Time Normalisation and Landmark Registration

Michele Gubian, Francisco Torreira and Lou Boves

July 20th, 2014

Additional material, related to
Michele Gubian, Francisco Torreira, Lou Boves, "Using Functional Data Analysis for investigating multidimensional dynamic phonetic contrasts", Journal of Phonetics

1 Introduction

Functional Data Analysis (FDA) offers a very powerful set of tools for analyzing phenomena that can be described as a continuous function of one variable. Invariably, the observations are available in the form of discrete samples of the phenomenon. In Phonetics research the independent variable is often *time*; the examples in the paper are $f_0(t), F_1(t), F_2(t)$, i.e., pitch, first and second formant frequency as a function of time. But the independent variable can also be *distance from a reference position*; here, the cross-section of the vocal tract as a function of distance to the glottis would be an example.

FDA aims to analyze the *shape* of the trajectories that characterize f_0 movements (or vocal tract cross-sections). Inevitably, and unsurprisingly, even extremely powerful tools come with some restrictions. For FDA the single most important restriction is that the trajectories under analysis must be defined on a common support. For f_0 movements this means the time from the start to the end of the movement, regardless of the exact number of milliseconds that it took to go from start to finish. For vocal tract cross-section this would mean the distance from the glottis to the lips, regardless of the actual number of millimeters.

The vocal tract cross-section example shows that *shape* may not provide sufficient information to allow for solving a specific problem. Vocal tract cross-section is sufficient for predicting the ratio between the frequencies of the first n formants, but if we also need to know the absolute value of the

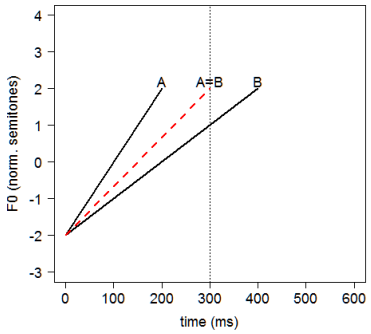

Figure 1: Application of linear time normalisation. A and B denote contour end points, black lines are the original contours, red dashed lines are the corresponding normalized (overlapping) contours.

formant frequencies, then we need to also know the actual length of the tract. Similar problems can occur with time normalisation. Figure 1 shows what happens to two (artificial) contours when linear time normalisation is applied. In the figure, two contours that have different duration and slope on their original time axis become identical after normalisation. This kind of information loss and potential ambiguity is inevitable any time a form of normalisation is introduced, and this is why we carried out an independent analysis of duration together with FDA.

2 Selecting the data

In the paper we advocate a data-driven approach to phonetic research and to building phonetic and phonological theory. However, there is no such thing as a completely theory-free data-driven analysis. There is always some theory (or if one prefers avoiding heavily loaded terms such as 'theory': there is always some knowledge or experience) that guides the selection of the data under analysis. In most phonetic research there is no reason for recording the colour of the eyes and hair of the participants, or their weight and height, or the temperature of the room, because there is no reason why these features could possibly affect the outcome of the experiment. Which is not to say that these features are never ever relevant. The research aimed at establishing

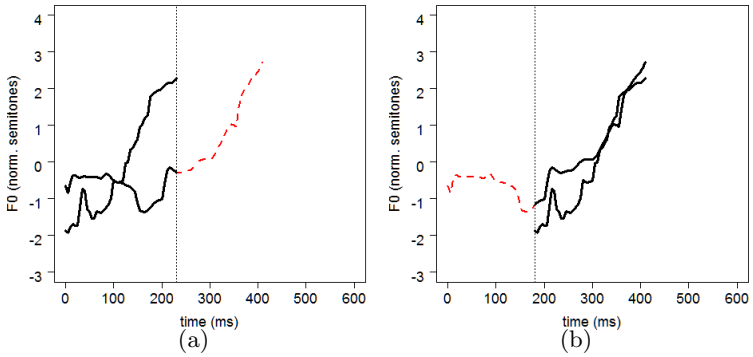

Figure 2: Application of procrustean bed from the start (a) and from the end (b) of the signal. The red dashed portion of curves is the one that is lost in the process.

whether a speaker's height can be estimated from speech features such as pitch and formant frequencies is a clear counter example van Dommelen (1993); van Dommelen and Moxness (1995); Gonzáalez (2004).

In our data-driven approach we cannot – and want not – escape the need for making justifiable, well-founded, decisions about how to select the data. For the analysis of f_0 contours we decided to extract complete accent-lending rises, even if this meant that we had to include part of an f_0 movement that extends into the subsequent, unstressed, syllable. We could have decided to only extract the f_0 movement within the accented syllable. However, this decision would not have taken away the need for doing some time normalisation, because not all accented syllables are equally long, and we would be left with a trajectory that would have lost its phonological meaning. The same would be true if we would have decided to extract a fixed time length from the start of the f_0 rise: doing that would have done away with the need for time normalisation, but again at an unsupportable cost, namely sacrificing the phonological and phonetic meaning of the data under analysis. The same arguments hold for the decision to extract the formant contours from the beginning to the end of the diphthongs and hiatuses. We give precedence to the phonetic and phonological integrity of the data, and accept the need for applying some form of time normalisation that this integrity incurs.

The need for extracting phonetically complete f_0 movements is illus-

trated in Figure 2. This figure clearly shows that extracting a fixed time interval would incur substantial loss of information. However the fixed time interval would be selected, relevant parts of the signal would be chopped off.

3 A common support

In phonetics and speech technology research a number of different techniques are being used for 'normalising' the time axis of observations, so as to bring all observations to a fixed length. A procedure that was quite successful in the early days of isolated word recognition was to represent a word in the form of a fixed number of samples of the short-time spectrum. In a long word the time distance between the samples would, by necessity, be longer than in a short word. Another obvious way for mapping observations to the same support on the time axis is linear scaling, where $f_i^s(\tau) = f_i(a_i \cdot t)$. For $t = 0$ all $f_i^s(\tau)$ are aligned at $\tau = 0$. The a_i are chosen such that $f_i^s(T) = C$ for all $t = t_{max}$ in the individual contours. Yet another popular method for time normalisation is Dynamic Time Warping Myers and Rabiner (1981).

The time normalisation applied in FDA differs from all the procedures mentioned above, but it is related to the first one, which was based on selecting a fixed number of samples to describe phenomena with different durations. In the FDA procedure that we propose in the paper time normalisation happens as an inevitable side-effect of the approximation of a variable length sequence of discrete samples by means of *B-spline* functions defined by a fixed number of *knots*. In effect, the knots in the B-spline approximation of a function of time take over the role of the tick marks on a mechanical stop watch. In the example of the vocal tract cross-section the knots would take over the centimeter marks on a ruler. As long as the number of knots is fixed, the approximating function is defined on a fixed support, that is the same for all individual observations.

From a strictly formal perspective, the independent variable in a B-splines approximation does no longer correspond to the physical dimension of the phenomenon under investigation. Note, however, that this holds for all methods for time normalisation, with the exception of extracting a fixed time window, a procedure that we have ruled out because it would result in meaningless curves. In all practical situations where spline interpolation can be justified the relation between the original physical variable and the mathematical variable in the spline function is so close that no harm is done if in talking about the spline function the independent variable is given the same name and interpretation as the original physical variable. Therefore,

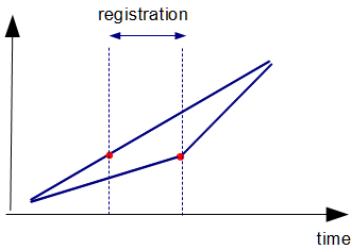

Figure 3: Introduction of an artifact caused by landmark registration. The straight line shows the result of linear time normalisation. The lower line show the result if the landmark, represented by the red dot, would be moved to the righthand position. This would introduce a knee that must be considered as an undesirable artifact.

it is acceptable to talk about the independent variable in our spline approximations as it it were time; to remind the reader that it is not strictly speaking time, we can replace the symbol t by the symbol τ.

In our analysis of f_0 and formant trajectories we have always used splines approximations based on equidistant knots. Doing so was a conscious decision, based on the fact that we had no strong theoretical arguments for moving specific knots to the location of specific 'events' in the f_0 curves or formant trajectories. However, there is no mathematical reason that would have prevented us from placing the knots so as to concur with specific events at non-equidistant positions, as long as we would have used the same positions for all trajectories under analysis. It goes without saying that the correspondence between the original physical dimension (time) and the independent variable in the spline function would be further compromised by using non-equidistant knots. In the stop watch analogy: where the knots are farther apart, 'time' must run faster to reach the next mark on schedule.

4 Landmark registration

Above we have said that there were no compelling arguments for marking specific events in the contours under analysis. This is true, with one exception: it is both phonetically meaningful and practically possible to reliably mark the boundary between the initial /l/ and the following diphthong or hiatus in the accented syllables. We decided to use this 'boundary event' for landmark registration in the analysis of the f_0 contours.

Landmark registration is a way to adjust curves to be able to align shape changes that occur as consequence of meaningful events (landmarks), which in our case are phone boundaries. Landmarks make it possible to distinguish between meaningful shape changes and differences between shape that are due to differences in the duration of segments. It should be obvious that it is the responsibility of the researcher to determine whether there are any such landmarks in the data under analysis. Thus, landmark registration is a tool that a researcher can apply to harness prior knowledge about the phenomena under analysis. In our analysis we decided to use only one landmark, namely the boundary between /l/ and the vowel sequence, because we were interested in referring the shape of the rising f_0 movement to the boundary between the segments. We did not attempt to put a landmark between /i/ or /j/ and /a/, since for the diphthong it is virtually impossible to annotate a boundary reliably, which is probably related to the fact such a boundary has no phonetic reality. Even if it would have been possible to define a 'landmark' in the diphthongs (for example on the basis of the derivative of the formant trajectories), we doubt that it would be safe to consider that 'landmark' in a diphthong as equivalent to the syllable boundary between /i/ and /a/ in a hiatus.

Landmark registration aims to avoid aligning shape elements that correspond to different underlying processes. However, this comes at a price, in the form of additional (over time normalisation) distortions of the original contours. In general, the larger the differences in duration of corresponding segments to be aligned, the higher the risk that some form of unwanted distortion may occur. In extreme cases landmark registration may introduce artifacts. This is illustrated in Figurefig:pitfall. Without landmark registration the time-normalized version of a straight line would remain a straight line. However, if a landmark located somewhere on the line is moved to a fixed position on the horizontal axis a 'knee' could be introduced. This illustration shows that it is mandatory that researchers who decide to apply landmark registration to make sure that no artifacts are introduced. An FDA analysis, which by definition is totally ignorant of the semantics of the

functions on which it operates, has no way for telling 'true' shape elements apart from artifacts.

4.1 Landmark registration in the analysis of f_0 curves

We carefully checked the effect of landmark registration on the shapes of the f_0 curves in our data set. Because the distance over which the landmarks needed to be shifted was always small, the registration had only a minor impact on the resulting shapes. Even in extreme cases, i.e., very long and very short utterances that require most distortion, illustrated in Figure 5 no artifacts can be seen. There was certainly no need to worry about artifacts. However, if the effects are minor, one may ask the question whether landmark registration had any effect at all. To answer this question we compared the output of FPCA analysis of the f_0 curves with and without landmark registration. Note that this means that the non-registered curves are linearly time-normalized as the result of the spline interpolation in the smoothing operation. The results of this comparison are shown in Figure 4. From this figure two interesting facts can be observed. First, PC curves look quite similar in the two conditions, which is a good sign in that it confirms that landmark registration did not introduce artifacts. In both cases, PC1 roughly describes slope variation, PC2 the presence of an elbow. The second fact is that PC2 score s_2 is clearly less effective in discriminating diphthongs from hiatuses when landmark registration is not applied. This is not only evident by comparing the separation of D and H points in Fig. 4(e) and (f), but it is also quantitatively supported by the fact that the linear model $s_2 = \beta_0 + \beta_1 \cdot x$ (cf. Eq. (3) in the paper) explains 22% of the variance when no registration is applied, while it explains 35% of the variance after landmark registration (cf. Table 2 in the manuscript). Even though the PC2 curves describe roughly the same effect on the shape of the curves, this effect is confounded by the different position of the boundary between /l/ and the vowel sequence. As a consequence, s_2 in the non-registered version accounts for both the D/H distinction and for the placement of the segmental boundary, while the latter effect is neutralized by registration.

References

Gonzáalez, J. (2004). Formant frequencies and body size of speaker: a weak relationship in adult humans. *Journal of Phonetics*, 32:277 – 287.

Myers, C. S. and Rabiner, L. R. (1981). A comparative study of several

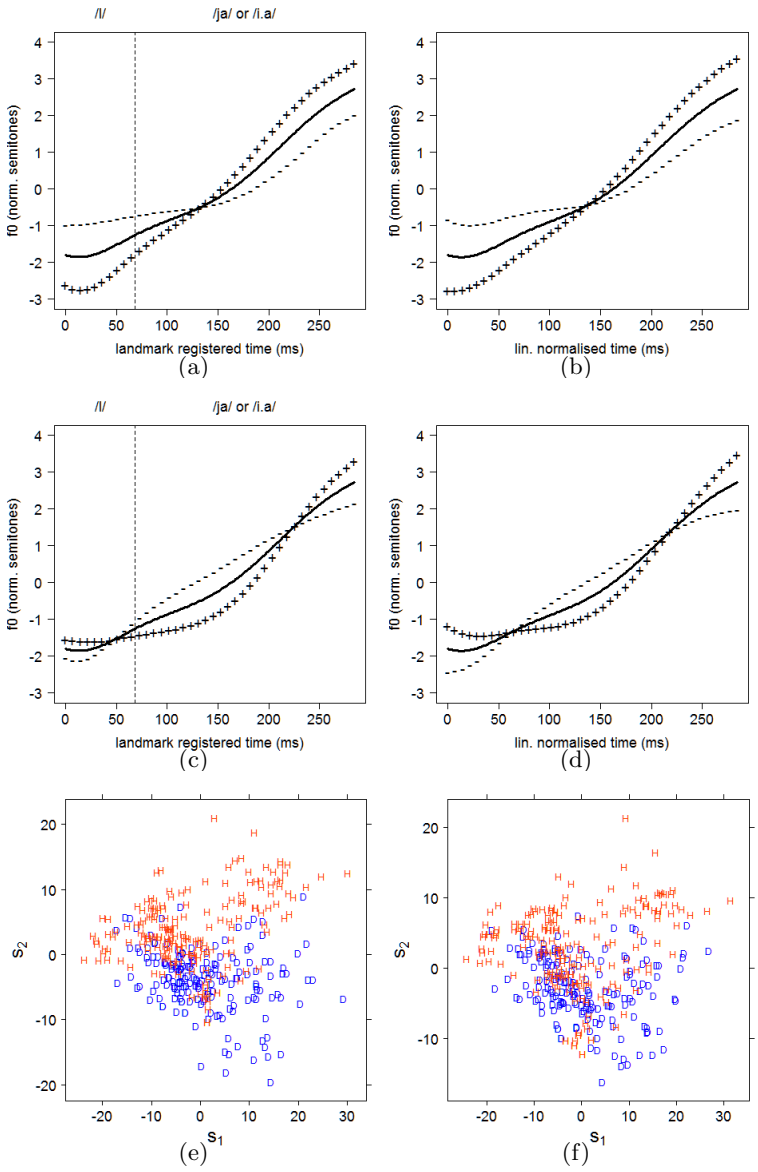

Figure 4: FPCA on f_0 contours with (left) or without (right) applying landmark registration. PC1 explains 61% / 75% of the variance resp. with (a) / without (b) registration, PC2 explains 26% / 23% of the variance resp. with (c) / without (d) registration. Scatter plots of PC scores correspond to the FPCA version with (e) / without (f) registration.

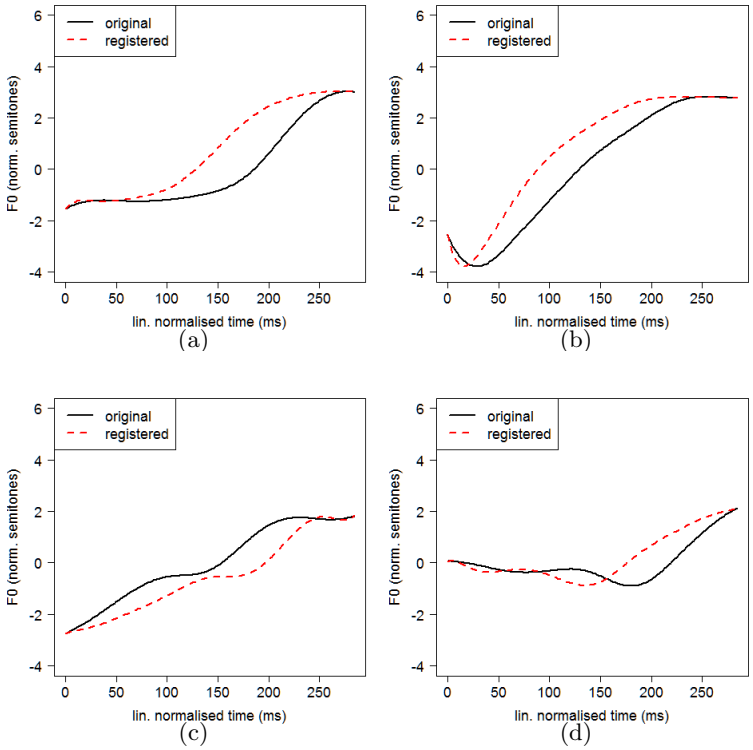

Figure 5: Some of the curves that underwent the largest shape modification as consequence of landmark registration.

dynamic time-warping algorithms for connected word recognition. *The Bell System Technical Journal*, 60(7):1389 – 1409.

van Dommelen, W. (1993). Speaker height and weight identification: A re-evaluation of some old data. *Journal of Phonetics*, 21:337 –341.

van Dommelen, W. and Moxness, B. (1995). Acoustic parameters in speaker height and weight identification: sex-specific behaviour. *Language and Speech*, 38:267 – 87.