Functional Data Analysis for Phonetics Research

Michele Gubian

Centre for Language & Speech Technology
Radboud University, Nijmegen, The Netherlands
M.Gubian@let.ru.nl

Abstract

This work introduces Functional Data Analysis (FDA) as a powerful methodology for speech analysis and re-synthesis. FDA allows one to carry out statistical analyses on a set of speech parameter contours in time, like f_0, formants, intensity, in isolation or jointly. FDA eliminates the intermediate step of (manual) extraction of shape descriptors, like peak and valley locations, slopes, and so on. All the information contained in the curve shapes is preserved and used in the analysis. A case study illustrates the potential of FDA for phonetics research. The author maintains a website where papers, didactic material and code samples can be freely downloaded.

Index Terms: Functional Data Analysis, Principal Component Analysis, Prosody

1. Introduction

In the analysis of the speech signal we are often confronted with the problem of how to summarise and quantify facts that have to do with contour *shapes*. A typical example comes from intonation research that analyses the f_0 contour as the main acoustic correlate of intonational phenomena. The widely employed framework of autosegmental-metrical theory [1] postulates the existence of high and low (H, L) tonal targets located at points in time (phonologically) associated to the segmental material and to specific functions (e.g. focus). In this framework, a quantitative analysis of f_0 contours starts with the search of those targets in the signal, and subsequently with the application of vector statistics (e.g. ANOVA, t-test) on numerical descriptors of them, typically their time-frequency coordinates and possibly simple shape indicators like slopes. This approach to f_0 contour analysis brings along three problems. One is an induced piecewise linear stylization, meaning that all that is preserved in the quantitative analysis of contours are the locations of the tonal targets (implicitly) connected by straight lines. The consequence is that other shape-related aspects are lost, while evidence is accumulating in favor of the perceptual relevance of aspects like peak vs. plateau [2] or concavity vs. convexity of a rising gesture [3]. The second problem is that the tone targets and possibly other relevant points have to be located in the signal, which is not an easy task in that for example a plateau can be found where a H target has to be marked, or a slowly varying slope where a precise ‘elbow’ point has to be located. A third problem is that both the tone sequence identification and their subsequent marking depend to an extent on the judgement of trained listeners. Thus, not only the risk of a biased analysis is inevitable, but also inter-annotator agreement has been shown to be highly variable [4], and finally relying on human intervention makes the analysis of large datasets too expensive.

Another field where contour shapes play a major role is the manipulation of f_0 contours (and possibly other speech parameters) for perceptual experiments employed in intonational research. This practice involves some combination of stylization of f_0 tracks measured in a corpus of spoken utterances, and perceptual experiments in which subjects judge resynthesized versions of the utterances with the manipulated f_0 contours [5, 6]. The experimental f_0 contours can be produced by some phonological or physiological model, or the contours are created manually. Difficulties arise when changes in the contour shape need to be applied globally and smoothly in the whole curve. Usually, assumptions and simplifications (e.g. stylization) are adopted in order to make the manipulation tractable by the experimenter. However, those simplifications may conceal subtle yet important dynamic variations that are used by the listener as discriminative cues, which ultimately will not be tested in the perception experiment.

Purpose of this work is to introduce a set of advanced statistical techniques collectively known as Functional Data Analysis (FDA) [7, 8, 9] as a way to alleviate most of the problems described above. FDA, proposed in the late 90’s by J. O. Ramsay and his group, extends well known statistical tools like Principal Component Analysis (PCA) and linear regression in such a way that their input elements become curves, appropriately represented in form of functions, rather than fixed length vectors. This brings three notable advantages. One is that all the information contained in the curves is preserved and used in the analysis. Second, the intermediate step of selecting and measuring shape descriptors, like in the way we have illustrated above for intonation, is eliminated. Third, the mathematical description of the curve dataset can be used to explore the space of shape variations and re-synthesize new curves that can be used for listening experiments.

My contribution is to bridge the gap between FDA as a general purpose statistical tool and the specific needs the analysis of the speech signal brings along. This gap is both technical and cultural. Ramsay and colleagues created and maintain two freely available software packages to perform FDA, one runs under R[1], the another under MATLAB[2]. I have developed speech-specific technical solutions in order to help making FDA a useful and complete tool for the community. These go from general methodologies to incorporate segment durations in the analysis [10] to more practical software solutions, e.g. to ease the interfacing between the FDA software and Praat. On the

The research of Michele Gubian is supported by the Marie Curie Training Network Sound2Sense (www.sound2sense.eu). The author thanks all the researchers who are collaborating to make FDA a valuable tool for phonetics research. These include Lou Boves, Francisco Torreira, Francesco Cangemi, Chierh Cheng, Xu Yi, Meg Zellers, Brechtje Post, Helmer Strik, Sarah Hawkins, Dino Seppi.

[1] http://cran.r-project.org/web/packages/fda/index.html

[2] ftp://ego.psych.mcgill.ca/pub/ramsay/FDAfuns/Matlab/

cultural side, I have written papers, tutorials as well as 'recipe' code and modified functions for the R version of the FDA software in order to make FDA more accessible for the linguistic community. Most of the aforementioned material can be freely downloaded from my website [11].

In the rest of this work, a case study is presented that will allow the reader to get familiar with the main steps to carry out FDA on a dataset of f_0 contours. The code written to perform the analysis is also available from the website [11].

2. Case study

2.1. The dataset

The data used in this case study is part of a larger corpus of read speech collected by F. Cangemi to study focus and question/statement modality in Neapolitan Italian. This material has been used in [10].

Starting with [12], many studies on various languages have shown that focused constituents (as is the Verb in the sentence "No, he LEAVES at 10", uttered as an answer to the question "Does John arrive at 10?") are acoustically characterized by greater f_0 movements, longer duration and, in some cases, higher overall intensity. Here we will analyze f_0 contours only, whereas in [10] also speech rate is considered.

Five speakers of Neapolitan Italian read three repetitions of three declarative sentences sharing the structure: $[\text{CV}\underline{\text{CV}}\text{CV}]_S$ $[\underline{\text{CV}}\text{CV}]_V$ $[\text{CV}\underline{\text{CV}}\text{CV}]_O$ (lexical stressed syllable is underlined, S(ubject), V(erb), O(bject) specify the syntactic role). All phones are voiced, S and O are proper names, as in the case of *Ralego vede Ladona* ('Ralego sees Ladona'). The data consist in $N = 132$ sampled f_0 contours (5 speakers × 3 repetitions × 3 sentences × 3 focus positions - 3 discarded). The f_0 samples were computed every 10 ms using Praat autocorrelation-based f_0 extractor with default settings [13]. The values are expressed in semitones, and each curve had its mean value subtracted in order to eliminate variation due mainly to speaker identity.

2.2. Sampled data smoothing

In order to perform FDA on a set of sampled curves, the first step is to obtain a functional representation of each curve. All FDA tools accept a set of functions as input that have to obey two rules. One is that all functions have to be defined on the same (time) interval (the reason will be illustrated in Sec. 2.4). The other is that functions are chosen from a basis, typically a B-splines basis, and considerable computational advantage is gained by using the same basis to represent all functions. The smoothing procedure I illustrate here follows the general recommendations of the FDA literature [7] with some adaptations. B-splines are generally a good choice for a basis, since they basically introduce no hypothesis on the contour shape. A B-spline basis is a set of adjacent polynomial functions defined on a finite (time) interval, where the number B and location of those functions have to be specified. Once a basis is chosen, we have to choose one function out of the basis that approximates the discrete sequence of samples y_i at time t_i, $i = 1, \ldots, S$ by satisfying a predefined optimality criterion. This criterion is the joint optimization of two contrastive goals, one is that the function resulting from the weighted sum should pass as near to the samples as possible, the other is that this function should have as little curvature as possible, i.e. being smooth. This is expressed by an optimization problem as follows:

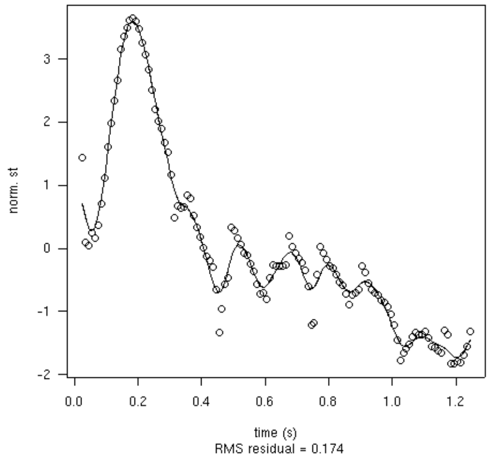


Figure 1: One of the 132 f_0 contours in the dataset. Points are the samples obtained from Praat. The solid curve is the smooth function obtained by following the procedure explained in the text.

$$\min\{\mathit{SSE} + \lambda \cdot \mathit{PEN}\} \tag{1}$$

where *SSE* is the sum of squared errors of the fitting function with respect to the original time samples, *PEN* is a measure of function roughness and $\lambda > 0$ is a coefficient that weights the importance between the two. Note that SSE and PEN are known only after the parameter λ and the function basis are specified.

While the solution of (1) is carried out by the software, we have to choose the number B and location of basis functions and the value of λ. An empirical approach is to use Generalized Cross Validation (GCV), which is available within the FDA software. More precisely, I recommend to use equidistant bases and to try several values of B and λ (the latter on a log scale). Then several candidate choices should be evaluated by eye inspection. The reason not to follow a purely quantitative approach (and not to apply Boor's theorem on the knots locations [7]) is that the experimenter may not be interested in modeling fine time scale fluctuations in the signal, which can be due to measurement error or, in the case of f_0, to microprosodic phenomena. In other words, the experimenter may have an idea of what to consider signal and what noise. An example of f_0 curve smoothing is shown in Figure 1.

2.3. Landmark registration

Once all curves have a functional representation, the actual FDA could start. However, we have to bear in mind that we had to use the same time interval to represent all curves. This means that each curve has been linearly time-normalized. This representation of the dataset may not be a good one when dealing with the speech signal. The reason is that FDA treats all curves as 'synchronized' on time t. To elaborate, sequences of comparable events or *landmarks*, like phone boundaries in a given spoken utterance, do not occur at the same time across different realizations, even if we allow linear time normalization. *Landmark registration*, on the other hand, allows us to align the input functions on those events as follows. If τ is the common adjusted time axis, for each function $f(t)$ a time distortion function $h(\tau)$ has to be determined that satisfies

$$t_l = h(\tau_l), \quad l = 0, \dots, L+1$$

where t_l are the landmarks for curve $f(t)$, τ_l their location on the common time axis τ (usually the average positions of landmarks across the dataset), $t_0 = \tau_0 = 0$, and $\tau_{L+1} = T$. Each function $h(\tau)$ is found by solving a regularization problem similar to (1).

The above procedure has been applied to our dataset by using each phone boundary as landmark. Even though the lexical material is not identical across the dataset, the sequences of C and V are (Sec. 2.1) The phone boundaries have been obtained using forced alignment carried out with an ASR trained on standard Italian made publicly available by D. Seppi[3]. From now on, all the curves in the dataset look like if they were synchronized on the sequence of phones. This takes away all the variation due to the asynchronicity of different utterances while leaving that coming from different realization of f_0 gestures with phone boundaries as reference. On one hand this makes the analysis of shapes meaningful. On the other hand all the information concerning phone duration is destroyed. In this case study we will not show how to recover the latter, but the reader is referred to [10], where a solution is proposed.

2.4. Functional Principal Component Analysis (FPCA)

PCA is a way to extract and display the main modes of variation of a set of multidimensional data. Starting from a data set in its original coordinates, a new coordinates system is found such that by expressing (projecting) the data points on it, the first projection accounts for the largest part of the variance in the data set, the second for the second most important part of the variance, etc. More formally, if the input data are N fixed size column vectors x_n, the k-th principal component (PCk) is the vector ξ_k of norm one that produces the largest possible variance of the scalar product $\xi_k \cdot x_n$ across the N vectors x_n, The vector ξ_k must be also orthogonal to the previous components ξ_1 to ξ_{k-1} obtained in the same way. Functional PCA extends PCA to accept input data in the form of functions $f_n(t)$ by defining the scalar product as

$$c_{k,n} = \int_0^T \xi_k(t) f_n(t) dt, \tag{2}$$

while keeping the remainder of the PCA math formally unchanged. The role of every PC function $\xi_k(t)$ is to amplify systematic shape variations that occur across the N input functions $f_n(t)$. As anticipated, (2) requires functions to be defined on a common interval $[0, T]$, since the integration (2) treats the variable t identically in all functions. Landmark registration introduced above provides a way to accommodate data that are not synchronized on t in their original form. In this way, shape variations induced by the random misalignment of curves cannot affect the maximization of the variance of (2).

Each input curve $f_n(t)$ can be approximatively reconstructed by using the first K PCs as follows:

$$\hat{f}_n(t) = \mu(t) + \sum_{k=1}^{K} c_{k,n} \xi_k(t), \tag{3}$$

where the $c_{k,n}$'s from (2) are called *PC scores* and $\mu(t) = N^{-1} \sum_n f_n(t)$ is the mean curve.

Figure 2 shows the first two principal components (PCs) obtained carrying out FPCA on the $N = 132$ registered contours $f_n(t)$. In solid line the mean contour $\mu(t)$ is shown, while the '+' and '-' curves visualize the effect of adding to/subtracting from the mean the first (a) or the second (b) PC multiplied by one standard deviation of the corresponding PC score, i.e. $\mu(t) \pm \mathrm{sd}(c_k) \cdot \xi_k(t)$, (a) $k = 1$, (b) $k = 2$. PC scores $c_{k,n}$ as in (2) for all contours are represented in a (c_1, c_2) plane in Figure 3, each score being marked with the focus condition S/V/O in the corresponding utterance n.

Figure 3 shows that the first two PCs, together capturing 58% of the variance, basically express the variability that focus position induces on the f_0 contours. The shape of a 'typical' focus condition S contour is obtained by applying reconstruction (3) using a negative c_1 and a positive c_2 (Figure 3). PC1 modifies $\mu(t)$ by rising the peak in correspondence to the first lexical stress and lowering the peak on the third one (Figure 2(a)), while PC2 lowers the peak of the second lexical stress and compensates the action on the third peak done by PC1 (Figure 2(b)). The resulting shape is in line with what expected from previous studies [2]. Similar considerations apply to focus conditions V and O.

2.5. Creating new f_0 contours by exploring the PC space

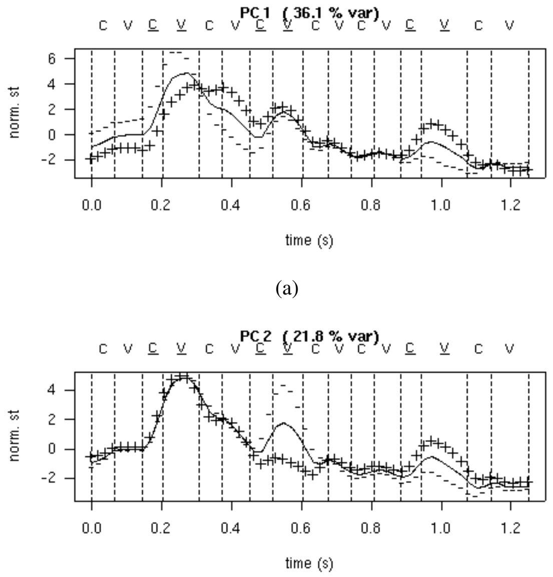
Equation (3) allows us to perform an approximate reconstruction of each original curve n by using only the mean $\mu(t)$, the first K PC functions $\xi_k(t)$ and the specific PC scores $c_{k,n}$, which can be read out from Figure 3. However, we can go beyond that. We can use (3) to produce new contours by simply choosing PC score combinations that do not correspond to any curve in the dataset. Since a small change in any PC score will result in a small and smooth change in the reconstructed curve, we expect that for example in PC score plane regions between two focus condition clusters (Figure 3) the resulting f_0 contours would be somewhat perceptually ambiguous. This has obvious potentials in the field of f_0 contours manipulation for perceptual experiments.

The following procedure is used to obtain a re-synthesized audio file: First choose an utterance which will be used as base signal upon which a new f_0 contour will be imposed. Then select a (preferably close) point in the PC space and construct a new f_0 contour using (3). Then apply the inverse of the landmark registration originally applied to the base signal, linearly re-expand it to its original duration, reconvert to unnormalized Hz, generate a set of samples from this last functional representation and finally use a synthesizer like Praat PSOLA [13] to apply the new f_0 contour to the base signal. All these operation are automatic and require only to write a script (available from my website [11]).

3. Conclusions

The case study I have shown contains many elements that are common to other scenarios where FDA is used for speech analysis. The input data consist of (i) f_0 sampled curves obtained using Praat, (ii) phone boundaries obtained using an ASR and (iii) the labels for the three focus conditions. Thus the analysis makes use of information that can be recovered automatically or it is available from the production data collection. The analysis does not require the user to decide what shape traits are important and what not beforehand (e.g. no stylization is required), save for some global consideration that can be applied to curve smoothing (Sec. 2.2), which to my experience has more impact

[3] http://www.esat.kuleuven.be/psi/spraak/demo/Italian/align.php



(a)

(b)

Figure 2: First (a) and second (b) principal component of the variation of f_0. Solid curves: $\mu(t) = N^{-1}\sum_n f_n(t)$; '+' and '-' curves: $\mu(t) \pm \mathrm{sd}(c_k) \cdot \xi_k(t)$, (a) $k = 1$, (b) $k = 2$.

on computational aspects (e.g. the number of bases) than on the outcome of the analysis. FPCA (like PCA) does not make use of labels. The clear separation that is visible in Figure 3 was *not* imposed but emerged from the unlabelled set of curves, thus ruling out any possible subjective bias.

FDA offers the possibility to use the same mathematical framework (and the same code) to analyze more than one speech parameter at the same time, e.g. more than one formant contour, or a joint analysis of f_0, intensity and local speech rate. This allows to capture correlations among different speech parameters across time automatically. An example of this is shown in [10].

We have seen that the FPCA results provide as a by-product a re-synthesis tool (Sec. 2.5). The guidance offered by the FPCA representation allows one to explore a highly reduced set of plausible contours, e.g. by 'moving' close to the borders between clusters in the PC score space (Figure 3) and generating the corresponding contours. This approach was first proposed in [14].

The application of FDA to speech research is recent and largely unexplored. For example, tools other than FPCA are available (e.g. linear models, canonical correlation analysis), which may contribute further in the development of a toolkit for speech analysis. Moreover, even though the presented case study was based on a dataset of modest size, large scale applications of FDA are not difficult to envision. FDA is a way to compare groups of contours, mostly helpful when those contours relate to comparable realizations of a given phenomenon, like f_0 measured on the same syllable, word or sentence spoken in different conditions, like focus in our case. Large annotated corpora can be searched automatically and comparable tokens can be extracted and processed with FDA.

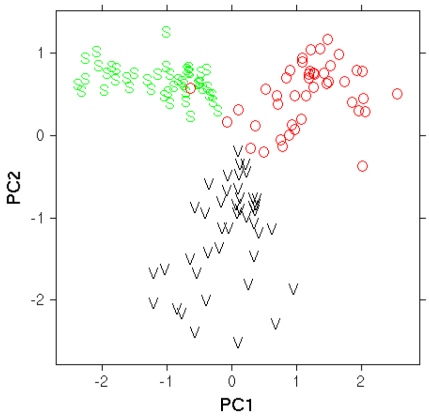


Figure 3: PC scores $c_{k,n}$ (2) of all $N = 132$ f_0 contours, each score is marked with the focus condition S/V/O in the corresponding utterance n.

4. References

[1] J. Pierrehumbert, "The phonology and phonetics of english intonation," *Doctoral dissertation, MIT*, 1980.

[2] M. D'Imperio, "Focus and tonal structure in neapolitan italian," *Speech Communication*, vol. 33, pp. 339–356, 2001.

[3] E. Dombrowski and O. Niebuhr, "Shaping phrase-final rising intonation in german," *Proceedings of Speech Prosody, 11–14 May 2010, Chicago, USA*, 2010.

[4] A.K. Syrdal and J. McGory, "Inter-transcriber reliability of tobi prosodic labeling," *Proceedings of International Conference on Spoken Language Processing, Beijing, China*, pp. 235–238, 2000.

[5] C. Gussenhoven and T. Rietveld, "The behaviour of H and L under variations in pitch range in dutch rising contours," *Language and Speech*, vol. 43, pp. 183–203, 2000.

[6] Keikichi Hirose, Yusuke Furuyama, Shuichi Narusawa, Nobuaki Minematsu, and Hiroya Fujisaki, "Use of linguistic information for automatic extraction of f0 contour generation process model parameters," *Proceedings of EUROSPEECH 2003, Geneva*, pp. 141 – 144, 2003.

[7] J. O. Ramsay and B. W. Silverman, *Functional Data Analysis - 2nd Ed.*, Springer, 2005.

[8] J. O. Ramsay and B. W. Silverman, *Applied Functional Data Analysis - Methods and Case Studies*, Springer, 2002.

[9] J. O. Ramsay, G. Hookers, and S. Graves, *Functional Data Analysis with R and MATLAB*, Springer, 2009.

[10] Michele Gubian, Francesco Cangemi, and Lou Boves, "Joint analysis of speech rate and f_0 with functional data analysis," *submitted to ICASSP 2011*.

[11] Michele Gubian, "Functional data analysis for speech research," *(online) lands.let.ru.nl/FDA*.

[12] S.J. Eady and W.E. Cooper, "Speech intonation and focus location in matched statements and questions," *J. Acoust. Soc. Amer*, vol. 80, pp. 402–416, 1986.

[13] Paul Boersma and David Weenink, "Praat: doing phonetics by computer (version 5.1.20) [computer program]," *online: http://www.praat.org/*, 2009.

[14] Michele Gubian, Francesco Cangemi, and Lou Boves, "Automatic and data driven pitch contour manipulation with functional data analysis," *Proceedings of Speech Prosody, 11–14 May 2010, Chicago, USA*, pp. 100954:1–4, 2010.