the other hand, it can be claimed that the acoustic expression of the word is the true object of interest, textual and phonetic representation being only the transcription used by linguists of the normative (or more careful) pronunciations of words. However, the use of speech recordings from actual speakers is not yet well established in historical linguistics, due to the complexity of speech as a data object, the theoretical challenges on how to deal with the variability within and between languages and the difficulties (or impossibility) of obtaining sound recordings of ancient pronunciations. A notable exception is the use of speech recordings in the field of language variation and change, a branch of sociolinguistic concerned with small scale variation within communities (for example, between younger and older members). Some of the techniques we describe might also be useful tools to address these kinds of sociolinguistics questions.

Indeed, the analysis of acoustic data highlights one of the fundamental challenges in comparative linguistics, namely that the definition of language itself is an abstraction that simplifies the reality of speech variability and neglects the somehow continuous, albeit with some clear edges, geographical spread of spoken varieties. For example, Grimes and Agard (1959) describe as "useful fiction" the definition of homogeneous speech communities, i.e. groups of speakers those linguistic pattern are alike. Given that for most of human history, most speakers of languages were illiterate, spoken characteristics are also likely to be of profound importance in historical development of languages. The complexity of the data object (speech) and the large amount of variation call for careful consideration from the statistical community and we hope this work will help in arouse attention on this relevant subject.

We use the expression "acoustic phonetic data" to refer to speech recordings associated with the same word (or other linguistic units) when pronounced by a group of speakers. In particular, we are interested in the case where multiple speakers from each language are included in the data set, since this allows one to better explore the phonetic characteristics of the language. This is very different from having only repetitions of a word pronounced by the same speaker, as common for example in speech recognition, and it calls for the development of a novel approach.

The aim of our work is to provide a framework where:

1. speech recordings can be analysed to identify features of the language,
2. the variability of speech within the language can be considered,
3. the acoustic differences between languages can be explored on the basis of speech recordings, taking into account intra-language variability.

Among other things, this will allow us to develop a model to explore how the sound produced by a speaker would be modified when moved towards the phonetic structure of a different language. More specifically we will take into account the variability of pronunciation within each language. This means we explore the variability of the speakers of the language so that we can then understand where a specific speaker is positioned with respect to the general population. This allows us to postulate a path that maps the sound produced by this speaker to that of a hypothetical speaker with the corresponding position in a different language. The idea here is to approximate the same kind of information we can extract when a speaker pronounces words in two different languages in which they are proficient even if we have only monolingual speakers. It is easy to understand that the observation (audio recordings) of many speakers from each group is essential to understand the intra-language variability and thus the relevance of the inter-language acoustic change. This idea has an immediate application in speech synthesis with the possibility to translate a recording from one language to another, with the translation preserving the speaker's voice characteristics. In the future, this approach could be also extended to modify synthesized speech in such a way that it sounds like the voice of a specific speaker (for example a known actress or a public person). This would be really interesting for many commercial applications, from computer gaming to advertising and it is only one example of the methods that can be developed in the general framework we provide.

The paper is structured as follows. Section 2 describes the acoustic phonetic data that are used to demonstrate our methods. In Section 3 a short introduction on the functional data analysis approach to surface data is given, because we choose to represent the speech recordings in a time-frequency domain using a local Fourier transform. The details of these representations, as well as the preprocessing steps needed to remove noise artifacts and time misalignment between the speech recordings are described in Section 4. Section 5 illustrates how to estimate some crucial functional parameters of the population of log-spectrograms and claims that the covariance structures are common across all the words in each language. Section 6 is devoted to the definition and exploration of cross-linguistic phonetic differences. The final section gives a discussion of the advantage of the proposed method and of how it is possible to extend it to even more complex situations, where the phonetic features depend continuously on historical or geographical variables.

2 The Romance digit data set

The methods in this paper will be illustrated with an application to a Romance digit data set of audio recordings. This data set has been compiled in the Phonetics Laboratory of the University of Oxford between 2012-2013. It consists of natural speech recordings of five languages; French, Italian, Portuguese, American Spanish and Castilian Spanish, the two varieties of Spanish being considered different languages for the purpose of the analysis. The speakers utter the numbers one to ten in their native language. The data set is inherently unbalanced; we have seven French speakers, five Italian speakers, five American Spanish speakers, five Iberian Spanish speakers and three Portuguese speakers, 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. As this data set consists of recordings made under non-laboratory settings, large variabilities may be expected within each group. This provides a real-world setting for our analysis, and allows us to build models which characterise realistic variation in speech recording, somewhat of a prerequisite for using this model in practice. The data set is also heterogeneous in terms of sampling rate, duration and even format. As such, before any phonetic or statistical analysis took place, all data were converted in *.wav files of 16 kHz. We indicate each sound recording as $x_{ik}^{L}(t)$, where L is the language, $i = 1, \ldots, 10$ the pronounced word and $k = 1, \ldots, n_L$ the speaker, n_L being the number of speakers available for language L and t the time. This data set has been collected within the scope of Ancient Sounds, an innovative research project with the aim of regenerating audible spoken forms of the (now extinct) earlier versions of Indo-European words, using contemporary audio recordings from multiple languages. More information about this project can be found on the website `http://www.phon.ox.ac.uk/ancient_sounds`.

Although the cross-linguistic comparison of spoken digits is interesting in its own right, this subset of words can also be considered as a representative of a language's vocabulary from a phonetic point of view, meaning that the words used for the numbers in the Romance languages were not chosen to possess any specific phonetic structure.

Consequently, we use the word "language" as a shorthand for the particular samples: the small samples of digit recordings. However, we view this analysis as a proof of concept, and will not focus on the problem of the representativeness of the sample of speakers or words. In view of a broad possible application of the approach which will be outlined, more structured choices of representative words could be taken or specific dialect choices made, but the approach would remain the same.

3 The analysis of surface data

Different representations are available in phonetics to deal with speech recordings. Many of them share the idea of representing the sound with the distribution of intensities over frequency and time. We choose in particular the power spectral density of the Local Fourier Transform, as detailed in Section 4. The output of this representation is a two-dimensional surface that describes the sound intensity for each time and for each frequency. Since we can represent each spoken word as a two dimensional smooth surface, it comes naturally to apply a functional data analysis approach. Good results have already been obtained applying functional data analysis techniques to acoustic analysis, although in the different context of a single language study, for example in Koenig et al. (2008) and Hadjipantelis et al. (2012). Functional data analysis is appropriate in this context because it addresses problems where data are observations from continuous underlying processes, such as functions, curves or surfaces. A general introduction to the analysis of functional data can be found in Ramsay and Silverman (2005) and in Ferraty and Vieu (2006). The central idea is that taking into account the smooth structure of the process helps in dealing with the high dimensionality of the data objects.

We focus here on the case where data are two dimensional surfaces on a bounded domain as in the case of acoustic phonetic data. Let X be a random surface so that $X : \Omega \to L^2([0, F] \times [0, T])$ and $E[||X^2||^2] < +\infty$. A mean surface can then be defined as $\mu(\omega, t) = E[X(\omega, t)]$ and the four dimensional covariance function as $c(\omega, \omega', t, t') = \text{cov}[X(\omega, t), X(\omega', t')]$.

In practice these surfaces are observed over a finite number of grid points and they are affected by noise. As noted by Ramsay and Silverman (2005), "the term *functional* in reference to observed data refers to the intrinsic structure of the data rather than to their explicit form". Thus a smoothing step is needed to recover the regular surfaces that reflect the properties of the underlying process. These surfaces are represented by means of a linear combination of basis functions which span the separable Hilbert space $L^2([0, F] \times [0, T])$. In particular, we choose the widely popular method of smoothing splines to estimate a smooth surface $\widetilde{X}(\omega, t)$ from the noisy observation on a regular grid $\mathfrak{X}(\omega_i, t_j),\ i = 1, \ldots, n_\omega,\ j = 1, \ldots, n_t$.

When analysing a sample of surfaces, we are implicitly assuming that the comparison of their values at the same coordinates (ω, t) is meaningful. However, this is often not the case when data are measurement of a continuous process such as human speech. For example, different speakers (or even the same speaker in different replicates) can speak faster or slower without this changing the acoustic information in the recordings.

The resulting sound objects are obviously not comparable though, unless this problem is addressed first. This situation is so common in functional data analysis that much work as been devoted to its solution and these techniques are referred to as functional registration (or warping or alignment). In the case of a two dimensional surface, the misalignment can in principle affect both coordinates, this is the case for example in image processing. A two dimensional warping function $h(\omega, t)$ is then needed to align each surface and this is a more complex problem than one-dimensional registration. However, even though we are considering data that are surfaces, the way they are produced, which will be detailed in Section 4, makes it sensible to adjust only for the misalignment on the temporal axis, this being due to different speech speeds, which are not relevant for our goals. On the contrary, we want to preserve the differences on the frequency axis which contain pieces of information about the phonetic characteristics of the speakers.

Thus, we apply a mono-dimensional warping to our surface data. If we aim to align a sample of surfaces $\widetilde{X}_1, \ldots, \widetilde{X}_N$, we look for a set of time-warping function $h_1(t), \ldots, h_N(t)$ so that the aligned surface will be defined as $X_1 = \widetilde{X}_1(\omega, h(t)), \ldots, X_N = \widetilde{X}_N(\omega, h(t))$. In the next section we will describe how to achieve this in practice for acoustic phonetic data.

Given the smooth and aligned surfaces $X_1, \ldots, X_N$, it is possible to estimate the functional parameters of the underlying process, for example

$$\widehat{\mu}(\omega, t) = \frac{1}{N}\sum_{i=1}^{N} X_i, \quad \widehat{c}(\omega, \omega', t, t') = \frac{1}{N-1}\sum_{i=1}^{N}(X_i(\omega, t) - \widehat{\mu}(\omega, t))(X_i(\omega', t') - \widehat{\mu}(\omega', t')).$$

However, the high-dimensionality of the problem makes the estimate for the covariance structure inaccurate or even computationally unfeasible. In Section 5 we introduce some modelling assumptions to make the estimation problem tractable.

4 From speech records to smooth spectrogram surfaces

As mentioned in the previous section, we choose to represent the sound signal via the power spectral density of the local Fourier transform. This means we first apply a local Fourier transform to obtain a two dimensional spectrogram which is function of time (the time instant where we centre the window for the local Fourier transform) and frequency. For the Oxford Romance Language data, we use a gaussian window function w with a window length of 10 milliseconds, defined as $w(\tau) = \exp(-\frac{1}{2}(\frac{\tau}{0.005})^2)$. Since the original acoustic data set was sampled at 16kHz, this result into a window size of 160 samples per frame and the maximal effective frequency detected is 8kHz, the Nyquist frequency of our sampling procedure.

We can compute the local Fourier transform as

$$X_{ik}^{L}(\omega, t) = \int_{-\infty}^{+\infty} x_{ik}^{L}(\tau) w(\tau - t) e^{-j\omega\tau} \mathrm{d}\tau.$$

The power spectral density, or spectrogram, is defined as the magnitude of the Fourier transform and the log-spectrogram (in decibel) is therefore

$$\mathfrak{S}_{ik}^{L}(\omega, t) = 10 \log_{10}(|X_{ik}^{L}(\omega, t)|^2).$$

Fig. 1 shows an example of a raw speech signals (top panel) and the corresponding log-spectrogram (bottom left panel), for the sound produced by a French speaker pronouncing the word *un*.

To deal with these objects as functional, we need to address the problems of smoothing and registration described in the previous section. Indeed, when data comes from real word recordings, as opposed to laboratory conditions, the raw log-spectrograms suffer from noise corruption. For this reason we apply a penalized least square filtering for grid data using discretized smoothing splines. In particular, we use the automated robust algorithm described in Garcia (2010), based on the discrete cosine transform, which allows for a fast computation in high dimensions when the grid is equally spaced.

The second preprocessing step consists of registration. This is due to the fact that speakers can speak faster or slower and this is particularly true when data are collected from different sources where the context is different. However, this difference in the speech speed is not relevant from a linguistic point of view and thus the alignment along the time axis is needed because of this phase distortion in the acoustic signals. First, we standardized the time scale so that each signal goes from 0 to 1. Then, we adapt to the case of surface data the procedure proposed in Tang and Müller (2008) to remove time misalignment from functional observation. Given a sample of functional data $f_1, \ldots, f_n \in L^2([0,1])$, this procedure look for a set of strictly monotone time warping function $h_1, \ldots, h_n$ so that $h_i(0) = 0$, $h_i(1) = 1$, $i = 1, \ldots, n$. In practice, these warping functions are modelled via piecewise linear functions and estimated by minimizing the pairwise difference between the observed curve while penalizing their departure from the identity warping $h(t) = t$. Hence, a pairwise warping function is first obtained as

$$h_{ij}(t) = \arg\min_{h} \int_0^1 (f_i(h(t)) - f_j(t))^2 + \lambda \int_0^1 (h(t) - t)^2,$$

where the minimum is computed over all the piecewise linear function on a chosen grid. Let now h_k, $k = 1, \ldots, n$, be the warping function from an individual specific time to the standardized time scale. Then, if $s = h_j^{-1}(t)0$, $h_i(s) = h_i(h_j^{-1}(t)) = h_{ij}(t)$. Under the assumption of the warping function to have average identity and thus $E[h_{ij}|h_j] = h_j^{-1}$, the estimator proposed by Tang and Müller (2008) is

$$h_j^{-1}(t) = \frac{1}{n} \sum_{j=1}^{n} h_{ij}(t).$$

When we want to apply this idea to the case of acoustic phonetic data, we need first to define the groups of log-spectrograms we want to align together. The idea being that the mean log-spectrogram is different from word to word, we decide to align the log-spectrograms corresponding to the same word. Then, we have to extend the procedure to two dimensional objects such as surfaces. As mentioned in the previous section, it is safe to assume that no phase distortion is present in the frequency direction, given the relatively small window used in the local Fourier transform. On the contrary, time misalignment can be a serious issue due to differences in speech rate across speakers. Therefore we modify the procedure in Tang and Müller (2008) so that we look for pairwise time warping function but minimizing the discrepancy between surfaces. For each word $i = 1$ in a group of log-spectrogram we want to align, we define the discrepancy between the log-spectrogram $\widetilde{\mathfrak{S}}_{ik}^L$ and $\widetilde{\mathfrak{S}}_{im}^L$ as

$$D_\lambda(\widetilde{\mathfrak{S}}_{ik}^L, \widetilde{\mathfrak{S}}_{im}^{L'}, g_{km}^{LL'}) = \int_{f=0}^{+\infty} \int_{t=0}^{1} (\widetilde{\mathfrak{S}}_{ik}^L(\omega, g_{km}^{LL'}(t)) - \widetilde{\mathfrak{S}}_{im}^{L'}(\omega, t))^2 + \lambda(g_{km}^{LL'}(t) - t)^2 dt d\omega, \quad (1)$$

where λ is an empirically evaluated non-negative regularization constant and $g_{km}^{LL'}(\cdot)$ is the pairwise warping function mapping the time evolution of $\widetilde{\mathfrak{S}}_{ik}^L(\omega, t)$ to that of $\widetilde{\mathfrak{S}}_{im}^{L'}(\omega, t)$. We obtain the pairwise warping function $\widehat{g}_{km}^{LL'}(\cdot)$ by minimizing the discrepancy $D_\lambda(\widetilde{\mathfrak{S}}_{ik}^L, \widetilde{\mathfrak{S}}_{im}^{L'}, g_{km}^{LL'})$ under the constraint that $g_{km}^{LL'}$ is piecewise-linear, monotonic and so that $g_{km}^{LL'}(0) = 0$ and $g_{km}^{LL'}(1) = 1$. Finally, the inverse of the global warping function for each pronounced word can be estimated as the average of the pairwise warping functions:

$$h_{ik}^{-1} = \frac{1}{\sum_{L'=1}^{5} n_L} \sum_{L'=1}^{5} \sum_{m=1}^{n_L} \widehat{g}_{km}^{LL'},$$

and the smoothed and aligned log-spectrogram for the language L, word i and speaker k is therefore $S_{ik}^L(\omega, t) = \widetilde{\mathfrak{S}}_{ik}^L(\omega, h_{ik}(t))$. In practice, warping functions are represented with a spline basis defined over a regular grid of 100 points on $[0, 1]$ and we look for the spline coefficients that minimize the discrepancies. The quantities in (1) are approximated by their discretized equivalent on a two-dimensional grid with 100 equispaced grid points on the time dimension and 81 equispaced grid points in the frequency dimension.

After this second preprocessing step, we are presented with 219 smoothed and aligned log-spectrograms. For example, the smoothed and time-aligned log-spectrogram from the sound produced by a French speaker pronouncing the word *un* can be found in the bottom right panel of Fig. 1.

5 Estimation of means and covariance operators

The process that generates the sounds (and thus their representation as log-spectrograms) is governed by unknown parameters that depend on the language, the word being pronounced and the speaker. However, we need to make some assumption to identify and estimate these parameters. We consider the mean of the random log-spectrograms as depending on the word, in each language, being pronounced. Indeed, the mean spectrogram is in general different for the different words, as would be expected. Let $i = 1, \ldots, 10$ be the pronounced words and $k = 1, \ldots, n_L$ the speakers for the language L. The smoothed and aligned log-spectrograms $S_{ik}^L(\omega, t)$ allow the estimation of the mean log-spectrogram $\overline{S}_i^L(\omega, t) = (1/n_L) \sum_{k=1}^{n_L} S_{ik}^L(\omega, t)$ for each word i of the language L.

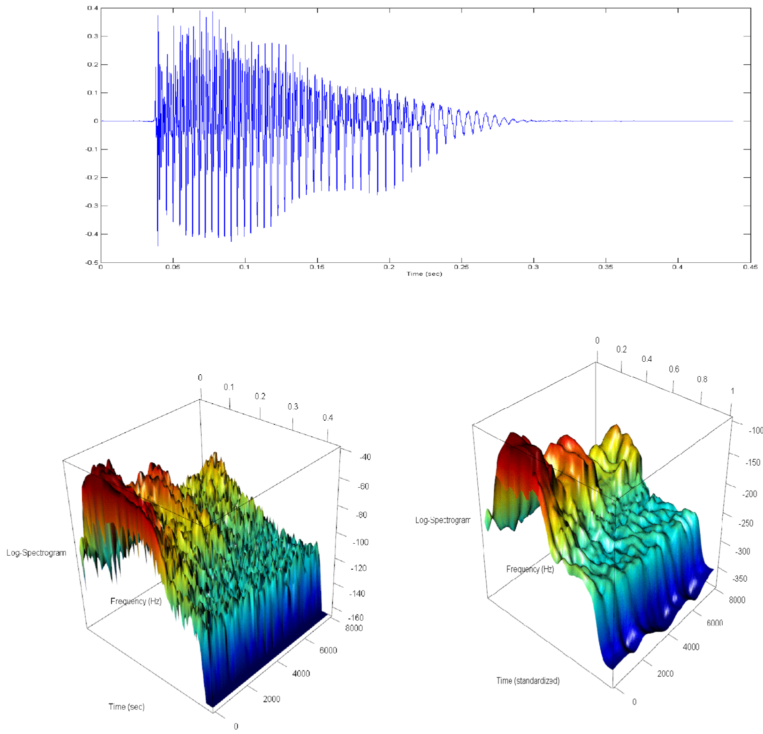

Figure 1: Raw record (top), raw log-spectrogram (bottom left) and smoothed and aligned log-spectrogram (bottom right) for a French speaker pronouncing the word "un" ("one").

Recent studies (Aston et al., 2010; Pigoli et al., 2014) show that significant linguistic features can be found in the covariance structure between the intensities at different frequencies. This can be considered as a summary of what a language "sounds like", without incorporating the differences at the word level. Thus, we first assume in our analysis that the covariance structure of the log-spectrograms is common for all the words in the language and we are going to estimate it using the residual surface obtained by removing the word mean effect. In Section 5.1 we develop a procedure to verify this assumption in the Oxford Romance Language data set.

Starting from the smoothed and aligned log-spectrograms $S_{ik}^L(\omega, t)$ of the records of the number $i = 1, \ldots, 10$ for the speakers $k = 1, \ldots, n_i$, we thus focus on the residual log-spectrograms $R_{ik}^L(\omega, t) = S_{ik}^L(\omega, t) - (1/n_i)\sum_{k=1}^{n_i} S_{ik}^L(\omega, t)$, which measure how each token differs from the word mean. In the following, we disregard in the notation the different speakers and words that originated the residual log-spectrogram and indicate with $R_j^L(\omega, t), j = 1, \ldots, n_L$ the set of observations for the language L including all speakers and words.

However, using standard covariance estimation techniques to find the full four-dimensional covariance structure is not computationally or statistically (in terms of sample size) feasible, thus we need some modeling assumptions. There are many ways to incorporate assumptions which allow such estimation, a common one being some form of sparsity. Rather than the usual sparsity ideas based on some elements being identically zero, we prefer to work on the principle that the covariance can be factorised.

In particular, we assume that the covariance structure $c^L(\omega_1, \omega_2, t_1, t_2) = cov(S^L(\omega_1, t_1), S^L(\omega_2, t_2))$ is separable in time and frequency, i.e. $c^L(\omega_1, \omega_2, t_1, t_2) = c^L_\omega(\omega_1, \omega_2) c^L_t(t_1, t_2)$. While we do not necessary believe this assumption to be true in general, a structure is needed to obtain reliable estimates for the covariance operators, and is a reasonable assumption that is frequently (implicitly) used in signal processing, particularly when constructing higher dimensional bases from lower dimensional ones.

Possible estimates for $c^L_\omega(\omega_1, \omega_2)$ and $c^L_t(t_1, t_2)$ are

$$\widehat{c}^L_r = \frac{\widetilde{c}^L_r}{\sqrt{\mathrm{trace}(\widetilde{c}^L_r)}},\ r = \omega, t, \tag{2}$$

where trace indicates the trace of the covariance function, defined as $\mathrm{trace}(c) = \int c(s, s)\mathrm{d}s$, while $\widetilde{c}^L_r$, $r = \omega, t$ are the sample marginal covariance functions

$$\widetilde{c}^L_\omega(\omega_1, \omega_2) = \frac{1}{n_L - 1} \sum_{j=1}^{n_L} \int_0^1 (R^L_j(\omega_1, t) - \overline{R}^L_{n_L}(\omega_1, t))(R^L_j(\omega_2, t) - \overline{R}^L_{n_L}(\omega_2, t))\mathrm{d}t,$$

and

$$\widetilde{c}^L_t(t_1, t_2) = \frac{1}{n_L - 1} \sum_{j=1}^{n_L} \int_0^{8\mathrm{kHz}} (R^L_j(\omega, t_1) - \overline{R}^L_{n_L}(\omega, t_1))(R^L_j(\omega, t_2) - \overline{R}^L_{n_L}(\omega, t_2))\mathrm{d}\omega,$$

$\overline{R}^L_{n_L}$ being the sample mean of the residual log-spectrogram for the language L. We introduce also the associated covariance operators as

$$\widehat{C}^L_r g(x) = \int_0^M \widehat{c}^L_r(x, x') g(x')\mathrm{d}x' \quad g \in L^2(\mathbb{R}),\ r = \omega, t,\ M = 8\mathrm{kHz},\ 0.$$

It is easy to see why we choose (2) to estimate the two separable covariance functions. Let $\widetilde{c}^L_\omega$ and $\widetilde{c}^L_t$ be the true marginal covariance functions, i.e.

$$\widetilde{c}^L_\omega(\omega_1, \omega_2) = \int_0^1 c^L(\omega_1, \omega_2, t, t)\mathrm{d}t, \quad \widetilde{c}^L_\omega(\omega_1, \omega_2) = \int_0^{8\mathrm{kHz}} c^L(\omega, \omega, t_1, t_2)\mathrm{d}\omega.$$

Then, if the full covariance function is indeed separable, their product can be rewritten as

$$\widetilde{c}^L_\omega(\omega_1, \omega_2)\widetilde{c}^L_\omega(\omega_1, \omega_2) = \int_0^1 c^L_\omega(\omega_1, \omega_2) c_t(t, t)\mathrm{d}t \int_0^{8\mathrm{kHz}} c^L_\omega(\omega, \omega) c_t(t_1, t_2)\mathrm{d}\omega =$$
$$= c^L_\omega(\omega_1, \omega_2)\mathrm{trace}(c_t) c_t(t_1, t_2)\mathrm{trace}(c_\omega).$$

Moreover, $\mathrm{trace}(\widetilde{c}_\omega^L) = \mathrm{trace}(c_\omega \mathrm{trace}(c_t)) = \mathrm{trace}(c_\omega)\mathrm{trace}(c_t)$ and the same is true for $\widetilde{c}_t$. Hence,

$$\frac{\widetilde{c}_\omega^L(\omega_1, \omega_2)}{\sqrt{\mathrm{trace}(\widetilde{c}_\omega^L)}} \frac{\widetilde{c}_t^L(t_1, t_2)}{\sqrt{\mathrm{trace}(\widetilde{c}_t^L)}} = c_\omega^L(\omega_1, \omega_2) c_t^L(t_1, t_2) = c(\omega_1, \omega_2, t_1, t_2)$$

and this suggests $\widetilde{c}_r^L$ as estimator for c_r^L, $r = \omega, t$.

Figures 2 and 3 show the estimated marginal covariance functions for the five Romance languages. As can be seen, the frequency covariance functions presents differences that appears to be language-specific, while the time covariances have similar structure, the dependence decreasing when time lag increases.

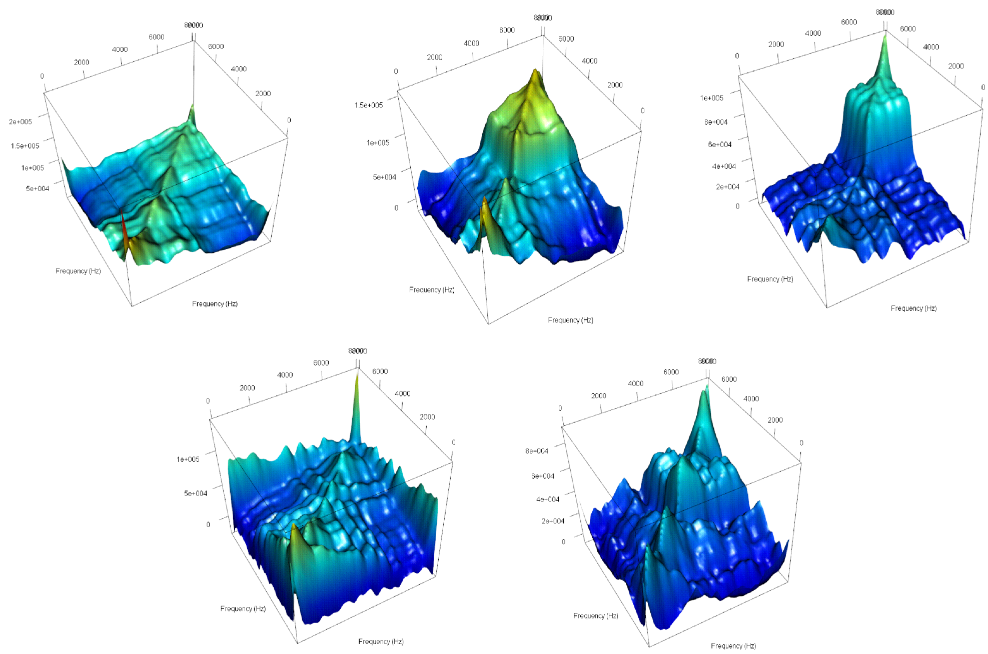

Figure 2: Marginal covariance function between frequencies for the five Romance languages. First row: Italian (left), French (center) and Portuguese (right). Second row: American Spanish (left) and Iberian Spanish (right).

5.1 A permutation test to compare means and covariance operators between groups

We made above the assumption that the covariance operators are common to all the words within each language, while the means are different between words. This assumption can be verified using permutation tests that look at the effect of the group factor on the parameters of the sound process.

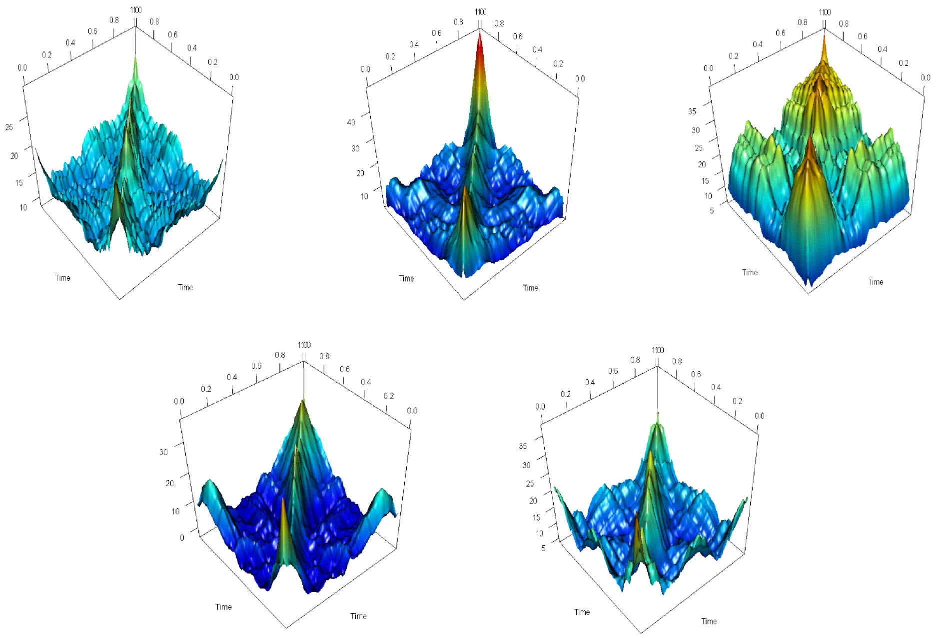

Figure 3: Marginal covariance function between times for the five Romance languages. First row: Italian (left), French (center) and Portuguese (right). Second row: American Spanish (left) and Iberian Spanish (right).

When an estimator for a parameter is available and it is possible to define a distance $d(.,.)$ between two estimates, a distance-based permutation test can be set up in the following way. Let $X_{1l}, \ldots, X_{nl}$ be a sample of surfaces from the l-th group under consideration and $K_l = K(X_{1l}, \ldots, X_{nl})$ be an estimator for an unknown parameter Γ_l of the process which generates the data belonging to the l-th group. In the case of acoustic phonetic data, this parameter can be for example the mean, the frequency covariance operator or the time covariance operator.

Permutation tests are non parametric tests which rely on the fact that, if there is no difference among experimental groups, the group labelling of the observations (in our case the log-spectrograms) is completely arbitrary. Therefore, the null hypothesis that the labels are arbitrary is tested by comparing the test statistic with its permutation distribution, i.e. the value of the test statistics for all the possible permutation of labels. In practice, only a subset of permutations, chosen at random, is used to assess the distribution. A sufficient condition to apply this permutation procedure is exchangeability under the null hypothesis. This is trivially verified in the case of the test for the mean. For the comparison of covariance operators, this means the groups having the same mean. If this not true, we can apply the procedure to the centred observations $\widetilde{X}_{il} = X_{il} - \overline{X}_l$, $i = 1, \ldots, n$, $l = 1, \ldots, G$, where $\overline{X}_l$ is the sample mean for the l-th group. This guarantees the observations to be asymptotically exchangeable due to the law of large numbers.

Indeed, if we want to test the null hypothesis that $\Gamma_1 = \Gamma_2 = \cdots = \Gamma_G$ against the alternative that the parameter is different for at least one group, we can consider as test statistic

$$T_0 = \frac{1}{G}\sum_{l=1}^{G} d(K_l, \overline{K})^2,$$

where $\overline{K}$ is the sample Fréchet mean of $K_1, \ldots, K_G$, defined as

$$\overline{K} = \arg\min_{K \in P} \frac{1}{G}\sum_{l=1}^{G} d(K_l, K)^2,$$

where P is the appropriate functional space where the parameters belong. This test statistic measures the variability of the estimator of the parameters across the different groups. If the parameter is indeed different for some groups, we expect that the their estimates from groups $1, \ldots, G$ show greater variability than those obtained from random permutations of the group labels in the data set. Thus, large values of T_0 are evidence against the null hypothesis.

Let us take M permutations of the original group labels and compute the test statistic for the permuted sample $T_m = \sum_{l=1}^{G} d(K_l^m, \overline{K}^m)^2$, where K_l^m, $L = 1, \ldots, G$ are the estimates of the parameters obtained from the observations assigned to the group l in the m-th permutation and $\overline{K}^m$ is their sample Fréchet mean. The p-value of the test will be therefore the proportion of permutations for those the test statistics is greater than in the original data set, i.e. $p = \frac{\#\{T_m > T_0\}}{M}$.

We apply now this general procedure to the three parameters of interest in our case, i.e. the mean, the frequency covariance operator and the time covariance operator, when the groups are the different words within each language and/or the different language.

Let us start considering the test to compare the means of the log-spectrograms across the words (digit) of each language. Here the natural estimator for the word-wise mean log-spectrogram is the sample mean, i.e.

$$K_l = \overline{S}_l^L(\omega, t) = \frac{1}{n_l}\sum_{k=1}^{n_l} S_{lk}^L(\omega, t)$$

and the distance can be chosen to be the distance in $L^2([0, 8\text{kHz}] \times [0, 1])$,

$$d(\overline{S}_l^L, \overline{S}_{l'}^L) = \int_0^{8\text{kHz}} \int_0^1 [\overline{S}_l^L(\omega, t) - \overline{S}_{l'}^L(\omega, t)]^2 \mathrm{d}\omega \mathrm{d}t.$$

Table 1 reports the results for the test for the difference of the means between the digit $l = 1, \ldots, 10$ for the five Romance language. It can be seen that a significant difference can be found for most of the considered language and thus we choose to account for this difference when modelling the sound changes.

We can apply the same procedure to the test for the covariance operators. First, we need to define a distance between covariance operators. Pigoli et al. (2014) show

Table 1: P-values of the permutation tests for H0: $\mu_1^L = \mu_2^L = \cdots = \mu_{10}^L$ *vs* H1: at least one is different, where μ_i^L is the mean log-spectrogram for the language L and word i, for the five Romance languages.

Language	French	Italian	Portuguese	American Spanish	Iberian Spanish
$p - value$	<0.001	0.02	0.96	<0.001	0.205

that when the covariance operator is the object of interest for the statistical analysis, a distance-based approach can be fruitfully used and the choice of the distance is relevant, different distances catching different properties of the covariance structure.

In particular, they propose a distance based on the geometrical properties of the space of covariance operators, the *Procrustes reflection size-and-shape distance*. This distance uses a map from the space of covariance operators to the space of Hilbert-Schmidt operators, i.e. compact operator with finite norm $||L||_{HS} = \text{trace}(L_i^* L_i)$. This being a Hilbert space, distances between the transformed operators can be easily evaluated. However, the map is defined up to an unitary operator and a Procrustes matching is therefore needed to evaluate the distance between the two equivalence classes. Let C_1 and C_2 the covariance operators we want to compare and L_1 and L_2 the Hilbert-Schmidt operators such that $C_i = L_i L_i^*$. Pigoli et al. (2014) prove that the Procrustes reflection size-and-shape distance has the explicit analytic expression

$$d_P(C_1, C_2)^2 = ||L_1||_{HS}^2 + ||L_2||_{HS}^2 - 2\sum_{k=1}^{\infty} \sigma_k,$$

where σ_k are the the singular values of the compact operator $L_2^* L_1$. A possible map is the square root $L_i = (C_i)^{1/2}$ and we use this choice in the following analysis, where we analyze the five selected Romance languages looking at the Procrustes distance between their frequency covariance operators.

For a given choice of the distance, a sample Fréchet mean and variance of a set of covariance operators $C^1, \ldots, C^L$ can be defined as

$$\overline{C} = \arg\inf_C \sum_{L=1}^{G} d(C^L, C)^2, \qquad \widehat{\sigma}^2 = \frac{1}{G}\sum_{L=1}^{G} d(C^L, \overline{C})^2.$$

These provide estimates for the centre point and the variability of the distribution with respect to the distance $d(.,.)$, which are needed for the test statistic in the permutation test.

Using this procedure, we can verify if the assumption that the covariance operators are the same across the words is disproved by data. Table 2 shows the p-values of the permutation tests for the equality of the marginal frequency covariance operator across the different words for the five Romance languages described in Section 2, obtained with Procrustes distance between sample covariance operators and $M = 1000$ permutations on the residual log-spectrograms. It can be seen that there is no evidence against the

Table 2: P-values of the permutation tests for H0: $C^L_{\omega,1} = C^L_{\omega,2} = \cdots = C^L_{\omega,10}$ *vs* H1: at least one is different, where $C^L_{\omega,i}$ is the marginal frequency covariance operator for the language L and word i, for the five Romance languages. The Procrustes distance is used for the test statistics

Language	French	Italian	Portuguese	American Spanish	Iberian Spanish
$p-value$	0.113	0.991	0.968	0.815	0.985

Table 3: P-values of the permutation tests for H0: $C^L_{t,1} = C^L_{t,2} = \cdots = C^L_{t,10}$ *vs* H1: at least one is different, where $C^L_{t,i}$ is the marginal time covariance operator for the language L and word i, for the five Romance languages. The Procrustes distance is used for the test statistics

Language	French	Italian	Portuguese	American Spanish	Iberian Spanish
$p-value$	0.02	0.422	0.834	0.683	0.17

hypothesis that the covariance operator is the same for all words. The same is true for the time covariance operator, as it can be seen in Table 3.

A possible concern is that the dimension of the data set becomes relatively small when it is split between the different words and language and therefore these testing procedure will have little power. On the other hand, this reasoning encourages us to simplify the model (assuming covariance operators constant across words) so that enough observations are available to estimate the parameters accurately. In the presence of a larger data set that allows to highlight differences between word-wise covariance operators, we would have also more information to estimate them accurately.

6 Exploring phonetic differences

We have now the tools to explore the phonetic differences between the languages in the Oxford Romance language data set. This can be of course done at different levels. A possible way to go would be to pair two speakers belonging to two different languages and look at their difference. However, this neglects the variability of the speech within the language and it would not be clear which part of the phonetic changes is to be credited to the difference between languages and which to the difference between the two individual speakers, unless we had available recordings from bilingual subjects. In this section we present a possible approach to the modelling of phonetic changes that takes into account the features of the speaker's population.

6.1 Modelling changes in the parameters of the phonetic process

We can start looking at the path that links the mean of the log-spectrograms between two words of different languages. These should be two words known to be related in the languages' evolution. This is the case for example for the same digit in two different Romance languages.

Considered as functional objects, the log-spectrograms means are unconstrained and integrable surfaces, thus interpolation and extrapolation can be simply obtained with a linear combination, where the weights are determined from the distance of the language we want to predict from the known languages. For example, if we want to reconstruct the path of the mean for the digit i from the language L_1 to the language L_2, we have

$$\overline{S}_i(x) = \overline{S}_i^{L_1} + x(\overline{S}_i^{L_2} - \overline{S}_i^{L_1}),$$

where $x \in [0, 1]$ provides a linear interpolation from language L_1 to language L_2, while $x < 0$ or $x > 1$ provides an extrapolation in the direction of the difference between the two languages, with $\overline{S}_i^L$ being the mean of the log-spectrograms from speakers of the language L pronouncing the i-th digit. Figure 9 shows for example a reconstructed path for the mean for "one" from French to Portuguese.

A natural question is if this can replicated for the covariance structure to be able to interpolate and extrapolate a more general description of the sound generation process. However, the case of the covariance structure is more complex. The experience with low dimensional covariance matrices (see Dryden et al., 2009) and the case of the frequency covariance operators illustrated in Pigoli et al. (2014) show that a linear interpolation is not a good choice for objects belonging to a non Euclidean space. We want therefore to use a geodesic interpolation based on an appropriate metric for covariance operator. Moreover, since we model the covariance structure as separable, we want also the predicted covariance structure to preserve this property. It is not possible to do this with geodesic paths in the general space of four-dimensional covariance structures and thus we define the new covariance structure as the tensor product of the geodesic interpolations (or extrapolations) in the space of time and frequency covariance operators,

$$C^x = C^x_\omega \otimes C^x_t,$$

where the geodesic interpolations (or extrapolations) C^x_ω, C^x_t depend on the chosen metric. In the case of the Procrustes reflection size and shape distance, the geodesic has the form

$$C^x_r = [(C^{L_1}_r)^{1/2} + x((C^{L_2}_r)^{1/2}\widetilde{R} - (C^{L_1}_r)^{1/2})][(C^{L_1}_r)^{1/2} + x((C^{L_2}_r)^{1/2}\widetilde{R} - (C^{L_1}_r)^{1/2})]^*$$

where $r = \omega, t$ and $\widetilde{R}$ is the unitary operators that minimizes $||(C^{L_1}_r)^{1/2} - (C^{L_2}_r)^{1/2}R||^2_{HS}$, see Pigoli et al. (2014). Other choices of the metric are of course possible, as long as they provide a valid geodesic for the covariance operator. However, some preliminary experiments reported in Pigoli et al. (2014) suggest that the Procrustes reflection size and shape geodesic performs better in the extrapolation of frequency covariance operators.

6.2 How would a speaker sound in a different language?

The framework we have set up allows also to observe how the sound produced by a speaker would be modified moving towards a different language. As mentioned in the introduction, we aim to project the sound produced by this speaker to that of a hypothetical speaker with the same position in a different language, with respect to the language variability structure. To do this, we need some additional specification to the statistical model which generates the log-spectrograms. For example, if we assume that the log-spectrograms of a spoken word is generated from a Gaussian process, its distribution is fully determined by the mean log-spectrogram (which is expected to be word-dependent) and the covariance structure. More in general, we identify the population of the pronunciations of a specific word of a language through its mean log-spectrogram, which is word-specific, and its time and frequency covariance functions, which are properties of the whole language. Thus, we identify here as speaker-specific residual what is left in the phonetic data once means and covariance information has been removed. Let us denote with $\mathfrak{F}_i^L$ this operation for the word i of the language L. Then, we can obtain a representation of the log-spectrogram for a speaker from a language L_1 in the language L_2 as

$$S_{ik}^{L_1 \to L_2} = [\mathfrak{F}_i^{L_2}]^{-1} \circ \mathfrak{F}_i^{L_1}(S_{ik}^{L_1}). \tag{3}$$

Here we choose to use the same word for both languages because in our data set words can be actually paired in a sensible way (the same digit in two Romance languages shares a common historical origin).

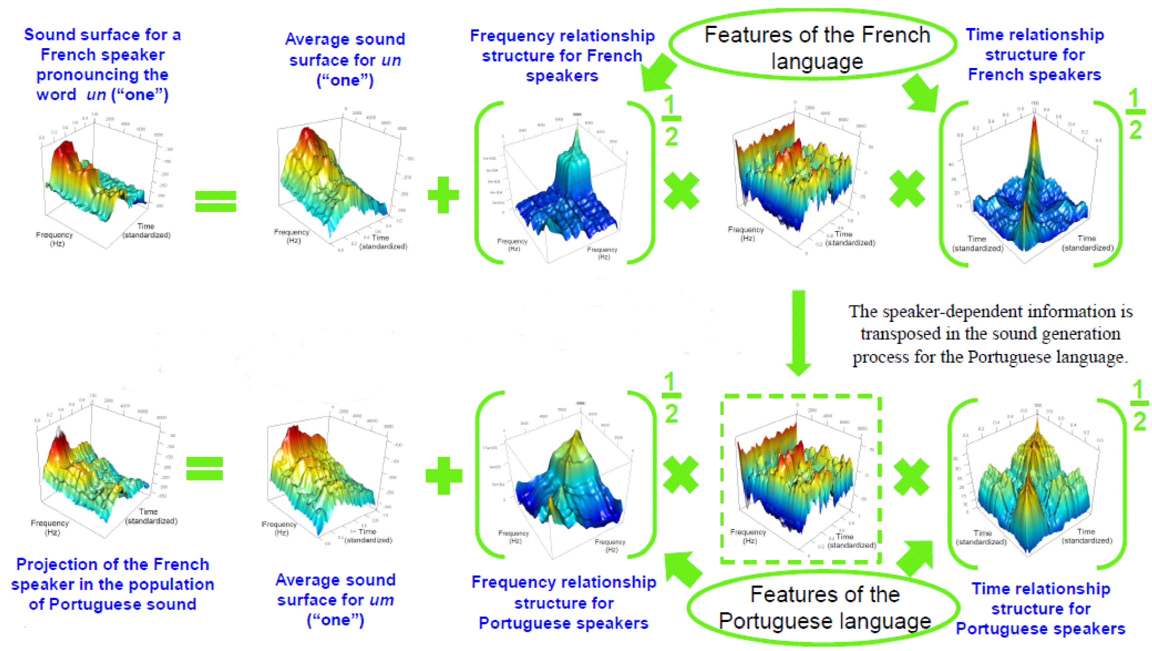

Figure 4: Graphical representation of the strategy to map a French speaker's log-spectrogram into the correspondent "position" in the Portuguese sound process.

The challenge now is how to define the transformation $\mathfrak{F}_i^L$. This is obtained considering both the characteristics of the sound populations in the two languages and the relative "position" of the speaker in their language. A graphical representation of this idea for the case of a French speaker mapped to the Portuguese language can be seen in Fig. 4. To define this transformation, we start from a speaker k from the language L_1 and we consider the residual log-spectrogram $R_{ik}^{L_1} = S_{ik}^{L_1} - \overline{S}_{i.}^{L_1}$. We would like to apply now a transformation that makes this residual uncorrelated, as generated by a white noise process. Let us consider the transformation from a finite dimensional white noise

$$Z = \sum_{i,j}^{p} z_{ij} v_i^{\omega} \otimes v_j^{t}, \qquad z_{ij} \sim N(0,1)$$

to a random surface with the same mean and covariance structure $(C_{\omega}^{L_1})_1^{1/2} \otimes (C_t^{L_1})^{1/2} Z$ of the sound distribution. We use here the notation for the application of a tensorized operator where

$$L_1 \otimes L_2 Z(\omega, t) = \int\int l_1(\omega, y) z(x, y) l_2(x, t) \mathrm{d}x \mathrm{d}y.$$

To obtain $\mathfrak{F}_i^L$, we would need to invert the transformation from Z to the sound process. This is not possible in general but we can restrict the inverse to work on the subspaces spanned by our data, thus defining $(C_l^L)^{-1/2} = \sum_{j=1}^{N} (\lambda_j)^{-1/2} \phi_j \otimes \phi_j$, $\phi_j, j = 1, \ldots, N$, $\{\lambda_j, \phi_j\}$ being eigenvalues and eigenfunctions for C_l^L. We then obtain

$$\mathfrak{F}_i^L(S_{ik}^L) = (C_{\omega}^L)^{-1/2} \otimes (C_t^L)^{-1/2} (S_{ik}^L - \overline{S}_{i.}^L)$$

and

$$[\mathfrak{F}_i^L]^{-1}(Z) = (C_{\omega}^L)^{1/2} \otimes (C_t^L)^{1/2} Z + \overline{S}_{i.}^L.$$

Figure 5 shows the log-spectrograms for the word "un" of the first French speaker S_{11}^{Fr}, its representation as Portuguese "um" $S_{11}^{Fr \to P}$ and the closest observed Portuguese "um", while Fig. 6 reports the result of the same operation applied to an Italian speaker towards the Iberian Spanish language.

6.3 Interpolation and extrapolation of spoken phonemes

The representation of a speaker in another observed language is interesting but is not enough for scholars to explore the changes that occur between two languages: a smooth estimate of the path of change is needed. This is also the case to extrapolate the sound process outside of the path connecting the two languages, which we recall to be the final goal of the Ancient Sounds project. Luckily, we can use the interpolated means and covariance operators described above to characterize the unobserved "languages" that are the intermediate steps in the phonetic changes. We thus obtain a smooth path between $S_{ik}^{L_1}$ and its representation in the language L_2 as

$$S(x) = [\mathfrak{F}_i^x]^{-1} \circ \mathfrak{F}_i^{L_1}(S_{ik}^{L_1}), \tag{4}$$

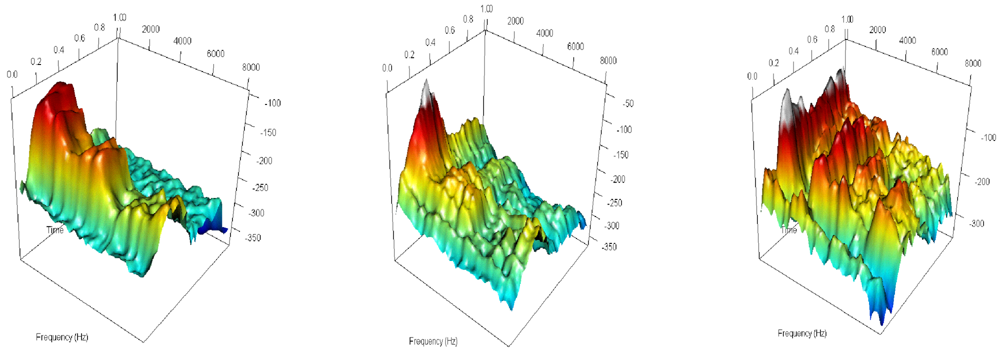

Figure 5: Log-spectrograms for the word “un” for a French speaker (left), its representation as word “um” in Portuguese using equation (3) (center) and the closest observed word “um” from a Portuguese speaker.

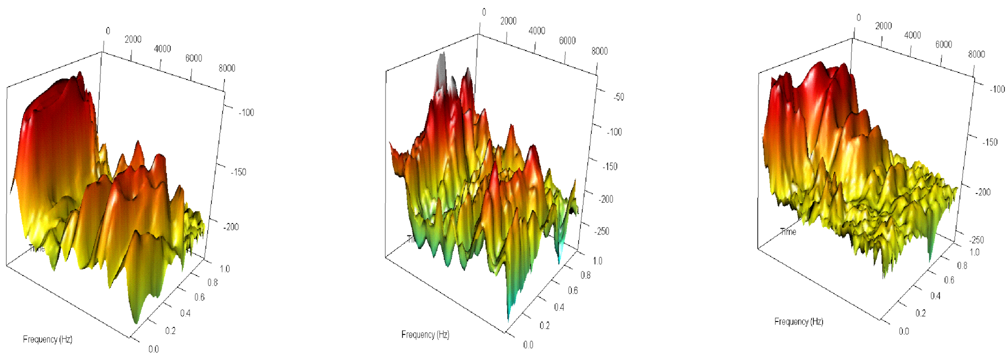

Figure 6: Log-spectrograms for the word “uno” for an Italian speaker (left), its representation as word “uno” in Spanish using equation (3) (center) and the closest observed word “uno” from a Spanish speaker.

$[\mathfrak{F}_i^x]^{-1} = (C_\omega(x))^{1/2} \otimes (C_t(x))^{1/2} Z + M(x)$ where $C_\omega(x)$ is the interpolated (or extrapolated) frequency covariance operator, $C_t(x)$ the correspondent time covariance operator and $M(x)$ the word-dependent mean. An example of a smooth path between the log-spectrogram for the word “un” for the same French speaker considered in the previous section and its representation in Portuguese can be seen in Fig. 7.

This strategy can also be used to reconstruct a smooth path between two observed log-spectrograms $S_{ik}^{L_1}$ and $S_{ik'}^{L_2}$, in this case the path being

$$S(x) = [\mathfrak{F}_i^x]^{-1}(x\mathfrak{F}_i^{L_1}(S_{ik}^{L_1}) + (1-x)\mathfrak{F}_i^{L_2}(S_{ik'}^{L_2})), \tag{5}$$

where a linear interpolation between the residuals takes the place of the residual of the single language. This can be useful when it is meaningful to pair two log-spectrograms in

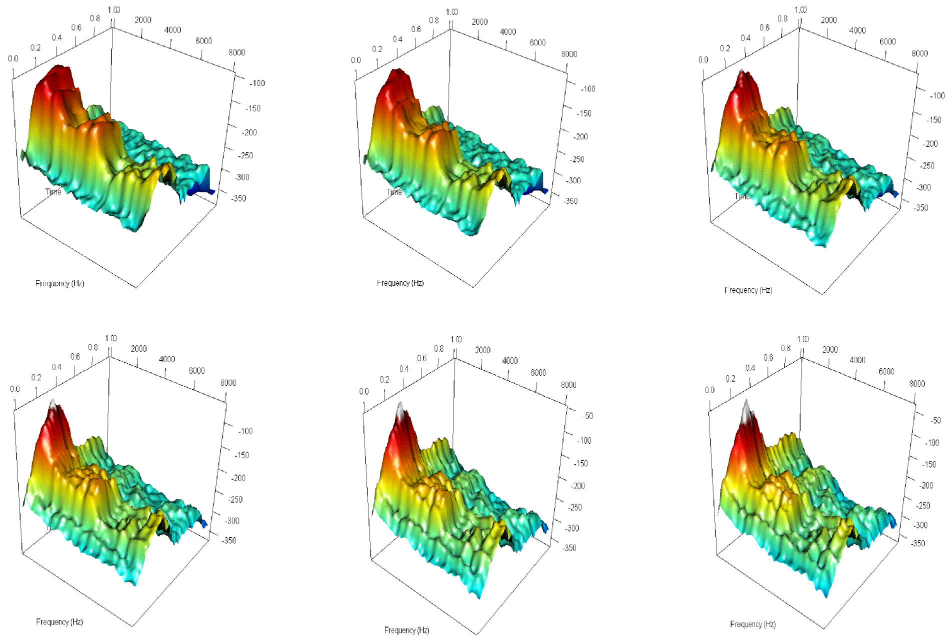

Figure 7: Smooth path between the log-spectrogram for the word “un” for a French speaker (top left) and its representation in Portuguese (bottom right).

different languages, for example because the same speaker is recorded in two languages. This is not the case in our data set but as a way of example we report in Fig. 8 the path between the log-spectrograms for the word “un” for a French speaker S_{11}^{Fr} and the word “um” for the Portuguese speaker which is closest to its representation $S_{11}^{Fr\to P}$. It is also interesting to compare this with the interpolated path between the two mean log-spectrograms in Fig. 9.

The possibility to extrapolate the sounds opens up to interesting possibility whenever two languages are known to be two stages of an evolutionary path. In this case extrapolating in the direction of the older language can provide an insight on the phonetic characteristics of the extinct ancestor languages. This of course will need to integrate in the model of sound change external information coming for example from textual analysis, history or anthropology. This is also due to the fact that the rate of change of languages is not constant and the path $S(x)$ can be go through at different speed for different branches of the languages evolution and it can be changed by events such as conquests, migrations, language contact, etc.

6.4 Back to sound reproduction

Even if observing the log-spectrograms (or other transformation of the recorded sounds) is often helpful, it is important to listen to the signals in the original domain. This is true also for the representation of a sound in a different language and the smooth paths

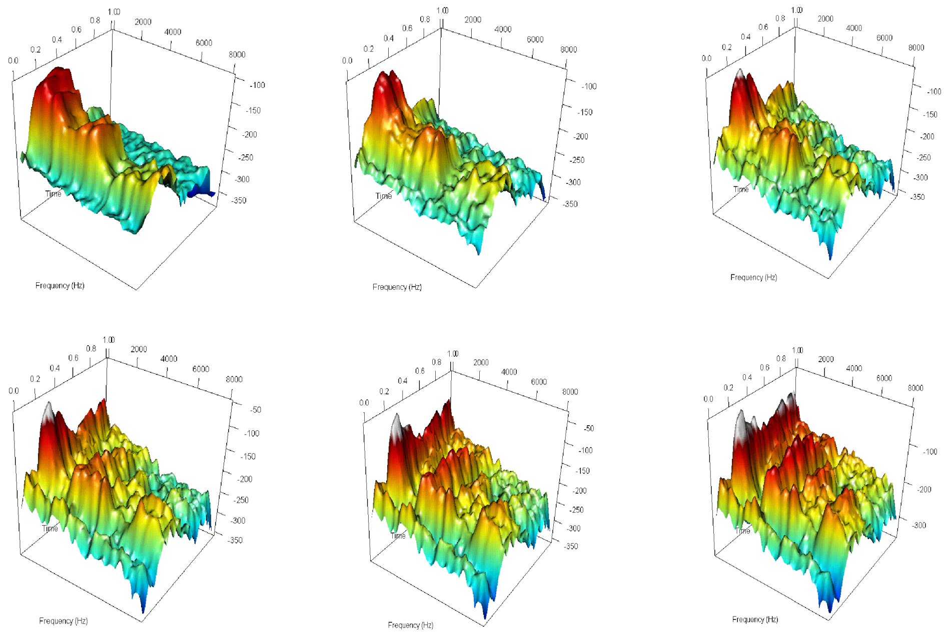

Figure 8: Smooth path between the log-spectrograms for the word “un” for a French speaker (top left) and the one for the word “um” closest to its representation in Portuguese (bottom right).

we have defined. Thus, we would like to reconstruct actual sounds from the estimated log-spectrograms. To do this, we would need also information about the phase that we have disregarded, since we have focused all our attention on the amplitude of the Fourier Transform, see Section 2. In principle, we could perform a parallel analysis on the phases to obtain representation of a phase in a different language, smooth path between phases and so on. However, this is tricky from a mathematical point of view, given the angular nature of the phases and there is unlikely to be much interesting additional information captured by the phase. Thus, in practice we use the phase associated to the log-spectrogram $S_{ik}^{L_1}$ to reconstruct the sounds all over the smooth path and the results appear quite satisfactory. An example of a reconstructed sound can be found in the Supplementary Material. In view of an interest in the quality of the reconstructed speech, a different representation may be chosen instead of the spectrogram, which is known to be problematic in this sense. Alternatives that have been shown to produce better reconstructions are linear predictive coding (LPC) coefficients (see,e.g, Bundy and Wallen, 1984) and mel frequency cepstral coefficient (MFCC, see Chazan et al., 2000, and reference therein). The general ideas of our approach can be extended to these alternative representations, although some of the details may need to be adapted to their specific mathematical properties.

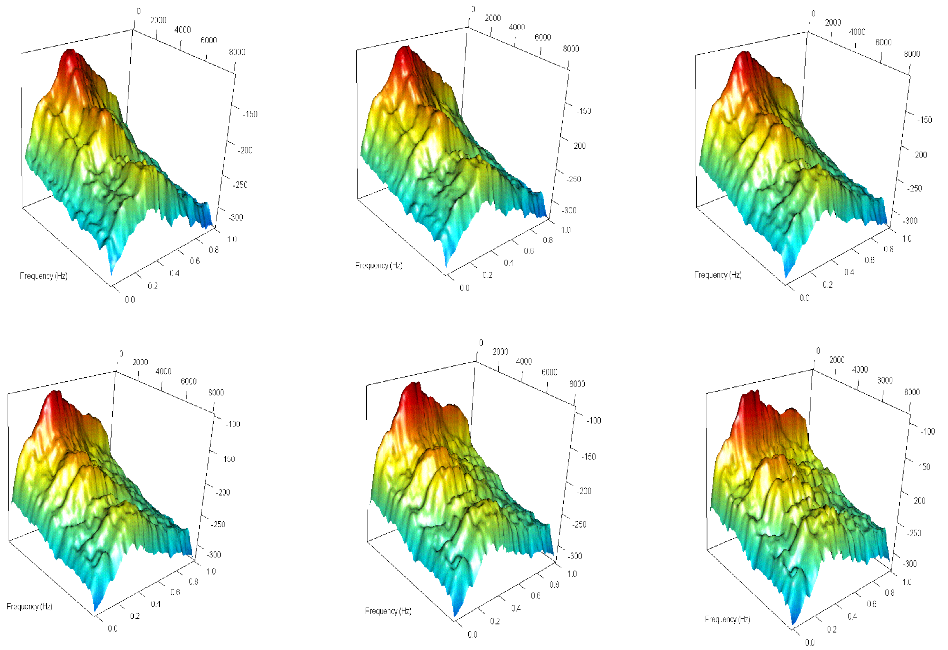

Figure 9: Smooth path $M(x)$ between the mean log-spectrogram for the word “un” in French (top left) and the one for the word “um” in Portuguese (bottom right).

7 Discussion

We have introduced a novel way to explore phonetic changes between languages that takes into account the characteristic of the sound population on the basis on actual speech recordings. The framework we introduced is useful to deal with acoustic phonetic data, i.e. samples of sound recordings of different words or other linguistic units from different groups (in our case, languages). We illustrate the proposed method with an application to the Romance digit data set, which includes the words corresponding to the numbers from one to ten pronounced by speakers of five different Romance languages. In particular, we verify in this data set the assumption that the covariance structure in the log-spectrograms is common for the different words within the language, thus increasing the sample available for its estimation. This is an interesting example of how the feature of a population (in this case the speakers from one language) may be captured in the second order structure and not only in the mean level. This in itself provides interesting information to linguists. It also fits within the recent development of the object oriented data analysis (see Wang and Marron, 2007), which advocates a careful consideration about what is the object of interest for the statistical analysis. Here, it seems that marginal covariance operators are promising features to represent phonetic structure at a language level.

We do not focus here on the representativeness or otherwise of the sample of speakers or words. In view of a broad use of this approach however, it is important to remember that the sample of speakers should reflect the population we are interested in and in particular a careful consideration should be given to regional stratification in the data set. Moreover, the considered words should be representative of the whole language. The digits studied here do have a wide ranging set of different phonemes present, indicating that the results are likely to be generalisable to some extent across a larger corpus, but,s of course, applying this to a comprehensive corpus of several languages would be most welcome, although challenging with respect to the amount of data required.

The proposed approach, using phonetic recordings in place of textual representation, allows us to explore the differences between different varieties of the same language, such as Spanish and American Spanish. Moreover, recent works (see The Functional Phylogenetic Group, 2012; Bouchard-Côté et al., 2013, and reference therein) focus on the reconstruction of the distribution of phonetic feature for ancestor languages. While the research in this field is still in its very earliest stages, when a good understanding of the historical evolution of sounds is available, this can be integrated in our methods to provide a reconstruction of how the speakers of extinct languages might have sounded. The final goal is therefore to integrate the modelling of the variability of speech within the language provided by our approach with the known dynamic of sound change established by linguistics research. We are confident that this will give a substantial contribution to the Ancient Sounds project whose goal is audible proto-language reconstruction.

We have illustrated the transformation of a speaker's speech from one language to another as a first example of application in speech generation but other problems can be addressed in this framework. For example, the proposed approach to model sound processes can be extended to take into account also discrete or continuous covariates associated to the mean and the covariance operators. These can be seen as function of the geographical coordinates or of antiquity when studying dialects. While we treated the language as a categorical variable, nothing prevents us seeing it as a continuous process in space and time. Indeed, the definition of the continuous path between two languages described in Section 6.3 can be seen as the first step in this direction, since the abscissa x of the path can be made dependent on external variables. While we do not claim this can straightforwardly reproduce the evolution branches in language history, it can still be a useful starting point for more complex models.

The application of the proposed method is not necessarily restricted to comparative linguistics. It can be useful whenever a comparison between groups of sounds is needed, or indeed other complex wavelike signals. In the future it will be interesting to explore micro-variation within a language (dialects, spoken language in different subgroups of the population) but also other types of sounds such as songs or even sounds different from human speech, for example bird calls.

References

Aston, J. A. D., Chiou, J.-M. and Evans, J. P. (2010) "Linguistic Pitch Analysis using Functional Principal Component Mixed Effect Models", *Journal of the Royal Statistical*

Society Series C - Applied Statistics, **59**, 297–317.

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

Bundy, A. and Wallen, L. (1984) "Linear Predictive Coding", in *Catalogue of Artificial Intelligence Tools*, Springer Berlin Heidelberg.

Chazan, D., Hoory, R., Cohen, G. and Zibulski, M. (2000) "Speech reconstruction from mel frequency cepstral coefficients and pitch frequency", *Proceedings of ICASSP 2000, IEEE International Conference on Acoustics, Speech, and Signal Processing*, **3**, 1299–1302.

Dryden, I. L. and Mardia, K. V. (1998) Statistical Shape Analysis. Wiley, Chichester.

Dryden, I. L., Koloydenko, A., Zhou, D. (2009), Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging, *Annals of Applied Statistics*, **3**, 1102–1123.

Ferraty, F. and Vieu, P. (2006) *Nonparametric Functional Data Analysis: Theory and Practice*, Springer, Berlin.

Garcia, D. (2010) "Robust smoothing of gridded data in one and higher dimensions with missing values", *Computational Statistics & Data Analysis*, **54**, 1167–1178.

Ginsburgh, V. and Weber, S. (2011) *How Many Languages Do We Need? The Economics of Linguistic Diversity*, Princeton University Press.

Grimes, J. E. and Agard, F. B. (1959) "Linguistic Divergence in Romance", *Language*, **35**, 598–604.

Hadjipantelis, P. Z., Aston, J. A. and Evans, J. P. (2012) "Characterizing fundamental frequency in Mandarin: A functional principal component approach utilizing mixed effect models", *The Journal of the Acoustical Society of America*, **131**, 4651.

Koenig, L. L., Lucero, J. C. and Perlman, E. (2008) "Speech production variability in fricatives of children and adults: Results of functional data analysis", *The Journal of the Acoustical Society of America*, **124**, 3158–3170.

Morpurgo Davies, A. (1998) *Linguistics in the nineteenth century*, Longman, London.

Nakhleh, L., Ringe, D. and Warnow, T. (2005) "A New Methodology for Reconstructing the Evolutionary History of Natural Languages", *Language*, **81**, 382–420.

Pagel, M. (2009) "Human language as a culturally transmitted replicator", *Nature Reviews Genetics*, **10**, 405415.

Pigoli D., Aston, J.A.D., Dryden, I.L. and Secchi, P. (2014) "Distances and Inference for Covariance Operators", *Biometrika*, **101**, 409–422.

Ramsay, J.O. and Silverman, B.W. (2005), *Functional Data Analysis (2nd ed.)*, Springer, New York.

Tang, R. and Müller, H.G. (2008) "Pairwise curve synchronization for high-dimensional data", *Biometrika*, **95**, 875–889.

The Functional Phylogenies Group. (2012) "Phylogenetic inference for function-valued traits: speech sound evolution", *Trends in Ecology & Evolution*, **27**, 160–166.

Torgerson, W. S. (1958) Theory and Methods of Scaling. Wiley, New York.

Wang, H. and Marron, J.S. (2007) "Object oriented data analysis: Sets of trees", *The Annals of Statistics*, **35**, 1849–1873.