Functional Data Analysis as a Tool for Analyzing Speech Dynamics A Case Study on the French Word *c'était*

Michele Gubian[1], *Francisco Torreira*[1,2], *Helmer Strik*[1], *Lou Boves*[1]

[1]Centre for Language & Speech Technology, Radboud University, Nijmegen, NL
[2]Max Planck Institute for Psycholinguistics, Nijmegen, NL
{M.Gubian@let.ru.nl, Francisco.Torreira@mpi.nl, w.strik@let.ru.nl, L.Boves@let.ru.nl}

Abstract

In this paper we introduce Functional Data Analysis (FDA) as a tool for analyzing dynamic transitions in speech signals. FDA makes it possible to perform statistical analyses of sets of mathematical *functions* in the same way as classical multivariate analysis treats scalar measurement data. We illustrate the use of FDA with a reduction phenomenon affecting the French word *c'était* /setɛ/ 'it was', which can be reduced to [stɛ] in conversational speech. FDA reveals that the dynamics of the transition from [s] to [t] in fully reduced cases may still be different from the dynamics of [s] - [t] transitions in underlying /st/ clusters such as in the word *stage*.

Index Terms: Functional Data Analysis, Principal Component Analysis, Categorical *vs.* gradual phenomena.

1. Introduction

It is well known that most of the information in speech signals is encoded in dynamic changes in formants, pitch, and power. Dynamic changes are best described in the form of some mathematical function, such as a piece of a second or third order polynomial. Nevertheless, more often than not measurements obtained in phonetic experiments are formulated in terms of absence or presence of some phenomenon, or of initial and final values of some dynamically changing parameter, such as pitch or formant frequencies. In other words: experimental data tend to be given in terms of nominal, ordinal, interval or ratio data, none of which are mathematical functions. The reason for this discrepancy is evident: we had increasingly more powerful methods for the statistical analysis of scalar data (or of sets of scalars in a vector), but there were no statistical methods that could operate on *functions*, instead of on numbers or labels.

Perhaps it is fair to say that phonetic research in a tradition where speech is represented as a sequence of discrete units (phonemes or feature vectors) was not hampered too much by the lack of statistical methods that can deal with functions. After all, the properties of discrete units can be summarized in scalar measurements, at least to a large degree. However, we are witnessing a growing interest in what has become known as 'fine phonetic detail', which is all about dynamic phenomena [2]. If dynamics is key, the need to map function-valued measurement data onto scalars or vectors to make them amenable to statistical analysis, becomes debilitating. After all, any mapping from functions to scalars is bound to destroy or distort essential information.

Fortunately, speech research is not the only field where dynamic processes play a central role. Therefore, it need not come as a surprise that there is a growing number of statistical techniques specifically developed for dealing with function-valued measurement, developed in fields like astronomy, but that might be applied to advantage in phonetic research. One such class of methods are known as *Functional Data Analysis* (FDA).

The use of FDA in phonetics is not completely new. It has been used to time align pitch-periods (for calculating harmonics to noise ratio) kinematic [1] and aerodynamic data [3] with an accuracy that is substantially better than conventional Dynamic Time Warping. In this paper we show that FDA can also be used for analyzing speech data that can only be described in terms of differences in dynamic transitions and where precise time alignment is difficult at best because the dynamic processes that we want to investigate may well represent different underlying phenomena.

The rest of this paper is structured as follows: in section 2 we briefly introduce the FDA technique used in our research. In section 3 we explain the phenomenon under study (the reduction of the vowel /e/ in French *c'était* /setɛ/ 'it was') in detail. Also, we explain the speech material and the measurements that we took to represent the dynamics of the transition from the initial [s] to the medial [t] of /setɛ/. Then we present the major results (section 4 and we finish with discussion and conclusions (section 5).

2. Functional Data Analysis

Functional Data Analysis [7, 8] is a suite of computational techniques that extend classic methods form statistics so that they can operate on functions instead of on scalars. Thus, they allow one to make quantitative inferences from sets of whole continuous functions (signals) without the need for an intermediate step in which functions are converted into scalars, a process that always causes information loss, and that makes inference from dynamic traits of signals problematic.

Analyzing sets of dynamic (function-valued) observations with FDA takes two steps. It must be emphasized that although FDA is applied to digital signals, this does not imply that functions are converted to scalars. The first step is data preparation, which consists in transforming the sampled signals into a functional form, usually employing basis functions like B-splines and standard least squares interpolation, often including a regularization term. In this process, all functions are normalized on the same time interval, to make them comparable across time. In cases when a set of landmarks can be reliably identified in all functions (e.g. a series of peaks with a clear physical interpretation) these landmarks can be used to produce a time registered version of the whole set of functions, making all corresponding landmarks coincide in (normalized) time (see e.g. [8], Chap. 7). The second part is data analysis. Many techniques from multivariate statistics have been extended to functions, including

functional Principal Component Analysis (fPCA) and different versions of functional linear modeling (c.f. [7] for a comprehensive overview).

In this study we will use fPCA. Classic PCA is a way to extract and display the main modes of variation of a set of multidimensional data [7]. Starting from a data set in its original set of coordinates, a new basis is found such that by expressing (projecting) the data points on this basis, the projection on the first dimension accounts for the largest part of the variance in the data set, the second for the next largest part, and so on. While in 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 functional data set. Fig.1 shows a typical way to display principal components in fPCA. The solid line shows the average signal, i.e. each point is the average of all the functions in the data set at that (normalized) time, while the '+' and '-' curves represent the effect of adding to or subtracting from it a multiple of the first principal component function (fPC1). Data points (i.e. functions) that get a high positive score when projected on fPC1 will then tend to look like the '+' curve, and vice versa for negative scores.

3. Experiments

3.1. Goals of the experiments

We demonstrate the application of fPCA with a study of vowel reduction affecting the French word *c'était* /setɛ/ 'it was'. In conversational speech the vowel /e/ can be reduced, even to the extent that it seems to be completely absent. The phonetic question that we want to address is whether vowel /e/ is gradiently reduced or categorically absent in the reduced pronunciations of *c'était*. To this aim, we investigated tokens of *c'était* with a vowel between /s/ and /t/, tokens where no voicing was present and tokens of underlying /st / clusters extracted from other words such as 'stage'.

We want to show that by taking scores of individual tokens on the first fPC as as descriptors of dynamic behavior, standard statistical techniques like ANOVA and k-means clustering can be used to quantitatively assess the nature of the reduction phenomenon. The difference between use of FDA and classic statistics is that the fPC1 score represents the actual dynamics of the signals, contrary to e.g. mean or variance, which are the result of time averaging. As a corollary, we will show that FDA can be successfully applied to sets of qualitatively heterogeneous signals (with or without a prominent maximum in the middle), which has not yet been described in detail in the literature [8].

If the dynamics of the [s] to [t] transition in *c'était* tokens with a completely deleted vowel display the same dynamics as tokens from words with underlying /st/, we will have found support for the hypothesis that the reduction process is categorical. If, on the other hand, the fully reduced tokens of *c'était* can still be distinguished from underlying /st/ clusters (because their dynamics are different), then the result seems to support the hypothesis that the reduction process is gradual.

In our experiments we will first investigate whether FDA can distinguish between between underlying /st/ clusters and *c'était* tokens that were annotated as fully reduced. Next, we will investigate if FDA can uncover differences between the dynamics of different forms of *c'était* and underlying /st/ clusters.

3.2. Materials

The materials used in this study were extracted from the Nijmegen Corpus of Casual French (NCCFr), which contains 35 hours of high-quality audio featuring casual conversations among French university students. A detailed description of the preparation, recording and contents of the NCCFr corpus can be found in [6]. The data set consists of 378 *c'était* pronunciations and 81 tokens of words starting with a /st/ cluster (e.g. *stage* 'internship'). In each token, we decided to take the beginning of [s] and the release of the [t] closure as the start and the end of the signal. Those events were manually marked by the second author by inspecting waveforms and spectrograms according to standard segmentation criteria. It should be noted that a considerable number of tokens exhibiting an incomplete [t] closure ($n = 100$) were discarded, since we preferred to have as clearly defined landmarks as possible. The presence of voicing between [s] and [t] was determined manually on the basis of voicing-like periodicity in the waveform. From the subset of *c'était* tokens, 191 contained voicing between [s] and [t], while 187 did not.

3.3. Feature Extraction

For our experiment we decided to characterize the dynamics of the [s] to [t] transition with a single signal feature that could be extracted in the exact same manner from all tokens in the set, viz. the log-energy contour of a low-pass filtered version of the acoustic signal, henceforth called *lowE*. First a low-pass filter with cut-off frequency 3250Hz is applied, then a 20 ms window is moved through the output signal at 5 ms steps and the average log-energy is calculated. We subtracted from each *lowE* sequence its average value, since a global difference in log-energy across tokens reflect random effects such as distance to the microphone and overall speaker volume.

We believe that *lowE* is a good index of the dynamics of the [st] transition because the total speech power is able to reveal opening and closing movements of the tongue related to the articulation of the [e] vowel. If the speaker made a gesture related to the production of a vowel, the constriction for the [s] should become less narrow, and consequently one would expect the power in the frication noise to drop. An intervening full vowel would cause a rise in the acoustic power after the release of the [s]. Even if the release of the [s] constriction does not result in a (voiced of voiceless) vowel, one would still expect a plateau or a somewhat gradual decrease of the acoustic power into the [t] closure. However, in the case of underlying /st/ clusters one would expect a fairly rapid and monotonic decrease of the acoustic power from the [s] into the [t] closure.

3.4. Data Preparation

All data processing from this point on was carried out using the `fda` library for the R software [4] available at [5]. In order to perform FDA all sampled contours have to be transformed into functions defined on the same time interval. Since each audio segment has a different duration, we proceeded as follow. Each sampled feature contour was first interpolated using a 4^{th} order B-spline basis with one knot per sample and a 2^{nd} order roughness penalty. The smoothing parameter λ was empirically set to 10. Then each function was re-sampled on 31 equally spaced points, and the obtained sampled contours were once again interpolated, this time using a 6^{th} order B-spline basis with one knot per sample and a 4^{th} order roughness penalty. The latter choice forces continuity up to the 2^{nd} order derivative, thus allowing a greater deal of smoothness. The smoothing parameter

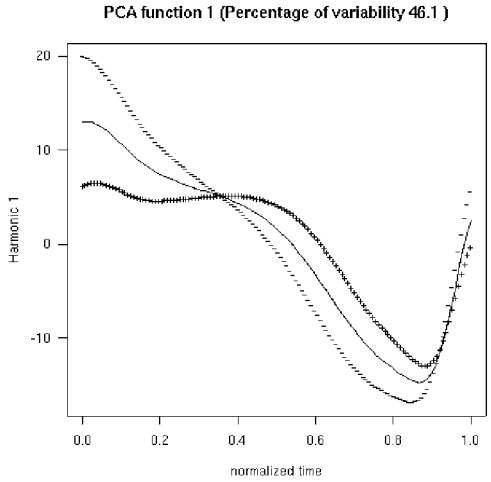

Figure 1: *First principal component (fPC1) of lowE contours of the fully reduced c'était token set and the underlying [st] clusters. Average signal (solid line) and +/- 2 standard deviations of fPC1 ('+' and '-' curves).*

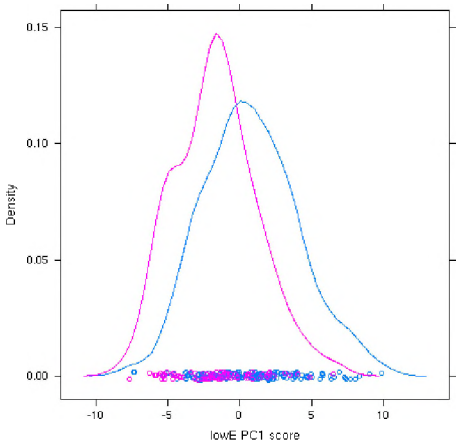

Figure 2: *Empirical densities for the fPC1 scores of lowE contours of the fully reduced c'était token set and the underlying [st] tokens.*

λ was empirically set to 10^{-12} (this value is very different from the previous one mainly because of a different representation of time, ms in the former case, normalized in [0,1] in the latter). We did not attempt to register data with landmarks, since on such short trajectories we were not able to identify reliable landmarks. A few audio tokens were discarded because of problems in signal processing. In the end we worked with 369 *c'était* tokens, of which 186 were annotated as containing voicing between [s] and [t], and 80 tokens of underlying /st/ clusters.

4. Results

4.1. Voiceless *c'était* vs. underlying /st/ clusters

We first applied fPCA to all voiceless tokens, which include the subset of 183 voiceless *c'était* realizations plus the set of 80 realizations of underlying /st/ clusters. The aim was to investigate traces of any difference in dynamics between those two sets that could distinguish [st] tokens resulting from vowel deletion from underlying /st/ tokens.

Using *lowE* as an indication of the dynamics of the [s] to [t] transition, fPCA suggests that fully reduced *c'était* and underlying /st/ clusters as in the word *stage* 'internship' are similar, but certainly not identical. Fig. 1 and 2 show fPC1, which explains 46% of variance, and the empirical densities [9] of the fPC1 scores of the tokens in the two subsets.

The solid line in Fig. 1 shows the average signal, i.e. each point is the average of all the 263 *lowE* contours at that (normalized) time, while the '-' curve (with rise-fall portion) and the '+' curve (without) represent the effect of ± 2 standard deviations of fPC1 on the average curve (like in classic PCA, signs have no intrinsic meaning). Qualitatively, we expect tokens with an associated negative fPC1 coefficient ('-' curve) to belong to the *c'était* subset, since those curves will tend to have a slower decrease in acoustic power between [s] and [t] than the underlying /st/ tokens. In accordance with the limited proportion of variance explained by fPC1, the distance between the '+' curve and the '-' curve is not very large. The '+' curve shows a sort of plateau that could be attributed to a gesture related to the articulation of an intervening [e]. It should be noted, however, that manual spectrographic inspection of the materials suggested that the energy in the plateau area might be attributable to [s] frication rather than to genuine formants.

A t-test on the two empirical distributions (cf. Fig. 2) yielded a statistically significant difference ($p < .0001$); thus, the two subsets must be considered as originating from different populations. However, by applying k-means with two clusters we could separate *c'était* tokens from underlying /st/ clusters only with 59% accuracy.

The results of the fPCA analysis suggest that fPC1 clearly cannot separate the populations effectively; yet the difference between the two distributions leaves open the possibility that the *c'était* tokens (several of which have fPC1 scores that are never reached by /st/ tokens) are from a population that is characterized by and underlying [e] in between [s] and [t].

4.2. Three-way analysis

We then applied fPCA analysis to the complete set of 449 tokens. The results are summarized in Figs. 3 and 4. fPC1 explains 63.4% of the total variance. The increase in the proportion of explained variance (compared to the analysis of the two sets of voiceless tokens) is not surprising. After all, the subset that contains clear traces of a vowel can be considered as structurally different from the completely voiceless tokens.

An ANOVA on the complete set of fPC1 scores, with three groups, i.e. *c'était* with a trace of a vowel, *c'était* without a trace of a vowel and underlying /st/. We obtained $F(2, 442) = 595.91, p < .0001$. A Tukey HSD post-hoc test revealed that the means of all three groups were statistically different from each other ($p < .0001$ in all cases). So, we see again that there is support for the hypothesis that we are dealing with three different underlying populations.

k-means clustering with two clusters yielded 94.5% agreement with the set annotated as containing voicing between [s] and [t] on the one hand and the union of the sets annotated as fully reduced *c'était* and underlying /st/ clusters. However, clustering with $k = 3$ did not yield meaningful results. This corroborates the previous finding that the two sets of voiceless tokens cannot easily be separated.

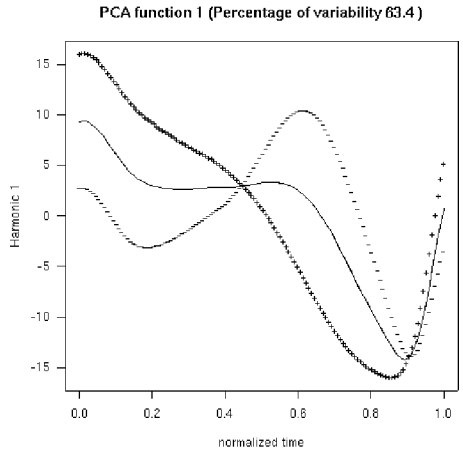

Figure 3: *First principal component (PC1) of lowE contours of the full token set (comprising three classes). Average signal (solid line) and +/- 2 standard deviations of fPC1 ('+' and '-' curves).*

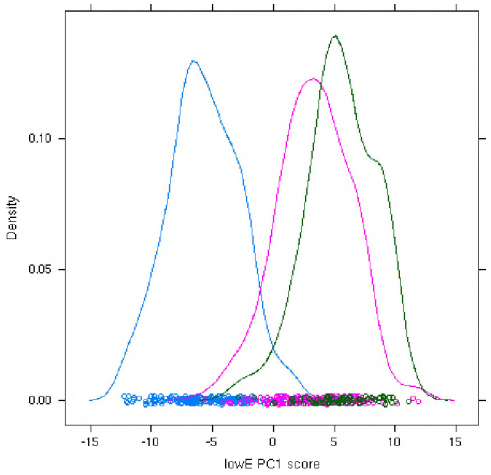

Figure 4: *Empirical densities for the fPC1 scores of lowE contours of the three token sets, separately computed for the two classes of c'était tokens and the /st/ clusters.*

5. Discussion and Conclusion

To the best of our knowledge this paper presents the first application of functional Principle Components Analysis for investigating the presence or absence of differences in the dynamics of the [s] to [t] transition between underlying /st/ clusters and [st] tokens that result from the deletion of an intervening vowel /e/ in French *c'était*. In the past, Functional Data Analysis has only been applied in phonetic research to obtain very accurate time alignments (cf. [1][3]). To represent the dynamics of speech phenomena that may or may not contain a vowel between [s] and [t] we used the contours of the speech power as function-valued observations.

Our results strongly suggest that fPCA can indeed be used for conducting statistical analyses on sets of data that are *functions* rather than scalars) or vectors. Both the fPC1 contours (relative to the average contours) and the empirical distributions of the fPC1 scores for the tokens in the three sub-sets indicate that there are three underlying populations, rather than two (the sub-set with vowel present and the union of the two fully voiceless sub-sets). Still, despite the fact that the difference between the distributions of the two voiceless sub-sets is statistically significant, the overlap is very substantial.

The long-term goal of our research is to investigate the contribution of fine phonetic detail to speech comprehension and its role in speech production. FDA (and fPCA in particular) has proved to be a powerful tool for analyzing speech dynamics. In ongoing research we are investigating whether FDA can also be applied to speech features that involve non-linear processing, so as for example the extraction of formants. It is interesting to see whether the heuristic strategies needed to decide whether some spectral peak is a formant or not interfere with FDA analysis.

The phonetic question that was at the basis of this study, viz. whether the reduction in French *c'était* is a gradual or rather a categorical process cannot be answered conclusively. While we have found statistically significant differences between the distributions of underlying /st/ clusters and [st] clusters that result from vowel deletion, there are several potentially relevant factors that were not controlled for in this corpus-based study. A study is under way to investigate the effects of prosody on the dynamics of [s] to [t] transitions. After all, French *c'était* tends to also differ from words with underlying /st/ clusters in that it is rarely stressed, and in that it is very often phrase initial.

In conclusion, we can say that FDA is an extremely promising tool in the study of fine phonetic detail. At the same time, the interaction between fine phonetic detail and other phonetic variables, notably prosody, is so strong that novel experimental designs may need to be developed to come to grips with the intricacies of speech dynamics.

6. Acknowledgements

The research of Michele Gubian is supported by the Marie Curie Research Training Network Sound-to-Sense [1].

7. References

[1] Byrd, D., Lee, S. and Campos-Astorkiza, R. (2008) Phrase boundary effects on the temporal kinematics of sequential tongue tip consonants. *J. Acoust. Soc. Am.*, Vol. 123, pp. 4456-4465.

[2] Carlson, R. and Hawkins, S. (2007) When is phonetic detail a detail? *Proc. ICPhS XVI*, pp. 211–214.

[3] Koenig, L. R., Lucero, J. C. and Perlman, E. (2008) Speech production variability in fricatives of children and adults: Results of functional data analysis. *J. Acoust. Soc. Am.*, Vol. 124, pp. 3158–3170.

[4] R Development Core Team (2008) *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org.

[5] Online: http://www.functionaldata.org.

[6] Torreira, F., Adda-Decker, M., and Ernestus, M. (submitted). The Nijmegen Corpus of Casual French.

[7] Ramsay, J. O. and Silverman, B. W. (1997) *Functional Data Analysis*, Springer-Verlag New York, Inc.

[8] Ramsay, J. O. and Silverman, B. W. (2002) *Applied Functional Data Analysis - Methods and Case Studies*, Springer-Verlag New York, Inc.

[9] Sarkar, D. (2008) *Lattice: Multivariate Data Visualization with R*, Springer.

[1] http://www.ling.cam.ac.uk/s2s/