

Using Functional Data Analysis for investigating multidimensional dynamic phonetic contrasts

Michele Gubian^{*,a}, 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

Cue trading in making phonetic distinctions is inextricably intertwined with within- and between-speaker variation, and several different phonetic features may be implicated. Most previous studies focused on a small number of pre-selected features, risking to attribute too much weight to features under analysis that appeared to differ significantly between categories. Simultaneous analysis of multiple features is even more susceptible to debatable conclusions if some of the features are represented as dynamic changes over time, for example in the case of f_0 contours and formant trajectories, which used to require additional decisions on how to represent the information in a fixed set of measurements. In this paper we introduce Functional Data Analysis as a tool that allows analysing multiple dynamically changing contours simultaneously. We demonstrate the use of the tools by applying them to a study of the diphthong – hiatus distinction in European Spanish, a phenomenon in which vowel duration, trajectories of the first two formant and alignment of f_0 movements with the vowel cluster have been implicated. Our results show that different speakers use different feature combinations to make the distinction. They also show that not all characteristics of formant trajectories and f_0 contours are important for making the distinction. The data and the software tools used in the study are available to the research

*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)

community.

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 analyzing 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 materials¹, and the web site on FDA maintained by the first author², this paper should 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,

¹code, data and plots not included in the text are available for download from this repository: <https://github.com/uasolo/FDA-DH/>

²<http://lands.let.ru.nl/FDA/>

some lexical exceptions have been noted that make Spanish a language with a phonological contrast between hiatuses and diphthongs (e.g. *diente* ['djen.te] vs. *cliente* [cli.'ente]; *italiano* [i.ta.lja.no] 'Italian' vs. *liana* [li.'a.na]).

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 patterns 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 were 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 (e.g. 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 come 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 wool, 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 typically displayed a rising f_0 contour throughout the target word, starting with a rising pitch accent associated to the lexically stressed syllable of the target word, and ending in a high boundary tone at its end. This intonation pattern was always used by the participants in the experiment.

Each speaker read a total 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. A few tokens were discarded due to reading errors, leaving a total of 365 tokens, 183 diphthongs and 182 hiatuses for analysis.

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 accentual rise associated to the stressed syllable of the target words and the final movement associated to the end of the intonational phrase. The reason for including parts of f_0 beyond the accented syllable was that details relevant for the interpretation of the accentual rise might be found in the contour immediately following it, which continued rising throughout the following syllable. 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 inside 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. 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. 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 the procedure to their own data should consult also Appendix A, while readers interested in a general and comprehensive 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. After that, a four-step procedure is carried out.

Smoothing (cf. Section 3.2) is the first operation applied to the preprocessed data. Smoothing transforms f_0 contours sampled at discrete points in time into continuous functions of time, so that f_0 values are defined for all values of time t . The smoothing earned its name from the fact that the procedure 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.

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 same 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 PCA to functional input. We can see FPCA as the ‘shape-to-numbers converter’ in the FDA work flow. FPCA provides a model of the input (smoothed and landmark-registered) 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 amount 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

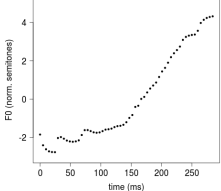
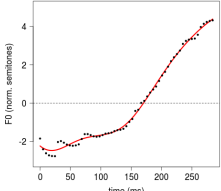
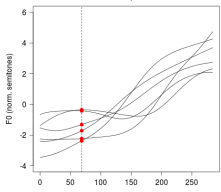
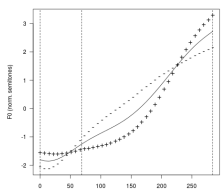
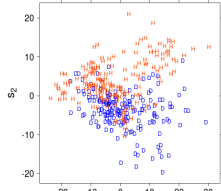
	<p>0. Raw data Input data is in the form of time sampled contours, e.g. the output of Praat pitch tracker.</p>
	<p>1. Smoothing Contours are represented in form of smooth continuous functions of time. Note: durations are linearly normalized.</p>
	<p>2. Landmark registration Time axis of each contour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

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

identify a small number of ingredients (PCs) with which we can reproduce all the dishes by only varying the ingredient dosages (PC scores).

Finally, *class analysis* (cf. Section 3.5) is a general term denoting 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 specified by the user. 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 curves that summed together approximate the desired contour. A B-spline is shown in Figure 1(a), where each polynomial corresponds to a ‘hill’. 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).

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 a smoother curve, we have to accept some amount of fitting error, i.e. the curve 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).

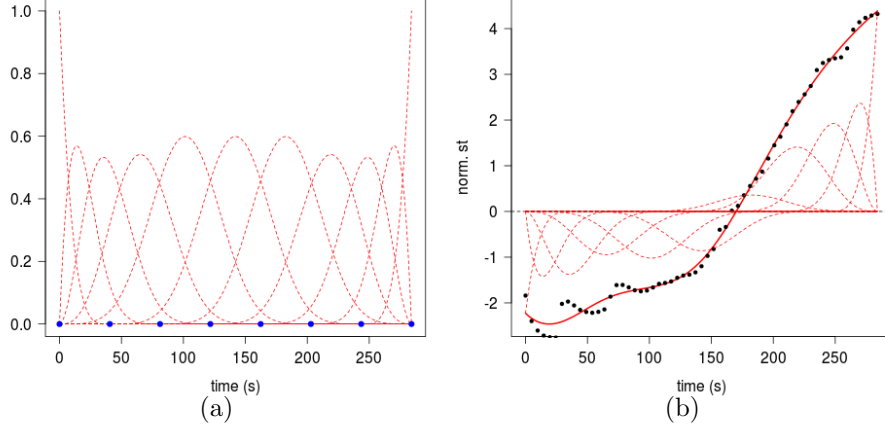


Figure 1: In (a), the B-splines basis used to represent the f_0 contours is shown. Blue dots indicate the points called *knots*, where spline ‘hills’ connect to each other (see text). In (b) an example of smoothing, where first each spline hill from (a) is multiplied by an appropriate coefficient, then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

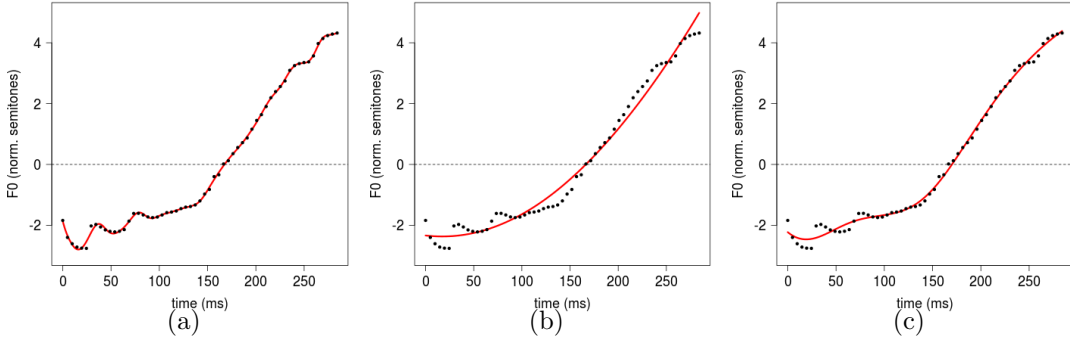


Figure 2: Three examples of f_0 contour smoothing. In (a) a case of overfitting, in (b) a case of underfitting, in (c) a good compromise.

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). When utterances are longer and also local segmental durations need to be analysed, procedures that jointly model duration and f_0 contours can be applied (see Gubian et al. (2011)).

3.3. Landmark registration

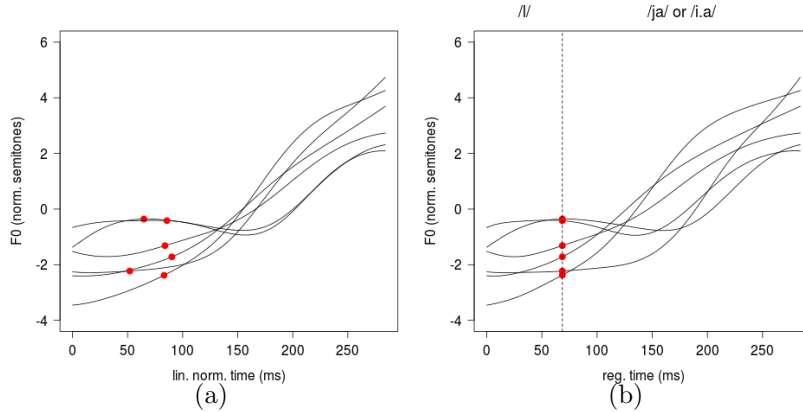


Figure 3: In (a) some smoothed f_0 contours, where the position of the only landmark (boundary between /l/ and vowel sequence) is marked with a red dot. In (b) the result of landmark registration applied on the curves in (a).

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 cluster across all f_0 contours, since we are interested in when the rise of the pitch accent starts 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.³ 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. (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’. In our case, we might want to know whether or not the f_0 rise starts already within the /l/.

A user has to specify where the landmarks on the input curve are and where they must be moved. 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 on the 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 distortions are gradual without 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.

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

³Strictly speaking, at this stage the time axis no longer represents physical time, because of the linear registration carried out in the smoothing process.

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).

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 on 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.

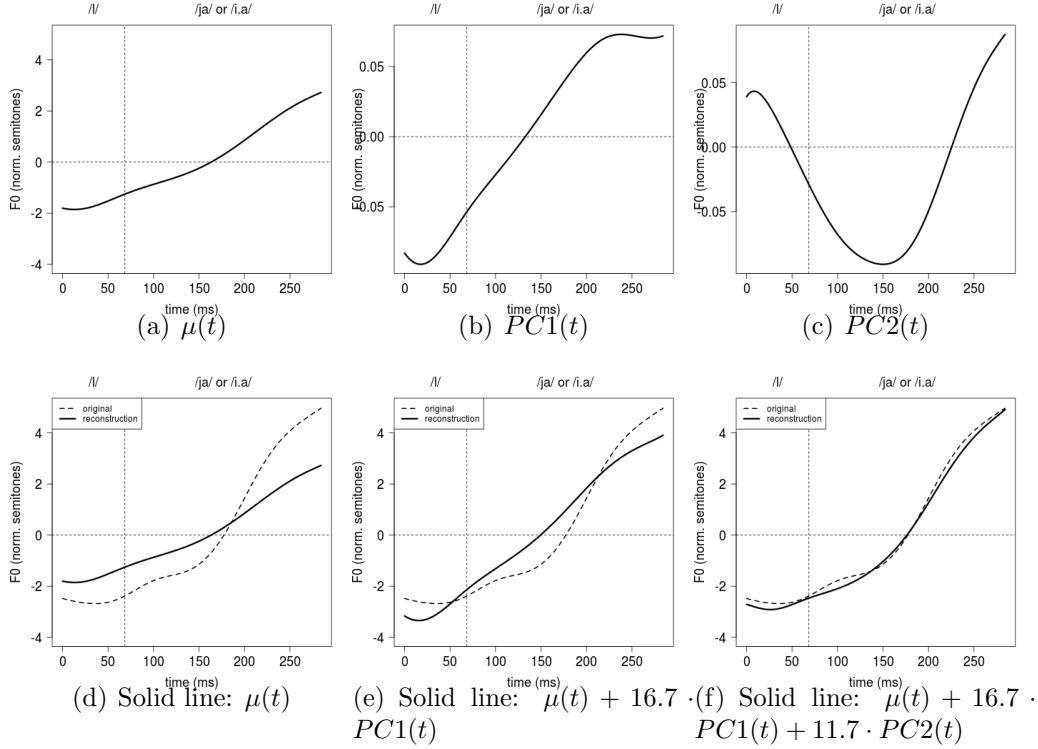


Figure 4: Upper row: (a) mean and (b,c) the first two principal component functions modeling the f_0 contour data set. Lower row: Dashed line is the input contour $f(t)$; solid lines: three approximations of $f(t)$ that use (d) no PCs, (e) one and (f) two PCs, according to Eq. (1).

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

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

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

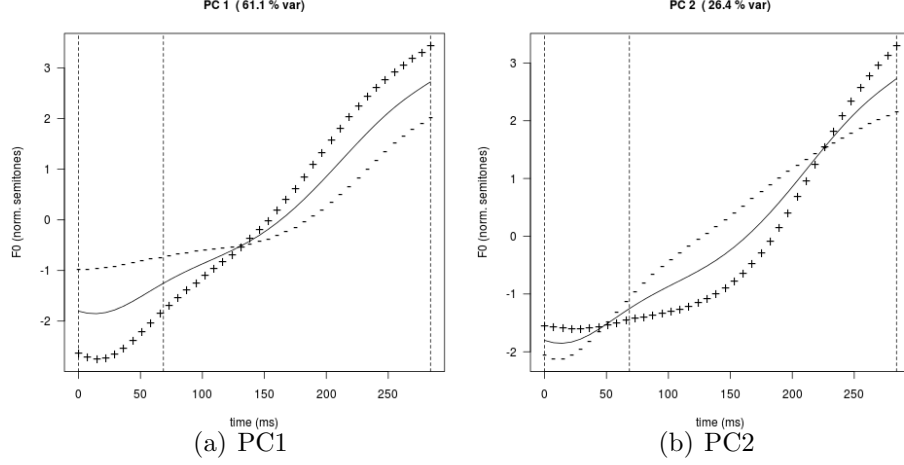


Figure 5: FPCA applied to f_0 contours according to Eq. (1). Each panel shows in solid the mean curve $\mu(t)$ and the \pm curves obtained by adding to or subtracting from $\mu(t)$ the curve (a) $\sigma(s_1) \cdot PC1(t)$ and (b) $\sigma(s_2) \cdot PC2(t)$, respectively, where σ denotes standard deviation ($\sigma(s_1) = 10.0$; $\sigma(s_2) = 6.5$). 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).

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.

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

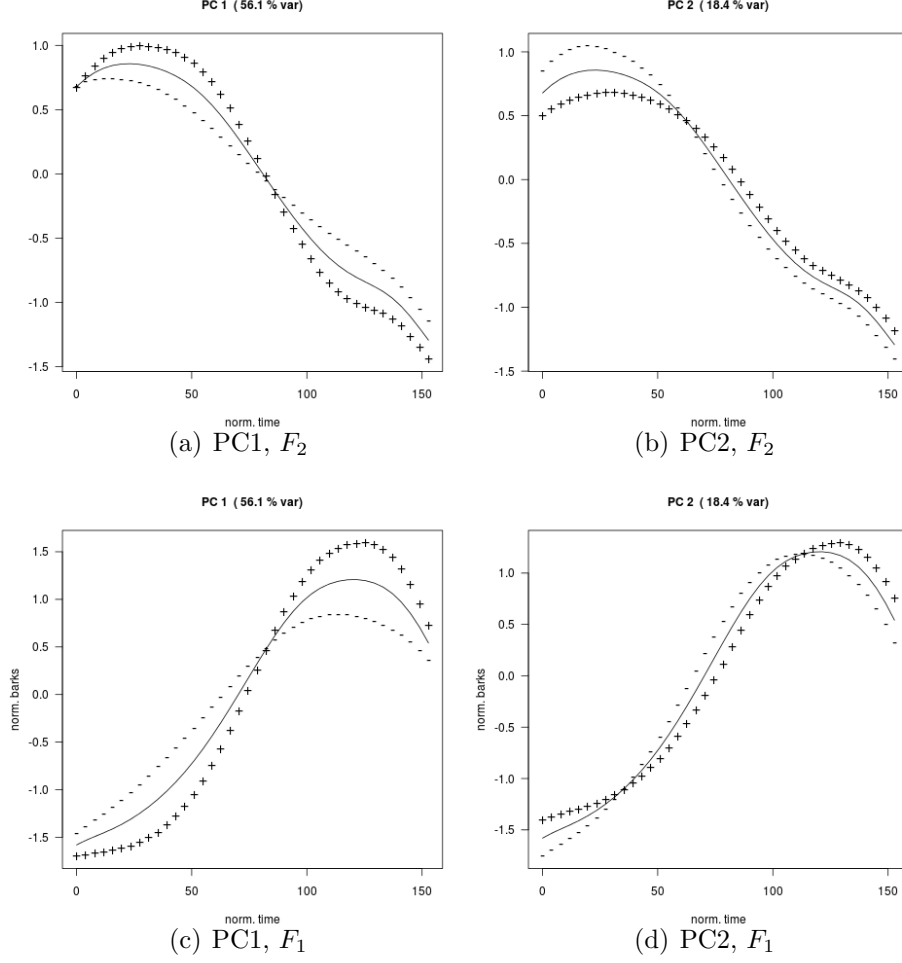


Figure 6: FPCA applied to formants F_1 and F_2 according to Eq. (2). Each panel shows in solid the mean curve for a given dimension, $\mu_{F_1}(t)$ or $\mu_{F_2}(t)$, and the \pm curves obtained by adding to or subtracting from the mean of a dimension a given PC curve multiplied by a PC score equal to 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.

by applying the smoothing procedure described in Section 3.2; landmark registration was not applied, since in our case formants span only the vowel sequence /ja/ or /i.a/, where we did not mark any internal boundary (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

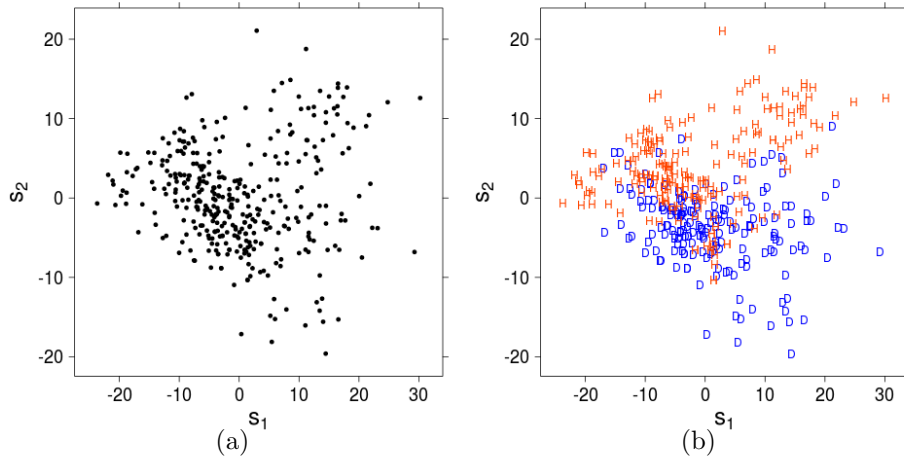


Figure 7: In (a) the PC scores s_1 and s_2 corresponding to the 365 f_0 contours in the D/H data set. In (b) the same points labeled according to the class D/H each contour belongs to.

Models like Eq. (1) and (2) are fully data-driven parametrisations 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.

The first step is shown in Figure 7(b), where the points of Figure 7(a) 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 FPCA procedures or other types of features, e.g. 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 will 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 in total 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).⁴

⁴Superscript notation ($s_1^{f_0}$ etc.) is used in the remainder of this paper to distinguish scores from different FPCA models.

The goals of the analysis 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) Is it used by all speakers? (III) Is it used consistently, (in the same direction by all speakers)? Next, we will take a look at all features 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 (in fact a one-way ANOVA) of the form

$$y = \beta_0 + \beta_1 \cdot x, \quad (3)$$

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 more relevant a feature is for the D/H contrast, the higher the explanatory power of the model. 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 t -tests, one for each speaker, on the values of y grouped by class, using Holm correction of p -values for multiple comparisons (Holm, 1979). Tests are one-sided, to check whether a speaker varies feature y according to the global trend, which is modeled by Eq. (3). For example, $\beta_1 > 0$ in Eq. (3) means that y is larger in hiatuses than in diphthongs. In this case, for each speaker the null hypothesis will be: $\text{mean}(y|H) \leq \text{mean}(y|D)$. A p -value smaller than the customary 0.05 level allows rejecting the null hypothesis in favour of the alternative hypothesis: $\text{mean}(y|H) > \text{mean}(y|D)$, i.e. the speaker adheres to the global trend. Therefore, an insignificant p -value shows that a speaker does not use feature y or possibly uses it in the opposite direction. This procedure provides a way to assess the consistency of the use of a feature across speakers.

After all features are examined separately, the individual features that appeared to be relevant for the realisation of the D/H contrast are investigated in combination (section 4.4). Question (IV) is tackled by building a generalised linear model, where the class of a token is predicted by a linear

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 y .

combination of the features. Contrary to the approach followed in Eq. (3), where the purpose is to assess the relevance of a feature, here features are used as predictors, in order to unveil interactions among them. Finally, question (V) will be addressed by building speaker-specific models that predict the class of a token from its features. The purpose is to discover whether different speakers exhibit different trends or preferences in the use of cues. In this case we employed conditional inference trees (Hothorn et al., 2006), a class of models that provides insight in the hierarchical structure of features.

In Appendix B parts of the analysis presented in section 4 are redone using linear mixed effects models. We decided to present the more conventional analyses in the main text, because they may be easier to follow. Moreover, the more advanced analyses did not uncover new insight.

4.1. Analysis of duration

The duration d in ms of the vowel cluster /ja/ for D and /i.a/ for H is analysed here. 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. The difference in the mean is highly significant and the model explains 45% of the variance, which suggests that duration is relevant for the D/H contrast.

Insight in per-speaker behavior is provided by an array of t -tests whose p -values are reported in the first column of Table 3. Each test is applied to durations for one speaker grouped by class. The t -tests show that all speakers

speaker	d	$s_1^{f_0}$	$s_2^{f_0}$	$s_1^{F_1-2}$	$s_2^{F_1-2}$
AM	***	n.s.	***	***	n.s.
CC	***	***	0.088	***	n.s.
CP	***	***	***	***	n.s.
CT	0.038	n.s.	0.001	***	n.s.
FM	n.s.	n.s.	n.s.	n.s.	n.s.
MC	***	0.007	***	***	n.s.
MM	***	n.s.	***	***	***
MS	***	0.003	***	***	n.s.
NM	***	n.s.	0.002	***	n.s.

Table 3: Speaker-dependent one-sided t -tests. Each cell shows the p -value obtained from a t -test where the null hypothesis is that the mean of the values of a specific feature (by columns) for a specific speaker (by rows) are the same for D and for H. The alternative hypothesis is that the difference is in the direction predicted by the global trend, i.e. the one reflected by the sign of the coefficient β_1 in Eq. (3) (cf. Table 2). A ‘***’ means p -value < 0.001 , n.s. means p -value > 0.05 . The Holm correction for multiple comparisons was applied on the entire table.

except for one (FM) produce significantly longer hiatuses than diphthongs. Visual inspection of the data (cf. box plots in Additional Material⁵) reveals that speaker FM is not using duration to produce the contrast.

4.2. Analysis of f_0 contours

In Section 3.4 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, which indicates that Eq. (1) truncated at the second component provides a rather accurate account of the f_0 contour shapes.⁶ 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 respective PC scores, $s_1^{f_0}$ and $s_2^{f_0}$. In this section we analyse each score separately following the same procedure applied for duration in

⁵<https://github.com/uasolo/FDA-DH/tree/master/paper/>

⁶PC3 was computed too, but it appeared not to contribute to the D/H contrast.

Section 4.1.

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) reflects a weight of 10.0 for PC1.

Additional investigation revealed that a large part of the variation of $s_1^{f_0}$ is related to speaker identity (cf. box- and curve plots in Additional Material); the f_0 contour slope mainly varies across speakers, irrespective of token class. The second column in Table 3 shows that five out of nine speakers seem not to vary $s_1^{f_0}$ with class consistently with the global trend modeled by Eq. (3) (second row of Table 2). This allows us to conclude that f_0 contour steepness is not a relevant feature for the D/H contrast. Therefore, the feature $s_1^{f_0}$ will be excluded from the global analysis in Section 4.4.

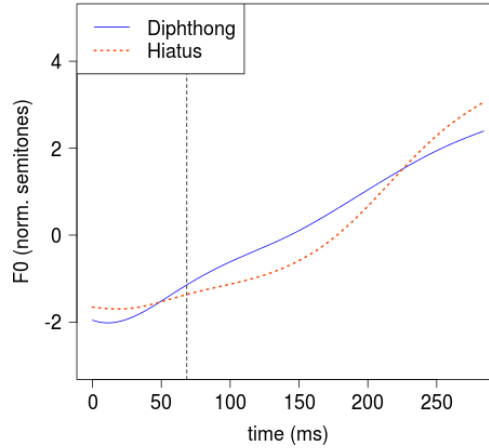


Figure 8: Two f_0 contours obtained by substituting the values of PC score $s_2^{f_0}$ predicted by the linear model $s_2^{f_0} = -3.8 + 7.6 \cdot x$ (cf. third row of Table 2) into the FPCA model in Eq. (1). The D curve is $\mu(t) - 3.8 \cdot PC2(t)$ ($s_2^{f_0} = -3.8$ is the value predicted for diphthongs 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).

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 8 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 $s_2^{f_0} = -3.8$ 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 8 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/.

The third column in Table 3 shows that seven out of nine speakers vary feature $s_2^{f_0}$ significantly and in accordance with the general trend. Speaker CT shows only a mildly significant effect, and speaker FM does not show any effect. Visual inspection of the data (cf. box- and curve plots in Additional Material) reveals that speaker FM is not varying the shape of f_0 contours to produce the contrast.

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, which indicates that Eq. (2) truncated at the second component provides a rather accurate account of the formant contour shapes.⁷ 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 $s_2^{F_1-2}$. Here, we analyse the scores separately, following the same procedure applied Section 4.1 and 4.2.

⁷PC3 was computed too, but it appeared not to contribute to the D/H contrast.

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 9 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 9 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.

The fourth column in Table 3 shows that eight out of nine speakers vary feature $s_1^{F_1-2}$ significantly and in accordance with the general trend, while speaker FM does not show any effect. Visual inspection of the data (cf. box- and curve plots in Additional Material) reveals that speaker FM does not vary the shape of formant contours to produce the contrast.

The last row of Table 2 reports the result of predicting score $s_2^{F_1-2}$ using 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. The fifth column in Table 3 shows that only one out of nine speakers (MM) seems to vary $s_2^{F_1-2}$ with class consistently with the global trend modeled by Eq. (3). 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.

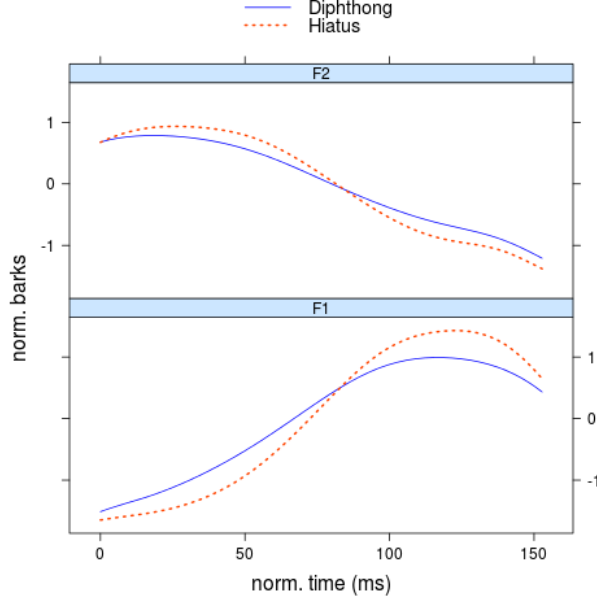


Figure 9: Two (F_1, F_2) contour pairs obtained by substituting the values of PC score $s_1^{F_1-2}$ predicted by linear model $s_1^{F_1-2} = -2.3 + 4.6 \cdot x$ (cf. fourth row of Table 2) into the FPCA model in Eq. (2). The D curves are resp. $\mu_{F_1}(t) + s_1^{F_1-2}|_{x=0} \cdot PC1_{F_1}(t)$ and $\mu_{F_2}(t) + s_1^{F_1-2}|_{x=0} \cdot PC1_{F_2}(t)$, 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.

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, which reflects a continuum between absence and presence of an elbow in the f_0 movement (cf. Section 4.2), and the PC1 score $s_1^{F_1-2}$ from FPCA jointly applied to formants, which reflects a continuum between a rapid transition between steady-state /i/ and /a/ vowels and a more gradual movement that spans the full duration of the diphthong (cf. Section 4.3). In this section we are interested in finding general patterns as well as individual differences among speakers in

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

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 generalised linear 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 4 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 5, where, to avoid confusion with Functional PCA, small letters are used for principal

	$pc1$	$pc2$	$pc3$
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 5: 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.

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 10 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.

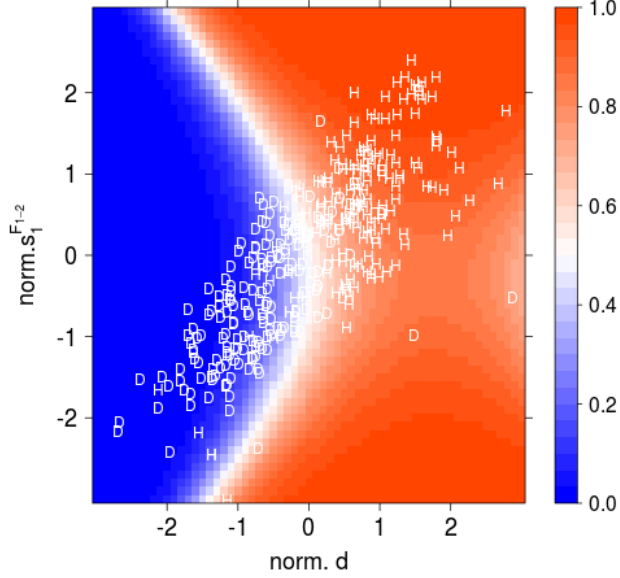


Figure 10: Representation of Eq. (4). Predictors $pc1$ and $pc3$ are converted back into their combinations of centered and normalised values of d (x-axis) and $s_1^{F_{1-2}}$ (y-axis), while $s_2^{f_0}$ is ignored (formally, it is set to $s_2^{f_0} = 0$). Each of the 365 tokens is represented by a letter indicating its class (D or H). In false colours, the probability of H as predicted by Eq. (4).

4.4.1. Differences between speakers in the use of the features

We built a set of speaker-specific models with the aim to investigate whether the speakers use idiosyncratic trade-offs between the cues in the realisation of the D/H contrast. For that purpose we built Conditional Inference Trees (CIT) (Hothorn et al., 2006), that estimate a regression relationship by means of a binary recursive partitioning in a conditional inference framework. CIT is conceptually identical to the better-known recursive partitioning algorithms CART (Breiman et al., 1984) and C4.5 (Quinlan, 1993), but the partitioning is based on statistical significance, which prevents overfitting and favours parsimonious models. For each speaker, a binary tree is computed that represents an optimal decision procedure for assigning a token, represented by the three features d , $s_2^{f_0}$ and $s_1^{F_{1-2}}$ (not normalised), to the D or the H class. Each tree is trained on the data belonging to one

speaker. A summary of the nine models is shown in Table 6. All trees are composed of only one decision rule, i.e. the root split criterion, except for the tree for speaker FM, where the trivial majority decision is the only criterion supported by data. All split criteria are significant (p-value < 0.001). The most interesting fact that emerges from Table 6 is that optimal decisions are based on different features for different speakers. Moreover, note that seven out of the eight speakers that realise the D/H contrast (i.e. all except for FM) seem to rely either on d or on $s_1^{F_{1-2}}$, while only speaker MM seems to rely consistently more on $s_2^{f_0}$. These observations confirm the interpretation of Eq. (4) provided above, namely that the variation of duration and formant shape are more powerful cues for the D/H contrast than the variation of f_0 contour shape.

speaker	root split criterion	tokens	err.
AM	$d > 163 \Rightarrow \text{H}$	39	1
CC	$s_1^{F_{1-2}} > -0.55 \Rightarrow \text{H}$	40	2
CP	$d > 146 \Rightarrow \text{H}$	40	0
CT	$s_1^{F_{1-2}} > -3.02 \Rightarrow \text{H}$	39	6
FM	H	48	23
MC	$d > 140 \Rightarrow \text{H}$	40	2
MM	$s_2^{f_0} > 4.61 \Rightarrow \text{H}$	39	0
MS	$d > 157 \Rightarrow \text{H}$	40	3
NM	$s_1^{F_{1-2}} > -2.60 \Rightarrow \text{H}$	40	3

Table 6: Summary of speaker-dependent conditional inference trees. The root split criterion, the number of tokens used for training and the classification error on the training set are indicated for each speaker. For example, the first row indicates that for speaker AM, the most reliable split criterion is based on duration. Given a token, if $d > 163$ then the token is classified as H, else as a D. The application of this rule produces only 1 classification error out of the 39 tokens produced by this speaker.

5. Discussion

As explained in the introduction of this article, previous phonetic research on D/H contrast in Spanish has identified several acoustic cues used by speakers to make this phonological distinction: the duration of the vocalic

sequence, its F_1 and F_2 dynamics, and the alignment of tonal targets within the vowel sequence. 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 scalars, and in that a joint statistical analysis has been performed. This procedure has allowed us to investigate our data in a more bottom-up manner than previous studies did, 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.

Contrary to more traditional methods, in which theoretical knowledge and hypotheses about the behaviour of specific features is required, a bottom-up approach to dynamic trajectory modelling such as FDA has allowed relevant patterns to emerge from the data. For instance, 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 two steady-state vowels, whereas a diphthong contains a glide and a steady-state vowel, and that a quadratic polynomial should be able to capture the relevant differences in their formant trajectories. In the same way, Torreira (2007) posited that if hiatus sequences contain a syllable boundary, whereas diphthongs do not, differences in the alignment of rising pitch accents within the vocalic sequence should be observed. As in Aguilar, Torreira hypothesised that such a difference should be captured by a single quadratic polynomial. In Section 3 and 4 we have shown how the same conclusions can be reached without prior assumptions about how specific dynamic features should behave. Operationally, we were able to start from a rich and generic family of functions (B-splines) to interpolate the raw data and end up with a graphically interpretable parametrisation of the data that did not require us to interpret the B-splines coefficients directly.

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 be-

havior nor focusing on any specific location in the utterance. The analysis revealed a previously unknown prosodic phenomenon in the post-focal region of the sentence, 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 Dutch learners of Italian revealed clear differences without requiring any prior knowledge about specific L1 or L2 pitch accent types.

This is not to say that FDA does not require any prior knowledge. 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. Smoothing and landmark registration are instances of the need to apply some degree of initial phonetic knowledge to guide automatic processing. During the smoothing step, which approximates sequence 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 required also in the set up of 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 our 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, those differences in general do play a role in the realisation of linguistic categories. In this paper we have taken care of this by carrying out a separate analysis of the duration of vowel sequences, but for longer utterances this approach 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); Turco and Gubian (2012).

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 using FLDA would 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 every PC corresponds to an easily interpretable shape characteristic, since more than one PC may be needed to compose recognizable shapes. 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 correlations 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 signals incurs the risk that one signal 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 has forced us to implement further measures to eliminate it (cf. Section 4.4).

Finally, 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 individual differences in the realisation of a multi-dimensional dynamic phonetic contrast. We observed that all but one speaker in our study distinguished hiatus and diphthongs through at least two of these cues. Interestingly, 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 four speakers favoured duration, three favoured the dynamics of F_1 and F_2 , and only one favoured f_0 alignment. This finding highlights the richness 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 FPCA-based method presented in this work can be extended with some adaptation to neighbouring domains of linguistic research, of which we mention three. 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 (cf. Parrell et al. (2013) for a recent application of FDA on EMA). 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 at least 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. Our results have exemplified the efficacy of FDA as automatic and data-driven ‘shape-to-numbers converter’ that allows researchers to reduce dynamic features into a small set of numerical parameters describing relevant variation, which can be then studied using traditional statistical tools, like linear models or t -tests. While FDA is a bottom-up approach to analysis, it still allows – and calls for – using prior knowledge. In fact there are clearly indentifiable points in the procedure where the investigator is required to make choices, which eventually must be grounded on prior knowledge, e.g. the level of detail in f_0 contour representation.

The data and software used in this paper are available at <https://github.com/uasolo/FDA-DH/>

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 software and the data that can be used to reproduce all the results reported in this article⁸.

⁸<https://github.com/uasolo/FDA-DH/> or direct link to zip bundle: <https://github.com/uasolo/FDA-DH/archive/master.zip>

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 these 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). 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 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 parameter. A larger value of k keeps more detail (corresponding to a higher time resolution), but it also causes a higher risk of overfitting (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 the cost function

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

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 of the 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 (e.g. due to measurement error). However, 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 which are of interest in a research question. In practice, minimizing GCV error with f_0 or formant contours yields curves that keep a level of details 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 weigh in. The strategy we adopt can be described as GCV-informed empirical judgement. First, we compute GCV errors on a grid of (k, λ) values by optimizing the parameters on a subset of the input contours (Figure A.1). Next, we select a promising combination (k', λ') on the basis of visual inspection of a number of curves from the held-out set. 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 (k'', λ'') with the curve corresponding to $k''' = 8$, $\lambda''' = 10^{-4}$. While GCV error in (k''', λ''') is almost the same as in (k'', λ'') (cf. Figure A.1), the low weight of the roughness penalty allows the curve to be ‘attracted’ by the isolated sample at the beginning, as shown in Figure 2(d).

Sometimes a full GCV-informed empirical judgement procedure may be extremely demanding. 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. For obvious reasons the best fit can be obtained by a large number of knots and a small weight of smoothness, 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 fit errors can be obtained with $\lambda = 10^6$, already in combination with $k = 8$.

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 an empirical linear relation 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 kinds of 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. Another has f_0 excursion as predictor and peak instantaneous f_0 change rate as dependent variable. The relations for a rising f_0 gesture (adapted from Xu and Sun (2002), Tables VI and VII,

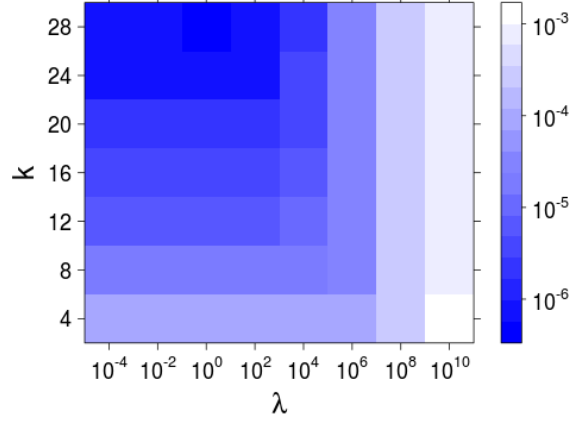


Figure A.1: A colour grid showing the generalised cross-validation (GCV) error for several (k, λ) combinations.

line ‘Mean’, column ‘Rise speed’) can be formulated as:

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

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

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 . 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, the average speed is around $1 \text{ st}/0.03\text{s} \approx 30 \text{ st/s}$, the excursion in the f_0 change enclosed by the box is around 1 st, so Eq. (A.2) is not satisfied. The same holds for the

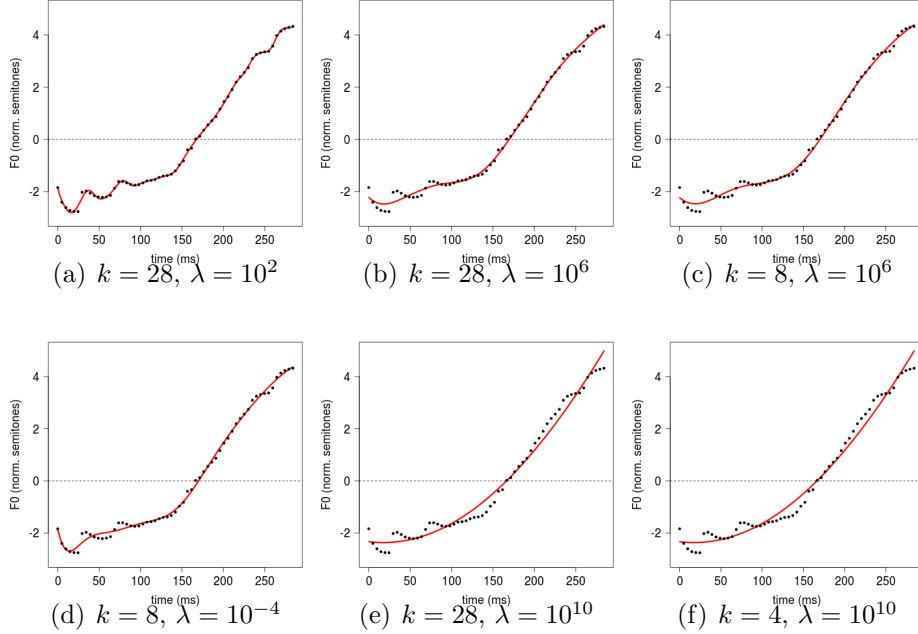


Figure A.2: An example showing the effect of the regularisation parameter λ and the number of B-splines knots k . Each panel shows a different (k, λ) value combination and the resulting smoothed curve, while the original f_0 samples are the same for all.

maximum speed of more than 50 st/s, indicated by an 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

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 $h(t)$ that specifies the mapping between the original and the registered time axis. The second stage applies

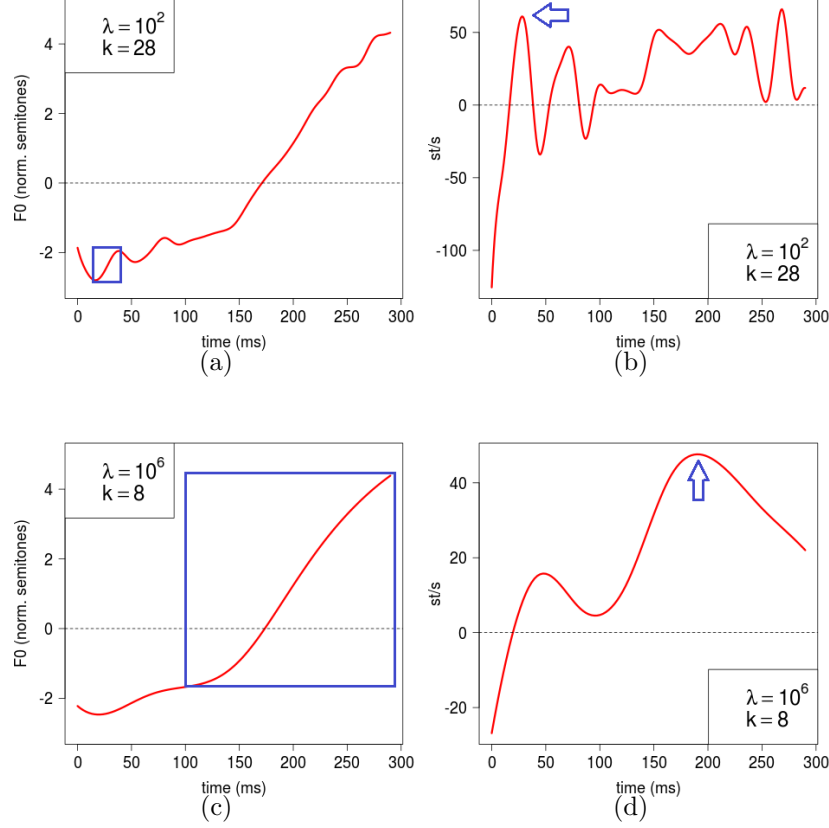


Figure A.3: In (a) an f_0 contour smoothed using parameters $k^* = 28$, $\lambda^* = 10^2$ (the same curve as in Figure 2(a), except that the time axis is not linearly normalised). In (b) the instantaneous velocity of the curve in (a), i.e. its first derivative with respect to time. In (c) an f_0 contour 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.

the warping functions $h(t)$ to the corresponding input curves $f(t)$ to obtain the registered curve $f(h(t))$.

Figure A.4 shows the warping functions $h(t)$ that transform the curves in Figure 3(a) (cf. Section 3.3) into those in Figure 3(b). While the landmark location on the input time axis (the Y-axis) of Figure A.4 varies considerably

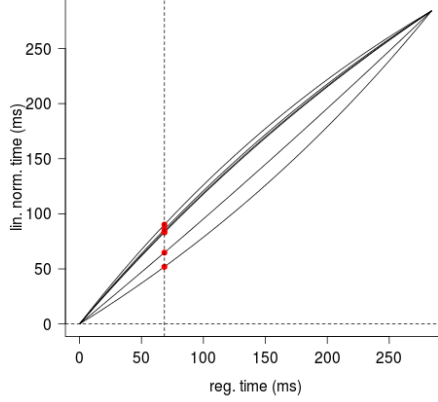


Figure A.4: The warping functions $h(t)$ that transform the curves in Figure 3(a) into those in Figure 3(b). The x -axis is the registered time axis, the y -axis is the original time axis, the dots show the position of the landmark of each curve (boundary between /l/ and vowel sequence).

across $h(t)$ curves, the location on the registered time axis (the X-axis) is almost the same for all curves. Moreover, the functions $h(t)$ show how time instants close to the landmark are ‘dragged’ more than others farther away from it. The starting and ending points of the $h(t)$ functions all coincide; this is due to the time normalisation that was part of the preceding smoothing operation.

The $h(t)$ 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 $h(t)$ must be monotonically increasing or decreasing; non-monotonic functions would correspond to local inversion of the time course. The more $h(t)$ departs from a straight diagonal line, the heavier the time warping is.

Obtaining reasonably smooth functions $h(t)$ 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. The reason is that we assume that the landmark locations are

error-free, i.e. we wish $h(t)$ to cross all landmark locations faithfully. As a consequence, the theorem by de Boor (de Boor, 2001) that states that the optimal knots location coincides with the sample location can be applied. 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. Here, we only give an intuitive feel of the underlying mathematical procedure. The first PC curve $PC_1(t)$ is computed by maximizing the variance of the integral

$$\int_0^T PC_1(t) f_n(t) dt, \quad (\text{A.4})$$

in the set of input curves $f_n(t)$, n is the index running from 1 to the number of curves; $[0, T]$ is the normalised time interval. The resulting function $PC_1(t)$ captures, on average, the largest possible proportion of the variance in the set of curves $f_n(t)$. Higher degree components PC_2, PC_3, \dots 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 $PC_m(t)$ 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, parts of the analysis carried out in Section 4 are 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 the 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 present.¹⁰

Table 7 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. Also, the values of \tilde{R}^2 and $\Delta\sigma(u_1)$ show the same pattern as R^2 in Table 2. Importantly, Table 7 confirms the discrepancy between the group of features d , $s_2^{f_0}$ and $s_1^{F_1-2}$ that benefit substantially from a fixed term encoding token class, and the group $s_1^{f_0}$ and $s_2^{F_1-2}$ that do not. A closer inspection of Table 7 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). In the group of features d , $s_2^{f_0}$ and $s_1^{F_1-2}$

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.

¹⁰This 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:

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

where σ denotes standard deviation.

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 7: 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).

we can observe a difference between $s_2^{f_0}$ and the other two. While for $s_2^{f_0}$ and $s_1^{F_{1-2}}$ the values of \tilde{R}^2 and $\Delta\sigma(u_1)$ are both high, $s_2^{f_0}$ shows a relatively low value of $\Delta\sigma(u_1)$, in combination with a high \tilde{R}^2 . While a high \tilde{R}^2 tells us that the model in general benefits from incorporating speaker information, a relatively low value of $\Delta\sigma(u_1)$ tells us that the way speakers differentiate D from H using feature $s_2^{f_0}$ is not as systematic as in the case of features d and $s_1^{F_{1-2}}$, where $\Delta\sigma(u_1)$ is much higher. This interpretation confirms the combined analysis in Section 4.4, which suggests that features d and $s_1^{F_{1-2}}$ play a more systematic role than $s_2^{f_0}$ in the realisation of the D/H contrast.

B.1.1. Consistency between speakers

In Section 4 speaker-specific t -tests were carried out 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,

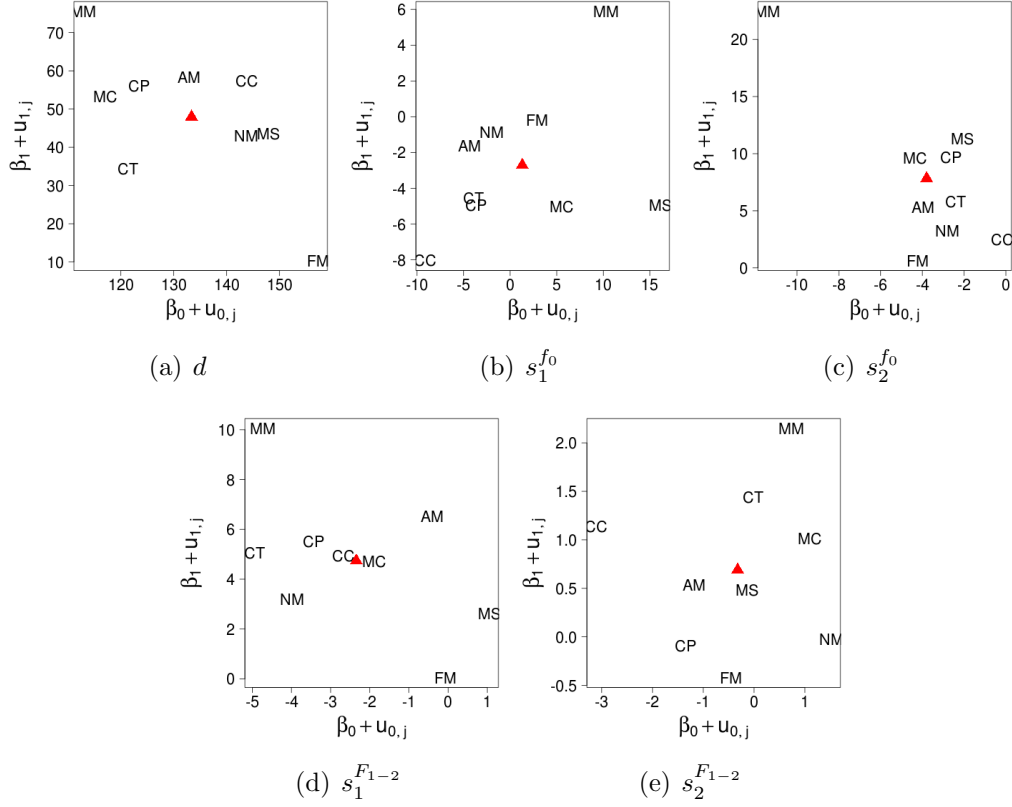


Figure B.1: By-speaker coefficients for models of the form of Eq. (B.2). Each panel shows the results for a different model, where y_{ij} is substituted with a different feature, indicated in the panel caption. In all panels, the x -axis shows the corrected speaker-specific term $\beta_0 + u_{0,j}$, the y -axis shows the corrected speaker-specific term $\beta_1 + u_{1,j}$. A red triangle locates the fixed terms (β_0, β_1) .

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 7 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 8). 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 $s_2^{F_1-2}$ 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 $s_2^{F_1-2}$ is small compared to the class-independent inter-speaker variation ($\beta_0 + u_{0,j}$).

B.2. Global Analysis

In this section, Eq. (4) 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 been already 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 10. 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. Interestingly, these are two of the three speakers for whom the Inference Trees in Section 4.4 showed that they rely more on formant trajectories than on duration for making the D/H contrast (cf. Table 6). Moreover, note that most of the data points for speaker FM lie in the undecided white 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. 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.

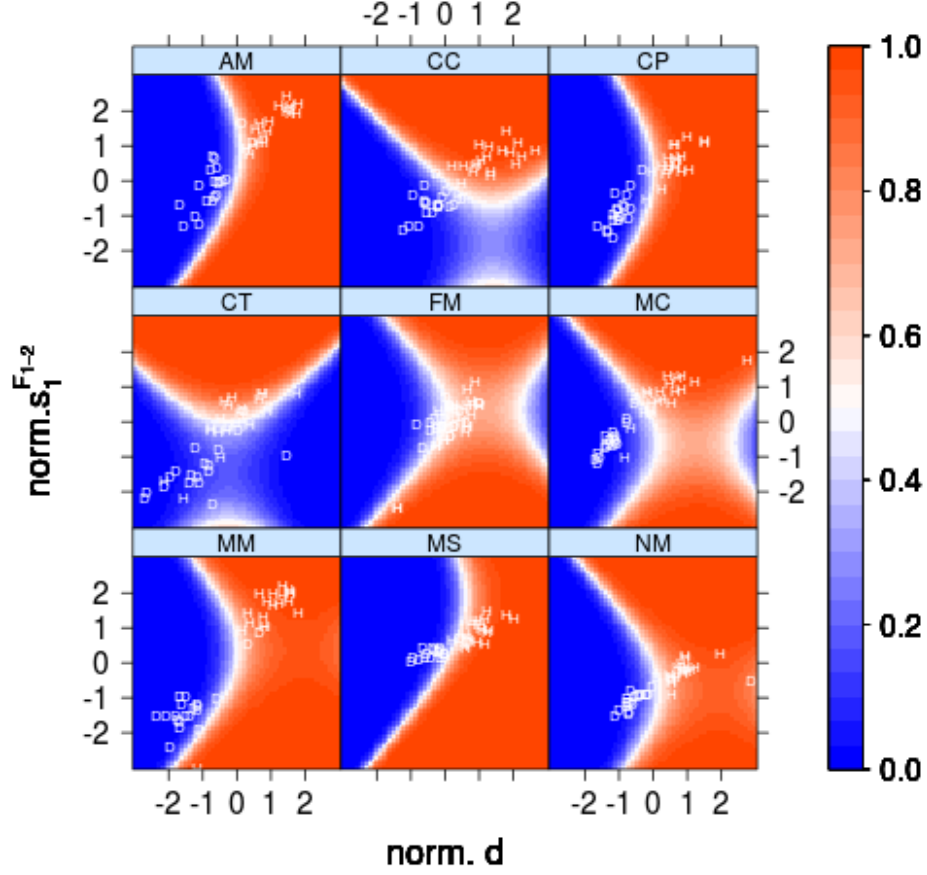


Figure B.2: Representation of Eq. (B.3). Predictors $pc1$ and $pc3$ are converted back into their combinations of centered and normalised values of d (x-axis) and s_1^{F1-2} (y-axis), while $s_2^{f_0}$ is ignored (formally, it is set to $s_2^{f_0} = 0$). Each of the 365 tokens is represented by a letter 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.

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/>.
- Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. The Wadsworth statistics/probability series, Belmont, California.
- 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.

- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70.
- Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3).
- 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.
- 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, 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.
- Quinlan, J. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Mateo, California.
- 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.