JOINT ANALYSIS OF F_0 AND SPEECH RATE WITH FUNCTIONAL DATA ANALYSIS

Michele Gubian, Lou Boves

Centre for Language & Speech Technology
Radboud University
Nijmegen, The Netherlands

Francesco Cangemi

Laboratoire Parole et Langage
University of Provence
Aix-en-Provence, France

ABSTRACT

In this work we propose the use of Functional Data Analysis (FDA) as a powerful methodology to tackle problems where multiple continuous speech parameters have to be analyzed jointly. A production study on contrastive focus placement in Neapolitan Italian is used as illustration. Two features are analyzed, viz. f_0 and relative speech rate, both expressed as continuous functions of time. The results show that known facts about the prosody of Neapolitan Italian emerge from the data, but also other interesting local or cross-feature relationships between contour traits appear. Thus, FDA results can be used as guidance in the exploration of speech feature contour shapes, an operation that used to be carried out manually in previous speech research. The capability of jointly analyzing multiple continuous features provides a valuable improvement not only for speech analysis but also for speech re-synthesis.

Index Terms— Functional Data Analysis, Principal Component Analysis, Speech re-synthesis, Prosody, Neapolitan Italian.

1. INTRODUCTION

It is generally assumed that basic prosodic speech parameters such as f_0 , energy and duration (or speech rate) are to a large degree mutually independent, and that these parameters can be analyzed and manipulated independently. Yet, experience in resynthesizing speech after changing one parameter shows that the result may sound somewhat unnatural, despite the fact that the changed parameter track is perfectly natural in its own right. For example, if local speech rate is very high, and sounds and syllables very short, there simply may not be enough time to fit in the f_0 contour that would theoretically be most appropriate given a certain syntactic and prosodic structure.

Treating prosodic parameters as if they were independent could be justified by the methodological and technical problems that had to be confronted. Thus, it is not surprising that there are several competing models of f_0 contours (e.g. [1, 2, 3]) as well as of duration (e.g. [4, 5]). Building models that account for possible interactions between parameters such as duration and f_0 has been virtually impossible, if only because f_0 and (relative) duration or local speech rate – although both functions of time – were expressed in seemingly incompatible terms: f_0 contours are stylized with continuous lines in time, while durations are expressed at the level of discrete phones. Recently, the development of Functional Data Analysis (FDA) [6] has allowed researchers in different areas to perform quantitative analysis on continuous data directly, i.e. without being forced to describe contours in discrete terms.

In this work we propose the use of FDA as a powerful method to tackle problems where multiple continuous speech parameters have to be analyzed jointly. A case study on contrastive focus placement in Neapolitan Italian is used to show how this can be done in the practice, as well as to let the reader appreciate the quality of the results. Two features will be analyzed, f_0 and relative speech rate, the latter expressed as a continuous function of time as proposed in [7]. The results show that known facts about the prosody of Neapolitan Italian emerged from the data, but also other interesting local or cross-feature relationships between contour traits appeared.

In [8] we have shown that FDA can be used also as a tool for speech re-synthesis. This approach still applies in the multidimensional case. In addition, the benefits of exploiting the dimensionality reduction produced by the FDA data description become more evident when more than one speech parameter has to be manipulated at the same time.

2. FUNCTIONAL DATA ANALYSIS (FDA)

Functional Data Analysis (FDA) is a set of statistical techniques proposed in the late 90's by J. O. Ramsay and his group [6]. FDA extends well known techniques 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, like in ordinary multivariate statistics. An FDA session consists of a sequence of steps broadly divided in two stages. The first stage is data preparation, which includes all the operations necessary to obtain a functional representation of the raw data, where by raw data we mean a set of sampled curves in time. The second stage is the actual FDA. In this work FDA will be carried out using the functional extension of PCA (FPCA).

2.1. Data preparation

Each input sampled curve $x(t_s)$, $s = 1, \dots, S$ has to be represented by a continuous function $x(t)$. This is usually done by first choosing a function basis, such as B-splines [6], and then by selecting the function $x(t)$ in the space spanned by the basis that solves the regularization problem

$$\min\{SSE + \lambda \cdot PEN\}, \quad (1)$$

where SSE is the sum of squared errors of the fitting function $x(t)$ with respect to $x(t_s)$, PEN is a measure of roughness of $x(t)$, usually $\int_0^T (\frac{d^2 x(t)}{dt^2})^2 dt$, and $\lambda > 0$ weights the importance attributed to the roughness.

In general, each of the N functions $x_n(t)$, $n = 1, \dots, N$ has a different duration T_n (for simplicity we assume they all start at $t = 0$). However, all FDA tools require input functions to be defined

The research of Michele Gubian and Francesco Cangemi is supported by the Marie Curie Training Network Sound2Sense (www.sound2sense.eu).

on the same interval, say $[0, T]$. Although time normalization can be effected by linearly mapping each interval $[0, T_n]$ to $[0, T]$, this may not be the best possible solution due to FDA treating all curves $x_n(t)$ as ‘synchronized’ on 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 $x_n(t)$ a time distortion function $h_n(\tau)$ has to be determined that satisfies

$$t_{l,n} = h_n(\tau_l), \quad l = 0, \dots, L+1,$$

where $t_{l,n}$ are the landmarks for curve $x_n(t)$, τ_l their location on the common time axis τ , usually taking $\tau_l = N^{-1} \sum_n t_{l,n}$, then $t_{0,n} = \tau_0 = 0$, $t_{L+1,n} = T_n$ and $\tau_{L+1} = T$. Each function $h_n(t)$ is found by solving a regularization problem similar to (1).

2.2. Functional Principal Component Analysis (FPCA)

In ordinary PCA the input data are N fixed size column vectors x_n , the k -th principal component (PC $_k$) 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 $x_n(t)$ by defining the scalar product $\xi_k \cdot x_n$ as

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

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 $x_n(t)$. 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).

FPCA allows also to perform joint PCA on multidimensional trajectories. For example, extending (2) for trajectories $(x_n(t), y_n(t))$ that take values in features x and y at every time t becomes

$$c_{k,n} = \int_0^T (w_x \xi_k^x(t) x_n(t) + w_y \xi_k^y(t) y_n(t)) dt, \quad (3)$$

where weights w can be set to balance the contribution of each feature to the global variance. Each input trajectory $(x_n(t), y_n(t))$ can be approximatively reconstructed by using the first K PCs as follows:

$$\hat{x}_n(t) = \mu^x(t) + \sum_{k=1}^K c_{k,n} \xi_k^x(t), \quad (4)$$

where the $c_{k,n}$ ’s from (3) are called *PC scores* and $\mu^x(t) = N^{-1} \sum_n x_n(t)$. The reconstructed signal $\hat{y}_n(t)$ is similarly defined. Note that PC scores are scalars that ‘control’ the reconstructed trajectories in both x and y dimensions at once.

3. CASE STUDY

3.1. General description

In this section a production study on contrastive focus placement in Neapolitan Italian is introduced. Starting with [9], 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.

3.2. Material

Five speakers of Neapolitan Italian read three repetitions of three declarative sentences sharing the structure: $[CVCVCV]_S [CVCV]_V [CVCVCV]_O$ (lexical stressed syllable is underlined, S(subject), V(erb), O(bject) specify the syntactic role). All phones are voiced, S and O are proper names, for example *Ralego vede Ladona* (‘Ralego sees Ladona’). Before uttering the target sentences, speakers silently read a contextualization paragraph which induced focus placement on either S, V or O. Since three utterances were discarded due to disfluencies, our data consist of 132 items (5 speakers $\times$ 3 repetitions $\times$ 3 sentences $\times$ 3 focus positions - 3 discarded).

3.3. Data preparation

Fundamental frequency samples $f_n(t_s)$ were computed every 10 ms using Praat autocorrelation-based f_0 extractor with default settings [10]. The average $S^{-1} \sum_s f_n(t_s)$ was subtracted from every sampled curve expressed in semitones, in order to eliminate variation mainly due to speaker identity. Each curve was smoothed using a function $f_n(t)$ selected by solving (1) on a B-splines basis defined on the interval $[0, T_n]$. (The details concerning the choice of the basis and λ in (1) are omitted, see [6], Chap. 5.) Then, a HMM-based ASR trained on standard Italian was used to perform forced alignment on the whole dataset [11]. As a result, the phone boundary positions $t_{l,n}$ for each utterance were obtained. Landmark registration (Sec. 2.1) based on $t_{l,n}$ was carried out obtaining new functions $f_n(t)$ (henceforth we use t for the normalized time axis) all defined on $[0, T]$ ($T = 1.25$ s) and smoothly synchronized on their phone boundaries. Time distortion functions $h_n(t)$, also defined on $[0, T]$, were then used to extract a continuous measure of (log) relative speech rate at t (cf. eq. (1) in [7])

$$r(t) = -\log \frac{dh(t)}{dt}. \quad (5)$$

According to this definition, if a phone is realized two times faster than its reference (average), then $r(t) = \log 2 = 0.7$ in the proximity of that phone, while $r(t) = -0.7$ for a two times slower realization.

3.4. Joint FPCA on f_0 and speech rate

Figures 1 and 2 show the results of FPCA applied to the $N = 132$ trajectories $(f_n(t), r_n(t))$ considered as functional input elements, where $f_n(t)$ are the f_0 contours and $r_n(t)$ are the speech rate curves defined in (5). To understand how to read Figure 1 consider panel (a) first, which represents the variation in f_0 contours captured by PC1. The mean curve $\mu^f(t) = N^{-1} \sum_n f_n(t)$ is plotted in solid line. By virtue of the reconstruction (4), function $\xi_1^f(t)$, i.e. the f dimension of PC1, acts on the mean as a ‘shape corrector’, since it can be added to or subtracted from it in a quantity determined by the PC1 score c_1 . To represent this effect, the ‘+’ and ‘-’ curves show the result when $c_1 = \pm \text{sd}(c_1)$, where $\text{sd}(c_1)$ is the standard deviation of c_1 calculated considering the N PC1 scores $c_{1,n}$ obtained from (3). Panel (b) in Figure 1 shows the speech rate counterpart of panel (a), i.e. the solid curve is the mean $\mu^r(t)$, the ‘+’ and ‘-’ curves are $\mu^r(t) \pm \text{sd}(c_1) \cdot \xi_1^r(t)$. Note that c_1 is the same as in (a) but $\xi_1^r(t)$ is not. The fact that the pair of functions $(\xi_1^f(t), \xi_1^r(t))$ is controlled

by the same coefficient c_1 tells us that the shape variations they represent are related in the original data set. The same mechanism is shown for PC2 in panel (c) and (d) for f_0 and speech rate, respectively, and likewise for PC3 in panels (e) and (f). Together the first three PCs explain 60% of the variance of the data set.

Figure 2 shows the distribution of PC scores (3) for the whole data set relative to the first three PCs. Each score is marked with the focus condition S/V/O (Sec. 3.2) of the corresponding utterance. First note how PC1 and PC2 alone achieve a considerable separation of the three focus categories. PC3 further helps to separate focus condition O from the other two. To relate PC scores to the contour shapes they determine by virtue of (4), let us consider focus condition S. Looking at the S clouds in Figure 2, typical scores are $c_1 < 0$, $c_2 > 0$ and $c_3 > 0$. A negative c_1 modifies the mean f_0 contour by raising the peak in correspondence to the first stressed position and lowering the peak on the third one (Figure 1 panel (a)), and at the same time slows down speech rate in the first 2/3 of the utterance (panel (b)). Similarly, a positive c_2 compresses f_0 movements in the area of the second stressed position (panel (c)) while jointly acting on speech rate, which is slowed down around the first stressed position and speeded up around the second position (panel (d)). A positive c_3 sharpens the right side of the first f_0 peak, levels the final part of the f_0 contour (panel (e)) and at the same time speeds up the first and the last syllable (panel (f)). As a result, for focus condition S the first stressed position has larger f_0 movements (higher peak) and longer duration (slower rate) than the two other positions, which is in line with previous studies on the phonetic marking of narrow focus in Italian [12]. Similar considerations apply to focus conditions V and O.

Fig. 2. PC1, PC2 and PC3 scores (3) obtained applying FPCA on all $N = 132$ f_0 and speech rate contours jointly. Each score is marked with the focus condition S/V/O of the corresponding utterance.

4. DISCUSSION

The analysis described above has been conducted by applying FPCA, a version of PCA adapted for functional input, to a data set consisting of f_0 contours provided with timing information relative to the position of phone boundaries. The use of FPCA, as opposed to e.g. linear discriminant analysis, allowed us to remain as theory-neutral as possible. In fact, even though the labels marking the three focus conditions present in the data were available, FPCA extracted the main contour shape variations in f_0 and speech rate without making use of those labels. Only after the analysis was completed, shape variations were related to the prior information about the three focus categories. The result was that around 60% of the variance in the data set can be related to the effects of the focus condition. A visual inspection allows us to confirm the overall matching of our data-driven analysis with previous linguistic studies.

The advantage of performing FPCA, as opposed to a discrete multivariate analysis, is that starting from a data set of sampled curves we can automatically extract the main relationships between continuous variations in contour shapes and visualize them easily. Clear and localized effects can be found also with traditional methods, for example by measuring f_0 peaks height and position and applying multivariate statistics. However, not only this requires to decide in advance which shape traits (like peaks, slopes, etc.) are potentially relevant in contours and which are not, but also it makes the discovery of more complex relations quite hard. These can be fine local details, like the way the shape of the first hump in Figure 1(a) modulates, or long range correlations, like the way the first and the last hump are related in the same figure. The added value of a joint analysis is that correlations across time and features are automatically retrieved and displayed too. For example, note how the localized variation of f_0 captured by PC2 is linked to a much wider range variation in speech rate (Figure 1(c) and (d)).

As we have shown in [8], FPCA can be used as a re-synthesis tool. Reconstruction (4) allows us to choose any PC score combination and to obtain the corresponding contours. These can be used as input to a speech re-synthesis tool (like PSOLA, available in Praat [10]) that allows to manipulate both f_0 and speech rate of a recorded utterance. Re-synthesis is used in perception studies as a way to locate the boundaries between linguistic categories (like focus conditions) in the continuum of possible modulations of f_0 , speech rate, and other features in time. 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 2) and generating the corresponding contours. On the other hand, the approach found in many studies is based on manual modification of contours, which becomes impractical when multiple time features have to be investigated, if only because the number of combinations of gradual shape variations grows exponentially with the number of features.

5. CONCLUSIONS

In this work we have shown one possible way to benefit of the power of FDA tools to tackle a problem of speech analysis involving multiple speech parameters (features) that evolve in time. By taking a theory-neutral approach we showed that FDA not only can reproduce results obtained using traditional (sometimes ad-hoc) methods, but also offers a way to extract regularities in contour shapes that are not immediate to find, especially when more dimensions are involved.

The application of FDA to speech research is recent and largely unexplored. Recent developments are collected in the web page

Fig. 1. First (a,b), second (c,d) and third (e,f) principal component of the joint variation of f_0 and speech rate. Panels (a,c,e) show the effect on f_0 , (b,d,f) the effect on speech rate (see Sec. 3.4 for details).

maintained by the first author (lands.let.ru.nl/FDA).

6. REFERENCES

- [1] H. Fujisaki, "Dynamic characteristics of voice fundamental frequency in speech and singing," in *The Production of Speech*, P.F. MacNeilage, Ed., pp. 39 – 47. Springer-Verlag New York Heidelberg Berlin, 1983.
- [2] D. R. Ladd, *Intonational Phonology*, Cambridge University Press, 1996.
- [3] J. 't. Hart, R. Collier, and A. Cohen, *A perceptual study of intonation: An experimental approach to speech melody*, Cambridge University Press, Cambridge, Mass., 1990.
- [4] D. Klatt, "Review of text-to-speech conversion for english," *J. Acoust. Soc. Amer.*, vol. 83, pp. 737 – 783, 1987.
- [5] J. van Santen, "Timing," in *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*, R. Sproat, Ed., pp. 115–140. 1998.
- [6] J. O. Ramsay and B. W. Silverman, *Functional Data Analysis - 2nd Ed.*, Springer, 2005.
- [7] S. Ohno, M. Fukumiya, and H. Fujisaki, "Quantitative analysis of the local speech rate and its application to speech synthesis," in *Proc. ICSLP '96*, Philadelphia, PA, Oct. 1996, vol. 4, pp. 2254–2257.
- [8] M. Gubian, F. Cangemi, and L. 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.
- [9] 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.
- [10] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 5.1.20) [computer program]," *online*: <http://www.praat.org/>, 2009.
- [11] F. Cangemi, F. Cutugno, B. Ludusan, D. Seppi, and D. Van Compernelle, "Assi - automatic speech segmentation for italian: tools, models, evaluation, and applications," in *proceedings of 7th AISV Conference*, Lecce, Italy, 2011.
- [12] M. D'Imperio, "Focus and tonal structure in neapolitan italian," *Speech Communication*, vol. 33, pp. 339–356, 2001.