ROYAL
STATISTICAL
SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

Appl. Statist. (2018)
67, *Part* 5, *pp.* 1103–1145

The statistical analysis of acoustic phonetic data: exploring differences between spoken Romance languages

Davide Pigoli,

King's College London, UK

Pantelis Z. Hadjipantelis,

University of California at Davis, USA

John S. Coleman

University of Oxford, UK

and John A. D. Aston

University of Cambridge, UK

[*Read before* The Royal Statistical Society *on Wednesday, April 18th, 2018,* Professor R. Henderson *in the Chair*]

Summary. The historical and geographical spread from older to more modern languages has long been studied by examining textual changes and in terms of changes in phonetic transcriptions. However, it is more difficult to analyse language change from an acoustic point of view, although this is usually the dominant mode of transmission. We propose a novel analysis approach for acoustic phonetic data, where the aim will be to model the acoustic properties of spoken words statistically. We explore phonetic variation and change by using a time–frequency representation, namely the log-spectrograms of speech recordings. We identify time and frequency covariance functions as a feature of the language; in contrast, mean spectrograms depend 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 these to define a phonetic transformation that models 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, enabling us to listen to the output of the statistical analysis. The approach proposed is demonstrated by using recordings of the words corresponding to the numbers from 1 to 10 as pronounced by speakers from five different Romance languages.

Keywords: Functional data analysis; Object data; Quantitative linguistics; Spectrograms

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 that is roughly similar to biological evolution is now generally accepted; see, for example, Cavalli-Sforza (1997) and Nakhleh *et al.* (2005). Pagel (2009) claimed that genes and languages have similar

Address for correspondence: John A. D. Aston, Department of Pure Mathematics and Mathematical Statistics, Statistical Laboratory, University of Cambridge, Wilberforce Road, Cambridge, CB3 0WB, UK.
E-mail: j.aston@statslab.cam.ac.uk

© 2018 The Authors Journal of the Royal Statistical Society: Series C (Applied Statistics) 0035–9254/18/671103
Published by John Wiley & Sons Ltd on behalf of the Royal Statistical Society.
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

evolutionary behaviour and offered an extensive catalogue of analogies between biological and linguistic evolution. This immediately gives rise to the notion of familial relationships between languages.

However, 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, and 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 in which they are spoken, is a topic of widespread interest for its cultural relevance. We all have our own experience with learning and using different languages (and different varieties within each language) and the effort to find quantitative properties of speech can shed some light on the subject.

The first step in exploring the language ecosystem is to choose how to analyse and measure the differences between languages. A language is a complex entity and its evolution can be considered from many points of view. The processes of change from one language to another have long been studied by considering textual and phonetic representation of the words (see, for example, Morpurgo Davies (1998) and references therein). This focus on written forms reflects a general normative approach towards languages: for cultural and historical reasons, the way in which we think about them is focused on the written expression of the words and their ‘proper’ pronunciations. However, this is more a social artefact than a reality of the population, as there is great variation within each language depending on socio-economic and physiological 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, change in the sounds associated with the pronunciations of words. On the one hand, these provide a complementary way to consider the difference between two languages which can be juxtaposed with the differences measured by using textual and phonetic representation. On the other hand, it can be claimed that the acoustic expression of the word is a more natural object of interest, textual and phonetic transcriptions being only the representation that is 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, because of the complexity of speech as a data object, the theoretical challenges of 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 sociolinguistics that is concerned with small-scale variation within communities (e.g. between younger and older members or particular social groups). Some of the techniques that we describe here 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 is an abstraction that simplifies the reality of speech variability and neglects the continuous geographical spread of spoken varieties, albeit with some clear edges. For example, Grimes and Agard (1959) described as a ‘useful fiction’ the definition of homogeneous speech communities: groups of speakers whose 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 the 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 that this work will help to bring attention to the subject.

In the remainder of this paper, we operationalize the term 'language' to mean a set of recordings of various words in a language or dialect, as spoken on various occasions by a group of speakers, without implying that the vocabulary is complete nor even necessarily large. However, the methodology proposed can be applied in a straightforward way to larger and more comprehensive *corpora*.

We use the expression 'acoustic phonetic data' to refer to sound recordings of the same word (or other linguistic unit) 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 enables better statistical exploration of the phonetic characteristics of the language. This is very different from having only repetitions of a word pronounced by a single speaker and it calls for the development of a novel approach.

The aim of our work is to provide a framework where

(a) speech recordings can be analysed to identify features of a language,
(b) the variability of speech within the language can be considered and
(c) the acoustic differences between languages can be explored on the basis of speech recordings, taking into account intralanguage variability.

Among other things, this will enable us to develop a model (in Section 6) to explore how the sound that is produced by a speaker would be modified when moved towards the phonetic structure of a different language. More specifically we shall take into account the variability of pronunciation within each language. This means that we explore the variability of the speakers of the language so that we can then understand where a specific speaker is positioned in a space of acoustic variation with respect to the population. This enables us to postulate a path that maps the sound that is 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 as we can extract when a speaker pronounces words in two different languages in which they are proficient even if we have only monolingual speakers. The observation (audio recordings) of many speakers from each group is essential to understand the intralanguage variability and thus the relevance of the interlanguage acoustic change. This model has an immediate application in speech synthesis, with the possibility of mapping a recording from one language to another, while preserving the speaker's voice characteristics. This approach could be also extended to modify synthesized speech in such a way that it sounds like the voice of a specific speaker (e.g. a known actress or a public person). This would be of interest for many commercial applications, from computer gaming to advertising, and it is only one example of the methods that can be developed in the framework that we provide. More generally, the framework that is given here addresses the problem of how to separate speaker-specific voice characteristics from language-specific pronunciation details.

The paper is structured as follows. Section 2 describes the acoustic phonetic data that are used to demonstrate our methods. We choose to represent the speech recordings in a time–frequency domain by using a local Fourier transform resulting in surface observations, known in signal processing as spectrograms. Therefore, a short introduction to the functional data analysis approach to surface data is given in Section 3. The details of these time–frequency representations, as well as the preprocessing steps that are needed to remove noise artefacts 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 and shows how the pronunciation of a word can be morphed into another language while preserving the speaker or voice identity. The final section gives a discussion of the advantages of the method proposed 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 digits data set

The methods in this paper will be illustrated with an application to a data set of audio recordings of digits in Romance languages. This data set was compiled in the Phonetics Laboratory of the University of Oxford in 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 from 1 to 10 in their native language. The data set is inherently unbalanced; we have seven French speakers, five Italian speakers, five American Spanish speakers, five Castilian Spanish speakers and three Portuguese speakers, resulting in a sample of 219 recordings, because not all words are available for all speakers. The sources of the recordings were either collected from freely available language training Web sites or standardized recordings 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 enables us to build models which characterize realistic variation in speech recording, which is somewhat a prerequisite for using this model in practice, as fieldwork recordings are often not recorded under laboratory conditions. The data set is also heterogeneous in terms of sampling rate, duration and format. As such, before any phonetic or statistical analysis took place, all data were converted to 16-bit pulse code modulation $*$.wav files at a sampling rate 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 time. This data set has been collected within the scope of 'Ancient sounds', which is a 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 Web site 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 representative of a language's vocabulary from a phonetic point of view, meaning that the words that were used for the numbers in the Romance languages were not chosen to have any specific phonetic structure. Consequently, we use the word 'language' as shorthand for these particular small samples of digit recordings. However, we view this analysis as a proof of concept, and we shall 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

Various representations are available in phonetics to deal with speech recordings (see, for example, Cooke *et al.* (1993)). Many of them share the idea of representing the sound with the distribution of intensities over frequency ω and time t. We choose in particular the power spectral density of the local Fourier transform (i.e. the narrow-band Fourier spectrogram), as detailed in Section 4. This widely used representation is a two-dimensional surface that describes the sound intensity for each time sample in each frequency band. Since we can represent each spoken word as a two-dimensional smooth surface, it is natural to employ a functional data analysis approach. Good results have already been obtained by applying functional data analysis techniques to acoustic analysis, although in the different context of single-language studies, e.g. 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.

In contrast, in most previous quantitative work on pronunciation variation, such as sociolinguistics or experimental phonetics, only one or a few acoustic parameters (one-dimensional time series) are examined, e.g. pitch or individual resonant frequencies. Variations in vowel qualities, for example, are typically represented by just two data points: the lowest two resonant frequencies (the first and second formants) measured at the mid-point of the vowel. Such a two-dimensional representation lends itself well to simple visualization of a large number of observations, in the form of a scatter plot. Although the validity of two-frequency representations of vowels or single-variable representations of pitch or loudness is motivated by decades of prior research, it clearly suffers from two limitations. First, almost all of the available time–frequency–amplitude information in the speech signal is simply discarded as if it were irrelevant. Second, we do not always know in advance which acoustic parameters are most relevant to a particular investigation; therefore, a more holistic approach to analysis of speech signals may be helpful. The methods that are presented in this paper, which take the entire spectrogram of each audio recording as data objects, enable us to examine and to manipulate a variety of properties of speech that are not easily reduced to a single low dimensional data point. By considering higher order statistical properties of the shape of spectrograms, it becomes possible to characterize such notions as the typical pronunciation of a word, what each speaker sounds like (in general, irrespectively of what words they are saying), how their pronunciation differs from that of other speakers and what it is that makes two languages sound different, beyond the differences in the words that they use and the speakers involved.

More formally, we consider here data objects that are two-dimensional surfaces on a bounded domain, as in the case of spectrograms. Let X be a random surface so that $X \in L^2([0,\Omega]\times[0,T])$ 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 $c(\omega,\omega',t,t')=\mathrm{cov}\{X(\omega,t),X(\omega',t')\}$.

In practice these surfaces are observed over a finite number of grid points and they are affected by noise; indeed they can be thought of as a noisy image. 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 that span the space $L^2([0,\Omega]\times[0,T])$. In particular, we choose the widely popular method of smoothing splines to estimate a smooth surface $\tilde{X}(\omega,t)$ 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 co-ordinates (ω,t) is meaningful. However, this is often not so when data are measurements 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 meaningful 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 has been devoted to its solution and these techniques are referred to as functional registration (or warping or alignment; see Marron *et al.* (2014) and references therein for details). In the case of a two-dimensional surface, the misalignment can in principle affect both co-ordinates; this is so for example in image processing. A two-dimensional transformation $h(\omega, t)$ 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 in which 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 speaking rates, which are not relevant for our goals. We necessarily want to preserve the differences on the frequency axis, which contains information about the phonetic characteristics of the speakers.

Thus, we apply a monodimensional warping to our surface data. If we aim to align a sample of surfaces $\tilde{X}_1, \ldots, \tilde{X}_N$, we look for a set of time warping functions $h_1(t), \ldots, h_N(t)$ so that the aligned surface will be defined as $X_1 = \tilde{X}_1\{\omega, h_1(t)\}, \ldots, X_N = \tilde{X}_N\{\omega, h_N(t)\}$. In the next section we shall 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, e.g.

$$\hat{\mu}(\omega, t) = \frac{1}{N} \sum_{i=1}^{N} X_i,$$

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

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 that we first apply a local Fourier transform to obtain a two-dimensional spectrogram that is a function of time (the time instant where we centre the window for the local Fourier transform) and frequency. For the Romance digit data, we use a Gaussian window function w with a window length of 10 ms (which is a reasonable length for the signal to be considered as stationary), defined as $\psi(\tau) = \exp\{-\frac{1}{2}(\tau/0.005)^2\}$. Since the original acoustic data were sampled at 16 kHz, this results in a window size of 160 samples per frame and the maximal frequency detected is $\omega_{\max} = 8$ kHz (see, for example, Blackledge (2006) for more details).

We can compute the local Fourier transform at angular frequency ω and time t as

$$X_{ik}^{L}(\omega, t) = \int_{-\infty}^{\infty} x_{ik}^{L}(\tau)\psi(\tau - t)\exp(-\mathrm{j}\omega\tau)\mathrm{d}\tau,$$

where $\mathrm{j} \equiv \surd(-1)$ denotes the imaginary unit. The power spectral density, or spectrogram, defined as the magnitude of the Fourier transform and the log-spectrogram (in decibels), is therefore

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

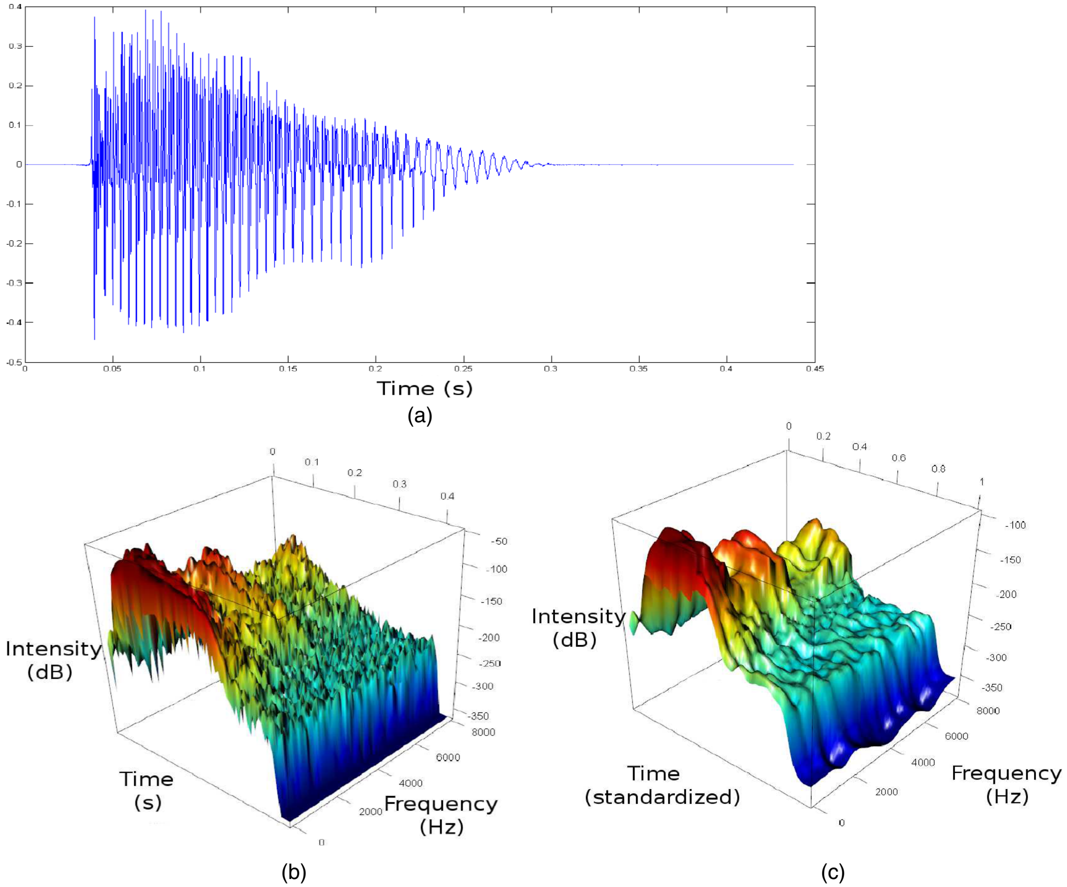


Fig. 1. (a) Raw record, (b) raw log-spectrogram and (c) smoothed and aligned log-spectrogram for a French speaker pronouncing the word ‘*un*’ (‘one’)

Fig. 1 shows an example of a raw speech signal (Fig. 1(a)) and the corresponding log-spectrogram (Fig. 1(b)) for the sound produced by a French speaker pronouncing the word *un* [œ̃].

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

The second preprocessing step consists of registration in time. This is necessary because speakers can speak faster or slower and this is particularly true when data are collected from different sources where the context is different. However, differences in the speech rate are normally not relevant from a linguistic point of view and thus alignment along the time axis is needed because of this time misalignment in the acoustic signals. First, we standardized the timescale so that each signal goes from 0 to 1. Then, we adapt to the case of surface data the procedure that was proposed in Tang and Müller (2008) to remove time misalignment from functional observations. Given a sample of functional data $f_1, \ldots, f_n \in L^2([0,1])$, this procedure looks for a set of strictly monotone time warping functions $h_1, \ldots, h_n$ so that $h_i(0)=0$, $h_i(1)=1$, $i=1,\ldots,n$. In practice, these warping functions are modelled via spline 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 spline functions on a chosen grid. Now let h_k, $k=1,\ldots,n$, be the warping function from a specific time to the standardized timescale. If $s=h_j^{-1}(t)$, then $h_i(s)=h_i\{h_j^{-1}(t)\}=h_{ij}(t)$. Under the assumption that the warping function has the identity on average and thus $E[h_{ij}|h_j]=h_j^{-1}$, the estimator that was proposed by Tang and Müller (2008) is

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

To apply this idea to acoustic phonetic data, we need first to define the groups of log-spectrograms that we want to align together. As the mean log-spectrogram is different from word to word, we decide to align the log-spectrograms corresponding to the same word. Then, we must extend the procedure to two-dimensional objects such as surfaces. As mentioned in the previous section, it is safe to assume that there is no phase distortion in the frequency direction, given the relatively narrow window that is used in the local Fourier transform. In contrast, 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 functions but minimize the discrepancy between surfaces. For each word $i=1$ in a group of log-spectrograms that we want to align, for every pair of languages L and L' and for every pair of speakers k and m, we define the discrepancy between the log-spectrogram $\tilde{W}_{ik}^{L}$ and $\tilde{W}_{im}^{L'}$ as

$$D_\lambda(\tilde{W}_{ik}^{L}, \tilde{W}_{im}^{L'}, g_{km}^{LL'})=\int_{\omega=0}^{\omega_{\max}}\int_{t=0}^{1}[\tilde{W}_{ik}^{L}\{\omega, g_{km}^{LL'}(t)\}-\tilde{W}_{im}^{L'}(\omega,t)]^2+\lambda\{g_{km}^{LL'}(t)-t\}^2\,\mathrm{d}t\,\mathrm{d}\omega, \qquad (1)$$

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 $\tilde{W}_{ik}^{L}(\omega,t)$ to that of $\tilde{W}_{im}^{L'}(\omega,t)$. We obtain the pairwise warping function $\hat{g}_{km}^{LL'}(\cdot)$ by minimizing the discrepancy $D_\lambda(\tilde{W}_{ik}^{L}, \tilde{W}_{im}^{L'}, g_{km}^{LL'})$ 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:

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

and the smoothed and aligned log-spectrogram for the language L, word i and speaker k is therefore $S_{ik}^{L}(\omega,t)=\tilde{W}_{ik}^{L}\{\omega, h_{ik}(t)\}$. 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 equation (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. In general, the number of grid points in the time axis needs to be chosen on the basis of the length of the sounds uttered but we have seen that 100 points provide an accurate reconstruction of the log-spectrograms in the Romance digit data set.

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 that was produced by a French speaker pronouncing the word *un* can be found in Fig. 1(c).

Other choices are, of course, possible in the preprocessing of the speech data. In particular, the time registration based on the minimization of the Fisher–Rao metric (Srivastava *et al.*, 2011) can be a computationally more efficient alternative when computing time is of concern. By way of example, we present as on-line supplementary material the analysis of the Romance digit data when the smoothing is performed with the thin plate regression splines implemented in the R package `mgcv` (Wood, 2003) and the time registration is obtained by minimizing the Fisher–Rao metric (R package `fdasrvf` (Tucker, 2014)). As can be seen there, the subsequent analysis is qualitatively similar to that reported below, giving credence to the idea that the results are not simply systematic misregistration by one technique *versus* another.

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 assumptions to identify and estimate these parameters. We consider the mean log-spectrogram as depending on the particular 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 words pronounced 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 $\bar{S}_i^L(\omega,t)=(1/n_L)\Sigma_{k=1}^{n_L}S_{ik}^L(\omega,t)$ for each word i of the language L.

Recent studies (Aston *et al.*, 2010; Pigoli *et al.*, 2014) have shown 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 estimate it by using the residual surface that is obtained by removing the mean effect of the word. In Section 5.1 we develop a procedure to verify this assumption in the Romance digit 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 $k=1,\ldots,n_L$, we thus focus on the residual log-spectrograms $R_{ik}^L(\omega,t)=S_{ik}^L(\omega,t)-\bar{S}_i^L(\omega,t)$, which measure how each speech token differs from the word mean. In what follows, we disregard in the notation the different speakers and words and for the residual log-spectrogram indicate by $R_j^L(\omega,t)$, $j=1,\ldots,n_L$, the set of observations for the language L including all speakers and words.

However, using standard covariance estimation techniques we find that the full four-dimensional covariance structure is computationally expensive or not statistically feasible (because of the small sample size); thus we need some modelling assumptions. There are many ways to incorporate assumptions that allow such estimation, a common assumption being some form of sparsity. Rather than the usual definition of sparsity that many elements are 0, we prefer to work on the principle that the covariance can be factorized.

We assume that the covariance structure $c^L(\omega_1,\omega_2,t_1,t_2)=\text{cov}\{S^L(\omega_1,t_1),S^L(\omega_2,t_2)\}$ is separable in time and frequency, i.e. $c^L(\omega_1,\omega_2,t_1,t_2)=c_\omega^L(\omega_1,\omega_2)c_t^L(t_1,t_2)$. Although we do not necessarily believe that this assumption is true in general, a structure is needed to obtain reliable estimates for the covariance operators, and it is a reasonable assumption that is frequently (implicitly) used in signal processing, particularly when constructing higher dimensional from lower dimensional bases. For more details about the use of separability assumptions for speech data, see Aston *et al.* (2017).

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

$$\hat{c}^L_r = \frac{\tilde{c}^L_r}{\surd \mathrm{tr}(\tilde{c}^L_r)}, \qquad r=\omega,t, \tag{2}$$

where 'tr' indicates the trace of the covariance function, defined as $\mathrm{tr}(c)=\int c(s,s)\,\mathrm{d}s$, whereas $\tilde{c}^L_r$, $r=\omega,t$, are the sample marginal covariance functions

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

and

$$\tilde{c}^L_t(t_1,t_2)=\frac{1}{n_L-1}\sum_{j=1}^{n_L}\int_0^{\omega_{\max}} \{R^L_j(\omega,t_1)-\bar{R}^L_{n_L}(\omega,t_1)\}\{R^L_j(\omega,t_2)-\bar{R}^L_{n_L}(\omega,t_2)\}\,\mathrm{d}\omega,$$

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

$$\hat{C}^L_r g(x)=\int_0^M \hat{c}^L_r(x,x')g(x')\,\mathrm{d}x', \qquad g\in L^2(\mathbb{R}),\ (r,M)\in\{(\omega,\omega_{\max}),(t,1)\}.$$

To see why we choose equation (2) to estimate the two separable covariance functions, let $\tilde{c}^L_\omega$ and $\tilde{c}^L_t$ be the true marginal covariance functions, i.e.

$$\tilde{c}^L_\omega(\omega_1,\omega_2)=\int_0^1 c^L(\omega_1,\omega_2,t,t)\mathrm{d}t,$$
$$\tilde{c}^L_t(t_1,t_2)=\int_0^{\omega_{\max}} c^L(\omega,\omega,t_1,t_2)\mathrm{d}\omega.$$

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

$$\begin{aligned}\tilde{c}^L_\omega(\omega_1,\omega_2)\tilde{c}^L_t(t_1,t_2)&=\int_0^1 c^L_\omega(\omega_1,\omega_2)c^L_t(t,t)\mathrm{d}t\int_0^{\omega_{\max}} c^L_\omega(\omega,\omega)c^L_t(t_1,t_2)\mathrm{d}\omega\\&=c^L_\omega(\omega_1,\omega_2)\,\mathrm{tr}(c^L_t)c^L_t(t_1,t_2)\,\mathrm{tr}(c^L_\omega).\end{aligned}$$

Moreover, $\mathrm{tr}(\tilde{c}^L_\omega)=\mathrm{tr}\{c^L_\omega\mathrm{tr}(c^L_t)\}=\mathrm{tr}(c^L_\omega)\mathrm{tr}(c^L_t)$ and the same is true for $\tilde{c}^L_t$. Hence,

$$\frac{\tilde{c}^L_\omega(\omega_1,\omega_2)}{\surd\mathrm{tr}(\tilde{c}^L_\omega)}\frac{\tilde{c}^L_t(t_1,t_2)}{\surd\mathrm{tr}(\tilde{c}^L_t)}=c^L_\omega(\omega_1,\omega_2)c^L_t(t_1,t_2)=c^L(\omega_1,\omega_2,t_1,t_2)$$

and this suggests $\hat{c}^L_r$ as an estimator for c^L_r, $r=\omega,t$.

Figs 2 and 3 show the estimated marginal covariance functions for the five Romance languages. As can be seen, the frequency covariance functions present differences that appear to be language specific (with peaks and plateaus in different positions), whereas the time covariances have similar structure, the dependence decreasing when time lag increases and most of the covariability is concentrated close to the diagonal.

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

We made the assumption above that the covariance operators are common to all the words within each language, whereas the means are different between words. This assumption can be

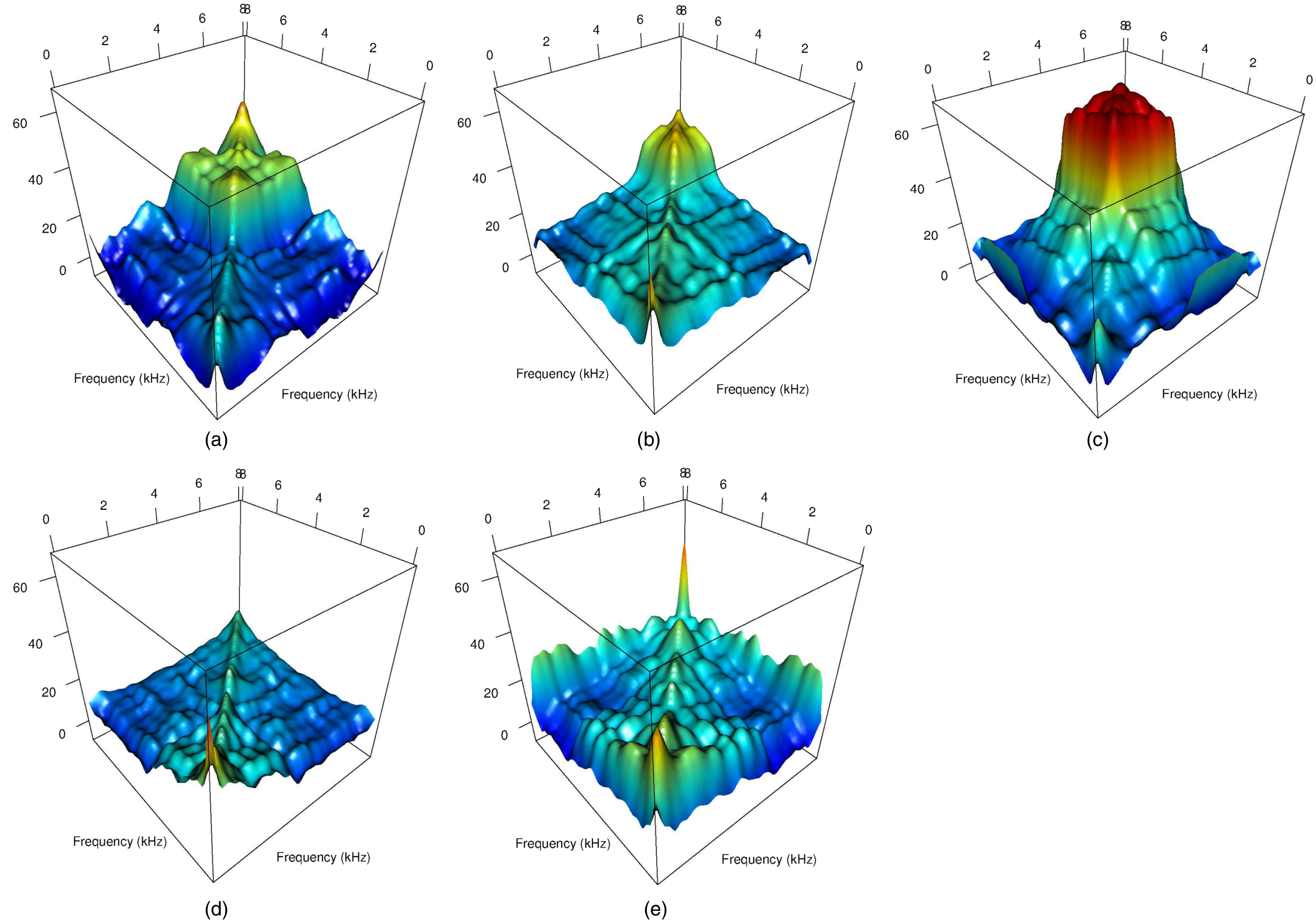


Fig. 2. Marginal covariance function between frequencies for the five Romance languages (a) Italian, (b) French, (c) Portuguese, (d) American Spanish and (e) Castilian Spanish

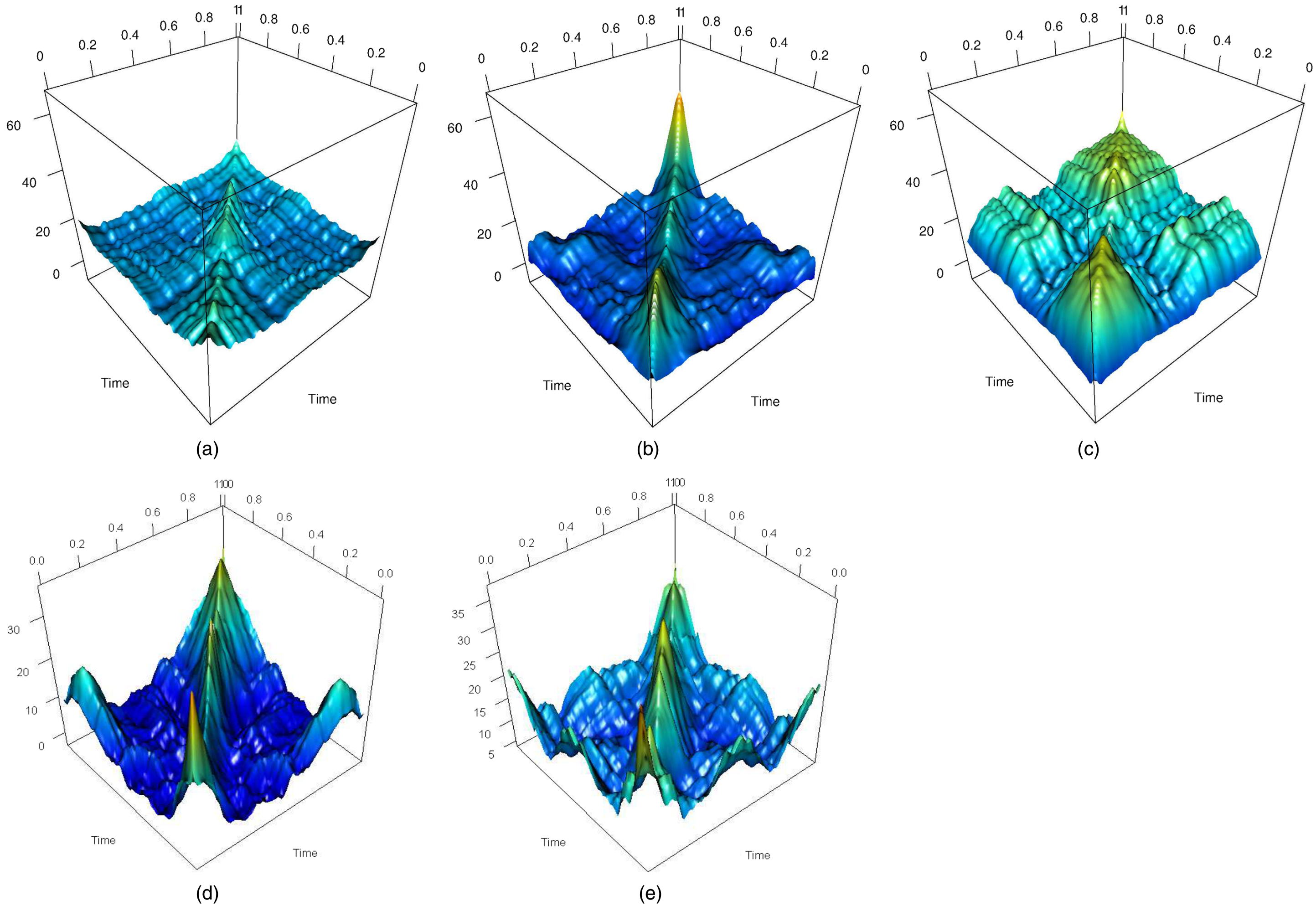


Fig. 3. Marginal covariance function between times for the five Romance languages (a) Italian, (b) French, (c) Portuguese, (d) American Spanish and (e) Castilian Spanish

verified by using permutation tests that look at the effect of the group factor on the parameters of the sound process.

When an estimator for a parameter is available and it is possible to define a distance $d(\cdot,\cdot)$ 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 lth 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 lth 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 that rely on the fact that, if there is no difference between 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 permutations 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 implies the groups having the same mean. If this is not true, we can apply the procedure to the centred observations $\tilde{X}_{il} = X_{il} - \bar{X}_l$, $i = 1,\ldots,n$, $l = 1,\ldots,G$, where $\bar{X}_l$ is the sample mean for the lth group. This guarantees that the observations are asymptotically exchangeable because of the law of large numbers.

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

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

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

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

where P is the appropriate functional space to which the parameters belong. This test statistic measures the variability of the estimator of the parameters across the various groups. If the parameter is indeed different for some groups, we expect the estimates from groups $1,\ldots,G$ to 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 $T_m = \Sigma_{l=1}^{G} d(K_l^m,\bar{K}^m)^2$, 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 mth permutation and $\bar{K}^m$ is their sample Fréchet mean. The p-value of the test will therefore be the proportion of permutations for which the test statistic is greater than in the original data set, i.e. $p = \#\{T_m > T_0\}/M$.

We now apply this general procedure to the three parameters of interest in our case, 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 languages.

Let us start by considering the test to compare the means of the log-spectrograms across the words (digit) of each language. Here the natural estimator for the wordwise mean log-spectrogram is the sample mean, $K_l = \bar{S}_l^L(\omega,t)$, and the distance can be chosen to be the distance in $L^2([0, 8\text{ kHz}] \times [0, 1])$:

Table 1. *p*-values of the permutation tests for $H_0: \mu_1^L = \mu_2^L = \ldots = \mu_{10}^L$ *versus* H_1: 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	Castilian Spanish
p-value	<0.001	0.02	0.96	<0.001	0.205

$$d(\bar{S}_l^L, \bar{S}_{l'}^L) = \sqrt{\int_0^{\omega_{\max}} \int_0^1 \{\bar{S}_l^L(\omega, t) - \bar{S}_{l'}^L(\omega, t)\}^2 \, \mathrm{d}\omega \, \mathrm{d}t}.$$

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 languages, using $M = 1000$ permutations. In the interpretation of these p-values, we need to account for the multiple tests that have been carried out. By applying a Bonferroni correction to the unadjusted p-values in Table 1, it can be seen that a significant difference can be found at least for French and American Spanish and thus we choose to account for this difference when modelling the sound changes. It may be surprising that for other languages there is little evidence to support a difference between word means but this might be ascribed to the small available sample of speakers.

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) showed 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, with different distances catching different properties of the covariance structure. In particular, they proposed 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: a compact operator with finite norm $||L||_{\mathrm{HS}} = \mathrm{tr}(L_i^* L_i)$. As this is a Hilbert space, distances between the transformed operators can be easily evaluated. However, the map is defined up to a unitary operator and a Procrustes matching is therefore needed to evaluate the distance between the two equivalence classes. Formally, let C_1 and C_2 be the covariance operators that 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) proved that the Procrustes reflection size-and-shape distance has the explicit analytic expression

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

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

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

$$\bar{C} = \arg \inf_C \sum_{L=1}^{G} d(C^L, C)^2.$$

Table 2. p-values of the permutation tests for $H_0: C^L_{\omega,1} = C^L_{\omega,2} = \ldots = C^L_{\omega,10}$ *versus* H_1: 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†

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

†The Procrustes distance is used for the test statistic.

Table 3. p-values of the permutation tests for $H_0: C^L_{t,1} = C^L_{t,2} = \ldots = C^L_{t,10}$ *versus* H_1: 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†

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

†The Procrustes distance is used for the test statistic.

This provides an estimate for the centre point of the distribution with respect to the distance $d(\cdot,\cdot)$, which is needed for the test statistics in the permutation test.

Using this procedure, we can verify whether the assumption that the covariance operators are the same across the words is disproved by the 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 that were described in Section 2, obtained with the 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 hypothesis that the covariance operator is the same for all words for all the languages considered. The same is true for the time covariance operator, as can be seen in Table 3, which reports the p-values of this second test.

A possible concern is that the dimension of the data set becomes relatively small when it is split between the different words and languages and therefore these testing procedures will have little power. However, this reasoning encourages us to simplify the model (assuming that covariance operators are constant across words) so that enough observations are available to estimate the parameters accurately. With a larger data set that enables us to highlight differences between wordwise covariance operators, we would have more information to estimate these operators accurately.

6. Exploring phonetic differences

We now have the tools to explore the phonetic differences between the languages in the Oxford Romance languages data set. This can be done at different levels. A possible way to go would be to pair two speakers belonging to two different languages and to look at their difference. However, this neglects the variability of the speech within the language and it would not be clear which aspects of the phonetic differences are 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 by looking at the path that links the mean of the log-spectrograms between two words of different languages. These should be two words that are known to be related in the languages' historical development. This is so for example for the same digit in any 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 that 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

$$\bar{S}_i(x) = \bar{S}_i^{L_1} + x(\bar{S}_i^{L_2} - \bar{S}_i^{L_1}), \tag{3}$$

where $x \in [0, 1]$ provides a linear interpolation from language L_1 to language L_2, whereas $x < 0$ or $x > 1$ provides an extrapolation in the direction of the difference between the two languages, with $\bar{S}_i^L$ being the mean of the log-spectrograms from speakers of the language L pronouncing the ith digit. For example, Fig. 4 shows six steps along a reconstructed path for the mean log-spectrogram of 'one', from French [œ̃] to Portuguese [ũ]. Indeed, this path has historical significance, as the sound change from Latin '*unus*' to French '*un*' probably went via the sound [ũ] (see, for example, Pope (1934), pages 176–177), which is still maintained in modern Portuguese (it should be noted that we are, of course, not implying that modern French is derived from Portuguese, but merely that a historical sound of modern French is maintained in Portuguese).

A natural question is whether this can be replicated for the covariance structure, to interpolate and extrapolate a more general description of the sound generation process. However, the case of the covariance structure is more complex. Experience with low dimensional covariance matrices (see Dryden *et al.* (2009)) and the case of the frequency covariance operators that were 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 the covariance operator. Moreover, since we model the covariance structure as separable, we also want 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,

$$C^x = C_\omega^x \otimes C_t^x,$$

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

$$C_r^x = [(C_r^{L_1})^{1/2} + x\{(C_r^{L_2})^{1/2}\tilde{R} - (C_r^{L_1})^{1/2}\}][(C_r^{L_1})^{1/2} + x\{(C_r^{L_2})^{1/2}\tilde{R} - (C_r^{L_1})^{1/2}\}]^*$$

where $r = \omega, t$ and $\tilde{R}$ is the unitary operator that minimizes $||(C_r^{L_1})^{1/2} - (C_r^{L_2})^{1/2}R||_{\text{HS}}^2$ (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 that were reported in Pigoli *et al.* (2014) suggest that the Procrustes reflection size-and-shape geodesic performs better in the extrapolation of frequency covariance operators than do existing alternatives.

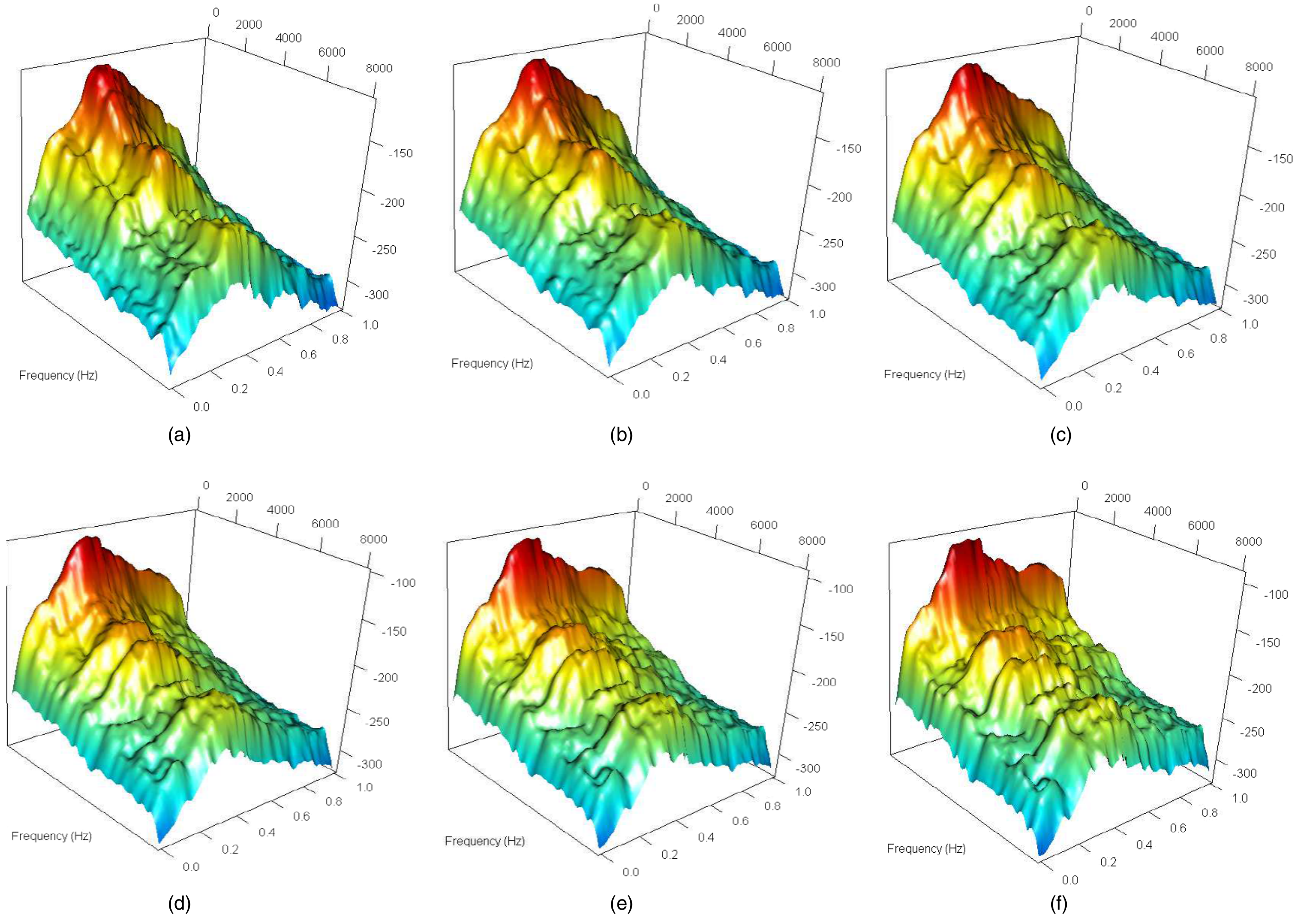


Fig. 4. Six steps along the smooth path between the mean log-spectrogram for the word (a) *un* (‘one’) in French and (f) the mean log-spectrogram for the word *um* (‘one’) in Portuguese: these are obtained from equation (3) for $x = 0, 0.2, 0.4, 0.6, 0.8, 1$

6.2. What would someone sound like speaking in a different language?

The framework that we have set up enables us also to observe how the sound that is produced by a speaker would be modified as we move to a different language. As mentioned in Section 1, we aim to map the sound that is produced by this speaker to that of a hypothetical speaker with the same position in the space of possible speakers in a different language, with respect to the language variability structure. To do this, we need some additional specification of the statistical model that generates the log-spectrograms. For example, if we assume that the log-spectrograms of a spoken word are 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 generally, we identify the population of possible 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 as a speaker-specific residual what is left in the phonetic data once means and covariance information have been removed. Let us denote with 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

$$S_{ik}^{L_1\to L_2}=[F_i^{L_2}]^{-1}\circ F_i^{L_1}(S_{ik}^{L_1}). \tag{4}$$

We choose to use the same word for both languages because in our data set words can be paired in a sensible way (the various pronunciations of the same digit in two Romance languages sharing a common historical origin).

The challenge now is how to define the transformation 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 *vis-à-vis* all the other speakers. A graphical representation of this idea for the case of a French speaker mapped to the Portuguese language can be seen in Fig. 5. To define this transformation, we start from a speaker k from the language L_1 and we consider the residual log-spectrogram $R_{ik}^{L_1}=S_{ik}^{L_1}-\bar{S}_{i.}^{L_1}$. We would now want to apply 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 defined via a linear combination of tensor basis functions $v_i^\omega\otimes v_j^t$, using p basis functions in each direction (time and frequency),

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

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

$$L_1\otimes L_2Z(\omega,t)=\int\int l_1(\omega,y)z(x,y)l_2(x,t)\,\mathrm{d}x\,\mathrm{d}y.$$

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

$$F_i^L(S_{ik}^L)=(C_\omega^L)^{-1/2}\otimes(C_t^L)^{-1/2}(S_{ik}^L-\bar{S}_{i.}^L)$$

and

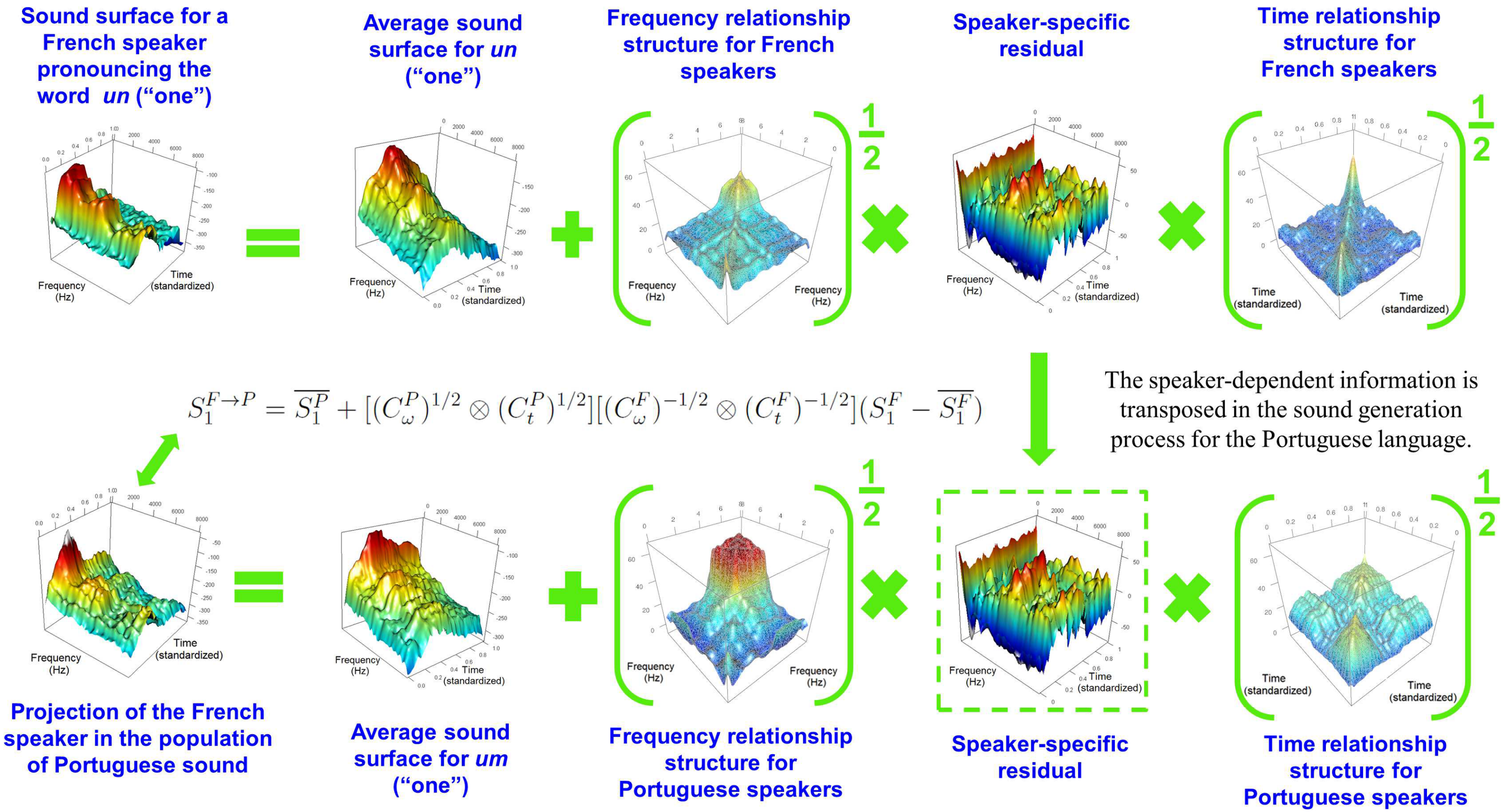

Sound surface for a French speaker pronouncing the word *un* ("one")
=
Average sound surface for *un* ("one")
+
Frequency relationship structure for French speakers
$\frac{1}{2}$
×
Speaker-specific residual
×
Time relationship structure for French speakers
$\frac{1}{2}$
$S_1^{F\to P} = \overline{S_1^P} + [(C_\omega^P)^{1/2} \otimes (C_t^P)^{1/2}][(C_\omega^F)^{-1/2} \otimes (C_t^F)^{-1/2}](S_1^F - \overline{S_1^F})$
The speaker-dependent information is transposed in the sound generation process for the Portuguese language.
Projection of the French speaker in the population of Portuguese sound
=
Average sound surface for *um* ("one")
+
Frequency relationship structure for Portuguese speakers
$\frac{1}{2}$
×
Speaker-specific residual
×
Time relationship structure for Portuguese speakers
$\frac{1}{2}$
Frequency (Hz)
Time (standardized)

Fig. 5. Example of the mapping of a French speaker's log-spectrogram to the same position in the space of Portuguese pronunciations for the corresponding word

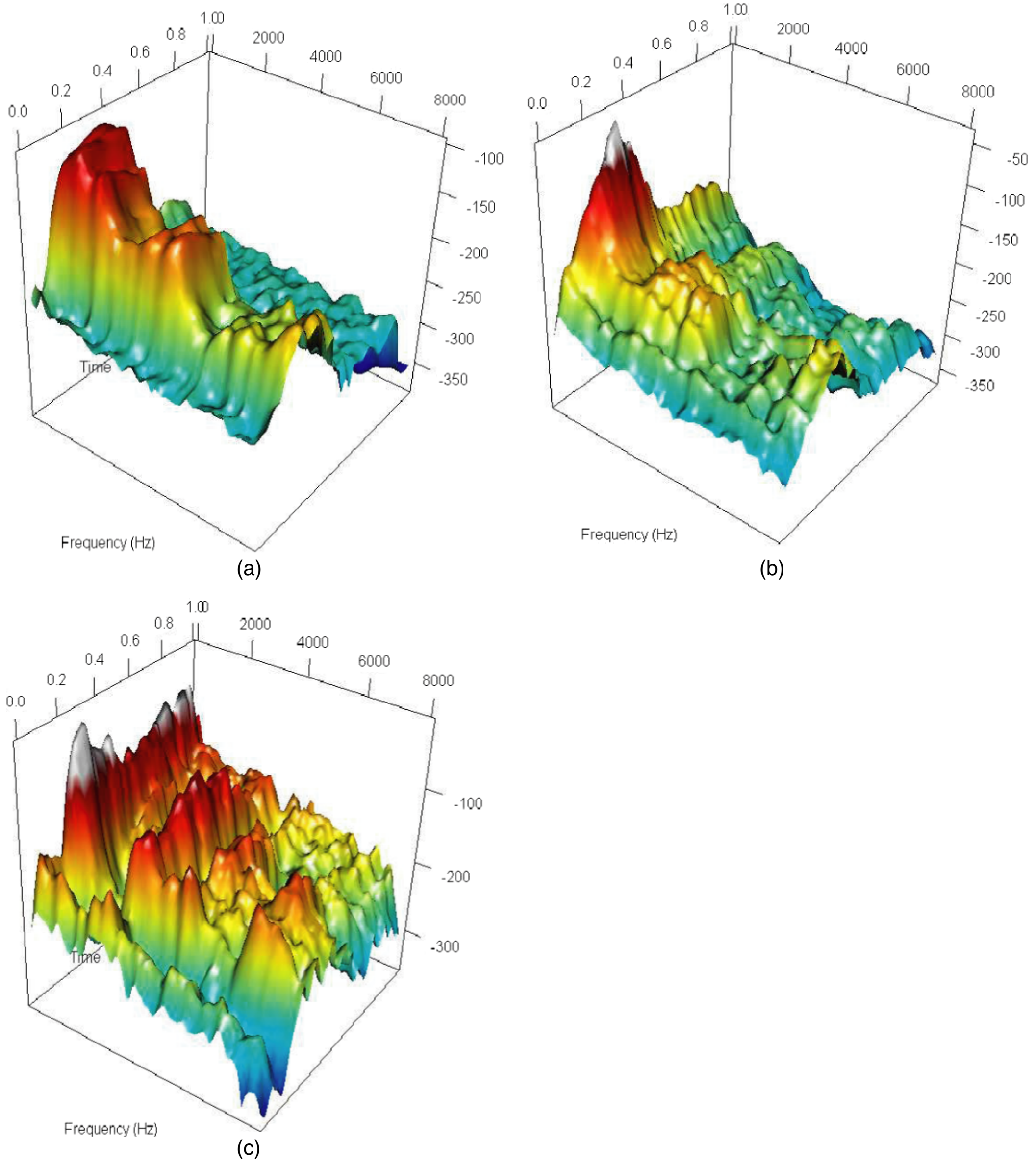


Fig. 6. Log-spectrograms for (a) the word *un* ('one') as spoken by a French speaker, (b) its representation as the word *um* ('one') in Portuguese by using equation (4) and (c) the closest observed word *um* ('one') spoken by a Portuguese speaker

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

Fig. 6 shows the log-spectrograms for the word *un* ('one') of the first French speaker S_{11}^{Fr}, its representation when mapped to Portuguese *um* ('one') $S_{11}^{\mathrm{Fr}\rightarrow\mathrm{P}}$ and the closest observed instance of Portuguese *um* as spoken by a Portuguese speaker, whereas Fig. 7 reports the result of the same operation applied to an Italian speaker, transforming Italian *uno* ('one') into Castilian Spanish *uno* ('one'). Though the spelling is the same in this case, the pronunciation of the word in the two languages is not identical, albeit similar.

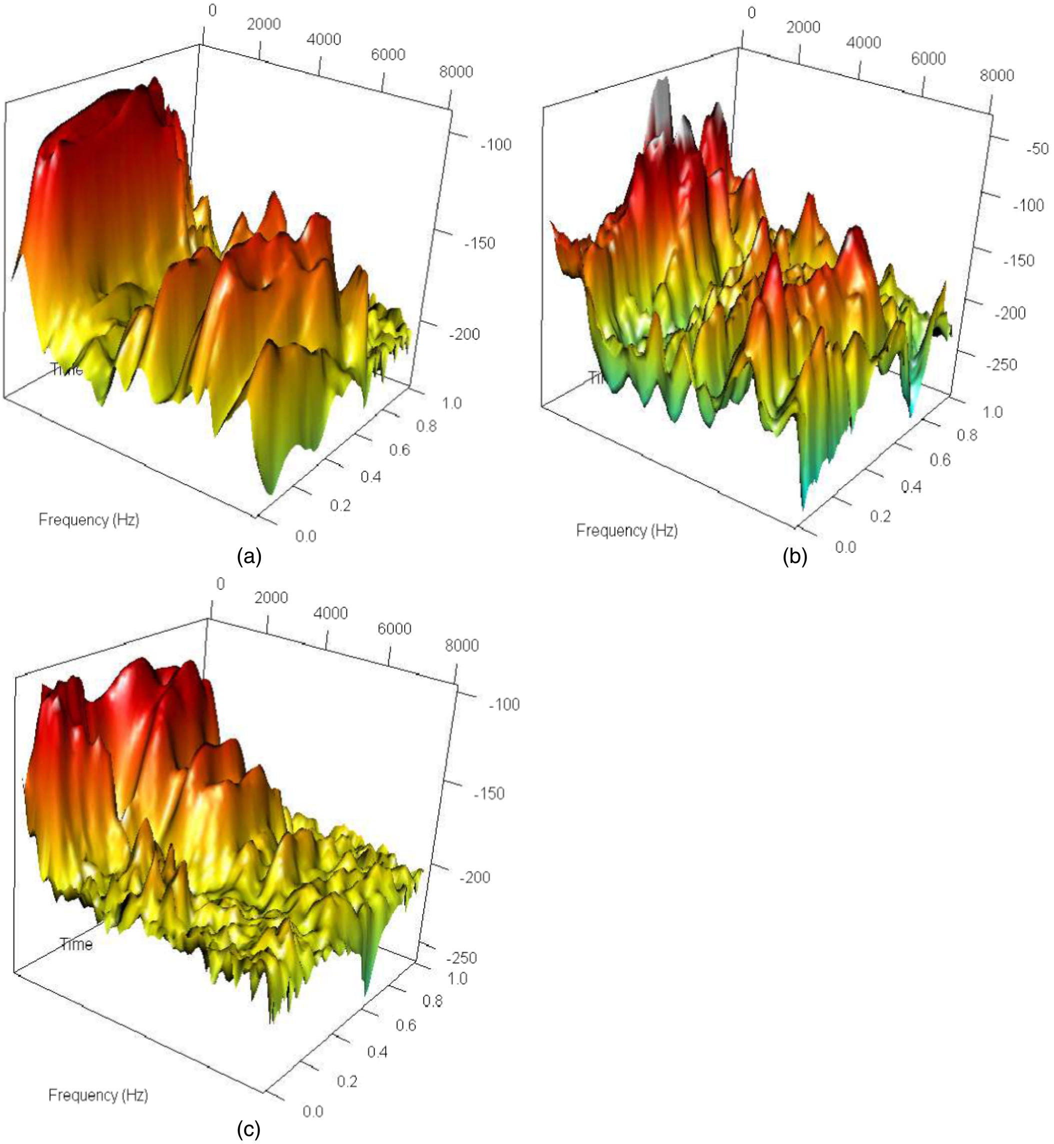


Fig. 7. Log-spectrograms for (a) the word *uno* ('one') as spoken by an Italian speaker, (b) its representation as the word *uno* ('one') in Spanish by using equation (4) and (c) the closest observed word *uno* ('one') spoken by a Spanish speaker

6.3. Interpolation and extrapolation of spoken phonemes

The representation of a speaker as they would sound when speaking another language is interesting but is not enough for scholars to explore the historical sequence of changes that occurred between two languages: a smooth estimate of the path of change is needed. This is also so if it is desired to extrapolate the sound transformation process beyond the path connecting the two languages, which we recall is a main goal of the 'Ancient sounds' project. Luckily, we can use the interpolated means and covariance operators that were described above to characterize the unobserved possible languages that are the intermediate steps in the phonetic path between two given languages. We thus obtain a smooth path between $S_{ik}^{L_1}$ and its representation in the language L_2 as

$$S_{ik}^{L_1 \to L_2}(x) = [F_i^x]^{-1} \circ F_i^{L_1}(S_{ik}^{L_1}), \qquad (5)$$

$[F_i^x]^{-1} = (C_\omega(x))^{1/2} \otimes (C_t(x))^{1/2} Z + M(x)$, 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* as spoken by the same French speaker as considered in the previous section and its corresponding acoustic representation in Portuguese can be seen in Fig. 8.

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

$$S_{ik \to ik'}^{L_1 \to L_2}(x) = [F_i^x]^{-1}\{x F_i^{L_1}(S_{ik}^{L_1}) + (1-x) F_i^{L_2}(S_{ik'}^{L_2})\}, \qquad (6)$$

where a linear interpolation between the residuals takes the place of the residual of the single language. This could be useful when it is meaningful to pair two log-spectrograms in different languages, for example because the same speaker is recorded in two languages. This is not so in our data set, but by way of example we report in Fig. 9 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 who is closest to the transformed $S_{11}^{\mathrm{Fr} \to \mathrm{P}}$. It is also interesting to compare this with the interpolated path between the two mean log-spectrograms in Fig. 4.

Being able to extrapolate the sounds opens up interesting possibilities whenever two languages are known to be at two stages of an evolutionary path. In this case extrapolating in the direction of the older (i.e. linguistically more conservative) language can provide an insight into the phonetic characteristics of the extinct ancestor languages. This, of course, will require some integration into a model of sound change with such information as that coming, for example, from textual analysis, history or archaeology (e.g. dating studies). This is also needed as the rate of change of languages is not constant and the path $S^{L_1 \to L_2}(x)$ can be travelled at different speeds for different branches of the language family's evolution, and it can be changed by events such as conquests, migrations and language contact. However, by having a path in the first place, addressing such questions is now a possibility.

6.4. Back to sound reproduction

Visualizing the log-spectrograms (or other transformation of the recorded sounds) is helpful but it is also important to listen to the signals in the original domain. This is also true for the representation of a sound in a different language and the smooth paths that we have defined. Thus, we would like to reconstruct actual audible sounds from the estimated log-spectrograms. To do this, we would also need information about the phase component that we have so far disregarded, since we have focused all our attention on the amplitude component of the Fourier transform (see Section 2). In principle, we could perform a parallel analysis on the phases to obtain a representation of phase in a different language, the 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 in any case there is no reason to believe that additional information is captured by phase (human hearing is largely insensitive to phase so it is quite normal practice in acoustic phonetics to disregard the phase component; see, for example, Kent and Read (2002)). In practice, we use the phase that is associated with the log-spectrogram $S_{ik}^{L_1}$ to reconstruct the sounds over the smooth path; the results are quite satisfactory. Some examples of reconstructed sound paths can be found in the on-line supplementary material. In particular, the audio file

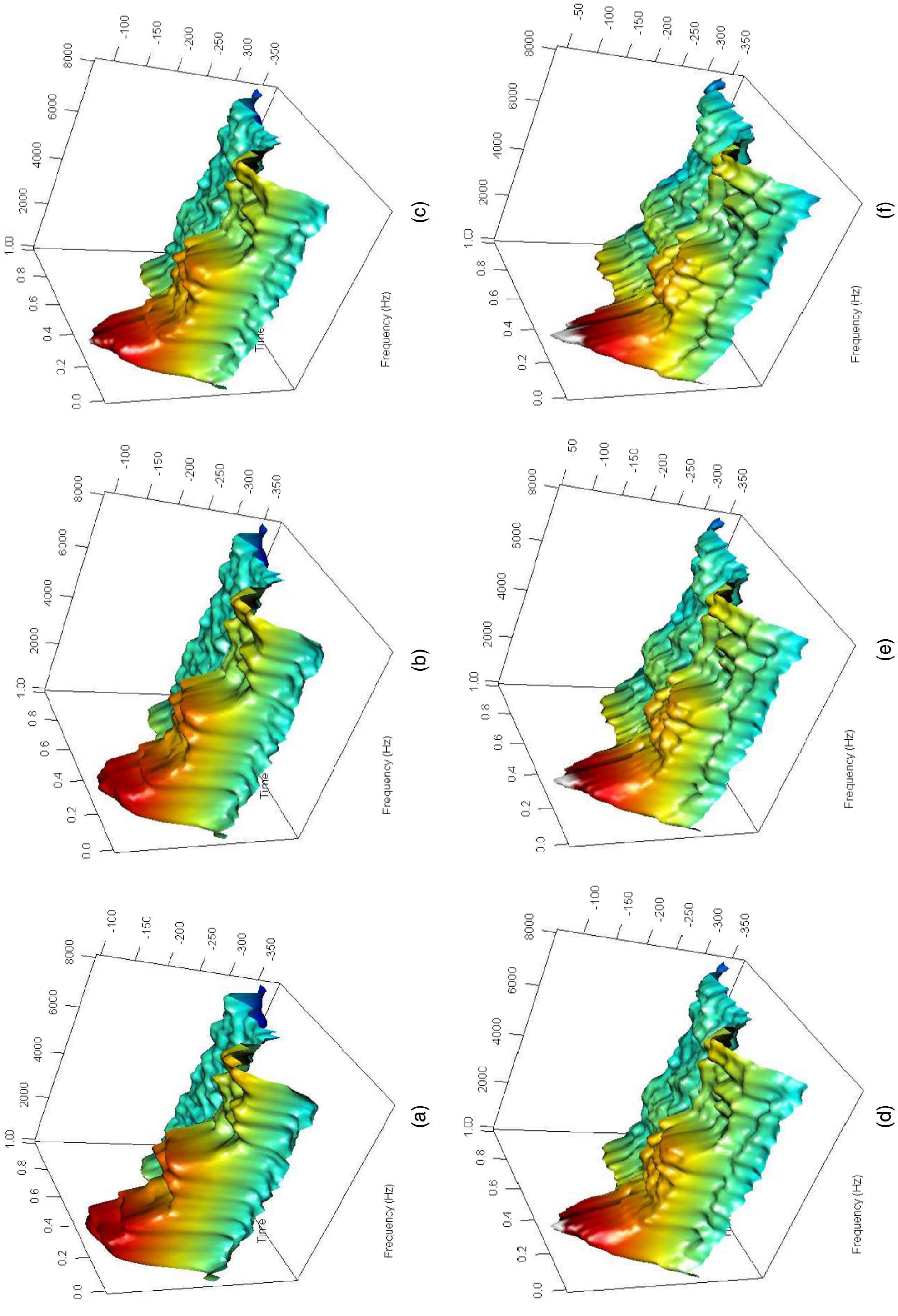
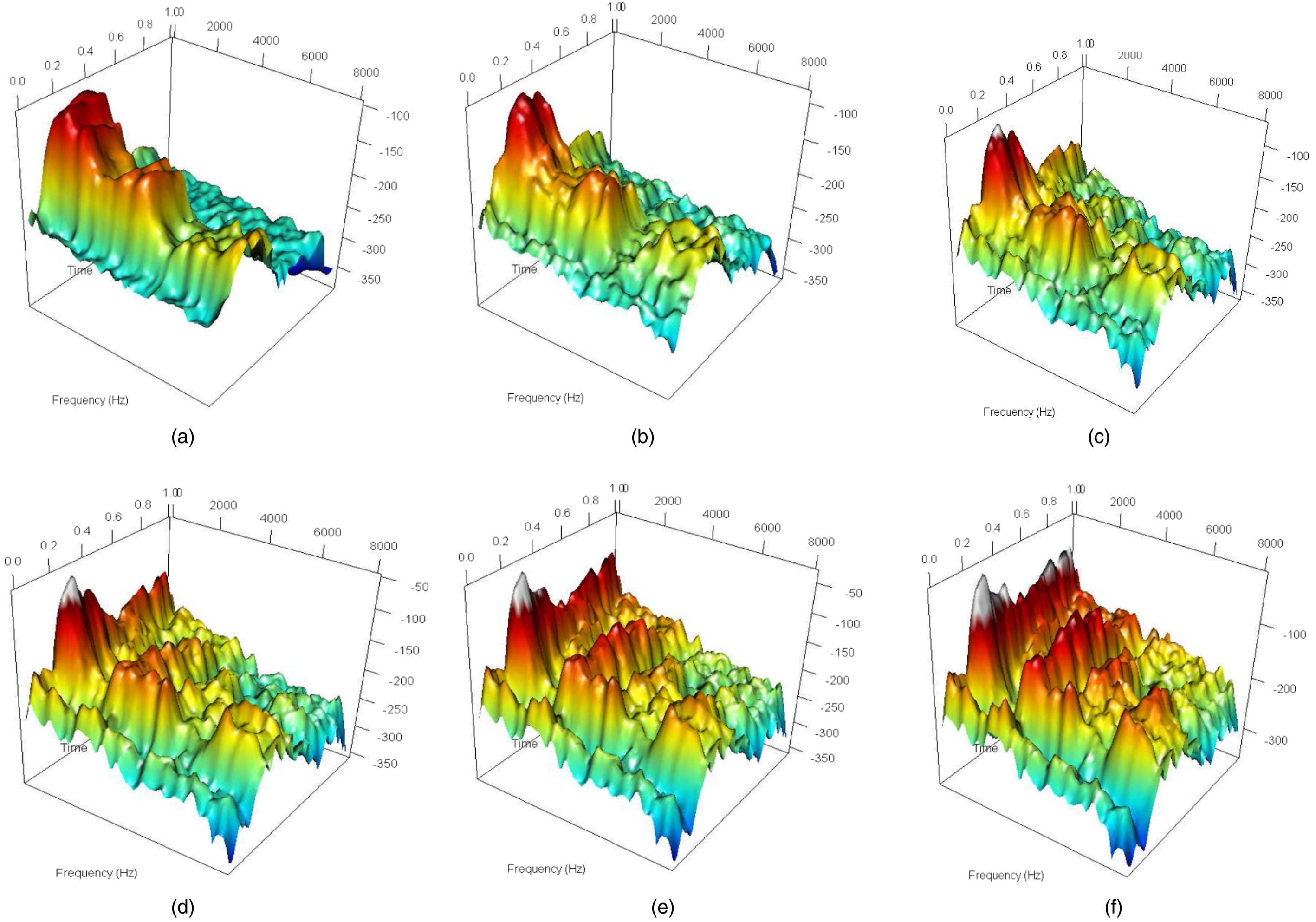


Fig. 8. Six steps along the smooth path between the log-spectrogram for (a) the word *un* ('one') as spoken by a French speaker and (f) its representation in Portuguese: these are obtained from equation (5) for $x = 0, 0.2, 0.4, 0.6, 0.8, 1$



(a) (b) (c)

(d) (e) (f)

Fig. 9. Six steps along the smooth path between the log-spectrograms for (a) the word *un* (‘one’) as spoken by a French speaker and for (f) the word *um* (‘one’) closest to its transformed representation in Portuguese: these are obtained from equation (6) for $x = 0, 0.2, 0.4, 0.6, 0.8, 1$

F2P_path_digit5.wav contains the reconstructed sound for the path $S_{5,1}^{\mathrm{Fr}\rightarrow\mathrm{P}}(x)$, $x=(0,1)$, connecting the spoken word *cinq* ('five') uttered by a French speaker with its projection into the Portuguese language. The audio file F2P_path_digit7.wav contains the corresponding path for the digit 'seven' (French *sept* to Portuguese *sete*). As mentioned above, in our data set we do not have meaningful connections between speakers of different languages. However, for comparison, we report in the audio file F2P_spk_digit5.wav the reconstructed sounds for the path $S_{5,1\rightarrow 5,2}^{\mathrm{Fr}\rightarrow\mathrm{P}}(x)$, $x=(0,1)$, which connects the spoken word for 'five' from a French speaker with that from a Portuguese speaker. Similarly, the audio file F2P_spk_digit7.wav contains the sound path between two different speakers (one French and one Portuguese) for the digit 'seven'. We leave it to the readers to form their own appraisal of the satisfactoriness or the plausibility of these audio transformations.

7. Discussion

We have introduced a novel way to explore phonetic differences and changes between languages that takes into account the characteristics of the sound population on the basis of actual speech recordings. The framework that we introduced is useful for dealing 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 method proposed with an application to a Romance digit data set, which includes the words corresponding to the numbers from 1 to 10 pronounced by speakers of five different Romance languages. In particular, we verify that the assumption that the covariance structure in the log-spectrograms is common for the different words within the language is tenable in this data set, thus increasing the sample that is available for its estimation. This is an interesting example of how the characteristics of a population (in this case the speakers of 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 as it captures the notion of 'the sound of a language'. It also fits within the recent development of object-oriented data analysis (see Wang and Marron (2007)), which advocates a careful consideration of the object of interest for a statistical analysis. Here it seems that marginal covariance operators are promising features to represent phonetic structure at the level of a language.

We do not focus here on the representativeness or otherwise of the sample of speakers or words in the data set. In view of a broad use of this approach, however, it is important to remember that the sample of speakers should reflect the population that we are interested in and, in particular, careful consideration should be given to regional and social stratification in the data set. Moreover, to speak properly of a 'language' (and not just of a small subset of words), the words that are considered should be representative of the whole language. The digits that are studied here do contain a wide ranging set of different vowels and consonants (for just a few words), indicating that the results are likely to be generalizable to some extent across a larger *corpus*, but, of course, applying this to a much more comprehensive *corpus* of several languages would be advisable.

The approach proposed, using audio recordings in place of textual representations, enables us to account for the differences between varieties of the same language, such as Castilian Spanish and American Spanish (Penny, 2000). Moreover, recent works (see Functional Phylogenies Group (2012), Bouchard-Côté *et al.* (2013), Coleman *et al.* (2015) and references therein) focus on the reconstruction of the distribution of phonetic features for ancestor languages. Although the research in this field is still in its very earliest stages, as a better understanding of the historical evolution of sounds becomes available, this can be integrated into our methods to provide a reconstruction of how the speakers of extinct languages might have sounded. The final goal is therefore to integrate our approach to the modelling of the variability of speech within the language with the dynamics of sound change established by other research both in linguistics and in statistics. We are confident that this will make a substantial contribution to the on-going project to create audible reconstruction of words in the protolanguages.

We have illustrated the transformation of a speaker's speech from one language to another as a first example 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 discrete or continuous covariates that are associated with the mean and the covariance operators. These can be seen as functions of the geographical co-ordinates or of time depth when studying dialects. Although we treated the language as a categorical variable, nothing prevents us from seeing it as a continuous process in space and time. Indeed, the definition of the continuous path between two languages that was 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. Although we do not claim that this can straightforwardly reproduce the evolutionary branches in language history, it can still be a useful starting point for more complex models.

The application of the method proposed 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 microvariation within a language (dialects; spoken language in different subgroups of the population) but also other types of sounds such as songs or even sounds that are different from human speech, e.g. animal calls.

8. Supplementary material

The file `Acoustic_Data_and_Code.zip`, which is available from `https://rss.onlinelibrary.wiley.com/hub/journal/14679876/series-c-datasets`, contains all the code and data that are required to reproduce the analysis in the paper. The file `README.txt` describes the purpose of all the files in the folder. The file `SupplementaryMaterial.pdf` reports the results of the analysis carried out with alternative methods for data preprocessing.

Acknowledgements

John Coleman appreciates the support of UK Arts and Humanities Research Council grant AH/M002993/1, 'Ancient Sounds: mixing acoustic phonetics, statistics and comparative philology to bring speech back from the past'. John Aston appreciates the support of UK Engineering and Physical Sciences Research Council grant EP/K021672/2, 'Functional object data analysis and its applications'.

References

Aston, J. A. D., Chiou, J.-M. and Evans, J. P. (2010) Linguistic pitch analysis using functional principal component mixed effect models. *Appl. Statist.*, **59**, 297–317.

Aston, J. A. D., Pigoli, D. and Tavakoli, S. (2017) Tests for separability in nonparametric covariance operators of random surfaces. *Ann. Statist.*, **45**, 1431–1461.

Blackledge, J. M. (2006) *Digital Signal Processing: Mathematical and Computational Methods, Software Development and Applications*. Amsterdam: Elsevier.

Bouchard-Côté, A., Hall, D., Griffiths, T. L. and Klein, D. (2013) Automated reconstruction of ancient languages using probabilistic models of sound change. *Proc. Natn. Acad. Sci. USA*, **110**, 4224–4229.

Cavalli-Sforza, L. L. (1997) Genes, peoples, and languages. *Proc. Natn. Acad. Sci. USA*, **94**, 7719–7724.

Coleman, J., Aston, J. and Pigoli, D. (2015) Reconstructing the sounds of words from the past. In *Proc. 18th Int. Congr. Phonetic Sciences, Glasgow*. Scottish Consortium for International Congress of Phonetic Sciences.
Cooke, M., Beet, S. and Crawford, M. (1993) *Visual Representations of Speech Signals*. Chichester: Wiley.
Dryden, I. L., Koloydenko, A. and Zhou, D. (2009) Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *Ann. Appl. Statist.*, **3**, 1102–1123.
Ferraty, F. and Vieu, P. (2006) *Nonparametric Functional Data Analysis: Theory and Practice*. Berlin: Springer.
Functional Phylogenies Group (2012) Phylogenetic inference for function-valued traits: speech sound evolution. *Trends Ecol. Evoln*, **27**, 160–166.
Garcia, D. (2010) Robust smoothing of gridded data in one and higher dimensions with missing values. *Computnl Statist. Data Anal.*, **54**, 1167–1178.
Ginsburgh, V. and Weber, S. (2011) *How Many Languages Do We Need?: the Economics of Linguistic Diversity*. Princeton: 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. *J. Acoust. Soc. Am.*, **131**, no. 6, article 4651.
Kent, R. and Read, C. (2002) *Acoustic Analysis of Speech*, 2nd edn. London: Singular.
Koenig, L. L., Lucero, J. C. and Perlman, E. (2008) Speech production variability in fricatives of children and adults: results of functional data analysis. *J. Acoust. Soc. Am.*, **124**, 3158–3170.
Marron, J. S., Ramsay, J. O., Sangalli, L. M. and Srivastava, A. (2014) Statistics of time warpings and phase variations. *Electron. J. Statist.*, **8**, 1697–1702.
Morpurgo Davies, A. (1998) *Linguistics in the Nineteenth Century*. London: Longman.
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. *Nat. Rev. Genet.*, **10**, 405–415.
Penny, R. J. (2000) *Variation and Change in Spanish*. Cambridge: Cambridge University Press.
Pigoli, D., Aston, J. A. D., Dryden, I. L. and Secchi, P. (2014) Distances and inference for covariance operators. *Biometrika*, **101**, 409–422.
Pope, M. K. (1934) *From Latin to Modern French with Especial Consideration of Anglo-Norman: Phonology and Morphology*. Manchester: Manchester University Press.
Ramsay, J. O. and Silverman, B. W. (2005) *Functional Data Analysis*, 2nd edn. New York: Springer.
Srivastava, A., Wu, W., Kurtek, S., Klassen, E. and Marron, J. S. (2011) Registration of functional data using the Fisher-Rao metric. *Preprint arXiv:1103.3817v2*. Florida State University, Gainesville.
Tang, R. and Müller, H. G. (2008) Pairwise curve synchronization for functional data. *Biometrika*, **95**, 875–889.
Tucker, J. D. (2014) fdasrvf: elastic functional data analysis. *R Package Version 1.4.2.* (Available from `https://CRAN.R-project.org/package=fdasrvf`.)
Wang, H. and Marron, J. S. (2007) Object oriented data analysis: sets of trees. *Ann. Statist.*, **35**, 1849–1873.
Wood, S. N. (2003) Thin plate regression splines. *J. R. Statist. Soc.* B, **65**, 95–114.

Discussion on the paper by Pigoli, Hadjipantelis, Coleman and Aston

Mario Cortina-Borja (*University College London Great Ormond Street Institute of Child Health, London*)
There is a long tradition of statistical analyses in linguistics. Well-known examples include the classification of language families based on lexical evidence (Nicholls and Gray, 2008), sociolinguistic studies (Maaegard *et al.*, 2013) and modelling word frequency distributions studied from both stylometric and linguistic points of view (Baayen, 1992). Recently, in a wider context, statistical models have been increasingly used in both natural language processing (Manninig and Schuetze, 1999) and text mining (Silge and Robinson, 2017). This paper deals with acoustic, rather than textual, phonological, grammatical or morphological aspects of language, and thus complements those approaches. It discusses methods based on functional data analysis over the frequency and time domains. Its main objective is to model changes and variations in recordings from the same word said by native speakers of different languages. Pigoli and his colleagues develop a modelling framework to investigate what language sounds like after accounting for individual and between-language acoustic differences. This in itself is quite complicated but, to judge it properly, their paper must be seen as part of an ambitious multidisciplinary project whose ultimate target is to '[mix] acoustic phonetics, statistics and comparative philology to bring speech back from the past'. In this paper,

Supporting information
Additional 'supporting information' may be found in the on-line version of this article:

'Supplementary material for the paper The statistical analysis of acoustic phonetic data: exploring differences between spoken Romance languages'.

they first use interpolation to construct a smooth path between two languages. This helps to define how one speaker would sound in another language. Besides, the paper explores extrapolation of the sound transformation beyond that connecting path. The idea is that, by moving in the direction of the differences between both connected languages, the model might reconstruct spoken forms of earlier versions of the words considered. This would require, however, models of sound change informed by textual analysis, archaeology and historical representations of language.

I would like to raise three specific concerns with the paper. The main one is about possible power issues for the permutation tests whose results appear in Tables 1–3. How confident are the authors in the results presented there, especially those related to the covariance structures? Secondly, the log-spectrograms seem to show a discrete effect on the frequency domain. If that were so, can this be accounted for in the models? Finally, which other *corpora* may be used for calibrating and applying this method to larger data sets? Do the authors plan to study larger samples from native speakers of a larger number of languages? Also, it would be fascinating to analyse in this context speech data from bilingual speakers.

I have two speculative questions. Firstly, have the authors thought of applying their methods to other types of data expressed along frequency and time domains? For example, epidemiological or clinical studies analysing data from wearable computing devices might benefit from this approach. Lastly, how does the current paper contribute to the ultimate goal of reconstructing not only sounds from ancestors in the Indo-European family but also from older versions of the same language?

Defining which language is one's own is not as straightforward as it may seem. For instance, the poet Octavio Paz said

> 'I consider myself a citizen of the Spanish language rather than a Mexican citizen; that's why it bothers me to hear speak of the Castilian language, because Castilian belongs to the Castilians and I am not one of them; I am a Mexican and as such I speak Spanish and not Castilian'

(cited in Salvador (1987), page 92). This reflection illustrates the ambiguities which appear from the very start of any attempt to analyse language. Pigoli, Hadjipantelis, Coleman and Aston have successfully dealt in this paper with considerable theoretical and modelling complexities derived from the key question of their research project. It gives me much pleasure to conclude by congratulating them on an impressive piece of work and to propose a vote of thanks for this thought-provoking paper.

Stephen Cox (*University of East Anglia, Norwich*)
I have several reservations about this work. The main one is that the authors' claim that they have discovered a metric (the frequency covariance matrix) that shows 'significant linguistic features' is mistaken. In my opinion, the effects that they have observed are due to variation within the speaker population, and possibly also to recording conditions.

The authors stress from the outset

> 'In particular, we are interested in the case where multiple speakers from each language are included in the data set, since this enables better statistical exploration of the phonetic characteristics of the language'.

Yet the data set that they use contains, in speech science terms, a tiny number of talkers, an average of five for each language they are using, and only three for one language. In the speech science world, we typically use hundreds of hours of data from hundreds of speakers for research, so that variation from different speakers, one of the banes of speech science, is averaged out.

The authors say

> 'The aim of our work is to provide a framework where ... the variability of speech within the language can be considered'.

The problem is that the variability between these five Romance languages when only the digits vocabulary is considered is very small and will be dwarfed by the acoustic variability between the speakers. This variability is caused primarily by physiological variations, and also by accent variations. Even when using means across speakers, the number of speakers is so small that any results are still dependent on the characteristics of the particular very small set of speakers who are selected in the recordings.

The recordings are another problem.

> 'The sources of the readings were either collected from freely available language training Web sites or standardized recordings made by university students. As this data set consists of recordings made under non-laboratory settings, large variabilities may be expected within each group'.

In fact, the most likely scenario is that the set of recordings for a particular language was recorded under the same conditions (microphone, distance to microphone, room, amplifier etc.), and the differences between these recording conditions may be substantial, as they were not recorded under controlled studio conditions. What the authors are identifying as ‘language differences’ may simply be differences in the recording channels, a mistake that was made in the early days of work in the technology of speaker verification. Deconvolving the effects of a channel has been the subject of years of research in that technology.

The authors’ acoustic analysis displays a lack of knowledge of speech processing techniques. To use the Fourier coefficients of a narrow-band analysis is naive, as

(a) these are linear in frequency, so low frequency coefficients cover a bandwidth of an octave or so but high frequency coefficients less than half a semitone, and
(b) they are dominated by the pitch of the speech, a speaker-dependent artefact that they need to exclude as much as possible in their analysis.

The use of Mel frequency cepstral coefficients alleviates both of these problems, and there are other alternatives.

The authors claim that this method can be ‘applied in a straightforward way to larger and more comprehensive *corpora*’. But the method relies on the languages under comparison having very similar words for the same object. In their case, it is clear that the digits in these five languages are very similar and have the same linguistic root. But, if one wanted to extend the technique to other words, who would decide whether a word was sufficiently similar or not to use? Would it work for, say, the digits in English and German?: what about English and Russian?

The authors claim to have discovered a measure (frequency covariance) that is invariant for, each language. However, I replicated their experiments by using English digits spoken by two different groups of six speakers, using as far as possible the same frequency analysis, word alignment (I used dynamic programming rather than the warping functions that they used, but the result is the same) and covariance estimation. The frequency covariance matrices that I produced were different for each group, suggesting that they were purely a result of the different speakers (recording conditions were identical here). However, the authors could test their claim that the covariance structure was representative of language by doing some language identification experiments that used this as a ‘feature’.

I have no comments on the statistics in the paper, but this very sophisticated analysis seems to me to have been used on data that are very limited and poorly processed to produce conclusions that are not substantiated. It does not seem to me to be possible to say anything meaningful about the acoustic differences in five languages based on about 150 s of data from a handful of speakers. I would encourage the authors to submit their work to a speech science journal to verify their techniques and findings and to obtain feedback from other workers in this field.

The vote of thanks was passed by acclamation.

Bernard W. Silverman (*University of Nottingham and University of Oxford*)
It is very exciting to see a paper which takes on board the philosophy of functional data analysis, trying to get to grips with the complexities of the kinds of data that we regularly collect nowadays. I would take gentle issue with the seconder of the vote of thanks, who asserted that speech science was distinct from statistics. Of course it is not, because speech science is functional data analysis writ large and a very interesting statistical area. It would be tragic if it were believed that it was somehow a completely different discipline.

The tapered Fourier series approach is one way of representing the data by using basis functions. An interesting question is whether one could use bases constructed specifically for the purpose, e.g. making use of known physiological properties of the speech-generating mechanisms.

The covariance hypothesis itself is one that could be tested by looking at the very large *corpora* of speech data rather than just the fairly small sample discussed here. Another possibility for future work is to analyse how native speakers of one language speak another. Assuming that the covariance hypothesis is correct, do they bring the covariance of their native language with them?

Finally, I have a question about Table 2. Three of the five languages yield p-values which are very close to 1. The most famous example of data which fit the null hypothesis suspiciously closely was the genetics publication of Mendel (1866), discussed in detail by Fisher (1936) who wrote

> ‘the data of most, if not all, of the experiments have been falsified so as to agree closely with Mendel’s expectation’.

See also Hartl and Fairbanks (2007). I am sure tonight's authors have not done anything like that, but it would be good if they could cast some light on their results!

J. P. Meagher and T. Damoulas (*University of Warwick, Coventry*), **K. Jones** (*University College London*) **and M. A. Girolami** (*Imperial College London*)

We offer our congratulation to Pigoli, Hadjipantelis, Coleman and Aston on this excellent contribution. The application of functional data analysis to a time–frequency representation of sound presents many exciting research directions in applied linguistics and beyond. A central motivation of the paper is to model the historical development of language through the statistical analysis of present day speech recordings. This process is presented by the authors as being analogous to biological evolution, and so we consider how phylogenetic Gaussian processes (Jones and Moriarty, 2013; Hadjipatelis *et al.*, 2013) could be applied to these smooth spectrogram surfaces for evolutionary inference.

Spectrogram surface over a phylogeny

For each language L, consider the sample of smoothed spectrogram surfaces representing the digit 1. Denote the language phylogeny, representing the evolutionary relationships between languages, as $\mathcal{P}$. Each language corresponds to a position on the phylogeny and so denote $S_{1k}(w,t,\mathbf{p}) \equiv S^L_{1k}(w,t)$, for $\mathbf{p} \in \mathcal{P}$.

The spectrogram surfaces can now be thought of as function-valued traits and modelled as phylogenetic Gaussian processes. This means that, subject to the simplifying assumption detailed by Jones and Moriarty (2013), we can write

$$S_{1k}(w,t,\mathbf{p}) = \mu(w,t) + \sum_{q=1}^{Q} \phi_q(w,t) f_q(\mathbf{p}) + \phi_q(w,t) z_q$$

where $\mu(w,t)$ is the mean spectrogram surface, $f_q(\mathbf{p}) \sim \mathcal{GP}\{0, k_q(\mathbf{p},\mathbf{p}')\}$, $z_q \sim \mathcal{N}(0,\sigma_q)$ and ϕ_q is the qth basis function.

Assume an Ornstein–Uhlenbeck process prior for $f_q(\mathbf{p})$, i.e.

$$k_q(\mathbf{p},\mathbf{p}') = \gamma_q \exp\left\{ -\frac{d_{\mathbf{P}}(\mathbf{p},\mathbf{p}')}{l_q} \right\}$$

where $d_{\mathcal{P}}(\mathbf{p},\mathbf{p}')$ is the distance between points $\mathbf{p}$ and $\mathbf{p}'$ over $\mathcal{P}$.

Given the set of hyperparameters $\theta_q = (\gamma_q, l_q, \sigma_q)$, $q = 1, \ldots, Q$, we have a posterior distribution for the smooth spectrogram representation of the digit 1 for any $\mathbf{p} \in \mathcal{P}$.

Application to bat echolocation

The model outlined above maps directly to a very different research question, that being, what did ancestral bat echolocation calls sound like? The model is illustrated in Fig. 10 where the analysis in Meagher *et al.* (2018) is repeated for smooth spectrogram surface representations of echolocation calls.

Looking ahead

In reconstruction ancestral acoustic phonetic states for language, it is not immediately clear what the appropriate unit for reconstruction should be. One approach would be to perform reconstructions word by word as outlined above; however, this would seem a haphazard approach. Perhaps ancestral reconstruction of the covariance structure would be more informative. In any case, we look forward to seeing how this work develops and informs research in linguistics, evolutionary biology or any domain where acoustic phonetic data are of interest.

Shahin Tavakoli (*University of Warwick, Coventry*)

I congratulate Pigoli, Hadjipantelis, Coleman and Aston for their inspiring paper. I have a couple of comments.

Within- and between-word sound covariance

The paper's finding is that the within-word sound covariance seems to be equal for all the words available, and that the within-word 'marginal covariance operators are promising features to represent phonetic structure at the level of a language'. From personal reflection, it seems that the within-word sound covariance (or more generally patterns of variations) is an important feature that helps us to *distinguish* between the sounds of different words in the same language. However, when it comes to what a language *sounds like*, it seems that the between-word covariance (or pattern of variation) is an important feature.

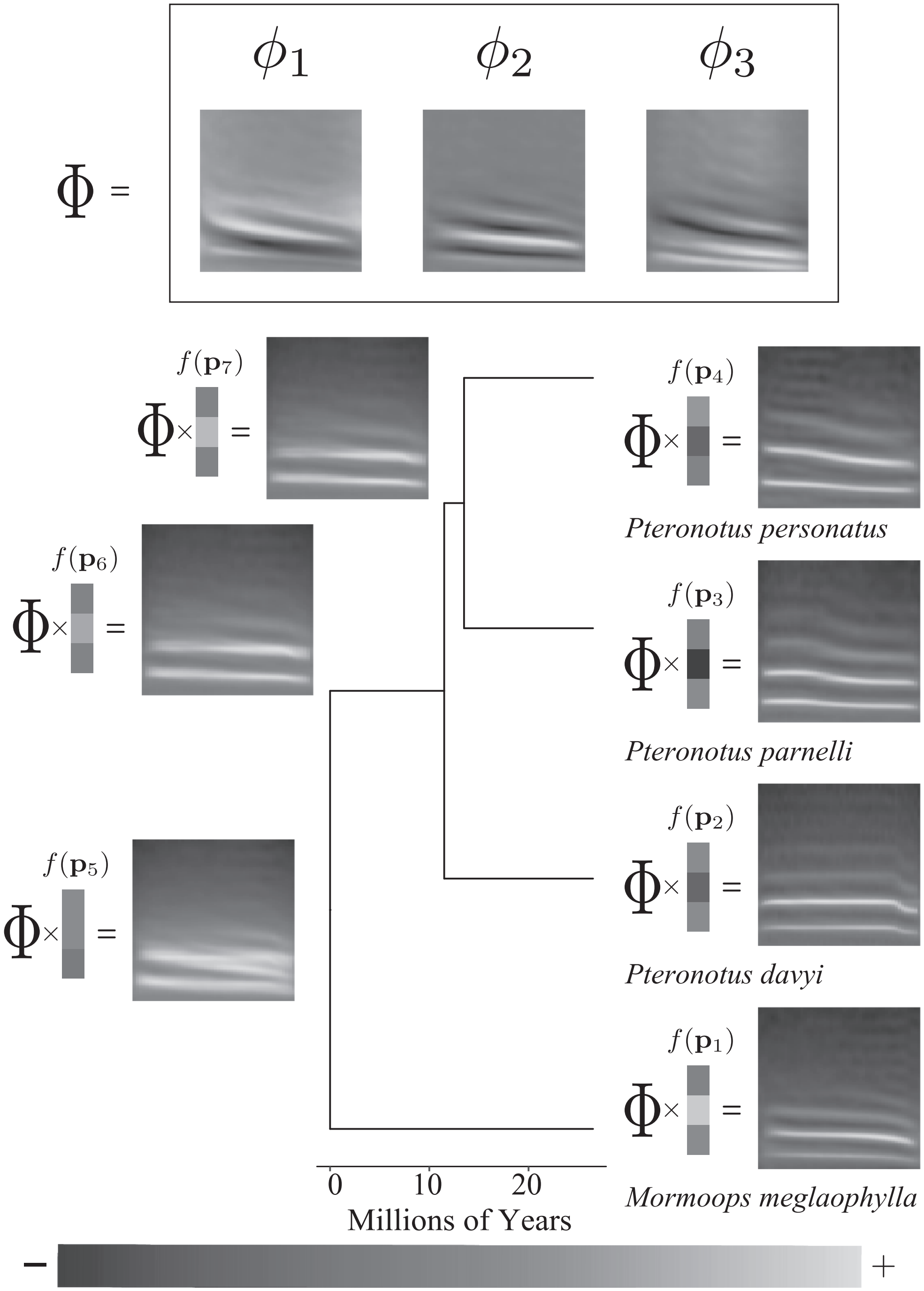
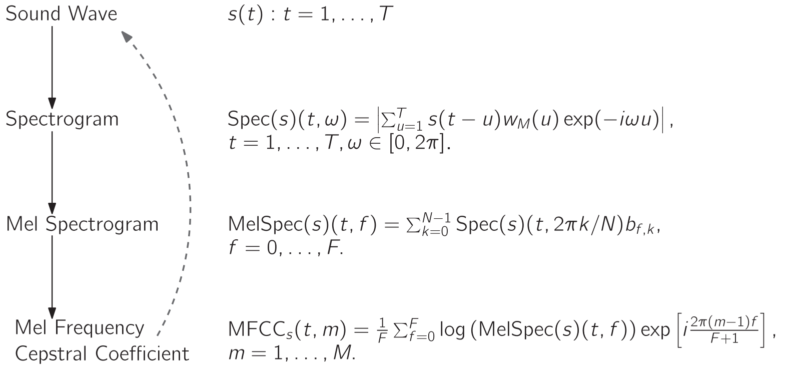


Fig. 10. An ancestral reconstruction for the echolocation call spectrogram for four species of bat from the *Molossidae* family: the set of ‘evolutionary basis’, $\mathbf{\Phi} = \{\phi_1, \phi_2, \phi_3\}$, has been inferred by a principal component analysis of echolocation spectrogram surfaces; $f(\mathbf{p}_n)$ denotes the phylogenetic Gaussian process of basis weights for the species at point $\mathbf{p}_n$ on the phylogeny, where the weight of each basis is modelled as an independent, univariate Ornstein–Uhlenbeck process; spectrogram surfaces at interval nodes represent maximum *a posteriori* estimates of the ancestral echolocation call spectrogram



$$s(t) : t = 1, \dots, T$$

$$\mathrm{Spec}(s)(t, \omega) = \left|\Sigma_{u=1}^{T} s(t-u) w_M(u) \exp(-i\omega u)\right|, \quad t = 1, \dots, T, \omega \in [0, 2\pi].$$

$$\mathrm{MelSpec}(s)(t, f) = \Sigma_{k=0}^{N-1} \mathrm{Spec}(s)(t, 2\pi k/N) b_{f,k}, \quad f = 0, \dots, F.$$

$$\mathrm{MFCC}_s(t, m) = \tfrac{1}{F} \Sigma_{f=0}^{F} \log\left(\mathrm{MelSpec}(s)(t, f)\right) \exp\left[i \tfrac{2\pi(m-1)f}{F+1}\right], \quad m = 1, \dots, M.$$

Fig. 11. Basic implementation of the MFCC of a sound wave: the $b_{f,k}$s are filter banks

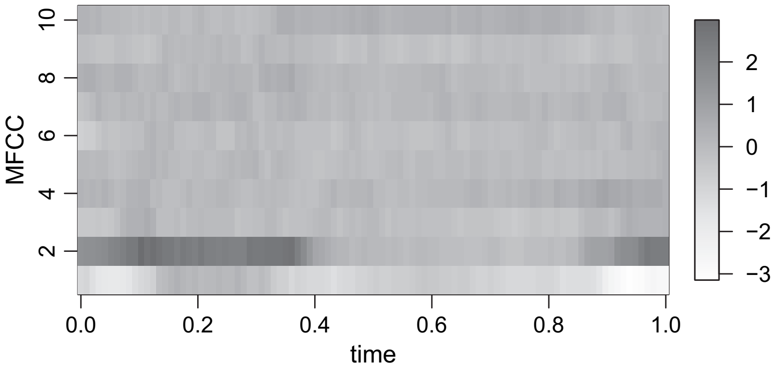


Fig. 12. MFCC of a sound of the word 'last'

Representation of sound signals

The two-dimensional representation of sound signals used by the authors is the spectrogram, which is the logarithm of the squared modulus of a local Fourier transformation of the sound wave $(s(t))_{t=1,2,\dots,T}$. Although spectrograms are valid representations of the sound wave for doing statistics (in particular taking averages of spectrograms, with the view of transforming the resulting spectrograms into sound waves), they suffer from the fact that energy peaks at high frequencies of two sounds of the same word are usually misaligned: taking averages smears out these peaks and, from personal experiments, the resulting sound waves become bland. An alternative two-dimensional representation of the sound wave is given by Mel frequency cepstral coefficients (MFCCs) (see Fig. 11 for a definition and Fig. 12 for an example). The advantages of working with MFCCs is that they are less prone to smearing the high frequency energies, while still allowing the MFCC to be transformed back into a sound wave. A drawback of working with MFCCs is that they are more difficult to interpret. Various implementations of the MFCC exist, Erro *et al.* (2011) being one of them that allows for high fidelity speech sound resynthesis.

Registration

The registration used in the paper is based on the discrepancy

$$D_\lambda(W_1, W_2, g) = \int_0^{\omega_{\max}} \int_0^1 [W_1\{\omega, g(t)\} - W_2(\omega, t)]^2 \mathrm{d}t \, \mathrm{d}\omega + \lambda \, \mathrm{penalty}(g)$$

between the spectrograms W_t and W_2. Although the L^2-metric is a natural measure of dissimilarity for registration of curves, the time registration of spectrograms could be done by using other choices of dissimilarity measure (e.g. a weighted L^2-metric, cross-correlation (Somervuo, 2019) or the distance between the relative volume of the two sound waves (Tavakoli *et al.*, 2018)).

Table 4. p-values for the permutation test against $H_0:\Theta_1=\Theta_2=\ldots=\Theta_5$†

Θ_i	*p-values for the following frequency bands:*				
	δ	θ	α	β	γ
Mean	< 0.001	< 0.001	< 0.001	0.002	0.037
Spatial covariance	< 0.001	< 0.001	0.410	0.134	0.660
Temporal covariance	< 0.001	< 0.001	0.429	0.113	0.785

†Θ_i can be the mean or the spatial or temporal covariance in odour group i. The tests were conducted across all the five frequency bands specifically.

The following contributions were received in writing after the meeting.

Xu Gao and Weining Shen (*University of California, Irvine*) **and Hernando Ombao** (*King Abdullah University of Science and Technology, Thuwal*)

We congratulate Pigoli, Hadjipantelis, Coleman and Aston (PHCA) for their work on acoustic phonetic data analysis. They proposed a general framework of modelling phonetic variation by time–frequency representations. They utilized the mean and second-order functions (covariance and spectrum) to identify specific word and language features and defined a phonetic transformation that enables exploring differences across languages.

We have some comments on this important work. First, it seems that the use of spectrograms might miss some of the features of speech signals (e.g. potential higher order information and narrow-band properties). We wonder whether the Mel frequency cepstral coefficients could be tailored for modelling the evolution of phonetic variation (Vergen and O'Shaughnessy, 1999). Second, a distinct advantage of spectrograms is interpretability. It would be interesting to see an interpretation of the evolutionary trends of frequency bands that led to the differentiation of these languages. Third, we wonder whether the method proposed can be used to predict how these languages might continue to evolve in the future. Of course, there are modern day challenges that are driven by technology such as sending written messages with emojis, which has been a preferred tool for communication for teenagers and young adults.

Finally, we extend the PHCA approach to brain signal analysis. Recently, Gao *et al.* (2016) proposed the evolutionary state space model (ESSM) for analysing multichannel local field potentials with spectra that evolve across replicates and epochs over the entire experiment in which rats were trained on a specific order of the presentation of five distinct odours (A–B–C–D–E). The PHCA method was applied to these local field potential data using the same preprocessing steps. The results are interesting: the evolution of the spectral power (at each of the standard frequency bands: δ (0–3 Hz), θ (4–7 Hz), α (8–12 Hz), β (12–38 Hz) and γ (greater than 32 Hz)) across the entire experiment were different for the five distinct odours by using the premutation test procedure of PHCA (see Table 4 for example). For the most part this is consistent with the findings from the ESSM approach. The main difference is that the use of the ESSM approach provides more interpretable results related to prespecified frequency bands of scientific interest, e.g. δ, α and γ. Researchers will directly observe the characteristic for bands of their interest.

Simon J. Greenhill (*Max Planck Institute for the Science of Human History, Jena, and Australian National University, Canberra*)

I thank Pigoli for their interesting contribution. One question: have they evaluated the fit of their data to the family tree of these languages? All the languages analysed are Romance languages, but there are degrees of relatedness within this grouping (for example Italian is the most divergent, whereas Portuguese is closer to Spanish than to French). Can this analysis correctly identify these nested patterns? Rather than averaging across words, are the authors able to align each word separately in a discrete manner and to use this to recover phylogeny?

There are some exciting consequences if this approach can indeed recover that, say, Portuguese is closer to Spanish (a task that a short list of 10 cognate number words would struggle to do). The basis of historical linguistics is sound change but, to date, this is used in a sadly attenuated way: transcribed word lists. However, transcriptions throw away all the detailed signal in speech that sociolinguistics has clearly shown to be of great use in identifying groups. Sociolinguistics has given us multivariate and detailed analyses on how sound change develops and percolates through speakers and how people use these tokens to demarcate important social groups. Historical linguistics has given us detailed knowledge of how sound changes play out over thousands of years, and how they demarcate languages over this deep timescale.

The approach of Pigoli and his colleagues may provide the natural link between the detailed microevolutionary processes happening at the sociolinguistic level and the broad scale macroevolutionary processes happening at the historical linguistics level. Linking these two disciplines will enable us to understand how sound change varies and accumulates within and between groups and over time, leading to better models of language evolution which can potentially track language history to a far greater degree than previously possible.

The ability to transform and extrapolate words between languages proposed here is also intriguing: can we now ask how *realistic* are proposed sound changes and historical reconstructions? Increasingly more cross-linguistic projects are recording rich audio and video data from the world's languages (e.g. `http://www.soundcomparisons.com`). It makes sense to analyse these data at the level—phonetics—where both speakers and selective evolutionary forces operate, rather than in the abstraction found in a transcript. Perhaps in future historical linguists will be listening to *Schleicher's Fable* rather than reading it.

Hyun Bin Kang and Matthew Reimherr (*Pennsylvania State University, State College*)

We congratulate Pigoli, Hadjipantelis, Coleman and Aston on their impressive work, providing a framework for analysing samples of sound recordings of different words in different languages pronounced by different speakers. They preprocess the sound recordings so that the data can be considered as the decibel functions in the time and frequency domain, which they call *log-spectograms*, and this enables the use of functional data analysis tools for analysis. This is a part of the recent movement in functional data analysis with more interest towards analysing complex data objects, whereas traditional functional data analysis has focused on the analysis of infinite dimensional stochastic processes on a single domain (Wang *et al.*, 2016).

Many assumptions are made for the framework, but one assumption to be noted is the separability assumption for the covariance structure of the log-spectograms. When estimating the covariance operator in Section 5, the authors make an assumption that the covariance structure is separable in time and frequency. Although the separability assumption reduces the dimension substantially and makes modelling and estimation easier, it can introduce bias when it does not hold. The results of the paper and the definition of a path and a transformation operator from one language to another in Section 6 appear to depend considerably on this separability assumption; thus it seems necessary to conduct a test for separability (see Aston *et al.* (2017) and Constantinou *et al.* (2017)). As Aston *et al.* (2017) have already found that the separability assumption does not hold for the acoustic phonetic data, it would be interesting to see how the results of this paper change without assuming a separable covariance structure (e.g. Kang *et al.* (2017)).

The framework proposed may also be more fruitful if more data are introduced. The Shtooka project (`http://shtooka.net`) is a free multilingual audio database of words and sentences of 18 languages and, if the recordings there can be turned into log-spectograms, it may be possible to give more detailed features of a population language and enable the construction of a more comprehensive relationship between different languages. For example, as Italian and Spanish are both Romance languages and have many similarities, a measure for the path from Italian to Spanish would be much shorter than the path from Italian to Chinese.

Adam B. Kashlak (*University of Alberta, Edmonton*)

Pigoli, Hadjipantelis, Coleman and Aston have made an excellent contribution to the statistical analysis of acoustic data. In accordance with the challenges of dealing with such complex data, they first preprocess them via smoothing and then curve registration. My curiosity lies in the latter tool. Namely the *pairwise warping function* from Tang and Müller (2008),

$$h_{ij}(i) = \arg\min_{h} \int_0^1 [f_i\{h(t)\} - f_i(t)]^2 \mathrm{d}t + \lambda \int_0^1 \{h(t) - t\}^2 \mathrm{d}t,$$

results from least squares alignment of the curves with an added penalization term to dissuade overwarping of the data. This leads to some questions. How is λ, the *empirically evaluated non-negative regularization* *constant* empirically evaluated in practice? Whereas it seems obvious that underwarping will be to the detriment of the methodology, will overwarping—i.e. λ much too small—also harm the methodology? And how sensitive is successful estimation of the mean and covariance dependent on a good choice of λ? Lastly, will the reduction in variation lead to trouble in the estimation of the covariance operator specifically?

Zeda Li (*City University of New York*) **and Scott A. Bruce** (*George Mason University, Fairfax*)
Pigoli, Hadjipantelis, Coleman and Aston have provided a careful treatment of the technical difficulties arising in the analysis of acoustic phonetic data through the time varying power spectrum. The approach considered can be cast in a more general mixed effects framework that can accommodate other dependence structures, forms of fixed effects and additional covariates. This more general framework also provides a model for the time series, whereas the paper provides a model for only the power spectra.

For a given language, words uttered corresponding to the numbers from 1 to 10 can be treated as fixed effects impacting the mean, whereas random effects capture the covariance in speech recordings between speakers. More specifically, let $U_{ik}=(u_{ik1},\ldots,u_{ikP})^{\mathrm{T}}$ and $V_{ik}=(v_{ik1},\ldots,v_{ikQ})^{\mathrm{T}}$ be vectors of covariates for the kth speaker uttering the ith word. The speech signal has a mixed effects Cramér representation

$$x_{ikt}=\int_{-1/2}^{1/2}A_i(t/T,\omega;U_{ik})A_{ik}(t/T,\omega;V_{ik})\exp(2\pi\mathrm{i}\omega t)\,\mathrm{d}Z_{ik}(\omega),$$

where Z_{ik} are independent, identically distributed, zero-mean orthogonal increment processes, and the time varying transfer function is represented as the product of two components, $A_i(u,\omega;U_{ik})$ and $A_{ik}(u,\omega;V_{ik})$, over scaled time $u\in[0,1]$ and frequency $\omega\in[-\frac{1}{2},\frac{1}{2}]$. The individual spectra for each speaker–word pair and the word level average spectrum are then defined as

$$f_{ik}(u,\omega;U_{ik},V_{ik})=|A_i(u,\omega;U_{ik})|^2|A_{ik}(u,\omega;V_{ik})|^2,$$

$$f_i(u,\omega;U_{ik})=|A_i(u,\omega;U_{ik})|^2.$$

This representation is an extension of existing functional mixed effects models (Guo, 2002; Qin and Guo, 2006; Krafty *et al.*, 2011) to non-stationary signals.

We then have the following mixed effects model for individual log-spectra:

$$\log\{f_{ik}(u,\omega;U_{ik},V_{ik})\}=U_{ik}^{\mathrm{T}}\beta(u,\omega)+V_{ik}^{\mathrm{T}}\alpha(u,\omega).$$

The covariance of the qth random effect is defined as $\Gamma_q(u_1,\omega_1,u_2,\omega_2)=E[\alpha_q(u_1,\omega_1)\alpha_q(u_2,\omega_2)]$.

Let $\Gamma(u_1,\omega_1,u_2,\omega_2)=\mathrm{diag}\{\Gamma_1(u_1,\omega_1,u_2,\omega_2),\ldots,\Gamma_Q(u_1,\omega_1,u_2,\omega_2)\}$ be a $Q\times Q$ diagonal matrix of these covariances; then we have the first two moments of the log-spectra

$$E[\log\{f_{ik}(u,\omega;U_{ik},V_{ik})\}]=U_{ik}^{\mathrm{T}}\beta(u,\omega),$$

$$\mathrm{cov}[\log\{f_{ik}(u_1,\omega_1;U_{ik},V_{ik})\},\ \log\{f_{il}(u_2,\omega_2;U_{il},V_{il})\}]=V_{ik}^{\mathrm{T}}\Gamma(u_1,\omega_1,u_2,\omega_2)V_{il}.$$

In the paper, the covariance function is assumed to be separable in time and frequency, so we have that $\Gamma(u_1,\omega_1,u_2,\omega_2)=\Gamma^{(u)}(u_1,u_2)\,\Gamma^{(\omega)}(\omega_1,\omega_2)$.

We can then define

$$\begin{pmatrix}U_{1k}\\U_{2k}\\U_{3k}\\\vdots\\U_{10k}\end{pmatrix}=\begin{pmatrix}1&0&0&\ldots&0\\0&1&0&\ldots&0\\0&0&1&\ldots&0\\\vdots&\vdots&\vdots&\vdots&\vdots\\0&0&0&\ldots&1\end{pmatrix},$$

$$\begin{pmatrix}V_{1k}\\V_{2k}\\V_{3k}\\\vdots\\V_{10k}\end{pmatrix}=\begin{pmatrix}1\\1\\1\\\vdots\\1\end{pmatrix}.$$

The fixed effects $\beta(u,\omega)=(\beta_1,\ldots,\beta_{10})^{\mathrm{T}}$ are the mean log-spectra for each word from 1 to 10, whereas $\Gamma_1(u_1,\omega_1,u_2,\omega_2)=\Gamma_1^{(u)}(u_1,u_2)\,\Gamma_1^{(\omega)}(\omega_1,\omega_2)$ accounts for the variation in across-speaker average log-spectra among words. Covariates can be incorporated in additional fixed and random-effect vectors to estimate their effect on the log-spectra. Estimation of such a model is challenging and requires more research in this direction. However, existing methodologies are available to explore the effect of covariates on time varying log-spectra for replicated non-stationary time series (Qin *et al.*, 2009; Fiecas and Ombao, 2016; Bruce *et al.*, 2018), which can used to analyse acoustic phonetic data.

J. S. Marron (*University of North Carolina, Chapel Hill*)
Thank you very much for a fascinating paper. To put it in some context, recently the term ‘big data’ has become ubiquitous, as data scale indeed presents important challenges. What is less well understood is that perhaps an even greater challenge is presented by *complex data*. This paper and the methods proposed provide a particularly deep example of research in this important direction. As noted, object-oriented data analysis, as first described in Wang and Marron (2007) and more carefully considered in Marron and Alonso (2014), provides a useful framework for the study of this and other types of complex data.

Particularly impressive is the ability to predict how a speaker in one language would sound in another, just using relative position in the population. It would be interesting to see whether this ability could be enhanced by focusing on common variation between the languages by using a data integration method such as *joint and individual variation*, as proposed by Lock *et al.* (2013), and recently substantially improved in Feng *et al.* (2018). That methodology could also be used for many other types of language analysis inference, such as deeper insights into how languages are similar and how they are different.

A related technology that may be worth comparing is the voice recognition software that one encounters for example when telephoning a major corporation. That was an early success story for the *neural network* method, which fell into some disrepute when it failed at many other tasks. Recently it has made a strong comeback under the name *deep learning*, where it has proven to be very effective, perhaps because of the availability of very large data sets and modern computational power. The present approach relies on carefully crafted data object choices, which seem essential for the small data sets that are considered here. But, as much more data become available, it will be interesting to see whether the apparent automatic feature selection abilities of a deep learning approach can provide similar good performance or not.

Jorge Mateu (*University Jaume I, Castellón*) **and Martha Bohorquez and Ruben Guevara** (*National University of Colombia, Bogotá*)
We congratulate Pigoli, Hadjipantelis, Coleman and Aston on this comprehensive and well-written paper. This work presents an important contribution to the analysis of acoustic phonetic data. The methodology is based on the theory of functional data analysis considering speech recordings as realizations of spectrograms. Although the theory of functional data allows objects in any dimension, the random surfaces require particular developments. The application field is certainly attractive but, from the statistical point of view, it can be considered as a motivating work to propose alternative methodologies to compare and estimate mean functions and covariance operators.

The tests based on permutations are computationally intensive, being necessary to explore alternative methods with both good statistical power and reduced computational costs, as shown by the concentration method that was proposed by Kashlak (2017) and Bagchi and Dette (2017). To test the difference of the means between digits, the generalization of the Mahalanobis distance to the Hilbert space L^2, as proposed by Galeano *et al.* (2015) and Ghiglietti *et al.* (2017), could be considered in two ways: as a distance in the statistical test, as considered by these authors, or as a new methodological set-up to develop a new test. We note that there is an implicit assumption of independence between speakers, but they could not be spatially independent. For people with the same language, considering spatial functional random fields instead of independent functional variables might be a good alternative. To avoid computational issues and additional assumptions, such as separability, there are alternative options to model the covariance structures. One possibility is to use the Karhunen–Loève expansion and the equivalence between the covariance of random functions and the covariance of the scores, obtained from the functional principal component analysis. Functional principal component analysis can be applied in one of the dimensions and the covariance model can be found along the other dimension (Horváth and Kokoszka, 2012; Bohorquez *et al.*, 2017). These methodologies enable us to extend the framework to the multivariate case; for example, we could consider several acoustic properties and some kind of spatial or temporal dependence, or even include some covariates. Classification methods, such as some generalization of the Mahalanobis distance extended to multivariate and correlated functional data analysis, would enable us to detect homogeneous speech communities. Finally, the analysis based on the additional information provided by the first-order derivatives could enrich the statistical analysis, as considered in Cuevas *et al.* (2007) and Claeskens *et al.* (2014).

Hernando Ombao (*King Abdullah University of Science and Technology, Thuwal*) **and Chee-Ming Ting** (*King Abdullah University of Science and Technology, Thuwal, and Universiti Teknologi Malaysia*)
In this paper, the starting point for studying similarities and differences in the acoustic and phonetic properties within and between the five Romance languages is the spectrogram. Spectrograms, which capture the time–frequency properties of the signals, are functionally coregistered to reduce misalignment across recordings due to variations in speaking rates. In addition, Pigoli, Hadjipantelis, Coleman and Aston look into the covariance function between observed spectrograms at different pairs of time and frequencies which are assumed here to be common to all words within a language. This is important because dual-frequency dependence could give potentially interesting features in various types of signals (see Lii and Rosenblatt (2002), Léskow (2012) and Gorrostieta *et al.* (2018)). Formal inference for comparing between languages and words was conducted through permutation testing.

It seems very natural to build on this important work though functional mixed effects models. Let $S_{k,L}(u,\omega)$ be the observed spectrogram for the kth speaker in language group L (ignoring digits for now). For each fixed frequency ω^*, decompose this into language-specific (fixed effect) log-spectra $\mu_L(u,\omega^*)$, subject-specific (random deviation) $b_{k,L}(u,\omega^*)$ and within-subject random variation $\epsilon_{k,L}(u,\omega^*)$ to account for variations across replicated recordings:

$$S_{k,L}(u,\omega^*) = \mu_L(u,\omega^*) + b_{k,L}(u,\omega^*) + \epsilon_{k,L}(u,\omega^*).$$

Of course one can replace the spectrograms by the data analogue of the cepstrum (see Bogert *et al.* (1963) and Childers *et al.* (1977)) since this is more commonly used in speech processing. The open questions of the proposed model above would be potential factorization into functions of frequency only and of time only, finding reasonable temporal basis functions (when frequency is fixed) for representing these objects, and the appropriate temporal covariance function. From this model, comparisons between languages can be performed through the coefficients and the variation across speakers through the corresponding random effects.

Victor M. Panaretos (*École Polytechnique Fédérale de Lausanne*)
I congratulate Pigoli, Hadjipantelis, Coleman and Aston for their fine contribution which furnishes considerable insights, while skilfully confronting multiple layers of functional complexity, and even interactions thereof. Here I focus on one such issue, namely the non-linear nature of the space of covariance operators. In Sections 5 and 6, the authors probe the variability and interrelationships of a collection of covariances $\{C_l\}_{l=1}^G$, corresponding to different subpopulations and estimated via n realizations from each group, say $\{X_{il}\}_{i=1}^n \sim^{\text{IID}} N(0, C_l)$, assumed centred Gaussian for simplicity (concretely, X_{il} represents the ith log-spectrogram in language l). To do so, the authors advocate the use of the *Procrustes* (*reflection size-and-shape*) *distance*. They then use this distance to define a Fréchet mean C of $\{C_1, \ldots, C_G\}$, and to measure variability around this centre point C.

I would like to remark that this choice of geometry implicitly imposes a concrete *hierarchical generative model*, with an appealing interpretation in the application's context, and potential implications on the analysis, either conceptual or concrete. Specifically, Masarotto *et al.* (2018) show that the Procrustes distance coincides with the *Wasserstein distance* between centred Gaussian processes with corresponding covariances; and, consequently, that the X_{il} can be seen to arise hierarchically, by first generating n latent realizations $\varepsilon_{il} \sim^{\text{IID}} N(0, C)$ from the 'Fréchet mean language' with covariance C, and then deforming them by G group-specific operators T_l, i.e. $X_{il} = T_l \varepsilon_{il}$. The maps T_l are non-negative linear operators that satisfy $\bar{T} = G^{-1} \Sigma_{i=1}^G T_l = I$, and they 'optimally transport' language l in that $C_l^t := \text{cov}\{t T_l \varepsilon_{il} + (1-t)\varepsilon_{il}\}$ is a unit speed geodesic from C to C_l in the Procrustes geometry. At the level of covariances, the model simply states that $C_l = T_l C T_l$ are conjugation perturbations of C, with $\bar{T} = I$ reflecting that the perturbation is the identity on average. The inverses T_l^{-1} provide an *optimal simultaneous registration* of different speakers to a latent template language. It may thus be argued that the resulting registered process $\varepsilon_{il} = T_l^{-1} X_{il}$ could serve as a more canonical definition of the speaker-specific residual 'left in the phonetic data once means and covariance information have been removed. The maps $\{T_l\}$ can be recovered explicitly from knowledge of C and C_l and essentially correspond to the log-images of the $\{C_l\}$ on the tangent space at the Fréchet mean C. In that sense, they can also be used to produce a tangent space principal component analysis of the main components of variation in the collection $\{C_l\}$, potentially revealing further structure. See Masarotto *et al.* (2018), sections 9 and 10.

Laura M. Sangalli, Piercesare Secchi and Simone Vantini *(Politecnico di Milano)*
We warmly congratulate Pigoli, Hadjipantelis, Coleman and Aston for this fascinating case-study in object-oriented data analysis, where some of the most advanced statistical methods are applied to the analysis of a non-trivial acoustic phonetic problem. The authors take a multistep approach, with various techniques used in sequence, with the output of the previous step being the input to the following step. Despite being commonly used, this multistep approach may present some issues, mostly related to the consistency of the various modelling assumptions made at the various steps and to the theoretical properties of the entire working pipeline. We thus encourage the authors to pursue a unified approach to the modelling of the problem under investigation following a fully object-oriented data analysis perspective. As a first move in this direction, a stronger point could be made by identifying the appropriate feature space offering the natural setting for the entire analysis; a space whose structure (Hilbert, Riemannian,...) must be directly suggested by the problem rather than by mathematical convenience. Secondly, warping could become an integrated part of the analysis of the variation in the data, instead of being performed as a preprocessing step. Given the nested block structure in the data set considered, with each block corresponding to records of a digit in one language, the authors could possibly devise a similar strategy to that considered in Bernardi *et al.* (2014), where the phase and the amplitude variabilities are simultaneously explored, complying with a block structure in the data. Again, the very same data structure suggests 'joint and individual variation explained' as an alternative approach for the exploration of variability; this extension of principal component analysis, proposed by Lock *et al.* (2013) and further explored in Feng *et al.* (2018), can decouple the variation that is common across different data blocks and the variation that is unique to each data block. For the case-study at hand, the 'joint and individual variation explained' approach could simultaneously extract the information common to the collection of words belonging to the same language and represent it in the common variation space pertaining to the language, and the information unique to each digit, represented in orthogonal digit spaces; phonetic differences between languages could then be explored by comparing the linear spaces capturing joint and individual variations, along the paradigmatic examples in Yu *et al.* (2017) for the analysis of functional magnetic resonance imaging and behavioural data.

Jian Qing Shi and Evandro Konzen (*Newcastle University*)
We congratulate Pigoli and his colleagues for introducing a very interesting approach to explore phonetic differences and changes between languages. The time and frequency covariance functions have been identified as a feature of the language. Therefore, the covariance function estimation may deserve some particular attention.

A popular way to make the estimation of the covariance structure accurate and computationally feasible is by assuming that the full covariance function $c(\omega_1, \omega_2, t_1, t_2)$ is separable. When the replicated two-dimensional curves are not densely and regularly sampled, non-parametric approaches (e.g. Yao *et al.* (2005)) require performing a non-parametric regression on four variables. Therefore, to reduce issues concerning computational costs, the curse of dimensionality and loss of asymptotic efficiency, the separability assumption seems unavoidable.

An alternative way is to use a parametric covariance function instead, selected from many families such as the powered exponential, rational quadratic and Matérn (Shi and Choi, 2011) families. In this way, although we may lose some flexibility in the sense that it would be difficult to model small-scale variability, a suitable parametric covariance function family can be used to model large-scale variability without assuming separability. In addition, by modelling via Gaussian processes, we do not even need multiple replications of the surface $X(\omega, t)$.

To visualize the main modes of variation of the data with a non-separable covariance structure, we can plot the associated eigensurfaces. In Fig. 13, we can see that the diagonal elliptical form of the contours represents a clear non-separable structure.

Milan Stehlík (*Johannes Kepler University in Linz and University of Valparaíso*) **and Mirtha Pari Ruiz** (*University of Valparaíso*)
Congratulations go to Pigoli, Hadjipantelis, Coleman and Aston for giving readers a challenging discussion on exploring differences between spoken Romance languages. Two important issues which we would like to point out are the oversmoothing and oversymmetrization of the data. Namely, regular surfaces and their corresponding covariance structures should be compared with semicontinuous covariances, introduced in Stehlík *et al.* (2017), which are more realistic.

In our work we studied different corpuses of *Nican Mopohua*, written in the original language (Náhuatl-Mexico) and we used two versions of the *corpus* from Lasso (1649) and Rojas (1978). If we compare

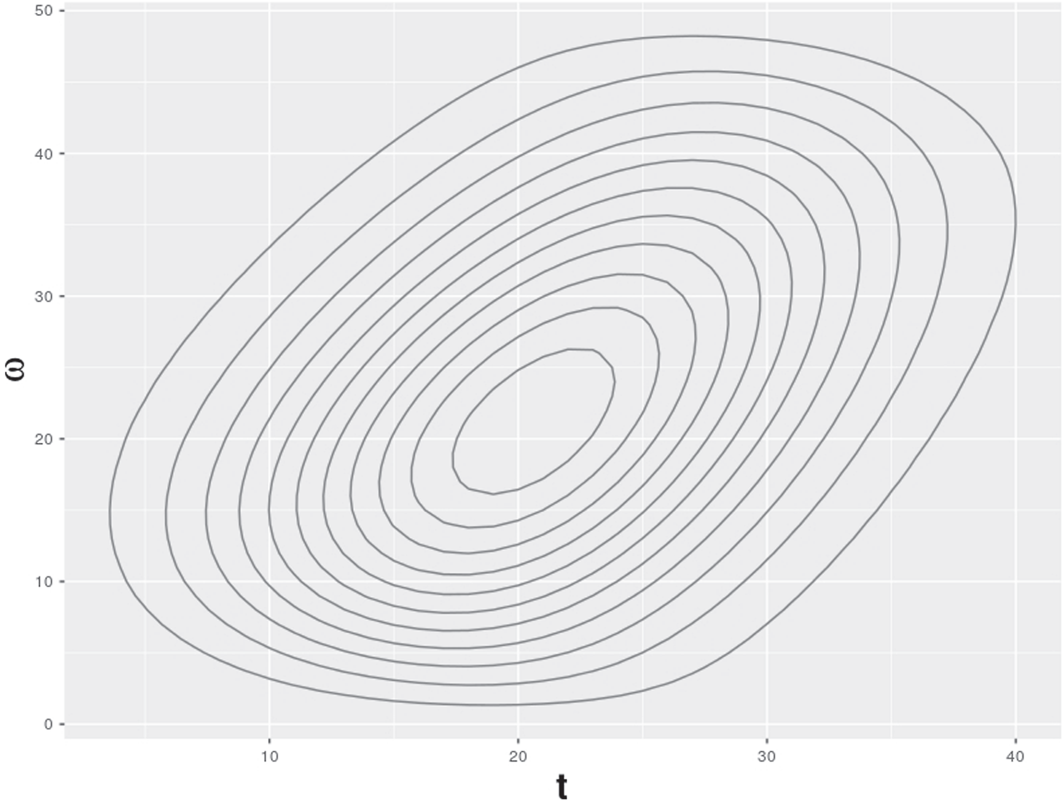


Fig. 13. Contour plot of an eigensurface of a non-separable covariance function

Kullback–Leibler divergences for the Náhuatl word '*xochitl*' (flowers) from one *corpus* to another, we obtain different divergences, namely 0.618 and 0.272. That empirically underlines the fact that we are not in the symmetrized world of distances, but asymmetric divergences. We recall here that Kullback–Leibler divergence can be symmetric for specific distributions also. This asymmetry supports the fact that language will naturally relate much more to the topology (see Stehlík (2016)) than to some metric, since language constructs go hand in hand with cognition. We may also find a topological basis (see, for example, Smith (1996)). Mereological and topological notions of connection, part, interior and complement are central to spatial reasoning and to the semantics of natural language expressions concerning locations and relative positions. Thus the exploration of the phonetic differences by the structure of the language based on the recording of the language will not be satisfactory to provide a reconstruction of an ancestral language.

The **authors** replied later, in writing, as follows.

We thank the discussants for their interesting comments and suggestions. Many of them proposed extensions and alternative approaches for the analysis of acoustic phonetic data, both from the point of view of models and methods and from the point of view of the data to be used. We wish first to highlight how the main purpose of our work was to point out an interesting and challenging problem which deserves attention from the statistical community and, although we think our analysis sheds some light on the problem, we do not claim it in any way to be the last word on the subject. Indeed, as we point out in the paper, we view this analysis as a 'proof of concept' to explore the potential of the application of functional data techniques to cross-linguistic speech data.

We thank Professor Cortina-Borja for his thoughtful comment that better contextualizes our work within existing statistical methods for linguistics and within the scope of the 'Ancient sounds' project. The contribution of this paper towards the final goal of the 'Ancient sounds' project lies mainly in the investigation of which features are essential to characterize the languages and how to model cross-linguistic sound change. These are the basic elements which any model for historical sound change needs to rely on. In response to the points Professor Cortina-Borja raises about the permutation tests, we think that the evidence about the existence of differences between languages is sufficiently compelling although, when broken down to the individual language level, the data set is not sufficiently large for the results to be trusted conclusively in terms of the between-words comparison. Concerning the presence of discontinuities in the frequency domain, we think these are mainly artefacts due to the quality of the recordings but they can in principle be accounted for by using less regular (and anisotropic) basis functions to represent the log-spectrograms. The extension to larger data sets would indeed be the next step for the project, both in the direction of including more speakers and more words and more languages in the analysis. The application of the proposed approach to other types of data is indeed intriguing, as also suggested by other discussants.

Professor Cox has some reservations about the paper that focus on the validity and limitations of the data set and on the lack of engagement with speech processing techniques. We agree that the data set we considered is too small (both in number of speakers and dictionary) to draw definitive conclusions about language change. At the same time, the fact that we focused on a small vocabulary makes the comparison with data sets with hundreds of hours of data from hundreds of speakers somewhat inappropriate, as these are used to build far more general models of languages and their pronunciation. Moreover, automatic analysis of large speech data sets to explore linguistic differences comes with its own problems, as some of us found out when working on the spoken part of the British National Corpus to explore dialectal variation (see Tavakoli *et al.* (2018)) and we preferred, for this proof of concept, a smaller but reliable data set. However, the variability between speakers does not appear to mask the variability between language completely, as feared by Professor Cox, since the permutation tests can indeed show significant differences between languages that are not simply the effect of the random grouping. Moreover, the recordings from each language were not collected under the same laboratory conditions, since each language was exemplified by multiple recordings from various on-line sources. It is an important concern, of course, but we can be confident that our characterization of different languages is in fact about languages and not recording channels, because the demonstrations of morphing the pronunciation of a word in one language into another are indeed successful; the morphed forms do not end up sounding like the same speaker but recorded under different channel conditions. Concerning the limitations of narrow-band fast Fourier transform, we have also experimented with using Mel frequency cepstral coefficients and also linear prediction coding reflection coefficients; these coding methods can produce somewhat more natural sounding synthetic reconstructions, but that does not imply that using fast Fourier transform is as hopeless as Professor Cox implies. Our demonstrations show that analysis, manipulation and synthesis using fast Fourier transform spectrograms are scientifically and auditorily reasonable, even granted that other signal representations may yield improved results. More generally, we agree that the cross-pollination with speech processing research will ultimately be the way to go forward but this will require additional work to understand how this can fit in the linguistics domain. The extension to other languages is a very interesting question, since the method clearly only makes sense when considering cognate words, i.e. those with a common ancestor, and therefore in the presence of some degree of acoustic similarity. Although we have not yet been able to explore the limits of how dissimilar the words compared may be for the statistical methods that are presented in this paper to be appropriate, in a separate set of demonstrations of morphing between related words in different languages, John Coleman has built pronunciation *continua* from English to German, Lithuanian, Elfdalian Swedish and Polish words (see `http://www.phon.ox.ac.uk/jcoleman/ancient-sounds-audio.html`). This gives us optimism that modelling continuous acoustic variation of a fairly wide spread of sometimes quite distantly related languages is scientifically reasonable.

Professor Silverman points to some additional directions that we agree would be very interesting to pursue in future works, such as using specifically developed basis functions to model speech data objects and extending the analysis to larger *corpora* and different experimental settings (non-native speakers). He also raises some concerns about high p-values in Table 2. However, if we consider them together with the p-values in Table 3 (tests for the time covariance), there is a good spread of p-values across the range. Therefore we do not think that this is sufficient reason to cast doubts on the experimental and inferential procedure (and some of the references suggested by Professor Silverman illustrate that Fisher may have been too hasty in judging Mendel as well!).

Many discussants also propose alternative ways to analyse the data, such as ‘joint and individual variation explained’ (Marron), joint and individual variation explained plus integration of time alignment in the main analysis (Sangalli, Secchi and Vantini) using Mel frequency cepstral coefficients in place of spectrograms and *ad hoc* metrics to measure dissimilarities (Gao, Shen and Ombao, and Tavakoli), different testing procedures (Mateu, Bohorquez and Guevara), or alternative modelling choices such as different structural assumptions for the covariance, both parametric (Shi and Konzen) and non-parametric (Kang and Reimherr, and Mateu, Bohorquez and Guevara), the use of mixed effects models (Li and Bruce, and Ombao and Ting), asymmetries between languages (Stehlík and Pari Ruiz) and spatial dependence and use of first-order derivatives (Mateu, Bohorquez and Guevara). Another notable comment that is recurrent among discussants is the desire to expand the current study to explore different aspects of the language variation, such as *between-words* covariance (Tavakoli), comparison with deep learning approaches (Marron), classification of languages (Mateu, Bohorquez and Guevara) and phylogenetic analysis (Greenhill). These are indeed beyond the scope of the present paper and richer data sets will need to be collected to answer many of these questions, so we thank the discussants who pointed us towards available sources of cross-linguistic data (Greenhill, and Kang and Reimherr).

In reply to Dr Kashlak's questions about the registration procedure, we followed the strategy that was proposed originally by Tang and Müller (2008) for the monodimensional case, where λ is chosen to be proportional to the average L^2-norm of the normalized curves and the degree of proportionality is such that it minimizes the variation of the curves. That being said, the results do not appear to be overly sensitive to the choice of λ.

Professor Greenhill also raises some important questions about how the approach proposed fits with the phylogenetic models of language evolution. This will indeed be the natural next step of the work but some open problems need to be addressed, such as how exactly the linguistic distance should be computed and, related to other discussants' comments, what kind of acoustic phonetic data would comprise the equivalent of the well-established linguistics *corpora* used for textual analysis. A first attempt in this direction can be found in Hadjipantelis (2013), chapter 6, We cannot definitively answer his questions at this stage but we indeed think that our analysis offers some insight into how the goal of bringing together the microevolutionary level of sociolinguistics and.the macroevolutionary level of historical linguistics can be reached but also indicates some of the challenges that are in front of us. In particular, an important *caveat* is that sound changes appear to happen very noticeably at certain moments in history and hardly at all at most other times; thus the intergenerational pronunciation differences observed in sociolinguistics rarely impact the historical evolution of the language and they may be best seen as small fluctuations around the overall trend.

Professor Panaretos's deep comment on the geometrical structure that is induced by the Procrustes metric on the space of covariance matrices indeed suggests a very interesting framework to explore their variation and may also help to interpret it better phonetically.

We found all these comments really stimulating. We shall try to consider as many of these directions as possible in our future work but we also hope that other researchers (maybe some of the discussants themselves!) may be motivated to apply these techniques to tackle phonetic data in linguistics.

We are also very pleased to see that the approach that we proposed for linguistic data can be fruitfully applied to other types of sound data (Meagher, Damoulas, Jones and Girolami) and to different instances of (hyper)surface data such as brain signals (Gao, Shen and Ombao).

References in the discussion

Aston, J. A., Pigoli, D. and Tavakoli, S. (2017) Tests for separability in nonparametric covariance operators of random surfaces. *Ann. Statist.*, **45**, 1431–1461.

Baayen, H. (1992) Statistical models for word frequency distributions: a linguistic evaluation. *Comput. Human.*, **20**, 347–363.

Bagchi, P. and Dette, H. (2017) A test for separability in covariance operators of random surfaces. *Preprint arXiv:1710.08388*.

Bernardi, M., Sangalli, L. M., Secchi, P. and Vantini, S. (2014) Analysis of Proteomics data: block K-mean alignment. *Electron. J. Statist.*, **8**, 1714–1723.

Bogert, B., Healy, M. and Tukey, J. (1963) The quefrency analysis of time series for echoes: cepstrum, pseudo autocovariance, cross-cepstrum and saphe cracking. In *Proc. Symp. Time Series Analysis* (ed. M. Rosenblat), ch. 15, pp. 209–243. New York: Wiley.

Bohorquez, M., Giraldo, R. and Mateu, J. (2017) Multivariate functional random fields: prediction and optimal sampling. *Stoch. Environ. Res. Risk Assessmnt*, **31**, 53–70.

Bruce, S. A., Hall, M. H., Buysse, D. J. and Krafty, R. T. (2018) Conditional adaptive Bayesian spectral analysis of nonstationary biomedical time series. *Biometrics*, **74**, 260–269.

Childers, D., Skinner, D. and Kemerait, R. (1977) The cepstrum: a guide to processing. *Proc. IEEE*, **65**, 1428–1443.

Claeskens, G., Hubert, M., Slaets, L. and Vakili, K. (2014) Multivariate functional halfspace depth. *J. Am. Statist. Ass.*, **109**, 411–423.

Constantinou, P., Kokoszka, P. and Reimherr, M. (2017) Testing separability of space-time functional processes. *Biometrika*, **104**, 425–437.

Cuevas, A., Febrero, M. and Fraiman, R. (2007) Robust estimation and classification for functional data via projection-based depth notions. *Computnl Statist.*, **22**, 481–496.

Erro, D., Sainz, I., Navas, E. and Hernáez, I. (2011) HNM-based MFCC+F0 extractor applied to statistical speech synthesis. In *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, pp. 4728–4731. New York: Institute of Electrical and Electronics Engineers.

Feng, Q., Jiang, M., Hannig, J. and Marron, J. S. (2018) Angle-based joint and individual variation explained. *J. Multiv. Anal.*, **166**, 241–265.

Fiecas, M. and Ombao, H. (2016) Modeling the evolution of dynamic brain processes during an associative learning experiment. *J. Am. Statist. Ass.*, **111**, 1440–1453.

Fisher, R. A. (1936) Has Mendel's work been rediscovered? *Ann. Sci.*, **1**, 115–137.

Galeano, P., Joseph, E. and Lillo, R. (2015) The Mahalanobis distance for functional data with applications to classification. *Technometrics*, **57**, 281–291.

Gao, X., Shen, W., Shahbaba, B., Fortin, N. and Ombao, H. (2016) Evolutionary state-space model and its application to time-frequency analysis of local field potentials. *Preprint arXiv:1610.07271*.

Ghiglietti, A., Ieva, F. and Paganoni, A. M. (2017) Statistical inference for stochastic processes: two-sample hypothesis tests. *J. Statist. Planng Inf.*, **180**, 49–68.

Gorrostieta, C., Ombao, H. and von Sachs, R. (2018) Time-dependent dual-frequency coherence in multivariate non-stationary time series. Submitted to *J. Time Ser. Anal.*

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

Hadjipantelis, P. Z. (2013) Functional data analysis in phonetics. *PhD Dissertation*. University of Warwick, Coventry.

Hadjipantelis, P. Z., Jones, N. S., Moriarty, J., Springate, D. A. and Knight, C. G. (2013) Function-valued traits in evolution. *J. R. Soc. Interfc.*, **10**, no. 82, article 20121032.

Hartl, D. L. and Fairbanks, D. J. (2007) Mud sticks: on the alleged falsification of Mendel's data. *Genetics*, **175**, 975–979.

Horváth, L. and Kokoszka, P. (2012) *Inference for Functional Data with Applications*. New York: Springer.

Jones, N. S. and Moriarty, J. (2013) Evolutionary inference for function-valued traits: Gaussian process regression on phylogenies. *J. R. Soc. Interfc.*, **10**, no. 78, article 20120616.

Kang, H., Reimherr, M., Shriver, M. and Claes, P. (2017) Manifold data analysis with applications to high-frequency 3D imaging. *Preprint arXiv:1710.01619*.

Kashlak, A. B. (2017) A concentration inequality based statistical methodology for inference on covariance matrices and operators. *PhD Dissertation*. University of Cambridge, Cambridge.

Krafty, R. T., Hall, M. and Guo, W. (2011) Functional mixed effects spectral analysis. *Biometrika*, **98**, 583–598.

Lasso, L. (1649) *Huey Tlamahuizoltica Omonexiti Ilhuicac Tlatoca Ihwapilli Sancta Maria*. Mexico City: Print Juan Ruyz.

Léskow, J. (2012) Cyclostationarity and resampling for vibroacoustic signals. *Acta Phys. Polon.* A, **121**, 160–163.

Lii, K. and Rosenblatt, M. (2002) Spectral analysis for harmonizable processes. *Ann. Statist.*, **30**, 258–297.

Lock, E. F., Hoadley, K. A., Marron, J. S. and Nobel, A. B. (2013) Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Ann. Appl. Statist.*, **7**, 523.

Maegaard, M., Jensen, T. J., Kristiansen, T. and Jørgensen, J. N. (2013) Diffusion of language change: accommodation to a moving target. *J. Socioling.*, **17**, 3–36.

Manining, C. and Schuetze, H. (1999) *Foundations of Statistical Natural Language Processing*. Cambridge: MIT Press.

Marron, J. S. and Alonso, A. M. (2014) Overview of object oriented data analysis. *Biometr. J.*, **56**, 732–753.

Masarotto, V., Panaretos, V. M. and Zemel, Y. (2018) Procrustes metrics on covariance operators and optimal transportation of Gaussian processes. *Sankhya*, to be published.

Meagher, J. P., Damoulas, T., Jones, K. E. and Girolami, M. (2018) Phylogenetic Gaussian processes for bat echo-location. In *Statistical Data Science* (eds N. Adams and E. Cohen), p. 111. Singapore: World Scientific Publishing.

Mendel, G. (1866) Versuche über Pflanzen-hybriden [Experiments in plant hybridization]. *Ver. Naturforsch. Ver. Abh. Brünn*, **4**, 3–47.

Nicholls, G. K. and Gray, R. D. (2008) Dated ancestral trees from binary trait data and their application to the diversification of languages. *J. R. Statist. Soc.* B, **70**, 545–566.

Qin, L. and Guo, W. (2006) Functional mixed-effects model for periodic data. *Biostatistics*, **7**, 225–234.

Qin, L., Guo, W. and Litt, B. (2009) A time-frequency functional model for locally stationary time series data. *J. Computnl Graph. Statist.*, **18**, 675–693

Rojas, M. (1978) *Nican Mophua*. Mexico City: Print Ideal.

Salvador, G. (1987) *Lengua Española y Lenguas de España*. Barcelona: Ariel.

Shi, J. Q. and Choi, T. (2011) *Gaussian Process Regression Analysis for Functional Data*. Boca Raton: Chapman and Hall–CRC.

Silge, J. and Robinson, D. (2017) *Text Mining with R: a Tidy Approach*. Sebastopol: O'Reilly Media.
Smith, B. (1996) Mereotopology: a theory of parts and boundaries. *Data Knowl. Engng*, **20**, 287–303.
Somervuo, P. (2019) Time–frequency warping of spectrograms applied to bird sound analyses. *Bioacoustics*, to be published, doi https://doi.org/10.1080/09524622.2108.1431958
Stehlík, M. (2016) On convergence of topological aggregation functions, *Fuzzy Sets Syst.*, **287**, 48–56.
Stehlík, M., Helperstorfer, Ch., Hermann, P., Šupina, J., Grilo, L. M., Maidana, J. P., Fuders, F. and Stehlíková, S. (2017) Financial and risk modelling with semicontinuous covariances. *Inform. Sci.*, **394–395C**, 246–272.
Tang, R. and Müller, H. G. (2008) Pairwise curve synchronization for functional data. *Biometrika*, **95**, 875–889.
Tavakoli, S., Pigoli, D., Aston, J. A. and Coleman, J. (2018) A spatial modeling approach for linguistic object data: analysing dialect sound variations across Great Britain. *J. Am. Statist. Ass.*, to be published.
Vergen, R. and O'Shaughnessy, D. (1999) Generalized Mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition. *IEEE Trans. Spch Audio Process.*, **7**, 525–532.
Wang, J. L., Chiou, J. M. and Mueller, H. G. (2016) Review of functional data analysis. *A. Rev. Statist. Appl.*, **3**, 257–295.
Wang, H. and Marron, J. S. (2007) Object oriented data analysis: sets of trees. *Ann. Statist.*, **35**, 1849–1873.
Yao, F., Müller, H. G. and Wang, J. L. (2005) Functional data analysis for sparse longitudinal data. *J. Am. Statist. Ass.*, **100**, 577–590.
Yu, Q., Risk, B. B., Zhang, K. and Marron, J. S. (2017) JIVE integration of imaging and behavioral data. *NeuroImage*, **152**, 38–49.