Automatic and Data Driven Pitch Contour Manipulation with Functional Data Analysis

Michele Gubian[1]*, Francesco Cangemi*[2]*, Lou Boves*[1]

[1]Centre for Language & Speech Technology
Radboud University, Nijmegen, The Netherlands
[2]Laboratoire Parole et Langage
University of Provence, Aix-en-Provence, France
[1]{M.Gubian,L.Boves}@let.ru.nl, [2]francesco.cangemi@lpl-aix.fr

Abstract

Creating stimuli for perceptual experiments in intonation research involves manipulation of pitch contours extracted from spoken utterances. Difficulties arise when changes in the contour shape need to be applied globally and smoothly in the whole pitch curve. Moreover, it is hard to relate a gradual modification in some contour trait to its perceptual counterpart. In this paper we propose a novel approach to stimuli manipulation that is based on an extension of Principal Component Analysis (PCA). Starting from a corpus of pitch curves a parametric description of the principal variation in the curve set is obtained. This allows to locate clusters in this parameter space that are related to linguistic categories. The search for pitch curves with desired perceptual characteristics is carried out by choosing convenient point locations with respect to the relevant clusters. We illustrate this approach in a case study on question/statement opposition in Neapolitan Italian.

Index Terms: Functional Data Analysis, Principal Component Analysis, pitch resynthesis

1. Introduction

Most intonation research involves some combination of stylization of pitch tracks measured in a corpus of spoken utterances, and perceptual experiments in which subjects judge resynthesized versions of the utterances with the manipulated pitch contours [1, 2, 3, 4]. The experimental pitch contours can be produced by some phonological or physiological model, or the contours are created manually. Both approaches have drawbacks, especially in research aimed at establishing the boundaries between putative discrete categories. In that type of research it is usually assumed that there exists some (possibly multidimensional) continuum of physical stimuli and that this continuum is broken up on the basis of phonological and linguistic knowledge of the subjects in an experiment. To conduct such experiments one must be able to create a set of stimuli that sample the continuum at equi-distance points. However, assumptions and simplifications (e.g. stylization) characterizing the model 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. Creating stimuli by hand runs into the same kind of problems.

In this paper we introduce a novel approach to producing a set of stimuli for intonation research that is based on principled processing of the pitch tracks of the original utterances. By virtue of the application of advanced statistical techniques it is now possible to cluster the pitch contours of the example utterances in a way that guards against unwarranted interpretations. In addition, it is now possible to generate a stimulus continuum by adding or subtracting the characteristics of a specific cluster to an 'average' pitch contour.

The approach proposed in this work is based on Principal Component Analysis (PCA). PCA is a way to extract the main variation dimensions, or Principal Components (PCs) underlying a set of samples, where the samples are vectors of real numbers. A recently proposed extension of PCA allows to extract the PCs from a set of *continuous functions*. This means that given a set of curves, like F_0 contours, one can obtain a compact description of the principal dimensions of the variation in the whole curve set. Such a description is compact in the sense that global and smooth changes in the curve shapes can be expressed by a few numerical parameters. By searching for relations between the parameter values and linguistic categories, one can infer the phenomena underlying the variability in the curve shapes. This approach was already shown in [5] for energy curves and goes under the umbrella term of Functional Data Analysis (FDA) [6].

Here we present a way to exploit the parametric description of the variations in pitch contours in order to generate new contours with desired characteristics, which in turn can be used for stimuli re-synthesis. The idea is based on reversing the process that brings from the acoustic signal to its pitch contour and then to its PCA parameters (henceforth *PC scores*). A linguistic category like the question/statement (Q/S) opposition can be realized in some language only (or mainly) by changing the pitch contour. Running a PCA on a set of the F_0 curves yields a picture like Fig. 3 (a), where each curve is represented by scores on two PCs (dimensions), where the Q/S separation is visible. The mathematical framework of PCA allows to reconstruct a curve starting form its PC scores. Hence one can explore the PC space and produce contours that were not present in the original data set, but that are located in interesting areas of the PC space, e.g. somewhere between the question and the statement cloud, where perceptually ambiguous stimuli are likely to be obtained. Then, by making use of an off-the-shelf F_0 synthesizer, like PSOLA available in Praat [7], one can impose the new F_0 contour on an existing waveform and obtain the desired stimulus.

This approach has several advantages. First, the parametric representation (PC scores) of the real F_0 contours is statistically grounded. Second, the search for stimuli with specified characteristics, like ambiguity, is guided by the PC representation that shows where the real samples are in the PC space. Third, the

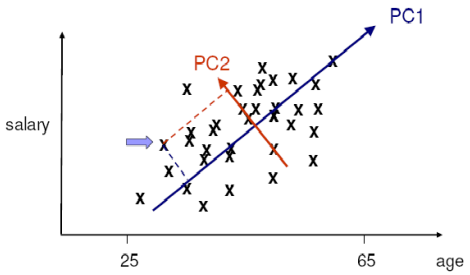


Figure 1: An example for classic PCA. A fictitious datasets collects people age and salary. The first two PC are shown as well as the projection of the point indicated by the arrow on the PC axes.

contour variations that can be obtained by moving in the PC space are smooth and global, i.e. they impact on the whole curve. Fourth, all the steps from a set of PC scores to a stimulus waveform are automatic.

The rest of the paper is organized as follows. Section 2 gives a global overview of the concepts underlying FDA and functional PCA (fPCA) in particular. In Section 3 fPCA and stimuli re-synthesis is applied to a case study based on Q/S opposition in Neapolitan Italian. Conclusions are in Section 4.

2. Functional Data Analysis

2.1. Doing statistics on curves

In empirical science data is often collected in the form of sampled functions, usually time series. The process of making inference out of these data sets involves questions like "Is the trend of temperature throughout the year in town A different from that in town B?". To answer this kind of questions usually global statistical indexes are first extracted from the time series (mean, variance, peaks, etc.) and then multivariate statistical techniques are applied on those indexes (ANOVA, linear regression, etc.). However, sometimes patterns are not easily revealed by simple (scalar) statistical indexes, because they reside in the dynamics of the signal in time. Functional Data Analysis (FDA) [6] is a suite of computational techniques that extend statistical techniques developed for scalar data to the function domain, offering the possibility to make inferences from sets of whole stretches of signals, without the need to reduce the dynamic variation to static statistical indexes, a process that may destroy essential information, and in practice makes inference on dynamic traits of signals problematic.

2.2. Functional Principal Component Analysis (fPCA)

Classic 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. Fig. 1 shows an example of PCA applied to a fictitious set of data about age and salary of people. The new set of coordinates (PC1, PC2) captures the "natural" modes of variation, since the data is mainly varying along the PC1 dimension, which captures the correlation between age and salary, while PC2 can be interpreted as relative wealth irrespective of age. Every point in the data set can now be re-expressed in terms of the PC coordinates. For example the point indicated by the arrow gets a negative score on the PC1 axis and a positive score on the PC2 axis. Note that PCA finds orthogonal coordinates such that the variations across the data set can be described by independent scores. In our example it means that while an older age is more likely associated with a higher salary (positive correlation), PC1 values do not bring any expectation on PC2 (no correlation).

While in classic PCA principal components are vectors of the same dimension as the data vectors, in fPCA principal components become functions defined on the same time interval as the original data set. Even though it is not possible to visualize orthogonal coordinate systems, the same concepts exposed in the example above still apply. Fig. 2 shows a way to represent functional PCs. Like in ordinary PCA, the percentage of explained variance is specified for every PC (top of every panel). Then, for each component the solid line shows the average signal, i.e. the function obtained by averaging each point in time across the whole data set, while the '+' and the '-' curves represent the effect of adding/subtracting a positive multiple of one of the PC functions to the average curve (the latter is the same in all panels). To make a parallel with Fig. 1, the average curve corresponds to the origin of the (PC1,PC2) coordinates, while the '+' curve in the PC1 panel corresponds to a point lying somewhere on the positive side of the PC1 axis, i.e. scoring zero on PC2.

3. Case study

3.1. General description

In Neapolitan Italian, as in other roman languages, the Question/Statement (Q/S) opposition can be expressed by intonational means alone. The same syntactic structure, with the same lexical content and displaying the same sequence of segments, can be uttered with two different intonational contours which lead to two different pragmatic meanings. A great body of research in the Autosegmental-Metrical framework for the study of the phonology of intonation has shown that contours can be successfully modeled as a sequence of discrete, local events. Perception experiments in this framework are usually based on the manipulation of individual events in F_0 contours. However, recent studies explore the hypothesis that dynamic proprieties of F_0 contours (e.g. global shapes) can be perceptually relevant too. In this perspective, re-synthesis should not be performed by manipulating discrete events in a pitch contour, but rather by manipulating longer stretches of the signal [8, 9].

3.2. Material

Two male native speakers of Neapolitan Italian ('AS' and 'SC') were recorded in a sound proof booth by means of a Roland Edirol UA 25EX sound card connected to a laptop and through a Sennheiser E 835 microphone. Three carrier sentences were used having the same syllable count and lexical stress positions and containing two accents also in the same relative positions ('Milẹna lo vuole amạro (?)' = *Milena drinks it (i.e. her coffee) black*, 'Valẹria viene alle nọve (?)' = *Valeria arrives at 9*, 'Amẹlia dorme da nọnna (?)' = *Amelia sleeps at grandma's*). Each sentence was pronounced five times in Q and five time in S mode by each speaker. Three out of the $2 \times 3 \times 5 \times 2 = 60$ utterances were discarded, thus leaving 57 utterances. After removing leading and trailing silence, the beginning and the end of the two accented vowels were manually marked by the sec-

ond author of this work. The duration of each sentence is around 1 second irrespective of the speaker/sentence. F_0 was extracted from each utterance using the Praat autocorrelation-based F_0 extractor with default parameter settings, which results in the computation of F_0 values every 10 ms.

3.3. Data preparation

In order to eliminate uninteresting but large variation in the signal due to speaker identity, F_0 was first converted into semitones, then the time average was subtracted from each F_0 curve. The first FDA step is to interpolate each sampled curve to obtain a function $y(t)$.This operation is essential because FDA tools work on functional representations, but also it allows the user to smooth the raw data.Although automatic procedures exist for determining the optimal degree of smoothing, like generalized cross-validation ([6], Par. 5.4.3), we decided to follow a more empirical guideline by smoothing up to a point where average and peak velocity in F_0 curves were within the experimentally observed limits found in [10].

Any FDA data processing requires that each curve (function) should have the same duration. The alignment we performed was not just a linear time normalization, but it made use of the so-called landmark registration ([6], Chap. 7). It is a nonlinear time warping procedure that warps the time axis of each curve in such a way to align all previously marked relevant points in time, called *landmarks*. In our case, we had four landmarks available from the manual annotation, i.e. onset and offset of each of the two accented vowels in the sentence, which were made to coincide in (normalized) time as much as possible, while keeping the curve distortion reasonable. The reason to do that is that FDA statistics are based on the hypothesis that a time instant means the same thing across all curves.

3.4. Functional PCA

Functional PCA was applied on the data prepared in the way described above. Fig. 2 shows that PC1 basically expresses the tendency of being ahead or behind in the region of the first accented vowel, and at the same time shows that a late peak in the first accented vowel region is associated with a high peak in the second accented vowel region. PC2 expresses mostly a difference in excursion. Note also that in all cases PCs stick to the mean in the middle region. This does not mean that there are no variations in that region, but that variations are not systematic enough to be captured. Fig. 3 shows the distributions of PC scores with respect to Q/S mode and speaker identity. Both speakers seem to use the two dimensions (factors) expressed by PC1 and PC2 in a different yet consistent way. PC1 coefficients are higher for questions than for statements for both speakers, but for speaker SC they are generally higher than for speaker AS (Fig. 3 (b)). PC1 coefficients express how similar a curve is to the '+' or '-' prototypes in the first panel of Fig. 2. Positive values correspond to curves whose peak corresponding to the first accented vowel is realized later, vice versa for negative values. This means that both speakers tend to realize questions with a later first peak than what they do for statements, but those shifts are relative to a characteristic of the speaker, since speaker SC tends to realize those peaks generally later than speaker AS. A similar story goes for PC2 (Fig. 3(b)). Questions consistently present a larger pitch excursion than statements, but this time AS is using generally larger excursion than SC does.

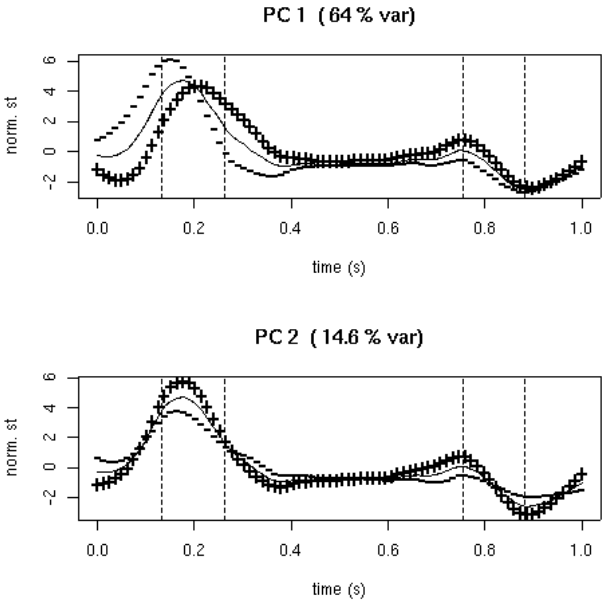


Figure 2: Functional PCA applied to the 57 landmark-registered pitch contours dataset. Solid line shows the average signal $\mu(t)$, while the '+' and the '-' curves represent the effect of adding/subtracting a multiple of the principal component function to $\mu(t)$. Vertical dashed lines indicate the landmarks..

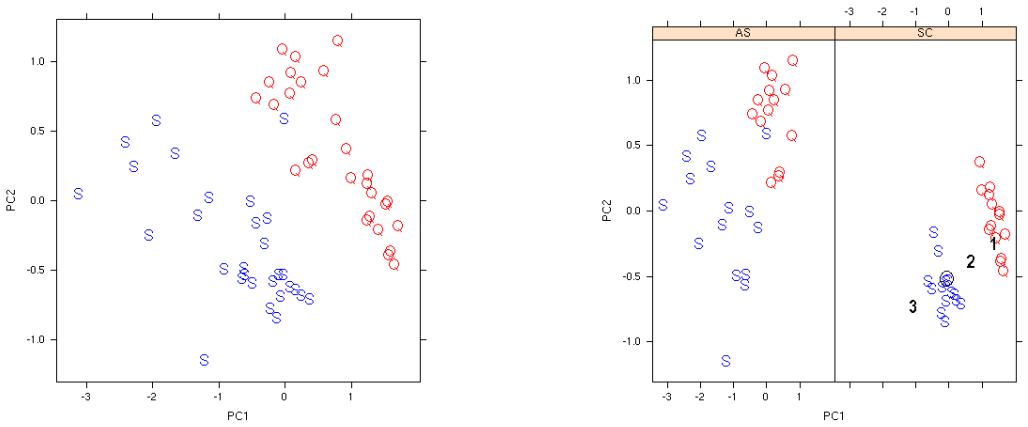
3.5. Creating new stimuli by moving in the PC space

(f)PCA can be considered as a dimensionality reduction or data compression technique. In our case, once we store the mean function $\mu(t)$ (the solid curve in Fig. 2) and the two PC functions $PC_{1,2}(t)$ (a positive multiple of which added to/subtracted from $\mu(t)$ gives the +/- curves in Fig. 2) we can reconstruct all of the original 57 F_0 curves $y(t)$ by only storing two PC scores s_1, s_2 for each of them. A reconstructed curve has the form:

$$y(t) \approx \mu(t) + \sum_{k=1}^{K} s_k PC_k(t) \qquad (1)$$

where $K = 2$ is the number of PCs you decide to use. This is an approximation since PC1 and PC2 together explain around 80% of the variance in our data set. To go beyond reconstruction of existing curves, one can observe that in eq. (1) any arbitrary set of s_k's could be used to obtain a new curve. To make this operation meaningful, one can look at the distribution of PC scores in Fig. 3 and decide to explore areas of potential interest. Since a small change in any PC score will result in a small and smooth change in the reconstructed curve, we expect that in the region between the Q and the S clusters we would find pitch contours with mixed characteristics, which are likely to be perceived as ambiguous by native speakers.

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 pitch contour will be imposed. Then select a (preferably close) point in the PC space and construct a pitch contour using eq. (1). 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



(a) Q/S modality

(b) Q/S modality and speaker

Figure 3: Scatter plots showing the distributions of PC scores with respect to Q/S mode and speaker identity.

Hz, generate a set of samples from this last functional representation and finally use a synthesizer like Praat PSOLA [7] to apply the new F_0 contour to the base signal. All these operation are automatic and require only some scripting.[1]

3.6. Quality of the manipulated stimuli

The second author, a native speaker of Neapolitan Italian, informally evaluated the perceptual impact of stimuli which were resynthesized using points in the (PC1,PC2) space. Starting from a base stimulus (circled in Fig. 3(b)), three kind of moves in the PC space were tested, viz. towards a target point (i) corresponding to an existing stimulus of the opposite category (Q vs. S), (ii) halfway the Q and S clusters and (iii) in the opposite direction with respect to the middle area between clusters (see Fig. 3(b), points 1, 2 and 3, respectively). The procedure was repeated for many base stimuli of each category and speaker. Stimuli re-mapped on an opposite category (i) generally sounded natural. Stimuli in the Q/S middle area (ii) do sound ambiguous. Subtle differences in the degree of ambiguity seem to be linked to speaker identity and the direction of the manipulation (Q towards S or vice versa). This leads us to hypothesize that along the continuum which joins two stimuli belonging to two different categories various points can be perceptually relevant, i.e. ambiguity is not necessarily restricted to the midpoint. Finally, going in the opposite direction (iii) we obtained overtly non ambiguous stimuli which sounded somehow emphatic, but still natural-sounding.

4. Conclusions

This work has shown that it is possible to adopt a data driven approach in the manipulation of stimuli for intonation research. The mathematical framework of functional PCA allows not only to recover a "geographic map" of the linguistic categories present in a data set of spoken utterances, but also to move inside this map in search for potential confusion areas where those categories get mixed up. Any point on the map generates automatically a pitch contour that can be used to re-synthesize a stimulus. The informal analysis of results presented above has shown that the fPCA maps geometry is indeed a useful guidance to retrieve perceptually interesting stimuli, whose quality was judged overall satisfactory.

[1] All data and scripts necessary to produce the results shown in this work are available at the url: http://lands.let.ru.nl/FDA.

5. Acknowledgements

The research of Michele Gubian and Francesco Cangemi is supported by the Marie Curie Research Training Network Sound-to-Sense (http://www.sound2sense.eu).

6. References

[1] S. Schötz, G. Bruce, and B. Granström, "Modelling compound intonation in dala and gotland swedish," *Proceedings FONETIK 2009*, pp. 48–53, 2009, ISBN 978-91-633-4893-8.

[2] J. 't. Hart, R. Collier, and A. Cohen, *A perceptual study of intonation: An experimental approach to speech melody*. Cambridge, Mass.: Cambridge University Press, 1990.

[3] 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.

[4] K. Hirose, Y. Furuyama, S. Narusawa, N. Minematsu, and H. Fujisaki, "Use of linguistic information for automatic extraction of f0 contour generation process model parameters," *Proceedings of EUROSPEECH 2003, Geneva*, pp. 141 – 144, 2003.

[5] M. Gubian, F. Torreira, H. Strik, and L. Boves, "Functional data analysis as a tool for analyzing pronunciation variation - a case study on the french word c' était," *Proceedings of INTERSPEECH 2009, 6–10 September 2009, Brighton, UK*, pp. 2199–2202.

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

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

[8] M. D'Imperio and D. House, "Perception of questions and statements in neapolitan italian," *in Kokkinakis, G., Fakotakis, N. & Dermatas, E. (eds.) Proceedings of Eurospeech '97*, vol. 1, pp. 251 – 254, 1997.

[9] M. D'Imperio and F. Cangemi, "The interplay between tonal alignment and rise shape in the perception of two neapolitan rising accents," *selected papers from PaPI conference (Las Palmas, June 2009)*, (to appear).

[10] Y. Xu and X. Sun, "Maximum speed of pitch change and how it may relate to speech," *J. Acoust. Soc. Am.*, vol. 111, no. 3, pp. 1399–1413, March 2002.