

Using Functional Data Analysis for investigating multidimensional dynamic phonetic contrasts

Michele Gubian^{*,a,1}, Francisco Torreira^{a,b}, Lou Boves^a

^a*Centre for Language and Speech Technology, Radboud University Nijmegen
Erasmusplein 1, 6525 HT, Nijmegen, The Netherlands*

^b*Max Planck Institute for Psycholinguistics
Wundtlaan 1, 6525 XD, Nijmegen, The Netherlands*

Abstract

The study of phonetic contrasts and related phenomena, e.g. inter- and intra-speaker variability, often requires to analyse data in the form of measured time series, like f_0 contours and formant trajectories. As a consequence, the investigator has to find suitable ways to reduce the raw and abundant numerical information contained in a bundle of time series into a small but sufficient set of numerical descriptors of their shape. This approach requires one to decide in advance which dynamic traits to include in the analysis and which not. For example, a rising pitch gesture may be represented by its duration and slope, hence reducing it to a straight segment, or by a richer coding specifying also whether (and how much) the rising contour is concave or convex, the latter being irrelevant in some context but crucial in others. Decisions become even more complex when a phenomenon is described by a multidimensional time series, e.g. by the first two formants.

In this paper we introduce a methodology based on Functional Data Analysis (FDA) that allows the investigator to delegate most of the decisions involved in the quantitative description of multidimensional time series to the data themselves. FDA produces a data-driven parametrisation of the main

*Corresponding author

Email addresses: `m.gubian@let.ru.nl` (Michele Gubian),
`francisco.torreira@mpi.nl` (Francisco Torreira), `l.boves@let.ru.nl` (Lou Boves)
URL: `http://lands.let.ru.nl/FDA/` (Michele Gubian)

¹Present address: School of Experimental Psychology, University of Bristol
Office 3D3, The Priory Road Complex, Priory Road, Clifton BS8 1TU, UK
Tel. +44 (0) 117 3317893, email: `mm14722@bristol.ac.uk`

shape traits present in the data that is visually interpretable, in the same way as slopes or peak heights are. These output parameters are numbers that are amenable to ordinary statistical analysis, e.g. linear (mixed effects) models. FDA is also able to capture correlations among different dimensions of a time series, e.g. between formants F_1 and F_2 . We present FDA by means of an extended case study on diphthong – hiatus distinction in Spanish, a contrast that involves duration, formant trajectories and pitch contours.

Key words: Functional Data Analysis, Cue Trading, dynamic trajectories, diphthong and hiatus, European Spanish.

1. Introduction

As early as the nineteen seventies it has been shown that the same phonetic phenomenon can be related to several different acoustic features (Slis and Cohen, 1969; Repp, 1981). Also, the information carried by some features, such as pitch or formants, is encoded in dynamical changes over time, rather than as fixed-length sequences of scalars that reflect the feature value at crucial points, such as the beginning and the end of a segment. Still, for perfectly understandable and legitimate reasons, most phonetic research has been based on the analysis of individual features represented with a small number of scalar values (e.g. formant values in the centre of vowels, minimum or maximum f_0 values in Hz or semitones, alignment of f_0 minima or maxima relative to the beginning of a segment). Probably, phoneticians prefer this approach because of technical and methodological constraints. For instance, conventional statistical methods require that all observations are expressed as a fixed-length sequence of numbers, and only one dependent variable can be investigated at a time. Such an approach requires phoneticians to decide in advance the points in time where a feature value is obtained, and therefore entails the risk of ignoring potentially relevant detail. For example, a rising pitch movement may be represented by its duration and slope, which effectively reduces the trajectory to a straight line, or it may be represented by a more complex coding that captures the concavity or convexity of the trajectory. Such details in the f_0 shape may be irrelevant in some context, but crucial in others (Dombrowski and Niebuhr, 2010). An alternative method to code pitch, formant or intensity contours consists in using the coefficients from a polynomial fit (Andruski and Costello, 2004; Grabe et al., 2007). This may be a good solution for short curves with a single extremum, but it leaves

us with the burden of interpreting the coefficients of the polynomial when fitting more complex trajectories.

In this paper, we introduce *Functional Data Analysis* (FDA; Ramsay and Silverman, 2005) as a new method for analysing phonetic phenomena involving dynamic changes across multiple acoustic parameters. In this sense, this paper extends previous presentations of FDA as tool for phonetic analysis (Gubian et al., 2011; Cheng et al., 2010; Zellers et al., 2010). We show that FDA allows us to apply familiar statistical methods (e.g. linear regression, principal component analysis) to dynamic features by representing them as continuous functions, and that this can be done for multiple features in a single joint analysis. To illustrate this, we revisit the diphthong – hiatus distinction in Spanish, a contrast in which several phonetic features have been implicated (duration, formant trajectories, f_0 alignment, see Aguilar, 1999; Hualde and Prieto, 2002; Torreira, 2007). In particular, we show how FDA can be used to reveal the trading relations between these features, and, more specifically, how different speakers trade the features in different ways.

The remainder of the paper is structured as follows. In Section 2, we describe the diphthong–hiatus distinction in Spanish, and the data that will be used in the following sections to illustrate the powers of FDA. Section 3 contains a practical introduction to using FDA for phonetic research. In Section 4 we first investigate the role of duration, formant trajectories and f_0 alignment in the diphthong–hiatus contrast by treating each feature separately. After this, we present a joint analysis of the three features, and show that features that seem to play a powerful role when observed in isolation may not be as important when other features are taken into account. Section 5 contains a discussion of the advantages and limitations of FDA in relation to traditional analysis, with particular attention on the role of prior knowledge, and a sketch of the opportunities offered by FDA to the study of different kinds of phonetic phenomena. Finally, Section 6 concludes the paper. In conjunction with the Appendices, the supplementary materials² ³ and the web site on FDA maintained by the first author⁴, this paper should

²Code, data and plots not included in the text are available for download from this repository: <https://github.com/uasolo/FDA-DH/>. Direct link to zip bundle: <https://github.com/uasolo/FDA-DH/archive/master.zip>

³Direct link to Additional Material pdf:https://github.com/uasolo/FDA-DH/blob/master/paper/FDA-DH-Additional_Material.pdf?raw=true

⁴<http://lands.let.ru.nl/FDA>

enable researchers to apply FDA to their own data.

2. Case Study

In Spanish, vowel sequences of rising sonority (e.g. /ie/, /ia/, /ua/) are said to be syllabified in a generally predictable manner, depending on the location of lexical stress. Hiatuses (i.e. /Ci.a/) occur when lexical stress is on the initial high vowel; otherwise, the vowel sequence is realized as a diphthong /Cja/. Despite the fair degree of generality achieved by this rule, some lexical exceptions have been noted that make Spanish a language with a phonological contrast between hiatuses and diphthongs (e.g. *diente* ['djen.te] 'tooth' vs. *cliente* [cli.'ente] 'client'; *italiano* [i.ta.lja.no] 'Italian' vs. *liana* [li.'a.na] 'liana').

Although the idea of a phonological contrast between diphthongs and hiatuses in Spanish (D/H contrast from now on) is not controversial, its distribution in the lexicon, and the consistency with which it is realized phonetically, appear to vary across dialects and speakers. Hualde (2005) mentions that hiatuses are much more common in Castilian Spanish than in Latin American dialects, and also points to the existence of idiolectal variation within dialects. Several phonetic studies have investigated the acoustic basis of this contrast in several varieties of Spanish. In Aguilar (1999), rising diphthongs and hiatuses were extracted from a Barcelona Spanish corpus of map-task conversations and read sentences. The durations of the vowel sequences were measured, and formants were modeled with second order polynomials capturing the slope and curvature of their trajectories. It was found that hiatuses have a longer average duration and a greater degree of curvature in the F_2 trajectory than diphthongs, both in conversational and read speech. Hualde and Prieto (2002) investigated the D/H contrast by asking Madrid Spanish speakers to syllabify and read series of words containing the vowel sequence *ia* and by measuring the duration of the produced vowel sequences. They found that, as reported in Aguilar (1999), speakers produced longer vowel sequences in cases categorized as hiatuses than in those categorized as diphthongs. However, they also found that the duration distributions of the diphthong and hiatus groups overlapped considerably in the case of some speakers. More recently, Torreira (2007) analyzed the alignment of rising pitch accents in Spanish segmental sequences involving similar gestural content but differing in syllabic structure, including the diphthongs and hiatuses investigated in the present study. In agreement with the syllabification pat-

terns proposed by Hualde and Prieto (2002), it was found that rising pitch accents in /ia/ hiatuses were aligned with the second vowel of the sequence, while in diphthongs, the start of rising accents was aligned earlier, presumably at the onset of the syllable containing the diphthong. Since the onset of rising pitch accents in Spanish have been reported to be aligned with the beginning of lexically stressed syllables (Prieto et al., 1995; Prieto and Torreira, 2007), this study concluded that the differences in f_0 alignment must be due to differences in syllabification of /ia/ vowel sequences (i.e. /Cja/ vs. /Ci.a/). Although intonational features are suprasegmental by definition, it is possible that their alignment with segmental features can be used as cues to the identity of the latter. For this reason, we will also consider f_0 alignment as a potential cue of the D/H contrast, along more straightforward vocalic features such as duration and formant trajectories.

2.1. Materials

Part of the materials analyzed in this study comes from a previous experiment (Torreira, 2007) examining the alignment of f_0 rises across different syllabic contexts in Spanish, in which data were collected from five speakers. Because one of the goals of the present study is to investigate inter-speaker variation, four additional speakers were recorded using the same procedure and equipment as in Torreira (2007). All the participants spoke European varieties of Spanish. Four speakers were native of Cádiz, while the remaining five speakers came from the towns of Almería, Seville, Granada, Murcia and Majorca. Of the nine speakers, four speakers were female, and five male. The recordings were conducted in a silent room using a Shure SM10A head-mounted microphone and an M-Audio 410 FireWire external sound card connected to a computer.

Speakers read a series of carrier sentences presented on a computer screen. The carrier sentences were of the type "X *no, tu* Y" ('not X, your Y'), where X contained the target diphthongs and hiatuses and Y was a random noun (e.g. *mi liana no, tu hilo* ('not my liana, but your thread')). The target diphthongs and hiatuses in the X word always have a /l/ as left context and a nasal (/n/ or /m/) as right context. In read speech, the first part of this carrier sentence is typically realised with a rising f_0 contour on the target word, with a rising pitch accent associated to the lexically stressed syllable of the target word, ending in a high boundary tone. This intonation pattern was always used by the participants in the experiment.

Each speaker was presented with a list of 100 sentences. From these, 20 contained a diphthong and 20 a hiatus. The diphthong occurred in one of two possible target words, the proper nouns *Emiliano* [e.mi.lja.no] and *Emiliana* [e.mi.lja.na], while the hiatuses always occur within the word sequence *mi liana* [mi.li.a.na] ‘my liana’. The remaining sentences corresponded to the other three contrasts investigated in Torreira (2007), and can therefore be considered as distracters for the purpose of the present article. Due to reading errors part of the sentences were presented a second time to some of the speakers. Thanks to these repetitions we ended up with 183 useful diphthong and 182 useful hiatus tokens. We decided to include the five extra tokens in the data for analysis; therefore, our data set contains a total of 365 tokens.

2.2. Feature extraction

The materials were manually annotated by the second author, who is a trained phonetician and a native speaker of European Spanish. Annotation was carried out on the segments /lja/ or /li.a/ by marking the onset of /l/ and the onset and offset of the target vowel sequence /ja/ or /i.a/ (the vowel sequence onset coincides with the end of /l/). This annotation was used to determine the relevant time intervals for the extraction of durations, f_0 contours and formant contours, as detailed below. The duration feature (d , in milliseconds) corresponded to the duration of the target vowel sequence, [ja] or [i.a]. f_0 was extracted from all the tokens using the pitch detection function in Praat (Boersma and Weenink, 2009) set to default parameters, except for a fixed time step of 5 ms. f_0 contours were measured from the beginning of /l/ up to the f_0 maximum located towards the end of the target word. f_0 contours therefore always encompassed the complete accentual rise associated to the stressed syllable of the target words. Parts of f_0 beyond the accented syllable were included because it is not possible to separate the part of the rise belonging to the pitch accent to the part that belongs to the upcoming boundary tone, since they were in contiguous syllables. If there are differences between hiatus and diphthongs, we should be able to see them in the complete f_0 rise. The first two formants, F_1 and F_2 , were extracted using the Burg method available in Praat, set to default parameters. In the case of formants, only the portion within the vowel sequences was extracted, since automatic formant estimation is known to be unreliable in the surrounding liquids and nasals.

3. Functional Data Analysis

In this section we introduce a procedure that allows carrying out a statistical analysis on a set of contours. The procedure is based on Functional Principal Component Analysis (FPCA), which is one of the tools available within the Functional Data Analysis (FDA) framework Ramsay and Silverman (2005). The procedure is illustrated on the data sets introduced in Section 2.2. Most of the presentation is based on the 365 f_0 contours extracted from the sequences /lja/ for diphthongs and /li.a/ for hiatuses, while in Section 3.4 the 365 pairs of formant contours F_1 and F_2 extracted from /ja/ or /i.a/ are used to introduce multidimensional trajectory analysis. Importantly, the f_0 contours and formant trajectories under analysis are the result of segmentations (manually or automatically) performed by a researcher before the analysis. In this section we limit the explanation of theoretical and technical topics to what is necessary to understand the results presented in Section 4. Readers interested in applying Functional Data Analysis to their own data should consult also Appendix A, while readers interested in an in-depth explanation of FDA are referred to the relevant literature (Ramsay and Silverman, 2005, 2002; Ramsay et al., 2009).

3.1. Procedure overview

The FDA procedure is schematically summarized in Table 1, where f_0 contours are used for illustration. Sampled f_0 contours obtained from Praat undergo two pre-processing operations. First, f_0 values in Hz are converted into semitones (st). Then, the mean value (in st) of all samples in a contour is removed, so that the mean st-level is zero for all contours. These pre-processing operations are aimed at reducing gender-related effects. The removal of the mean value per utterance is motivated by the fact that there is no theoretical justification of why average pitch levels could differ between diphthongs and hiatuses. To make absolutely sure that there is no such effect, we checked that the means did indeed not differ; this was even the case when we analysed mean values separately for female and male speakers, to prevent the variance caused by the difference between the genders from obscuring within-gender effects. Similar –conventional– pre-processing operations may be called for when analysing other features than f_0 . The pre-processed data are subjected to a four-step procedure.

Smoothing (cf. Section 3.2) is the first operation applied to the pre-processed data. Smoothing transforms f_0 contours sampled at discrete points

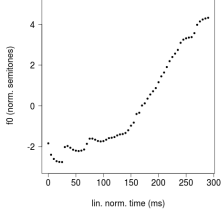
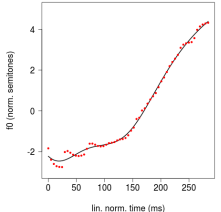
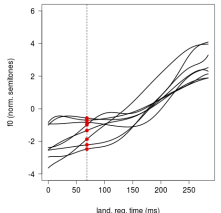
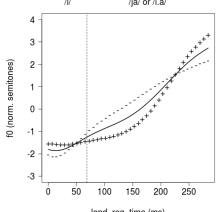
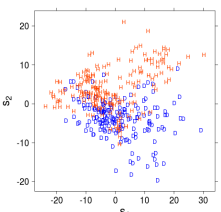
	<p>0. Raw data</p> <p>The input data is in the form of time sampled contours, e.g. the output of Praat pitch tracker.</p>
	<p>1. Smoothing</p> <p>Contours are represented in the form of smooth continuous functions of time. Note: durations are linearly normalized.</p>
	<p>2. Landmark registration</p> <p>The time axis of each contour is warped in order to synchronize the position of events across contours.</p>
	<p>3. Functional PCA</p> <p>The main shape variations across the set of contours are extracted. Each contour is parametrised by a set of <i>PC scores</i>.</p>
	<p>4. Class analysis</p> <p>The class membership information is correlated with PC scores.</p>

Table 1: The four-step FDA procedure applied on f_0 contours. The small figures on the left have iconic purpose and are reproduced in larger format in the text.

in time into continuous functions of time, so that f_0 values are defined for all values of time t . The procedure earned its name 'smoothing' from the fact

that it removes undesired detail from the sampled contours. By varying the smoothness of the f_0 curves one can decide to what extent microprosodic detail is to be incorporated in the representation. A side effect of the smoothing procedure is that all input curves are projected onto a fixed time interval.

Landmark registration (cf. Section 3.3) is a warping of the time axis that allows the user to smoothly modify the time axis to align corresponding events, typically syllable or phone boundaries. Different realisations of a segmental pattern exhibit variation in the duration of the corresponding phones or syllables. Landmark registration synchronises all time axes on the segmental boundaries indicated by the user.

Functional Principal Component Analysis (FPCA) (cf. Section 3.4) is the extension of conventional Principal Component Analysis (Jackson, 1991; Baayen, 2008) to input that is represented in the form of continuous functions. We can see FPCA as the ‘shape-to-numbers converter’ in the FDA work flow. FPCA provides a model of the (smoothed and landmark-registered) input curves in terms of a mean curve and a small number of Principal Component curves (PCs). Each PC curve represents a different deformation of the mean curve. Each input curve is associated with parameters called *PC scores*, each one determining the weight with which the corresponding deformation (PC) has to be applied in order to approximate that curve as closely as possible. With an analogy, we can think of each input curve as a different dish, and the purpose is to identify a small number of ingredients (PCs) with which we can reproduce individual dishes by only varying the ingredient dosages (PC scores).

Finally, *class analysis* (cf. Section 3.5) denotes any (ordinary) statistical analysis that we may carry out by combining the output of FPCA with the class membership information (D/H). At this stage we are liberated from the complexity of using contours as elements of the analysis, since FPCA has provided a numerical description of each contour in terms of PC scores. For example, we can apply a *t*-test on PC scores grouped by class in order to determine which score, and consequently which shape variation, correlates most with the D/H contrast. The four operations described above are explained in detail in the remainder of this section.

3.2. Smoothing

Smoothing transforms a sampled contour into a smooth continuous function of time. The target function is chosen from a set of possible functions. In the case of features like f_0 or formants, whose contours can assume a

wide range of shapes, it is customary to adopt *B-splines* as the function set (de Boor, 2001). A B-spline is a sequence of polynomial functions that, multiplied by appropriate weights and summed together, approximate a sampled data contour. A set of B-spline functions is shown in Figure 1(a), where each ‘hill’ corresponds to a polynomial function. B-splines approximation consists of adjusting the position and excursion (positive or negative) of each of the adjacent ‘hills’ so that their sum is a curve that is close to the original samples, as shown in the example in Figure 1(b).

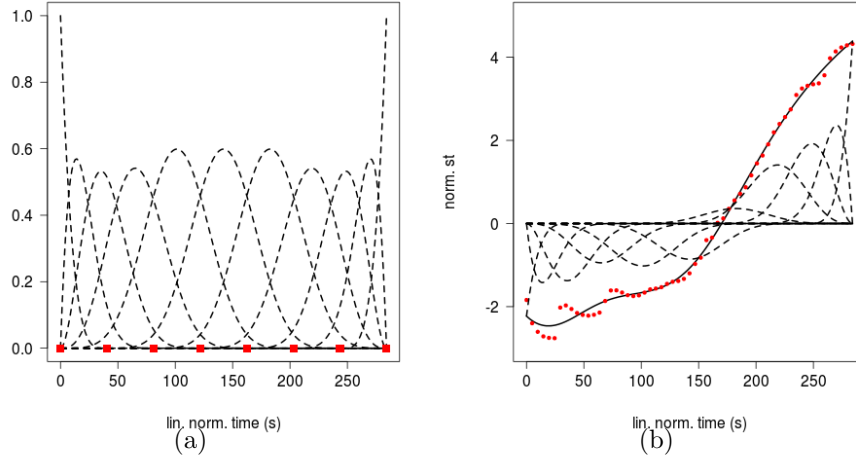


Figure 1: In (a), the B-splines basis used to represent the f_0 contours is shown. Dots indicate the points called *knots*, where spline ‘hills’ connect to each other (see text). In (b) an example of smoothing, where first each spline hill from (a) is multiplied by an appropriate coefficient, then all the resulting curves are summed together to obtain the continuous function (solid line), which approximates the f_0 samples (dots).

Although it is possible to construct a curve that coincides exactly with all the samples, this is usually a bad idea, since the result is likely to be a wiggly curve that is not a phonetically meaningful representation of the data, like the curve in Figure 2(a). There are at least two reasons to prefer a smoother curve. One is that we want to prevent overfitting, which in our case means that we do not want to reproduce all the erratic oscillations in the output of the f_0 (or formant) tracker, which may be due to inaccurate measurements. The other is that we are typically interested in a time resolution that is coarser than the one suggested by the richness of detail of Figure 2(a). For example, we may not be interested in microprosodic effects. If we opt for

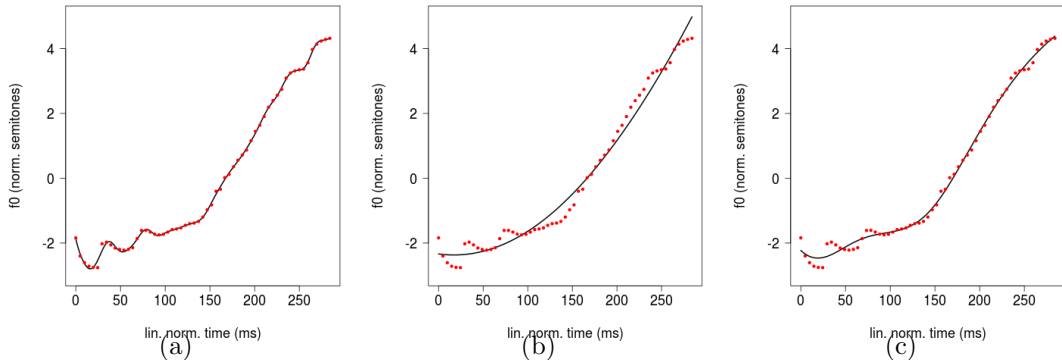


Figure 2: Three examples of f_0 contour smoothing. In (a) a case of overfitting, in (b) a case of underfitting, in (c) a good compromise.

a smoother curve, we have to accept some amount of fitting error, i.e. the function will not coincide with all the input samples exactly. Clearly, by smoothing too much we run the risk of removing potentially relevant detail, like in Figure 2(b). An example of good compromise between smoothing too much and too little is shown in Figure 2(c).

The degree of smoothing is controlled by two parameters. The first determines the number of hills; the second determines the balance between fitting error and roughness of the curve. Empirical methods exist that help to find the optimal values for those parameters. However, to account for the (often not quantitatively stated) preferences with respect to time resolution and degree of detail which are of interest in a research question, automatic parameter optimization should always be accompanied by visual inspection of the result. In Appendix A.1 a detailed example based on f_0 contours explains two different approaches to smoothing. Smoothing incorporates also a linear time registration that scales all contours to a common duration, a requirement of the tools downstream in the FDA procedure. The fact that duration needs to be normalised must be taken into account in interpreting the results. In this work, we have carried out a separate analysis of duration (Sec. 4.1), which eventually will be combined with the f_0 and formant shape analysis (Sec. 4.4). Other more sophisticated approaches to time normalisation are available, like the procedure proposed in Gubian et al. (2011) to jointly model duration and f_0 contours. However, those approaches add complexity to data analysis and resulting interpretation, which generally pays off only when utterances are longer and many local segmental durations need

to be analysed. In this paper, where f_0 trajectories are relatively short, we opted for a simpler approach in order to focus the exposition on the basic FDA operations.

3.3. Landmark registration

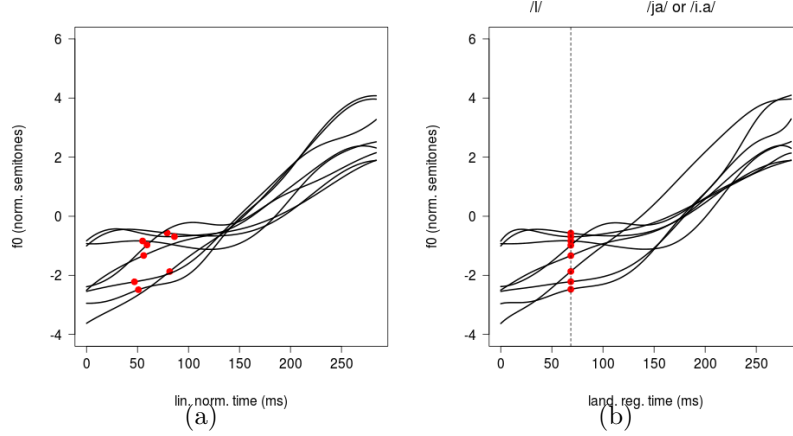


Figure 3: In (a) some smoothed f_0 contours, where the position of the only landmark (boundary between /l/ and vowel sequence) is marked with a dot. In (b) the result of landmark registration applied on the curves in (a).

The purpose of landmark registration is to time-align points on different contours that correspond to the same event. In our case, we align the boundary between /l/ and the vowel sequence across all f_0 contours, since we are interested in the position of the rise of the pitch accent relative to the segment boundaries, which in our case is the onset of /l/, and not in the starting point in terms of physical time from the beginning of the speech signal.⁵ Landmark registration prevents that, in the analysis steps that follow, parts of f_0 contours occurring before and after the offset of /l/ get mixed up, which would blur the results and make segment-related interpretation harder. In general there can be multiple landmarks, possibly at every phone boundary on a f_0 contour spanning a whole sentence, e.g. see Gubian et al.

⁵Strictly speaking, at this stage the time axis no longer represents physical time, because of the linear registration carried out in the smoothing process. If it is useful to distinguish between ‘normalised time’ and ‘physical (stopwatch) time’ we will use the symbol τ instead of t .

(2011). The alignment of corresponding events will produce results that can be interpreted in terms of those events, rather than in terms of potentially meaningless ‘stopwatch time’.

A user has to mark the position of the landmark events in the input curves. Next, (s)he must decide where the landmarks must be located on the registered time axis. Usually, the desired landmark positions, which are the same for all curves, are chosen to be at the mean position of the corresponding landmark positions in the time-normalised input curves (assuming all curves start at $t = 0$). Registration is carried out by a smooth time warping function, i.e. a function that maps the original time axis of a curve on its new time axis, such that landmarks are at the desired positions. The underlying smoothing procedure guarantees that curve deformation will be distributed along the time axis proportionally to the vicinity of landmarks, which ensures that the warping does not introduce jumps or discontinuities and that distortions are larger in the parts that require a larger displacement. Figure 3 shows a subset of our f_0 contours before and after registration. In Appendix A.2 more detail on the procedures involved in landmark registration is provided. In the document entitled *Time normalisation and landmark registration* in the additional material (see the web site referred to in footnote 3) we show that omitting landmark registration in the analysis of f_0 blurs the results substantially, even if the overall results go in the same direction.

3.4. Functional PCA

Functional Principal Component Analysis (FPCA) provides a model of the set of input contours in terms of combinations of a small number of curves, namely the mean curve and the principal component curves, plus weights for the principle component curves. The mean curve $\mu(t)$ is obtained by computing the mean of all input curves at each instant in time. The mean of the 365 smoothed and landmark-registered f_0 contours from our data set is shown in Figure 4(a). The principal component curves (PCs) are numbered from 1 onwards and are computed by the FPCA algorithm based on the same principles as ordinary PCA (Jackson, 1991; Baayen, 2008). The rank of the PCs reflects the decreasing percentage of variance in the input data that the PCs explain. Figure 4(b) and 4(c) display the first two PCs modeling the f_0 data set. Note that $PC1(t)$ and $PC2(t)$ do not look like the input curves, nor like the mean curve $\mu(t)$. This is because PCs are *shape modifiers*, i.e. they are added in a certain amount to the mean $\mu(t)$ so as to reproduce each input curve as faithfully as possible. Given an input curve

$f(t)$, FPCA provides the weights s_1, s_2 , etc., called *PC scores*, which produce the best approximation of $f(t)$ according to the formula:

$$f(t) \approx \mu(t) + s_1 \cdot PC1(t) + s_2 \cdot PC2(t) + \dots \quad (1)$$

This principle is illustrated in Figure 4(d), 4(e) and 4(f), where a curve $f(t)$ from the f_0 data set is plotted as a dashed line. The first two PC scores associated to this particular curve are $s_1 = 16.7$ and $s_2 = 11.7$. Figure 4(d) compares the mean curve $\mu(t)$ (solid line) and $f(t)$; clearly, $\mu(t)$ alone is a poor approximation of $f(t)$. Figure 4(e) shows the improvement obtained by approximating $f(t)$ with $\mu(t) + s_1 \cdot PC1(t)$, i.e. using only the first PC in Eq. (1). Figure 4(f) shows the result using the first two PCs in Eq. (1).

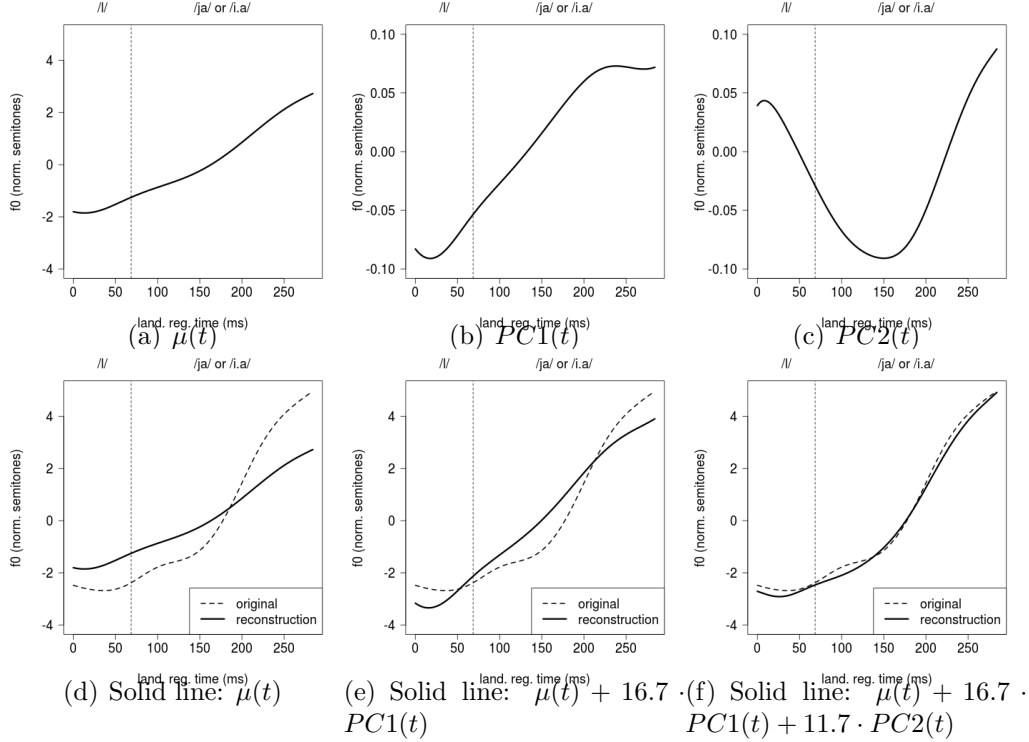


Figure 4: Upper row: (a) mean and (b,c) the first two principal component functions modeling the f_0 contour data set. Lower row: Dashed line is the input contour $f(t)$; solid lines: three approximations of $f(t)$ that use (d) no PCs, (e) one and (f) two PCs, according to Eq. (1).

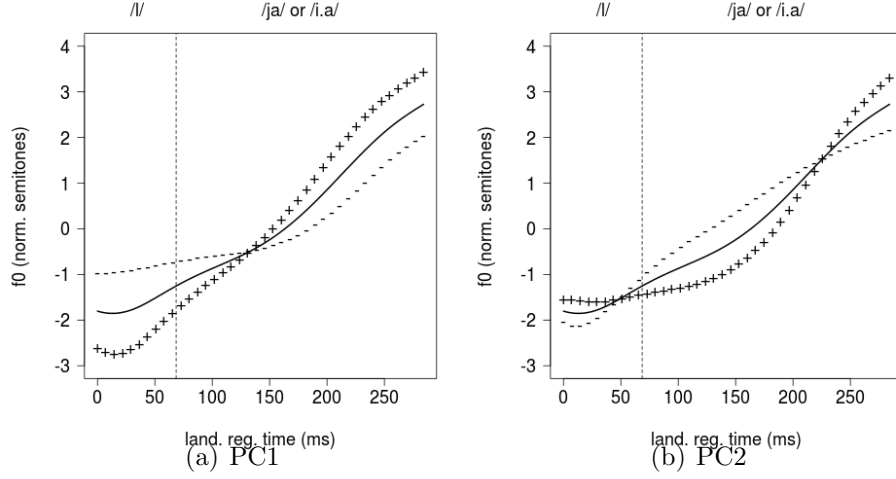


Figure 5: FPCA applied to f_0 contours according to Eq. (1). Each panel shows in solid the mean curve $\mu(t)$ and the \pm curves obtained by adding to or subtracting from $\mu(t)$ the curve (a) $\sigma(s_1) \cdot PC1(t)$ and (b) $\sigma(s_2) \cdot PC2(t)$, respectively, where σ denotes standard deviation ($\sigma(s_1) = 10.0$; $\sigma(s_2) = 6.5$). The x-axis reports registered time in ms, the y-axis frequency values in semitones, where the mean value from each curve was removed (thus corresponding to the zero level).

So, FPCA yields two very different types of output: the PC curves that serve for all contours in the data under analysis, and the set of weights (PC scores) that must be used to approximate the individual contours. The weights can be used in subsequent statistical analyses. The PC curves make it possible to understand the phonetic effect of adding a PC with a specific weight to the mean curve.

Figure 5 shows a convenient way to display FPCA curves, where the action of each PC on the mean curve $\mu(t)$ is displayed independently on a different panel. Each panel contains three curves, namely $\mu(t)$, which is drawn as solid line and is the same for both panels, and two other curves, drawn with “+” and “−” symbols, that represent the result of adding to or subtracting a PC curve from $\mu(t)$. For example, the “−” curve in Figure 5(a) represents the curve $\mu(t) - \sigma(s_1) \cdot PC1(t)$, i.e. the mean curve minus PC1 multiplied by the standard deviation of the score s_1 computed on the whole data set. Figure 5(a) suggests that PC1 mainly alters the slope of the f_0 contours, with positive/negative s_1 scores making the curve more/less steep than $\mu(t)$. Figure 5(b) suggests that PC2 acts as an elbow somewhere in the middle of the vowel sequence, where positive s_2 scores accentuate this elbow,

while negative scores make it less prominent or even eliminate it altogether. As with conventional multivariate procedures, for example Factor Analysis, it cannot be guaranteed that the PCs which result from an FPCA will always be easy to interpret.

3.4.1. Joint analysis of multiple contours

FPCA can be applied to multi-dimensional curves or trajectories, provided that they share the same time axis. An example is provided by applying FPCA jointly on formants F_1 and F_2 . The FPCA model for this set of curve pairs $(F_1(t), F_2(t))$ is

$$F_1(t) \approx \mu_{F_1}(t) + s_1 \cdot PC1_{F_1}(t) + s_2 \cdot PC2_{F_1}(t) + \dots \quad (2a)$$

$$F_2(t) \approx \mu_{F_2}(t) + s_1 \cdot PC1_{F_2}(t) + s_2 \cdot PC2_{F_2}(t) + \dots \quad (2b)$$

where each of the two equations models one of the formants in the same way as Eq. (1) does for f_0 . Crucially, though, Eq. (2a) and (2b) *share the same PC scores*. This means that PCs, which are pairs of functions taking values in F_1 and F_2 , act jointly on the mean formant contour pair $(\mu_{F_1}(t), \mu_{F_2}(t))$. For example, if $s_1 = 3$ for a given input pair of formant contours $(F_1(t), F_2(t))$, then the mean contour of F_1 has to be altered by adding three times the PC1 curve for F_1 to it $(\mu_{F_1}(t) + 3 \cdot PC1_{F_1}(t))$, and at the same time the mean contour of F_2 has to be altered by adding three times the PC1 curve for F_2 to it $(\mu_{F_2}(t) + 3 \cdot PC1_{F_2}(t))$. Thus, while $PC1_{F_1}(t)$ and $PC1_{F_2}(t)$ are different and act on different mean contours, they are applied in the same ‘dosage’ s_1 to the mean contours $\mu_{F_1}(t)$ and $\mu_{F_2}(t)$. The advantage of building a joint FPCA model, as opposed to applying separate FPCA procedures to F_1 and F_2 , is that temporal dependencies across formants are captured automatically. If we were to carry out one independent FPCA for each formant we would eventually find their respective PC scores to be strongly correlated. However, that numerical correlation would not provide us with explicit information on where on the (common) time axis the curves change their shape together. In fact, we expect F_1 and F_2 to move together in time, since we know that they are two parameters describing a common entity, namely the spectrum of a vowel.

Figure 6 displays the first two pairs of PCs obtained by applying FPCA on the 365 formant contour pairs in our data set. Formant curves are obtained by applying the smoothing procedure described in Section 3.2; we did not apply landmark registration, since there is no principled way for locating a

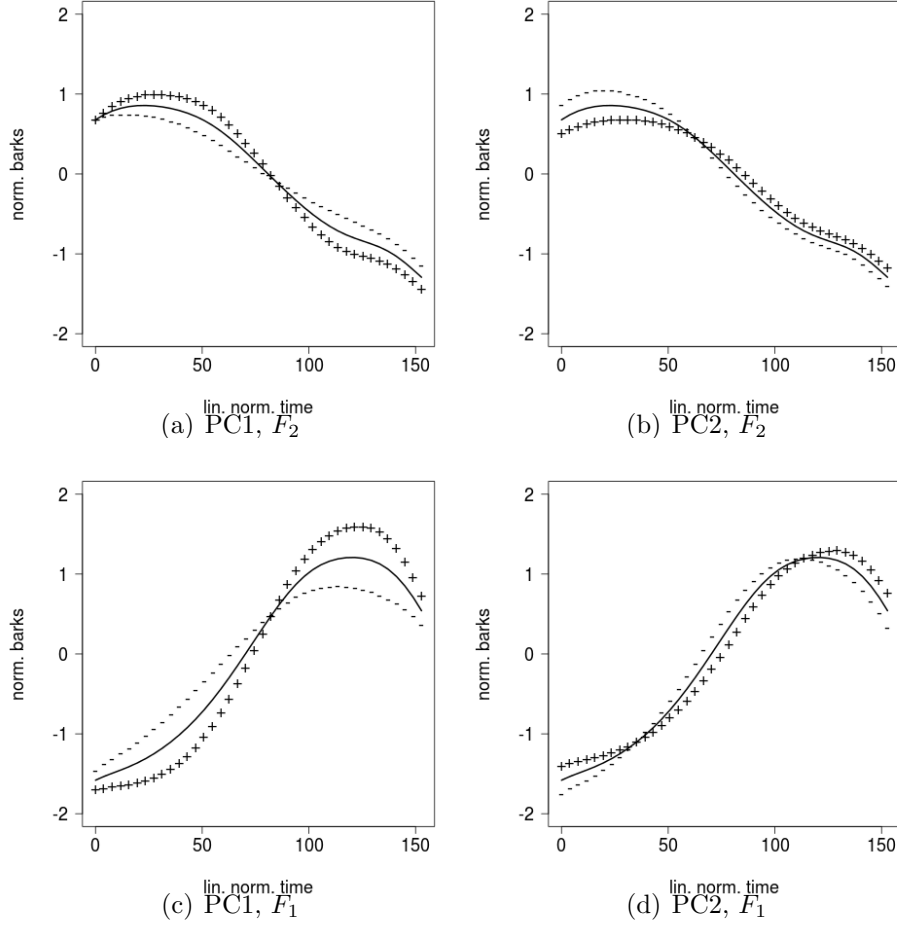


Figure 6: FPCA applied to formants F_1 and F_2 according to Eq. (2). Each panel shows in solid the mean curve for a given dimension, $\mu_{F_1}(t)$ or $\mu_{F_2}(t)$, and the \pm curves obtained by adding to or subtracting from the mean of a dimension a given PC curve multiplied by a PC score equal to one standard deviation of the distribution of that score ($\sigma(s_1) = 4.0$; $\sigma(s_2) = 2.3$). The x-axis reports linearly normalised time in ms, the y-axis frequency values in barks, where the mean value from each curve was removed (thus corresponding to the zero level). F_2 panels are placed above F_1 panels to help the reader recognizing familiar vowel patterns observed by looking at spectrograms.

landmark event in a diphthong (cf. Section 2.2). Figure 6(a) and 6(c) suggest that the PC1 curves capture the difference between a wide and rapid formant movement (+ curves) and a flatter and a more gradual transition (− curves). Considering the PC2 curves, Figure 6(b) looks like Figure 6(a) with + and − curves reversed, while Figure 6(d) suggests a slight time shift in F_1 .

3.5. Class analysis

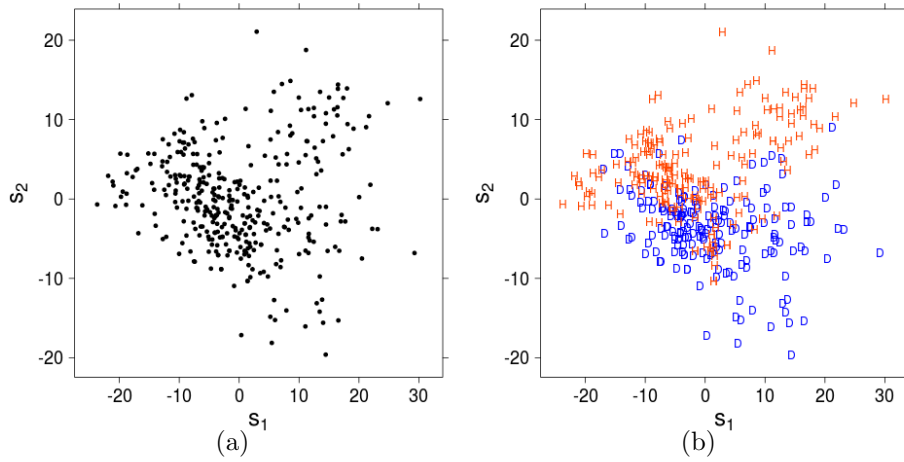


Figure 7: In (a) the PC scores s_1 and s_2 corresponding to the 365 f_0 contours in the D/H data set. In (b) the same points labeled according to the class D/H each contour belongs to.

Models like Eq. (1) and (2) are data-driven parameterisations of a set of contours or trajectories, where PC scores are the numerical parameters. A scatter plot of the first two PC scores of the 365 f_0 contours modeled by FPCA is shown in Figure 7(a). Each point corresponds to the values of s_1 and s_2 in Eq. (1) for a specific contour. PC scores were obtained by running FPCA without using class membership information, i.e. in our case D and H contours were not distinguished at the input. Class analysis reintroduces this information, so that it becomes possible to discover quantitative relations between PC scores and the linguistic categories or classes under study (D/H). This will enable us to interpret those relations in terms of contour shape properties by virtue of the link between PC scores and PC curves.

Figure 7(a) shows the locations of the 365 contours in the s_1, s_2 plane in the form of black dots; the same symbol is used for H and D tokens. In

Figure 7(b) the dots are replaced by the class membership labels D or H. Note that in this case the second PC score appears to be strongly correlated with the D/H contrast, while the first one is not. This suggests that f_0 slope, mostly captured by PC1 (cf. Section 3.4) does not play an important role in the realisation of the D/H contrast, while the presence or absence of an elbow, described by PC2, seems to be the main correlate of the contrast in the f_0 movement. A complete account on f_0 is given in Section 4.2, where a number of statistical models and tests are applied on PC scores, and links to f_0 gestures are provided.

PC scores can play the role of variables in statistical models where other numerical variables are present. The latter may be either PC scores from other FDA procedures or conventional numerical types of features, such as segment durations. In this way, features describing the dynamic shape of contours are treated like any other numerical feature and analysed by applying standard statistical methods, e.g. linear (mixed effects) models. In Section 4.4 and Appendix B we present several ways for analysing the joint behavior and interactions of vowel sequence duration, f_0 and formant movement in the context of the D/H contrast realisation.

4. Data analysis

In this section we investigate how vowel sequence duration and the shape of f_0 and formant contours vary in the realisation of the D/H contrast in Spanish. We present a detailed analysis based on the data set composed of 365 utterances, spoken by nine speakers described in Section 2. While the duration of the /ja/ or /i.a/ vowel sequence is a scalar feature (d) that can be directly used in a statistical test or model, the shape of f_0 and formant trajectories will be represented in the form of numerical parameters obtained from the application of Functional PCA. In particular, the shape of f_0 contours will be parameterised by the first two PC scores ($s_1^{f_0}$ and $s_2^{f_0}$) obtained from the application of FPCA to f_0 contours spanning /lja/ or /li.a/, and the shape of formant trajectories will be parameterised by the first two PC scores ($s_1^{F_{1-2}}$ and $s_2^{F_{1-2}}$) obtained from the application of FPCA to formants F_1 and F_2 spanning /ja/ or /i.a/ (cf. Section 3.4). This leaves us with five numerical features (d , $s_1^{f_0}$, $s_2^{f_0}$, $s_1^{F_{1-2}}$ and $s_2^{F_{1-2}}$) on which we will carry out

statistical analysis.⁶

The goals of the analyzes are expressed in a number of questions concerning the realisation of the D/H contrast. For each feature we want to know:

- (I) Is the feature relevant for the D/H contrast?
- (II) If so, is it used by all speakers?
- (III) Is it used consistently, i.e. in the same direction by all speakers?

Next, in Section 4.4, we will take a look at those features that proved to be relevant in combination and ask:

- (IV) Are there global trends in the joint use of features, for example in terms of trade-offs (i.e. cue trading)?
- (V) Can we characterise differences between speakers in terms of different trade-offs in the use of the features?

In sections 4.1, 4.2 and 4.3 we address the first three questions by analysing each feature separately. Let y be one of the numerical features introduced above (e.g. $y = s_1^{f_0}$). Question (I) is addressed by building a linear regression model of the form

$$y = \beta_0 + \beta_1 \cdot x, \tag{3}$$

where y is the dependent variable and x is a binary variable encoding class ($x = 0$ if class is D, $x = 1$ if class is H). This model is used to assess how much of the variation of y is explained by the class of each token. The relevance of feature y will be evaluated by examining the significance of the coefficient β_1 and the percentage of variance explained by the model (R^2). Questions (II) and (III) are addressed by running nine speaker-specific models for each of the relevant features. These models have the same form of Eq. (3) but are based on the data from only one speaker. By examining β_1 coefficients and respective two-sided confidence intervals we will be able to determine which speakers use which feature in the expected direction.

⁶Superscript notation ($s_1^{f_0}$, etc.) is used in the remainder of this paper to distinguish scores from different FPCA models.

y	β_0	β_1	F-statistics	p -value	R^2
d [ms]	134	47	$F(1, 363) = 297.1$	$p < 0.001$	0.45
$s_1^{f_0}$	1.3	-2.7	$F(1, 363) = 6.9$	$p = 0.01$	0.016
$s_2^{f_0}$	-3.8	7.6	$F(1, 363) = 192.4$	$p < 0.001$	0.35
$s_1^{F_{1-2}}$	-2.3	4.6	$F(1, 363) = 186.9$	$p < 0.001$	0.34
$s_2^{F_{1-2}}$	-0.3	0.6	$F(1, 363) = 7.3$	$p = 0.007$	0.02

Table 2: Summary of the models of the general form: $y = \beta_0 + \beta_1 \cdot x$ in Eq. (3). In all models, $x = 0$ when a token is in class D, $x = 1$ when it is in class H. Rows correspond to the individual features.

Questions (IV) and (V) are tackled in Section 4.4 by building logistic regression models, where the class of a token is predicted by a linear combination of the features that were deemed relevant in the analysis of the individual features. Contrary to the approach followed in Eq. (3), where the purpose is to assess the relevance of a feature, the features will be used as predictors of class membership. This will allow us to investigate interactions between the relevant features. Question (V) will be addressed by building logistic regression models for all nine individual speakers. The difference between the global and the particular models will help to discover speaker-specific trends or preferences in the use of cues.

We realise that building a large number of simple linear models according to Eq. (3), instead of more powerful mixed effects models that would allow us to combine the questions about global and speaker-specific effects in a single analysis, may seem to be in discord with the advanced FDA procedures advocated above. Yet, we decided to postpone the presentation of the results of mixed effect models to Appendix B, for two reasons. The first reason is didactic. By limiting ourselves to presenting the results of well-known statistical analyses we hope to make the FDA analyses easier to digest. The second reason is that Appendix B shows that for our data the more complex mixed effects models only confirm the conclusions drawn from the simpler models.

4.1. Analysis of duration

The first row in Table 2 shows the result of using d as dependent variable in Eq. (3). The predicted duration of D is 133 ms, while for H this is 181 ms, which is in line with the findings in Aguilar (1999) and in Hualde and Prieto

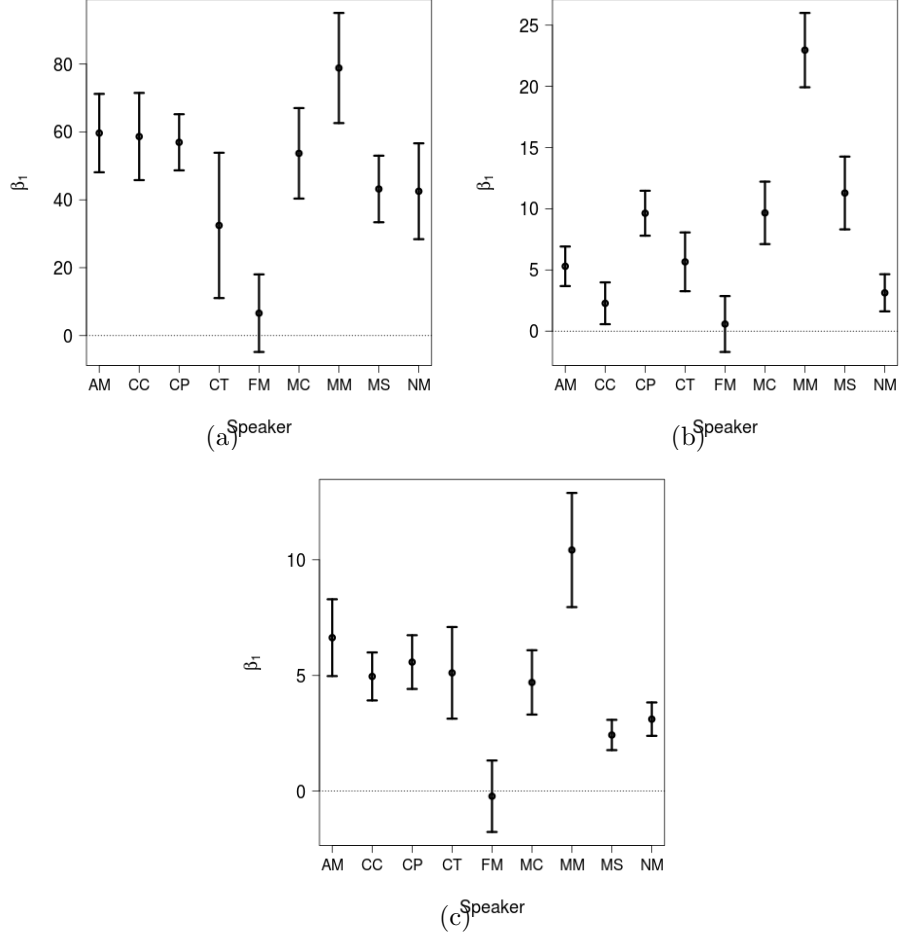


Figure 8: Values of β_1 coefficient and corresponding 95% two-sided confidence intervals computed applying Eq. (3) on data from each speaker separately (speakers are indicated on the x-axis). Results for d (a), $s_2^{f_0}$ (b) and $s_1^{F_1-2}$ (c) are shown. Confidence intervals that do not cross the zero line correspond to speakers whose realisation of H is significantly different from that of D with respect of the given feature.

(2002). The difference between the means is highly significant and *class* (i.e. x) explains 45% of the variance, which suggests that duration is relevant for the D/H contrast.

Insight in per-speaker behavior is provided in Figure 8(a), where Eq. (3) is applied to the values of d for each speaker separately. Confidence intervals for β_1 coefficients show that all speakers except for FM are using the duration

cue and in the expected direction, i.e. H is longer than D. Visual inspection of the data (box plots in supplementary material, cf., footnote 3) reveals that speaker FM is not using duration to produce the contrast.

4.2. Analysis of f_0 contours

In Section 3.4 it was explained how Functional PCA was applied to the 365 f_0 contours spanning /lja/ for D and /li.a/ for H, respectively. The first two PCs explain respectively 61.0% and 26.3% of the variance of the contour data set. PC3 was computed too, but it appeared to explain little variance and it did not contribute to explaining the D/H contrast. Figure 5 portrays the effect of the two PCs. PC1 mainly modulates slope, while PC2 modulates the sharpness of an elbow in the middle of the vowel sequence. Those modulations are controlled by the PC scores $s_1^{f_0}$ and $s_2^{f_0}$.

The second row of Table 2 reports the result of predicting score $s_1^{f_0}$ from token *class*. An effect of *class* seems to be present, since $s_1^{f_0}$ is significantly larger for D, which means that f_0 contours are steeper for D (as can be seen from the + curve in Figure 5(a)). However, *class* explains only 1.6% of the variance of $s_1^{f_0}$. Moreover, a predicted difference of 2.7 units of $s_1^{f_0}$ between D and H is small in terms of f_0 contour shape variation. This can be inferred from the fact that the distance between the mean f_0 curve and the curve displayed as + signs in Figure 5(a) corresponds to a difference of 10.0 PC score units for PC1. Additional investigation revealed that a large part of the variation of $s_1^{f_0}$ is related to speaker identity (see the box- and curve plots in supplementary material, cf., footnote 3). It can be concluded that the f_0 contour slope mainly varies across speakers, irrespective of token class. This allows us to conclude that f_0 contour steepness is not a relevant feature for the D/H contrast, which is in line with the previous findings Torreira (2007). Therefore, the feature $s_1^{f_0}$ will be excluded from the global analysis in Section 4.4.

The third row of Table 2 contains the result of predicting score $s_2^{f_0}$ using Eq. (3). The value of $s_2^{f_0}$ is larger for H, which in this case means that f_0 contours present a sharper elbow for H (cf. + curve in Figure 5(b)). The effect is not only significant but also relevant, since *class* alone explains 35% of the variance of $s_2^{f_0}$, which qualifies $s_2^{f_0}$ as a relevant feature for the D/H contrast. Figure 9 shows how the values of $s_2^{f_0}$ for D and for H predicted by Eq. (3) translate into predicted f_0 contours. The two curves are obtained by using only PC2 in Eq. (1). The D curve is $\mu(t) - 3.8 \cdot PC2(t)$, where $s_2^{f_0} = -3.8$

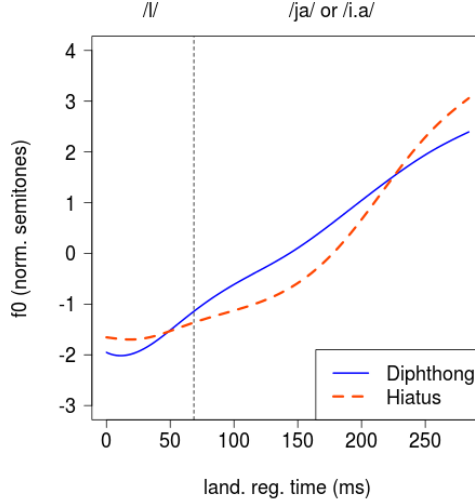


Figure 9: Two f_0 contours obtained by substituting the values of PC score $s_2^{f_0}$ predicted by the linear model $s_2^{f_0} = -3.8 + 7.6 \cdot x$ (cf. third row of Table 2) into the FPCA model in Eq. (1). The D curve is $\mu(t) - 3.8 \cdot PC2(t)$ ($s_2^{f_0} = -3.8$ is the value predicted for diphthongs by the linear model). The H curve is $\mu(t) + 3.8 \cdot PC2(t)$ ($s_2^{f_0} = 3.8$ is the value predicted for hiatuses).

corresponds to the value predicted for D, i.e. when $x = 0$ (cf. third row of Table 2), while the H curve is $\mu(t) + 3.8 \cdot PC2(t)$, where $s_2^{f_0} = -3.8 + 7.6 = 3.8$ corresponds to the value predicted for H, i.e. when $x = 1$. Figure 9 allows us to say that the typical H curve contains a clear elbow in the middle of /i.a/, while the typical D curve does not vary its slope within /ia/. This result is in line with the findings on the alignment of f_0 movements and syllable boundaries in Torreira (2007).

Figure 8(b), which displays confidence intervals for β_1 in Eq. (3) applied on the values of $s_2^{f_0}$ for each speaker separately, shows a trend similar to the one found for duration, yet with some differences. Like in Figure 8(a), all speakers except for FM make use of the cue in the expected direction ($\beta_1 > 0$ means H has an elbow, D has not). However, some speakers (e.g., CC and NM) seem to use this cue much less than other speakers. Visual inspection of the data (see the box- and curve plots in supplementary material, cf. footnote 3) reveals that speaker FM is not varying the shape of f_0 contours to produce the contrast, and that for speakers CC and NM the presence/absence of the elbow is hardly visible.

4.3. Joint analysis of formant contours F_1 and F_2

In Section 3.4 Functional PCA was applied to the 365 (F_1, F_2) contour pairs spanning /ja/ for D and /i.a/ for H. The first two PCs explain 56.1% and 18.4% of the variance of the contour data set. PC3 was computed too, but it explained only a very small proportion of the variance and it appeared not to contribute to the D/H contrast. Figure 6 shows the effect of the two PCs. PC1 changes the shapes of the formant trajectories from a sharp transition between /i/ and /a/, with flatter regions at the extremes (+ curves in Figure 6(a) and 6(c)) to a shallower and more gradual movement without clear plateaus at the beginning and end (− curves in the same figures). PC2 acts mainly on F_2 . Its effect on the shapes is similar to the effect of PC1, but with opposite sign (compare +/− curves in Figure 6(a) with −/+ curves in Figure 6(b)). Those changes are determined by the PC scores $s_1^{F_1-2}$ and $s_2^{F_1-2}$.

The fourth row of Table 2 contains the result of predicting score $s_1^{F_1-2}$ using Eq. (3). The value of $s_1^{F_1-2}$ is larger for H, which in this case means that formant contours tend to show a wider movement between vowels /i/ and /a/ for H and a more gradual movement for D. The effect is not only significant but also relevant, since *class* alone explains 34% of the variance of $s_1^{F_1-2}$, which qualifies $s_1^{F_1-2}$ as a relevant feature for the D/H contrast. Figure 10 shows how the values of $s_1^{F_1-2}$ for D and for H predicted by Eq. (3) translate into predicted formant contours. The two curve pairs are obtained by using only PC1 in Eq. (2). The D curves are $\mu_{F_1}(t) - 2.3 \cdot PC1_{F_1}(t)$ for F_1 and $\mu_{F_2}(t) - 2.3 \cdot PC1_{F_2}(t)$ for F_2 , where $s_1^{F_1-2} = -2.3$ corresponds to the value predicted for D, i.e. when $x = 0$ (cf. fourth row of Table 2), while the H curves are $\mu_{F_1}(t) + 2.3 \cdot PC1_{F_1}(t)$ for F_1 and $\mu_{F_2}(t) + 2.3 \cdot PC1_{F_2}(t)$ for F_2 , where $s_1^{F_1-2} = -2.3 + 4.6 = 2.3$ corresponds to the value predicted for H, i.e. when $x = 1$. Figure 10 allows us to say that the typical H formant curves indeed exhibit a movement that suggest a rapid transition between a stable /i/ and a stable /a/, while the typical D curves are more gradual, with no clear stable regions. This finding is in line with Aguilar (1999), among others.

Figure 8(c), which displays confidence intervals for β_1 in Eq. (3) applied to the values of $s_1^{F_1-2}$ for each speaker separately, shows a trend similar to the one found for duration in Figure 8(a), namely all speakers, except FM, use the cue in the expected direction.

The last row of Table 2 reports the result of predicting score $s_2^{F_1-2}$ using

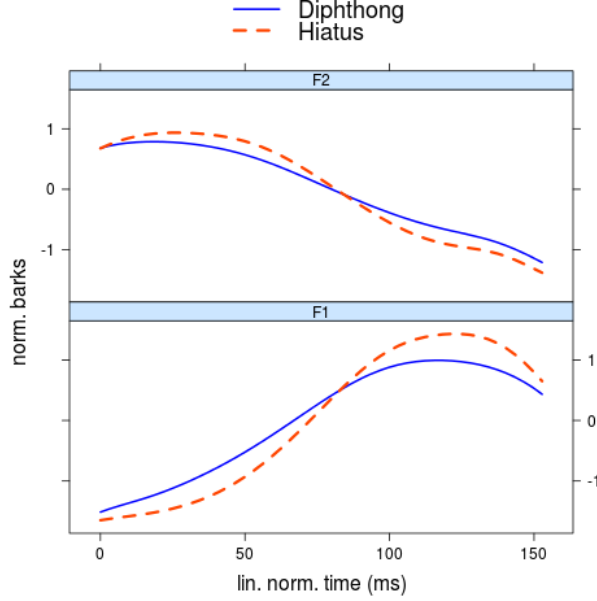


Figure 10: Two (F_1, F_2) contour pairs obtained by substituting the values of PC score $s_1^{F_1-2}$ predicted by linear model $s_1^{F_1-2} = -2.3 + 4.6 \cdot x$ (cf. fourth row of Table 2) into the FPCA model in Eq. (2). The D curves are resp. $\mu_{F_1}(t) + s_1^{F_1-2}|_{x=0} \cdot PC1_{F_1}(t)$ and $\mu_{F_2}(t) + s_1^{F_1-2}|_{x=0} \cdot PC1_{F_2}(t)$, where $s_1^{F_1-2}|_{x=0} = -2.3$, i.e. the value predicted for diphthongs by the linear model. Similarly, the H curves are resp. $\mu_{F_1}(t) + s_1^{F_1-2}|_{x=1} \cdot PC1_{F_1}(t)$ and $\mu_{F_2}(t) + s_1^{F_1-2}|_{x=1} \cdot PC1_{F_2}(t)$, where $s_1^{F_1-2}|_{x=1} = 2.3$, i.e. the value predicted for hiatuses.

Eq. (3). The value of $s_2^{F_1-2}$ is larger for H, which in this case roughly means that F_2 tends to be flatter for H, which is the opposite of the effect found for $s_1^{F_1-2}$, while F_1 changes very little. However, the effect is rather small, since the model explains only 2% of the variance of $s_2^{F_1-2}$. Moreover, a predicted difference of 0.6 units of $s_2^{F_1-2}$ is small in terms of formant contour shape variation, as it translates into a distance between curves that is roughly four times smaller than the one between the + and the solid curves in Figure 6(b) and 6(d), whose $s_2^{F_1-2}$ parameters differ in 2.3 units. This allows us to conclude that $s_2^{F_1-2}$, mainly a correction on F_2 contours, is not a relevant feature for the D/H contrast. Therefore, the feature $s_2^{F_1-2}$ will be excluded from the global analysis in Section 4.4.

	d	$s_2^{f_0}$	$s_1^{F_{1-2}}$
d	1	0.43	0.63
$s_2^{f_0}$		1	0.41
$s_1^{F_{1-2}}$			1

Table 3: Spearman correlations between the features d , $s_2^{f_0}$ and $s_1^{F_{1-2}}$.

4.4. Combined features analysis

In this section we carry out a combined analysis of the three features that were found to be relevant for the realisation of the D/H contrast. These features are the duration of the vowel sequence or diphthong d (cf. Section 4.1), the PC2 score $s_2^{f_0}$ from FPCA applied on f_0 contours, (cf. Section 4.2), and the PC1 score $s_1^{F_{1-2}}$ from FPCA jointly applied to formants (cf. Section 4.3). Here, we are interested in finding general patterns as well as individual differences among speakers in the coordination of gestures that are used in the production of D and H (cf. questions IV and V at the beginning of this section).

General patterns are investigated by building a logistic regression model, where the class of a token is the predicted variable, and the three features d , $s_2^{f_0}$ and $s_1^{F_{1-2}}$ are the predictors. First, the correlation structure between features is analysed, in order to assess collinearity. Table 3 shows the Spearman correlations between all pairs of features. The high correlation values, as well as a condition number as high as 15.6 (Belsley et al., 1980), discourage the use of the original features in a linear model without applying a suitable decorrelation transformation. We eliminate collinearity by centering and scaling each feature on its mean and standard deviation and then performing an ordinary Principal Component Analysis. The loadings of the original features on the principal components are shown in Table 4, where, to avoid confusion with Functional PCA, small letters are used for principal components. The first pc , which explains most of the variance, is basically a combination of the three original features with equal loadings and signs. The loadings of $pc2$ produce a dimension that is basically the sum of d and $s_1^{F_{1-2}}$ minus two times $s_2^{f_0}$; the loadings of $pc3$ represent the difference between d and $s_1^{F_{1-2}}$.

At this point we have a new set of features, $pc1$, $pc2$, and $pc3$, which can be used safely as predictors in a generalised linear model. To identify a parsimonious model, we start with a generalised linear model in which the

variables $pc1$, $pc2$, $pc3$ predict (the logit of) the probability that a token is an H. This redundant model is then pruned by applying fast backward variable selection (Lawless and Singhal, 1978). The result is the model:

$$\text{logit}(Pr(H)) = 1.97 \cdot pc1 + 1.20 \cdot pc3 - 1.36 \cdot pc1 \cdot pc3, \quad (4)$$

where all coefficients are significant (p-value < 0.001) and Somers' $D_{xy} = 0.86$, which indicates a high predictive power (Siegel and Castellan, 1988). Note that there is no intercept, which is a consequence of the fact that the data set is well balanced between the two classes.

From the analysis of the individual features in the previous sections we know that the values of d , $s_2^{f_0}$ and $s_1^{F_1-2}$ tend to be higher for H than for D (cf. β_1 in Table 2). Since $pc1$ is the sum of those features, this term says that the higher that sum is, the higher the probability that the token is a H. More interesting insight comes from the other terms, both of which contain $pc3$, which expresses a trade-off between d and $s_1^{F_1-2}$. Interestingly, the equation does not contain terms in $pc2$, which expresses a trade-off between $s_2^{f_0}$ and the other two terms. These observations suggest two things: that $s_2^{f_0}$ might be less systematically related to the token class than the other two features and that d and $s_1^{F_1-2}$ may be in some trade-off relation in determining the class.

Figure 11 provides a representation of Eq. (4), where $pc1$ and $pc3$ are converted back into their combinations of centered and normalised values of d (x-axis) and $s_1^{F_1-2}$ (y-axis), while $s_2^{f_0}$ is ignored. The plot shows a separation region between D and H that is curved (because of the interaction term $pc1 \cdot pc3$) but tending to be vertical, which suggests that duration, more than formant shape, is the most globally reliable cue to distinguish D from H.

	$pc1$	$pc2$	$pc3$
d	0.59	0.37	0.72
$s_2^{f_0}$	0.55	-0.84	-0.02
$s_1^{F_1-2}$	0.59	0.41	-0.70
var.	0.79	0.14	0.07

Table 4: Ordinary PCA computed on features d , $s_2^{f_0}$ and $s_1^{F_1-2}$, which were previously centered on their mean value and divided by their standard deviation. Each column shows the loadings of the original features on the new pc coordinates. In the last line, the resp. fraction of explained variance.

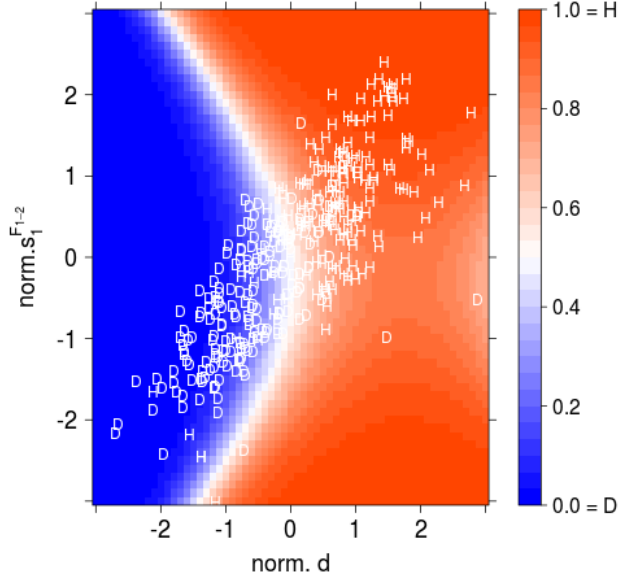


Figure 11: Representation of Eq. (4). Predictors $pc1$ and $pc3$ are converted back into their combinations of centered and normalised values of d (x-axis) and $s_1^{F_{1-2}}$ (y-axis), while $s_2^{f_0}$ is ignored (formally, it is set to $s_2^{f_0} = 0$). Each of the 365 tokens is represented by a letter indicating its class (D or H). In false colours, the probability of H as predicted by Eq. (4).

Insight in speaker-specific trade-offs is gained by building nine logistic regression models like Eq. (4) on data from individual speakers. Although none of the models yielded significant terms, the trends they exhibit are insightful. Figure 12 shows the result in a way similar to Figure 11. It can be seen that the separation regions for most speakers are different than in Figure 11. Tokens from speakers CC and CT seem to be best classified by relying on the formants feature $s_1^{F_{1-2}}$, since the separation is clearly closer to a horizontal than to a vertical line. Moreover, the tokens from speaker FM are all concentrated in the undecided region where the probability of H is close to 0.5, which confirms that FM is not producing the contrast (cf. similar trends in Figure B.2).

5. Discussion

As explained in the Introduction, previous phonetic research on D/H contrast in Spanish has identified several acoustic cues that are involved in

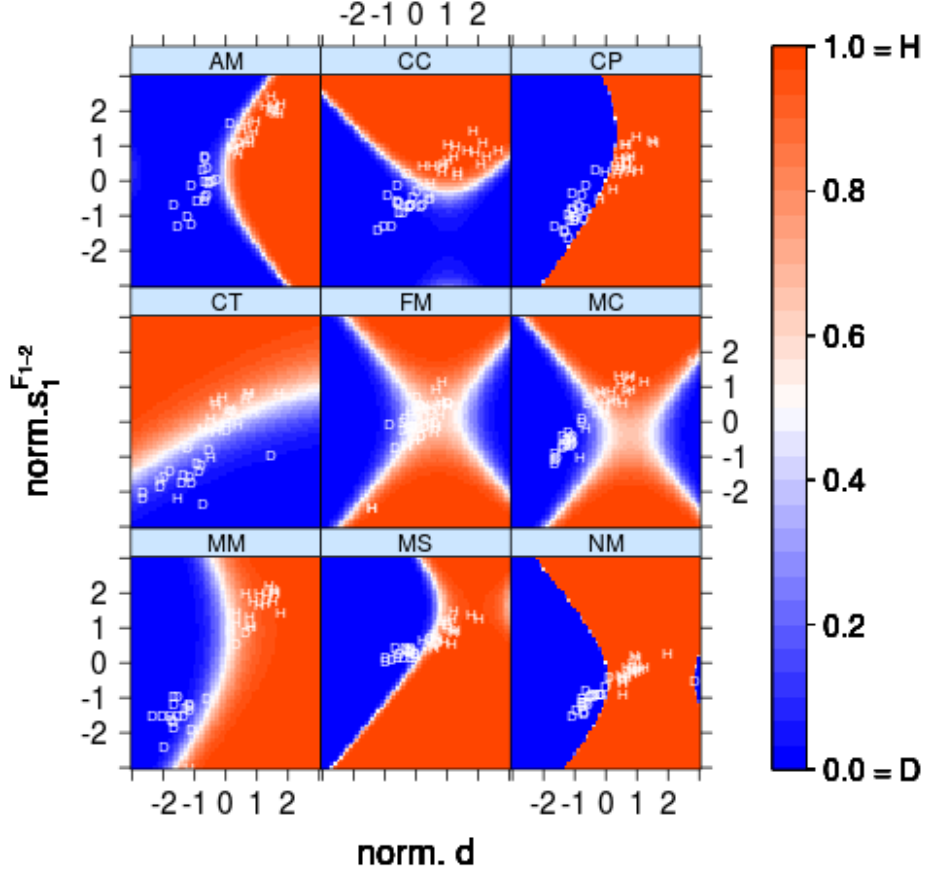


Figure 12: Representation of Eq. (4) applied to each speakers’ data separately. Predictors $pc1$ and $pc3$ are converted back into their combinations of centered and normalised values of d (x-axis) and s_1^{F1-2} (y-axis), while $s_2^{f_0}$ is ignored (formally, it is set to $s_2^{f_0} = 0$). Each of the 365 tokens is represented by a letter indicating its class (D or H). In false colours, the probability of H as predicted by Eq. (4).

this phonological distinction: the duration of the vocalic sequence, the F_1 and F_2 dynamics, and the alignment of f_0 movements with syllable boundaries (Aguilar, 1999; Prieto, 2002; Torreira, 2007). The current study differs from previous studies in that it treated dynamic features such as f_0 , F_1 and F_2 trajectories as continuous functions rather than as sets of scalars, and in that a joint statistical analysis has been performed. This procedure has allowed us to investigate our data in a more data-driven approach than

previous studies, and also to assess the relevance of specific features in the context of the other features present in the signal. In this manner, we were able to uncover the relevant information in the acoustic data, while avoiding the need for deciding a priori, on theoretical grounds, which details in the dynamic features are relevant.

By doing so, we have shown that by combining independent FPCA results and conventional measurements into a comprehensive ordinary statistical analysis it is possible to characterise global trends as well as between-speaker differences in the realisation of a phonological contrast that has been associated with several dynamically-changing phonetic features. We observed that all but one speaker in our study distinguished hiatus and diphthongs by means of at least two of these features. We also showed that not all speakers that produced the D/H contrast use all features in the same manner. In particular, we found that six speakers favoured duration, two favoured the dynamics of F_1 and F_2 , while f_0 alignment appeared to be a side effect rather than a cue that is used for signaling the distinction between H and D. Still, the findings related to between-speaker differences highlight the importance of individual differences in gesture coordination that is present even in the context of a short-duration and well-understood phenomenon like the D/H contrast in Spanish. We believe that the level of cue trading observed in our data calls for a re-analysis of other phonological distinctions in which multiple phonetic features are involved. More often than not, phonetic features change dynamically over time. The FDA toolbox makes it possible to investigate these features without the need for representing the feature dynamics in the form of sets of discrete numbers that may obscure or even destroy potentially relevant information.

Previous studies of dynamic features have attempted to avoid the need for converting contours into a small set of discrete measurements. Andruski and Costello (2004), Grabe et al. (2007), Aguilar (1999) and Torreira (2007) fit polynomial functions to the sampled trajectories and use the coefficients of these polynomials as numerical data that can be submitted to statistical analysis. Aguilar (1999) hypothesised and verified that the curvature of F_1 and F_2 trajectories would differ between diphthong and hiatus. This was based on the idea that a hiatus consists of a sequence of two steady-state vowels connected by a very rapid transition, whereas a diphthong contains a glide and a steady-state vowel. It was verified that a quadratic polynomial are able to capture the differences in the formant trajectories. In the same way, Torreira (2007) verified that the presence of the syllable boundary in

the hiatus sequences and the absence of such a boundary in the diphthongs corresponded with statistically significant differences between the coefficients of quadratic polynomials. However, in general it is difficult to relate the coefficients of a polynomial to underlying articulatory processes. In Section 3 and 4 we have shown that the results of an FPCA analysis confirm the conclusions of Aguilar and Torreira about the shapes of the trajectories, and that the principle component functions allow for direct interpretation in terms of articulatory mechanisms.

The fact that the findings from a bottom-up analysis using FDA agree with those of previous studies suggest that FDA is an excellent method for situations in which little or no theoretical prior knowledge is available. We therefore believe that FDA is especially well suited for exploratory studies that deal with poorly understood dynamic phenomena (e.g. prosodic variation in spontaneous speech or in undescribed languages). Along these lines, Turco et al. (2011) used FDA to explore the prosodic marking of *verum focus* in Italian. FDA was applied on a set of short sentences elicited in semi-spontaneous dialogues without hypothesising any specific linguistic behavior nor focusing on any specific part of a sentence. The analysis revealed a previously unknown prosodic phenomenon in the post-focal region of the sentences, which suggested a process of pitch compression and sentence phrasing reorganization. Similarly, in Turco and Gubian (2012) an FDA-based exploration of the production of pitch accents by native Italians, native Dutch and Dutch learners of Italian revealed clear differences without requiring any prior knowledge about specific L1 or L2 pitch accent types.

The fact that FDA can uncover structure in a set of dynamic contours does by no means imply that it is a miracle procedure that can be applied without any prior knowledge about the phenomena under investigation. FDA is a procedure that consists of several steps that must be taken in a fixed sequence. Each step can, eventually, be performed completely automatically, but each step also requires expert intervention to get it started. In this paper we focused the attention on the processing of a set of data; in doing so, we have taken for granted that the collection of the data already involves decisions that require domain knowledge. We extracted complete f_0 movements, rather than just the part of a movement contained in a single syllable. And we restricted formant extraction to the vocalic nuclei of the accented syllable. These decisions were based on phonological and phonetic knowledge and experience. In the preceding sections, and even more in Appendix A and in the supplementary material, we also emphasised that it is

necessary to check the output of the four steps that make up a complete FDA analysis for possible artifacts. Smoothing and landmark registration clearly require some degree of initial phonetic knowledge to guide automatic processing. During the smoothing step, which approximates sequences of discrete sample points by a continuous function, phonetically informed decisions must be made about the degree of temporal detail that is relevant in a specific research question. In general, underfitting, i.e. smoothing too much, is a greater risk than overfitting, i.e. keeping too much detail (cf. Figure 2), since detail that is removed in the first stage of the processing can never be recovered, while spurious detail may well be ignored in subsequent processing stages.

Domain knowledge is also required in landmark registration, since the researcher has to specify which events (if any) in complex dynamic processes correspond to the same underlying phenomenon. The type of events that can serve as landmarks will depend very much on the research at hand. In most previous research in which we used FDA, we opted for landmarks that are relatively close in time (e.g., landmarks at syllable or word onset). When analysing ‘slower’ processes it may very well be acceptable to set landmarks further apart. Once it has been decided which events to use as landmarks, marking them can be carried out manually, like in the research in this paper, or by using an automatic alignment procedure (Gubian et al., 2011; Turco and Gubian, 2012). Landmark registration eliminates differences in duration between corresponding segments. However, duration does play a role in the realisation of many linguistic categories. In this paper we have taken care of this by carrying out a separate analysis of the duration of vowel sequences. For longer utterances independent analysis of shape and duration may not be optimal. A principled solution to recover and use the lost duration information within the FDA procedure is proposed in Gubian et al. (2011) and used in Turco et al. (2011) and in Turco and Gubian (2012). It is doubtful whether such a more complex analysis would have yielded additional insights for the fairly simple set of data in this study.

The FDA toolbox contains function-based equivalents of many of the multivariate statistical analysis procedures that are available for processing discrete data. One of those tools is Functional Linear Discriminant Analysis (FLDA). In the analysis of a two-way contrast, FLDA might have seemed the obvious choice. However, we opted for using Functional PCA instead. The most important reason for preferring a non-discriminative criterion in computing the principle component functions (the PC functions) is that we

wanted to obtain a rich and unconstrained representation of the shapes of the f_0 contours and the formant trajectories. In fact, in our analysis of f_0 we were able to interpret PC1 and PC2, as well as to recognise that PC1, accounting for the largest part of the variance in the raw data, mainly represented between-speaker variation that had nothing to do with the D/H contrast. On the other hand, using FLDA would have returned only one component, namely the one that best separates the two classes.

In general, it is not guaranteed that the principle component functions returned by an FDA procedure are easy to interpret in phonetic terms. FDA is not different from conventional multi-dimensional data processing techniques, which may also result in dimensions that are difficult to interpret in terms of the loadings of the features that represent the data. It may happen that more than one PC is needed to compose shapes that can be linked to some phonetic process. Moreover, PCs represent quantitative, rather than qualitative aspects of curves. For example, there is no explicit notion of ‘elbow’ encoded in the mathematical form of PC2 from the analysis carried out on f_0 . The ‘elbow’ is the result of phonetically informed visual interpretation. Also, FPCA may bring to light characteristics of contours that are difficult to detect in the raw sampled data, e.g. see Turco et al. (2011).

In this paper we decided to perform two independent FPCA-based procedures for modeling f_0 and formants and a conventional analysis of duration. After that, the output of those parallel procedures was combined into a comprehensive analysis that did not involve FDA. In Section 4.3 we showed that it is possible to carry out FPCA on two formants F_1 and F_2 , which allowed us to capture relations between those two trajectories automatically. In principle, f_0 could have been included to obtain a three-dimensional joint analysis (f_0 , F_1 and F_2). However, we decided to analyse f_0 and formants in different time intervals; f_0 included the /l/ preceding the vowel cluster, while the formant trajectories were limited to the vowels. In general, there are no restrictions on the number of features that can be analyzed together, provided that they span the same time interval. However, combining highly heterogeneous features incurs the risk that one feature obscures the contribution of the others. At the same time, the results of a high-dimensional FPCA might become hard to interpret. On the other hand, processing different features of the same signal independently, rather than combining them from the beginning, may result in highly correlated data, like it was in our case, where duration, f_0 and formant features exhibited a high correlation, which complicated the statistical analysis (cf. Section 4.4). Still, we preferred these complications

in well-understood statistical analysis procedures over the difficulty in interpreting principle component functions that mix up several features. In other words: we preferred phonetic and phonological insight over parsimony. It is for the same reason that we limited statistical analysis to fairly simple procedures, realising that caution must be exercised in interpreting the results, because these may be misleading. The results of more advanced (and probably more appropriate) statistical analysis in Appendix B confirm that the results of the simple analyses hold.

In a similar vein, we recommend to start with the simplest FDA approach that can be justified by the data and the goals of an experiment. If the results of a 'simple' analysis suggest that there may be more information in the data than what has been uncovered, more complex procedures can be attempted. This paper contains two clear cases where a more complex procedure might be useful. In one case, the interaction of duration on the one hand and the shapes of the f_0 and formant trajectories, we decided that the complex procedure for jointly analysing duration and shape proposed in Gubian et al. (2011) did not add additional insight beyond what we could learn from independent treatment of shapes and durations. In the second case, landmark registration in the analysis of the f_0 contours, we did decide to include the additional –complicating– procedure, because it did yield a crisper separation of D and H tokens (see also the document *Time normalisation and landmark registration* in the additional material pointed to in footnote 3.)

We believe that the FPCA-based method presented in this work can be extended with some adaptation to neighbouring domains of linguistic research, of which we mention four. One is the study of prosody and intonation, in particular when phenomena spanning an entire utterance are involved. The FDA procedure described here, especially in combination with a suitable representation of segment durations (Gubian et al., 2011), is capable of discovering long range correlations within and across features, e.g. f_0 contour variations that co-occur in different parts of a sentence. Another is the field of Electromagnetic Articulometry (EMA), where two- or three-dimensional trajectories of speech articulators are obtained and synchronised with the speech signal. EMA-based studies often investigate the timing of articulator movements in relation to the phonetic segmental boundaries, which involves the analysis of multidimensional trajectories sharing common landmarks, e.g. see Parrell et al. (2013) for a recent application of FDA on EMA. Investigations are under way to explore FDA as an alternative for averaging EEG signals in Evoked Response Potential (ERP) experiments (Morris and Carroll, 2006).

A similar scenario is also found in studies of head and body gestures based on motion capture technology (Beskow et al., 2010). The application of FDA would be valuable in revealing coordination patterns between head and hand movements, represented as three-dimensional trajectories, and speech features like f_0 and intensity, all being referred to common landmarks given by the speech segments.

6. Conclusions

In this paper we have introduced Functional Data Analysis as a toolbox that can facilitate the investigation of phonetic contrasts, especially when more than one feature is involved, and when some of the features are dynamical changes of phonetic parameters, such as f_0 and formant frequencies. We have used the contrast between diphthongs and hiatuses in continental Spanish, which according to previous research involves at least duration, formant frequency trajectories and f_0 alignment, as a case study in cue trading and individual differences. We have found substantial differences between speakers in their phonetic implementation of the phonological contrast.

The case study served as a platform for introducing and explaining the ways in which FDA can discover structure in a set of dynamically changing contours. FDA eliminates the need for choosing in advance a representation of the contour shapes in the form of a number of discrete measurements, which might obscure relevant information. We have shown how FDA acts as a data-driven ‘shape-to-numbers converter’ that respects the integrity of the contours, but still yields a representation in the form of a small set of numerical features that can be studied using conventional statistical tools.

The data and software used in this paper are available for download, cf. footnote 2.

Acknowledgements

We thank the editor of the journal and two anonymous reviewers for their comments on a previous version that helped to improve the paper substantially.

Appendices

A. Inside the FDA procedure

This appendix complements Sections 3.2, 3.3 and 3.4 with theoretical background on FDA and with detailed illustrations of a number of practical procedures involved in FDA. Most of those procedures are explained with a reference to f_0 contours. The presentation should help readers who consider to apply FDA on their own data. Those readers are also invited to download the code and the data that can be used to reproduce all the results reported in this article (see the link in footnote 2 in the main text). The code is based on the *R* package `fda` (Ramsay et al., 2009).⁷

A.1. Smoothing

In Section 3.2 we introduced the principles of smoothing. Here we provide more detail on the parameters governing B-splines-based smoothing and explain two methods for optimising those parameters.

A.1.1. Smoothing parameters

B-splines are a set of partially overlapping hill-shaped polynomial curves. The number and location of those curves have to be specified by the user. Technically, the curve locations are determined by the position of their connecting points called *knots*, represented as dots on the time axis in Figure 1(a) in the main text. Knots are not constrained to be equally spaced, thus we could choose their locations one by one. A theorem by de Boor (2001) states that the optimal knot locations coincide with the input samples. However, this principle may yield unsatisfactory solutions, because de Boor's theorem takes into account neither noise nor prior knowledge about what is considered relevant detail in the curves. Hence, unless there are reasons to keep a different amount of detail in different parts of curves, it is convenient to set the knots uniformly spaced along the time axis and consider the number of knots k a single parameter. A larger value of k keeps more detail (corresponding

⁷It is strongly recommended to execute the script `FDA.R` step-by-step. That should help in understanding the procedure, and then in adapting the code to the specific requirements posed by other data. The steps in the script `FDA.R` implement the procedures that are explained in detail below in the order of presentation. Therefore, it is recommended to carefully study Appendix A before executing the script.

to a higher time resolution), but it also increases the risk of overfitting (i.e. keeping irrelevant detail).

The optimal trade-off between overfitting and underfitting can be found by a procedure called *smoothing with roughness penalty* (Ramsay and Silverman, 2005). The procedure minimises a cost function composed of two terms

$$\min\{SSE + \lambda \cdot ROUGHNESS\}, \quad (\text{A.1})$$

where SSE is the sum of squared errors, which quantifies fitting error, $ROUGHNESS$ is a measure of how rapidly the curve changes direction, which is a measure that decreases with increasing smoothness, and the regularization parameter λ determines the importance of smoothness relative to fitting error. Each pair of values (k, λ) produces a different result in terms of fitting error and roughness. In what follows we use the smoothing of f_0 contours to explain how to obtain values for k and λ such that the smoothed curves are optimal for phonetic research.

A.1.2. *Smoothing with generalised cross-validation (GCV)*

The smoothing problem as formulated in Eq. (A.1) can be approached by using empirical model selection techniques, such as generalised cross-validation (GCV) proposed by Ramsay and Silverman (2005). The procedure consists in exploring several (k, λ) value combinations, compute the GCV error for each of them, and choose the combination (k^*, λ^*) that yields the smallest error. The GCV error provides an unbiased estimate of the fitting error that factors out the effect of random noise in the curves (for example due to measurement error). However, the straightforward application of GCV does not allow one to impose constraints on the solution that are related to the degree of detail in the curves that are of interest in a research question. In practice, minimising GCV error with f_0 or formant contours yields curves that keep a level of detail that cannot be linked to underlying articulation.

Here, we propose a way for using GCV in such a way that the degree of smoothing can be controlled. The strategy we adopt can be described as GCV-informed empirical judgement. First we define a grid of (k, λ) values that we want to explore and a randomly selected subset S of the input contours. For each (k, λ) pair, we smooth all the curves in S applying Eq. (A.1) and we compute the corresponding GCV error (Figure A.1). Next, we select a promising combination (k', λ') on the basis of visual inspection of a

number of curves from a held-out set disjoint from S . Finally, we select the combination (k'', λ'') with the smallest k and the highest λ from the pairs that yield (almost) the same GCV error as (k', λ') . By giving preference to a smaller number of knots we reduce the number of degrees of freedom of the minimisation (A.1), which in turn minimizes the risk of overfitting. This procedure amounts to finding the least complex element in a set of empirically equivalent solutions. By choosing a large λ we increase the chances of getting smooth curves.

An example of this procedure is shown in Figure A.1 and A.2. The minimum GCV error in the grid of Figure A.1 corresponds to $k^* = 28$, $\lambda^* = 10^2$. A curve obtained by solving Eq. (A.1) using those values is shown in Figure 2(a), where we see how the microprosodic ripples at the beginning have been faithfully reproduced. Since we are not interested in modeling microprosody, we have to look for another solution yielding smoother curves. A possible combination could be $k' = 28$, $\lambda' = 10^6$, corresponding to Figure 2(b), where a good compromise seems to be reached. Looking at this combination on the grid of Figure A.1, we see that by reducing the number of knots down to $k'' = 8$ we obtain almost the same GCV error, as well as almost the same curves, as shown in Figure 2(c). According to the minimum complexity principle, the solution $k'' = 8$, $\lambda'' = 10^6$ is selected and applied to all f_0 contours. The effect of selecting the highest value of λ when two values yield approximately the same GCV error is illustrated by comparing the curve obtained with $(8, 10^6)$ with the curve corresponding to $k''' = 8$, $\lambda''' = 10^{-4}$. While GCV error in $(8, 10^{-4})$ is almost the same as in $(8, 10^6)$ (cf. Figure A.1), the low weight of the roughness penalty allows the curve to be ‘attracted’ by the isolated sample at the leftmost part of the curve, as shown in Figure 2(d).

Sometimes a full GCV-informed empirical judgement procedure may be quite demanding, especially during preliminary data explorations. To quickly select an ‘acceptable’ (k, λ) pair, one must keep in mind that overfitting generally occurs when λ is too small and k is too large, while underfitting occurs mainly when λ is too large, irrespective of k . Moreover, overfitting and underfitting have a different impact on the rest of the FDA procedure. Overfitting introduces irrelevant detail that may make the results harder to interpret and that increases computational costs. Underfitting deletes useful information that cannot be recovered later on. Figure A.2 illustrates the general relation between k and λ . From the figure it can be seen that the parameters k and λ have opposite effects on the resulting fitting error.

Clearly the best fit can be obtained by a large number of knots and a small weight of smoothness, i.e. the recipe for a continuous function that visits all samples in the original sampled data representation. However, it can also be seen that reasonably small fitting errors can be obtained with $\lambda = 10^6$, already in combination with $k = 8$.

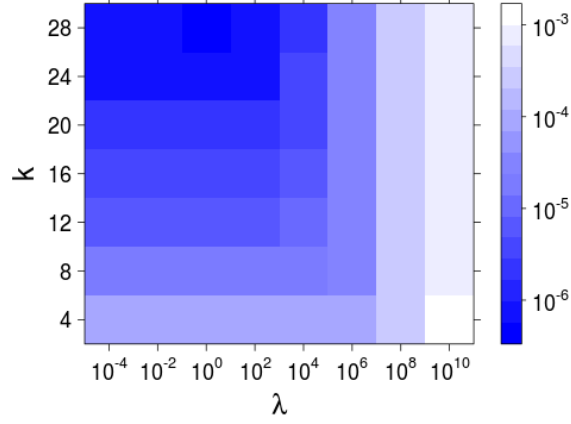


Figure A.1: A colour grid showing the generalised cross-validation (GCV) error for several (k, λ) combinations.

A.1.3. Using domain knowledge in optimizing k and λ

In some (perhaps exceptional) cases we have quantitative information about the upper bound of the speed with which a phonetic parameter can change. f_0 happens to be such a parameter. Xu and Sun (2002) derived empirical linear relations between voluntary prosodic gestures and the maximum speed at which an f_0 excursion can be produced. More precisely, given an observed voluntary f_0 gesture (a rise or a fall) elicited in such a way that the maximum controllable speed in f_0 change is used by a subject, two linear inequalities were derived. One has f_0 excursion as predictor and average rate of f_0 change (i.e. the f_0 excursion divided by the time required to achieve it) as dependent variable. The other has f_0 excursion as predictor and peak instantaneous f_0 change rate as dependent variable. The relations for a rising

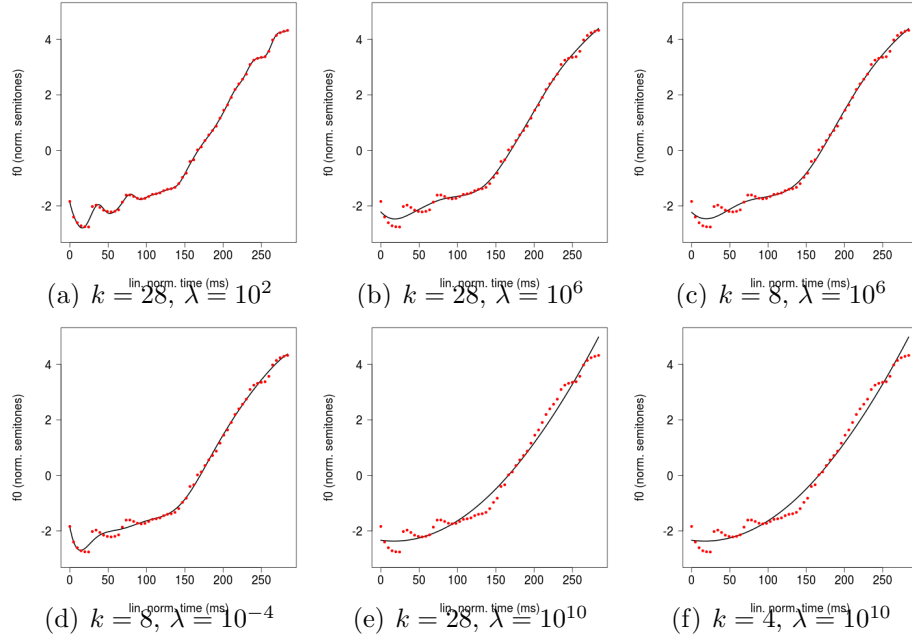


Figure A.2: An example showing the effect of the regularisation parameter λ and the number of B-splines knots k . Each panel shows a different (k, λ) value combination and the resulting smoothed curve, while the original f_0 samples are the same for all.

f_0 gesture can be formulated as⁸:

$$ave. \ speed \leq 10.8 + 5.6 \cdot excursion \quad (A.2)$$

$$max. \ speed \leq 12.4 + 10.5 \cdot excursion \quad (A.3)$$

where both *speed*'s are in semitones/s and *excursion* is in semitones. Eq. (A.2) and (A.3) say that the maximum average or instantaneous speed at which speakers can voluntarily produce a change in f_0 depends on the f_0 excursion involved in the gesture: larger excursions can be produced at faster speed.

Figure A.3 investigates whether the solutions $(k^* = 28, \lambda^* = 10^2)$ and $(k'' = 8, \lambda'' = 10^6)$, obtained in Section A.1.2, comply with the constraints formulated in Eq. (A.2) and (A.3). Figure 3(a) shows a curve with several small ripples. Figure 3(b) shows the corresponding speed of change of f_0 .

⁸Eq. (A.2) and (A.3) are adapted from Xu and Sun (2002), Tables VI and VII, line ‘Mean’, column ‘Rise speed’.

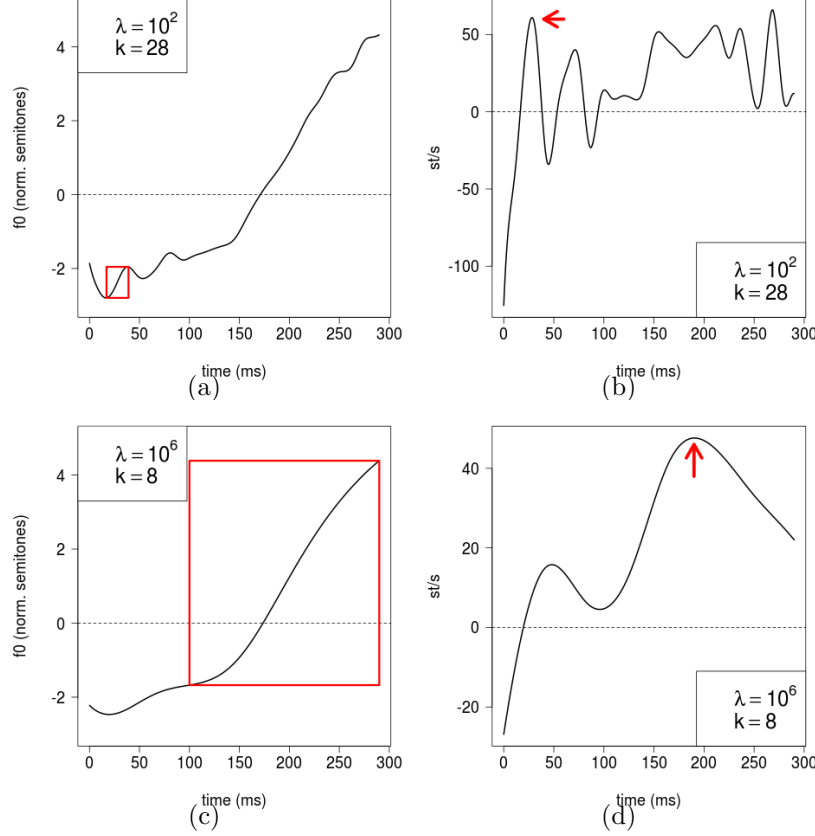


Figure A.3: In (a) an f_0 contour smoothed using parameters $k^* = 28$, $\lambda^* = 10^2$ (the same curve as in Figure 2(a), except that the time axis is not linearly normalised). In (b) the instantaneous velocity of the curve in (a), i.e. its first derivative with respect to time. In (c) an f_0 contour smoothed using parameters $k'' = 8$, $\lambda'' = 10^6$ (the same curve as in Figure 2(c), except that the time axis is not linearly normalised), and in (d) its instantaneous velocity. In (a) and (c) a rectangle isolates a rising gesture. In (b) and (d) an arrow points at the peak instantaneous velocity reached within the gesture.

According to Xu and Sun (2002), if the f_0 changes are the result of a voluntary gesture, then Eq. (A.2) should be satisfied. However, from the box in the lower left part of Figure 3(a) it can be seen that the f_0 change enclosed by the box is around 1 st, in 30 ms, which amounts to an average speed of approximately 30 st/s; obviously, Eq. (A.2) is not satisfied. The same holds for the maximum speed of more than 50 st/s, indicated by the arrow in Figure 3(b). The situation is different in Figure 3(c) and 3(d), where a wide

gesture of around 6 st excursion realised in around 200 ms can be identified, which satisfies both Eq. (A.2) ($6/0.2 = 30 \leq 10.8 + 5.6 \cdot 6 = 44.4$) and Eq. (A.3) ($45 \leq 12.4 + 10.5 \cdot 6 = 75.4$).

A.2. Landmark registration

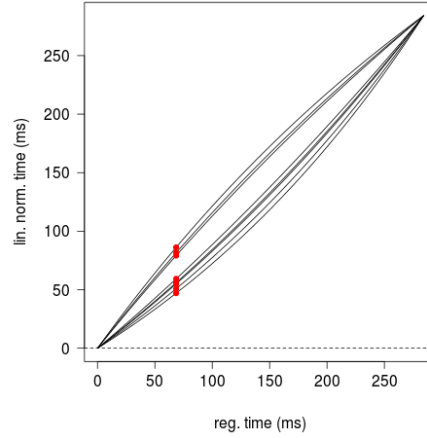


Figure A.4: The warping functions $t = h(\tau)$ that transform the curves in Figure 3(a) in the main text into those in Figure 3(b). The x -axis is the landmark-registered time axis τ , the y -axis is the linearly normalised time axis t , the dots show the position of the landmark of each curve (boundary between /l/ and vowel sequence).

This section complements Section 3.3 by providing a more in-depth description of the smoothing procedure involved in landmark registration. Landmark registration takes place in two stages, and it is applied to each curve $f(t)$ separately. The first stage operates solely on the landmarks provided by the user and produces a time warping curve $t = h(\tau)$ that specifies the mapping between the registered time axis τ and the original time axis t . The second stage applies the warping functions $h(\tau)$ to the corresponding input curves $f(t)$ to obtain the registered curves $f(h(\tau))$.

Figure A.4 shows the warping functions $h(\tau)$ that transform the curves in Figure 3(a) in the main text (cf. Section 3.3) into those in Figure 3(b). While the landmark location on the input time axis t (the y -axis) of Figure A.4 varies considerably across $h(\tau)$ curves, the location on the registered time axis τ (the x -axis) is almost the same for all curves. Moreover, the functions $h(\tau)$ show how time instants close to the landmark are ‘dragged’ more than others farther away from it. The starting and ending points of the $h(\tau)$

functions all coincide; this is due to the time normalisation that was part of the preceding smoothing operation (see also the in-depth discussion about time normalisation in the additional material on the web site mentioned in footnote 3 in the main text).

The $h(\tau)$ curves are obtained by fitting a B-splines approximation to the landmark locations indicated by the dots in Figure A.4, plus the extreme points marking the beginning and the end of all curves. As in the smoothing procedure explained above, it is not necessary that the fitted curve passes through all points, resulting in a zero fit error. As an additional constraint, the functions $h(\tau)$ must be monotonically increasing or decreasing; non-monotonic functions would correspond to local inversion of the time course. The more $h(\tau)$ departs from a straight diagonal line, the heavier the time warping is.

Obtaining reasonably smooth functions $h(\tau)$ is easier than in Sec. A.1, since it is safe to place the B-spline knots at the landmark positions on the x -axis, as prescribed by de Boor's theorem (de Boor, 2001). The reason is that we assume that the landmark locations are error-free and that their placement reflects the time resolution of interest. This eliminates the parameter k , leaving only λ from Eq. (A.1) to be set. The latter can be done by picking the largest λ yielding an alignment error (i.e. difference between the desired and obtained landmark positions) below a threshold specified by the user.

A.3. Functional PCA

Functional PCA extends the idea of ordinary PCA to input elements that are functions defined on a normalised (time) interval, as opposed to numerical vectors of fixed size. The first PC curve $PC1(t)$ is the function solving the following maximisation:

$$\max \left\{ var_n \left(\int_0^T PC1(t) f_n(t) dt \right) \right\}, \text{ subject to } \int_0^T PC1^2(t) = 1, \quad (\text{A.4})$$

where var_n is the variance with respect to the index n running from 1 to the number of curves, $[0, T]$ is the normalised time interval and $f_n(t)$ are the smoothed curves. The resulting function $PC1(t)$ captures, on average, the largest possible proportion of the variance in the set of curves $f_n(t)$. Higher degree components $PC2, PC3, \dots, PCm$ are obtained in the same way, after subtracting the contribution of the lower degree components from the curves $f_n(t)$. See Ramsay and Silverman (2005) for a complete account on FPCA.

The functions $PCm(t)$ are also given in the form of B-splines. This implies that one must determine the degree of smoothness of the PC curves by specifying a number of knots (the parameter k) and the regularisation parameter λ which penalises rapidly varying PCs, which tend to increase the variance of the integral (A.4). Setting k and λ to the same values used for smoothing the input curves usually provides good results, since PC curves should reflect the time resolution and dynamics of the input curves.

B. Data analysis with linear mixed models

In this appendix, the analysis carried out in Section 4 is repeated applying linear mixed effects models (West et al., 2007; Baayen, 2008). In Section B.1 the analysis of individual features carried out in Section 4.1, 4.2 and 4.3 is repeated. In Section B.2 we repeat the analysis of the combined features (complementing section 4.4).

B.1. Analysis of individual features

In Section 4 five scalar features (d , $s_1^{f_0}$, $s_2^{f_0}$, $s_1^{F_1-2}$ and $s_2^{F_1-2}$) were analysed separately. The first part of the analysis consisted in predicting each feature using the model

$$y = \beta_0 + \beta_1 \cdot x, \quad (\text{B.1})$$

where y stands for one of the five features and x encodes class ($x = 0$ if class is D, $x = 1$ otherwise). The purpose was to assess how much of the variability in each feature is explained by class alone. Eq. (B.1) does not take into account that the 365 tokens are not a random sample from a single population. Rather, the tokens are grouped according to the nine speakers. Here we extend Eq. (B.1) by introducing *speaker* as random factor. For all five features we will adopt the same model structure, in which speaker-related random variables are included. Formally, we have:

$$y_{ij} = \beta_0 + u_{0,j} + (\beta_1 + u_{1,j}) \cdot x_{i,j} + \epsilon_{i,j}, \quad (\text{B.2})$$

where $j = 1, \dots, 9$ is the index of speakers, $i = 1, \dots, n_j$ is the index of the individual tokens, and n_j the number of tokens uttered by speaker j . The β terms are the fixed effects (cf. Eq. (B.1)), the u terms are the speaker-related corrections to the fixed effects, modeled as independent Gaussian variables with zero mean, and $\epsilon_{i,j}$ is the residual model error, modeled as a Gaussian variable of zero mean and independent of the u variables.

In order to determine whether the complexity of Eq. (B.2) is justified by the data, likelihood tests were employed (West et al., 2007; Baayen, 2008). For each of the five features used as dependent variable in place of y_{ij} , a chain of models was built, starting from a so-called Null Model of the form: $y_{ij} = \beta_0 + u_{0,j} + \epsilon_{i,j}$, and subsequently adding terms until the full form of Eq. (B.2) is reached. At every step, a likelihood test is performed. In all cases, the tests proved that including the random correction $u_{1,j}$ is justified, and in all cases except for $s_1^{f_0}$ including the fixed effect for class β_1 is justified. In the same way, we checked all models (B.2) for heteroscedasticity with respect to class, i.e., whether the variance of the residual error ϵ is different for D and H. In the case of dependent variables d , $s_2^{f_0}$ and $s_1^{F_1-2}$ no such effect was found. A small difference between prediction errors for D and H was found for variables $s_1^{f_0}$ ($\sigma(\epsilon|H) = 0.85 \sigma(\epsilon|D)$), and $s_2^{F_1-2}$ ($\sigma(\epsilon|H) = 1.28 \sigma(\epsilon|D)$). However, we decided to ignore these small differences among models, which would call for modeling different features with different model structures, in favour of adopting Eq. (B.2) for all features, in order to facilitate comparisons and simplify the presentation.

In Section 4 the relevance of each feature for the D/H contrast was established by observing the explained variance R^2 . However, there is no straightforward equivalent of R^2 for Eq. (B.2). As substitute, we computed two relevance criteria suggested in Baayen (2008). The first is \tilde{R}^2 , which quantifies the gain in explained variance between model (B.2) and the Null Model.⁹ The other is $\Delta\sigma(u_1)$, which quantifies how much of the standard deviation of the random variable u_1 disappears if the fixed effect term for class β_1 is

⁹ \tilde{R}^2 is defined as:

$$\tilde{R}^2 = 1 - \left(\frac{\rho(y_{ij}, y_{\text{predicted by Null Model}})}{\rho(y_{ij}, y_{\text{predicted by model (B.2)}})} \right)^2,$$

where $\rho(\cdot, \cdot)$ denotes Pearson correlation.

y	β_0	β_1	$\sigma(u_0)$	$\sigma(u_1)$	$\sigma(\epsilon)$	\tilde{R}^2 (%)	$\Delta\sigma(u_1)$ (%)
d [ms]	133	48	16	19	21	82	62
$s_1^{f_0}$	1.3	-2.7	8.2	4.3	4.5	8	11
$s_2^{f_0}$	-4.0	8.2	3.4	6.9	3.8	84	34
$s_1^{F_{1-2}}$	-2.3	4.7	2.2	2.8	2.4	70	48
$s_2^{F_{1-2}}$	-0.03	0.7	1.5	0.9	1.7	15	16

Table 5: Summary of models in form of Eq. (B.2). The first column reports the predicted feature (y in Eq. (B.2)). The second and third column report the values of the fixed effect terms, to be compared with the analogous terms in Table 2. Columns from four to six report standard deviations of speaker-related random terms and residual error. The last two columns report respectively the gain in explained variance and the reduction in standard deviation of $\sigma(u_1)$ due to the fixed effect of class (see text).

present.¹⁰

Table 5 summarises the results of modeling the five features with Eq. (B.2). The values of the fixed terms β_0 and β_1 are very similar to their counterparts in Table 2 in the main text. Also, the values of \tilde{R}^2 and $\Delta\sigma(u_1)$ show the same pattern as R^2 in Table 2. Importantly, Table 5 confirms the discrepancy between the group of features d , $s_2^{f_0}$ and $s_1^{F_{1-2}}$ on one hand, which benefit substantially from a fixed term encoding token class, and the group $s_1^{f_0}$ and $s_2^{F_{1-2}}$ on the other hand that do not. A closer inspection of Table 5 reveals that $s_1^{f_0}$ varies substantially between speakers: The speaker-related random variation around the fixed term β_0 , quantified by $\sigma(u_0)$, is much larger than β_0 itself (contrary to, for example, the corresponding values for d). This suggests that f_0 slope, captured by $s_1^{f_0}$, varies mainly between speakers, irrespective of the token class (see also box- and curve plots in Additional Material referred to in footnote 3 in the main text). In the group of features d , $s_2^{f_0}$ and $s_1^{F_{1-2}}$ we

¹⁰ This time model (B.2) is compared against a different Null Model:

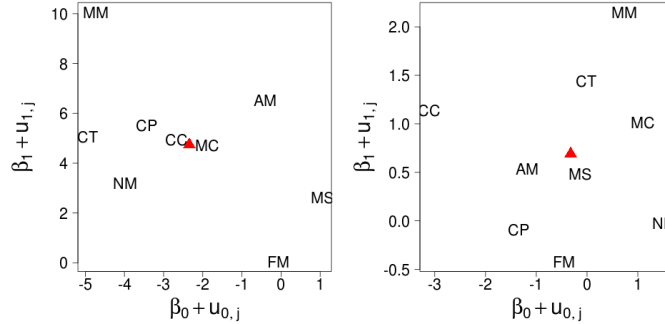
$$y_{ij} = \beta_0 + u_{0,j} + u_{1,j} \cdot x_{i,j} + \epsilon_{i,j},$$

where class-related variation is modeled only as a speaker-related random factor. $\Delta\sigma(u_1)$ is defined as:

$$\Delta\sigma(u_1) = 1 - \frac{\sigma(u_1|\text{model (B.2)})}{\sigma(u_1|\text{Null Model})},$$

where σ denotes standard deviation.

(c) $s_2^{f_0}$


$$(e) \ s_2^{F_1-2}$$

48

B.1.1. Consistency between speakers

In Section 4 speaker-specific regression models were built in order to verify whether speakers use features consistently. Here, we will address the same questions by inspecting the results of models (B.2). In addition, we will investigate whether mixed models can provide insight in between-speaker differences that were not uncovered by a more conventional analysis. Each model produces by-speaker adjusted values $\beta_0 + u_{0,j}$ and $\beta_1 + u_{1,j}$, which adapt the global fixed terms β_0 and β_1 to each speaker. Figure B.1 displays five panels, one for each feature used as dependent variable in Eq. (B.2), reporting $\beta_0 + u_{0,j}$ on the x -axis and $\beta_1 + u_{1,j}$ on the y -axis, together with the value of the fixed effects alone (β_0, β_1) for reference purpose (red triangles).

Figure B.1(a) shows that all speakers –except FM– vary d consistently, since the values of $\beta_1 + u_{1,j}$ (y -axis) are positive for all speakers. Speaker FM seems not to differentiate D from H, since a difference of 10 ms is clearly too small to be perceivable. From Table 5 it can be seen that the predicted values for FM (157 ms for D, 167 ms for H) are between the global average for D (133 ms) and for H (181 ms). Therefore, FM does not seem to produce durations that are representative for either class. At the other extreme, speaker MM appears to produce a much larger contrast than any of the remaining speakers. In general, the between-speaker variation of β_1 is quite small and it seems to be insensitive to the larger variation of β_0 , which suggests that a somewhat fixed extra duration (around 50 ms) is added to a token to realise an H.

Figure B.1(b) shows that the effect of class on $s_1^{f_0}$ (-2.7 for H, which means that H is flatter than D), is small compared to the class-independent between-speaker variation ($\beta_0 + u_{0,j}$). Three speakers (AM, NM and FM) show a variation that may be too small to be perceivable, and speaker MM appears to produce D flatter than H. We may interpret this small effect on f_0 contour slope as a side effect of the fact that H curves tend to present an elbow, while D curves do not (cf. Figure 9). This may alter the slope of the curve, captured by PC1 for f_0 , as a whole (cf. Figure 5(a)).

Figure B.1(c) shows that all speakers –except FM– vary $s_2^{f_0}$ consistently, since the values of $\beta_1 + u_{1,j}$ are positive for all speakers, while the average variation for FM is around +0.6 for H, which translates into a negligible correction of the shape of f_0 (cf. Figure 5(b), where the value of $s_2^{f_0}$ for the +curve is 6.5 higher than for the solid curve). As in the case of duration, speaker MM produces a larger contrast than the other speakers. Differently

from Figure B.1(a), between-speaker variation on β_1 is substantial, compared to the variation of β_0 . This makes it impossible to conclude that the ‘elbow correction’ for H is fixed; rather, it is applied differently by different speakers.

Figure B.1(d) shows a pattern that is very similar to Figure 1(a) in all respects. Finally, Figure B.1(e) shows that the small effect on formants ($\beta_1 = 0.7$, cf. Figure 6(b) and (d), where the value of s_2^{F1-2} for the +curve is 2.3 higher than for the solid curve) is applied by six speakers, while two do not apply it (CP and NM) and one (FM) does it in the opposite direction. Moreover, the class effect on s_2^{F1-2} is small compared to the class-independent inter-speaker variation ($\beta_0 + u_{0,j}$).

B.2. Global Analysis

In this section, Eq. (4) in the main text is extended to include random effect terms. The result is in the following generalised linear mixed model:

$$\text{logit}(Pr(H))_{ij} = (2.77 + u_{0,j}) \cdot pc1_{i,j} + (2.00 + u_{1,j}) \cdot pc3_{i,j} - 1.88 \cdot pc1_{i,j} \cdot pc3_{i,j} + \epsilon_{i,j}, \quad (\text{B.3})$$

where the fixed terms have already been substituted. First, note that the values for the fixed terms are similar to their corresponding values in Eq. (4). More interesting insight comes from the analysis of the speaker-specific adjustments from the u terms. In this case, rather than looking at their values in isolation, like we have done in the previous section, we will directly show their impact on the speaker-adjusted prediction of class. Figure B.2 shows the result. Each panel shows a different prediction surface for a different speaker as well as the location of the respective data points marked with their class. The plots show that not all speakers exhibit the same trend as predicted in Eq. (4) and represented in Figure 11. In particular, speakers CC and CT seem to depart from the trend in that for them the formant feature seems to be more relevant than duration in determining token class. Moreover, note that most of the data points for speaker FM lie in the undecided region where class probability is around 0.5. This confirms the fact that this speaker does not realise the D/H contrast.

B.2.1. Conclusion

With our data about the D/H contrast in continental Spanish the application of advanced statistical models (generalised linear mixed effects models) did not uncover insights that could not also be discovered when applying more conventional statistical analysis methods, such as ANOVA and t -tests.

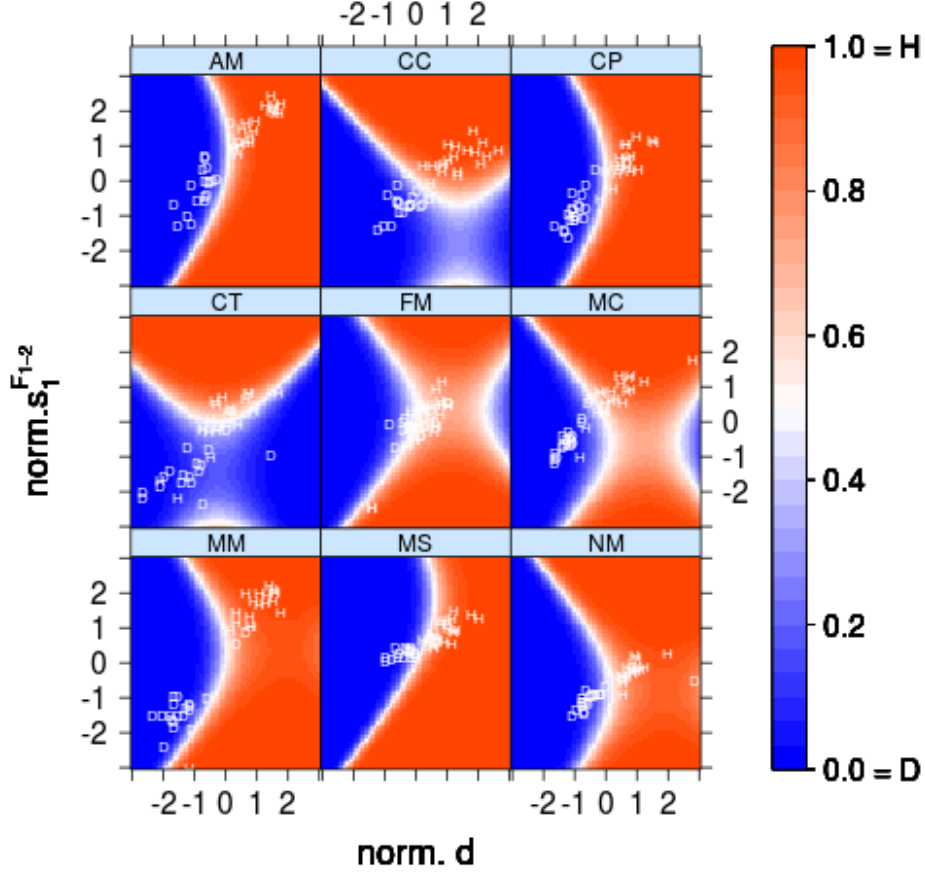


Figure B.2: Representation of Eq. (B.3). Predictors $pc1$ and $pc3$ are converted back into their combinations of centered and normalised values of d (x-axis) and s_1^{F1-2} (y-axis), while $s_2^{f_0}$ is ignored (formally, it is set to $s_2^{f_0} = 0$). Each of the 365 tokens is represented by a letter indicating its class (D or H). In false colours, the probability of H as predicted by Eq. (4). In each panel, speaker-dependent corrections corresponding to the u terms in Eq. (B.3) are applied.

Perhaps the most attractive advantage of the advanced models for our case is the fact that they allow making visual representations of the results that are somewhat easier to interpret. Having said this, it must be added that the number of potentially relevant factors in the D/H contrast data is relatively small. With a much larger number of predictors a combination of ANOVAs and t -test might have become extremely cumbersome. Then, a well-designed

application of generalised linear mixed effects models might have been much faster.

References

- Aguilar, L. (1999). Hiatus and diphthong: Acoustic cues and speech situation differences. *Speech Communication*, 28(1):57–74.
- Andruski, J. E. and Costello, J. (2004). Using polynomial equations to model pitch contour shape in lexical tones: an example from green mong. *Journal of the International Phonetic Association*, 34:125–140.
- Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge University Press, Cambridge, UK.
- Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). *Regression Diagnostics. Identifying Influential Data and Sources of Collinearity*. Probability and Mathematical Statistics. Wiley, Hoboken, NJ.
- Beskow, J., Edlund, J., Granström, B., Gustafson, J., and House, D. (2010). Face-to-face interaction and the KTH Cooking Show. In *Development of Multimodal Interfaces: Active Listening and Synchrony*, pages 157–168. Springer.
- Boersma, P. and Weenink, D. (2009). Praat: Doing phonetics by computer (version 5.1.20) [computer program]. *online: <http://www.praat.org/>*.
- Cheng, C., Xu, Y., and Gubian, M. (2010). Exploring the mechanism of tonal contraction in taiwan mandarin. In *Proceedings of INTERSPEECH 2010*, pages 2010 – 2014, Chiba, Japan.
- de Boor, C. (2001). *A Practical Guide to Splines, Revised Edition*. Springer, New York.
- Dombrowski, E. and Niebuhr, O. (2010). Shaping phrase-final rising intonation in german. In *Speech Prosody 2010*, pages 100788:1–4.
- Grabe, E., Kochanski, G., and Coleman, J. (2007). Connecting intonation labels to mathematical descriptions of fundamental frequency. *Language and Speech*, 3(50):281–310.

- Gubian, M., Boves, L., and Cangemi, F. (2011). Joint analysis of f_0 and speech rate with Functional Data Analysis. In *Proceedings of ICASSP 2011*, pages 4972–4975, Prague, Czech Republic.
- Hualde, J. I. (2005). *The Sounds of Spanish*. Cambridge University Press, Cambridge, U.K.
- Hualde, J. I. and Prieto, M. (2002). On the diphthong/hiatus contrast in spanish: some experimental results. *Linguistics*, 40:217–234.
- Jackson, J. (1991). *A User’s Guide to Principal Components*. Wiley, Hoboken, NJ.
- Lawless, J. F. and Singhal, K. (1978). Efficient screening of nonnormal regression models. *Biometrics*, 34:318–327.
- Morris, J. S. and Carroll, R. J. (2006). Wavelet-based functional mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(2):179–199.
- Parrell, B., Lee, S., and Byrd, D. (2013). Evaluation of prosodic juncture strength using functional data analysis. *Journal of Phonetics*, 41(6):442–452.
- Prieto, M. (2002). *Phonetic Correlates of the Syllable: Evidence from Spanish*. PhD thesis, University of Illinois at Urbana-Champaign.
- Prieto, P. and Torreira, F. (2007). The segmental anchoring hypothesis revisited: syllable structure and speech rate effects on peak timing in spanish. *Journal of Phonetics*, 35:473–500.
- Prieto, P., van Santen, J., and Hirschberg, J. (1995). Tonal alignment patterns in Spanish. *Journal of Phonetics*, 23:429 – 451.
- Ramsay, J. O., Hookers, G., and Graves, S. (2009). *Functional Data Analysis with R and MATLAB*. Springer Verlag, New York, NY.
- Ramsay, J. O. and Silverman, B. W. (2002). *Applied Functional Data Analysis - Methods and Case Studies*. Springer Verlag, New York, NY.
- Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis - 2nd Ed.* Springer Verlag, New York, NY.

- Repp, B. H. (1981). Two strategies in fricative discrimination. *Perception & Psychophysics*, 30(3):217–227.
- Siegel, S. and Castellan, N. J. (1988). *Nonparametric statistics for the behavioral science (2nd ed.)*. McGraw-Hill, New York, NY.
- Slis, I. and Cohen, A. (1969). On the complex regulating the voiced-voiceless distinction I. *Language and Speech*, 12:80 – 102.
- Torreira, F. (2007). Tonal realization of syllabic affiliation in spanish. In *16th International Congress of Phonetic Sciences, ICPhS XVI*, pages 1073–1076, Saarbrücken, Germany.
- Turco, G. and Gubian, M. (2012). L1 prosodic transfer and priming effects: A quantitative study on semi-spontaneous dialogues. In *Proceedings of Speech Prosody 2012*, Shanghai, China.
- Turco, G., Gubian, M., and Schertz, J. (2011). A quantitative investigation of the prosody of Verum Focus in Italian. In *Proceedings of INTERSPEECH 2011*, Florence, Italy.
- West, B. T., Welch, K. B., and Galecki, A. T. (2007). *Linear Mixed Models, A Practical Guide Using Statistical Software*. Chapman & Hall/CRC, Boca Raton, FL.
- Xu, Y. and Sun, X. (2002). Maximum speed of pitch change and how it may relate to speech. *Journal of the Acoustical Society of America*, 111(3):1399–1413.
- Zellers, M., Gubian, M., and Post, B. (2010). Redescribing intonational categories with functional data analysis. In *Proceedings of INTERSPEECH 2010*, pages 1141 – 1144, Chiba, Japan.