Measurement of nonlinear distance in data derived from linguistic corpora

Hermann Moisl
School of English Literature, Language, and Linguistics
Newcastle University
Newcastle upon Tyne NE1 7RU
United Kingdom
Email: Hermann.moisl@ncl.ac.uk

Abstract. Most science and engineering disciplines recognize that application of linear analytical methods to data containing nonlinearities can distort results, and in response have developed mathematically and statistically based methods for dealing with nonlinearity. In linguistics, however, there has thus far been little recognition of the possibility that there might be nonlinearity in data abstracted from speech and text corpora or, where found, what the implications for analysis are. The present paper addresses this issue in three main parts. The first part outlines the nature of data nonlinearity, the second reviews existing methods for detection of nonlinearity and proposes a way of measuring nonlinear relationships between data objects, and, using these methods, the third identifies and quantifies the degree of nonlinearity is present in data abstracted from the *Diachronic Electronic Corpus of Tyneside English*, a dialect speech corpus.

Keywords: Corpus linguistics, corpus data analysis, nonlinearity, geodesic distance, graph distance

Introduction

Most science and engineering disciplines recognize that application of linear analytical methods to data containing nonlinearities can distort results, and in response have developed methods for dealing with nonlinearity [1]. In linguistics, however, there has thus far been little recognition of the possibility that there might be nonlinearity in data abstracted from speech and text corpora or, where found, what the implications for analysis are [2]. The present paper addresses this issue in three main parts. The first part outlines the nature of data nonlinearity, the second reviews existing methods for detection of nonlinearity and proposes a way of measuring nonlinear relationships between data objects, and, using these methods, the third identifies and quantifies the degree of nonlinearity is present in data abstracted from the *Diachronic Electronic Corpus of Tyneside English*, a dialect speech corpus [3].

1. Nonlinearity

1.1 Nonlinearity in Natural Processes

In natural processes there is a fundamental distinction between linear and nonlinear behavior. Linear processes have a constant proportionality between cause and effect. If a ball is kicked x hard and it goes y distance, then a $2x$ kick will appear to make it go $2y$, a $3x$ kick $3y$, and so on. Nonlinearity is the breakdown of such proportionality. In the case of our ball, the linear relationship increasingly breaks down as it is kicked harder and harder. Air and rolling resistance become significant factors, so that for, say, 5*x* it only goes 4.9 *y*, for 6*x* 5.7*y*, and again so on until eventually it bursts and goes hardly any distance at all. Such nonlinear effects pervade the natural world and gives rise to a wide variety of complex and often unexpected --including chaotic—behaviours [4].

1.2 Nonlinearity in Data

Data is a description of objects involved in a natural process of interest in terms of a set of variables. Given *m* objects described by *n* variables, a standard representation of data for computational analysis is a matrix M in which each of the *m* rows represents a different object, each of the *n* columns represents a different variable, and the value at $M_{i.j}$ describes object *i* in terms of variable *j*, for *i* = 1..*m*, *j* = 1..*n*. The matrix thereby makes the link between the researcher's conceptualization of the process in terms of the semantics of the variables s/he has chosen and the state of the world, and allows the resulting data to be taken as a representation of the process based on empirical observation. Assuming that the representation is a faithful one, any nonlinearity in the process will be reflected in the data.

M is linear when the functional relationships between all its variables, that is, the values in its columns, conform to the mathematical definition of linearity. In mathematics, a linear function *f* is one that satisfies the following properties, where *x* and *y* are variables and *a* is a constant [5]:

- Additivity: $f(x+y) = f(x) + f(y)$ - adding the results of *f* applied to *x* and *y* separately is equivalent to adding *x* and *y* and then applying *f* to the sum.
- Homogeneity: $f(ax) = af(x)$ - multiplying the result of applying *f* to *x* by a constant is equivalent to multiplying *x* by the constant and then applying *f* to the result.

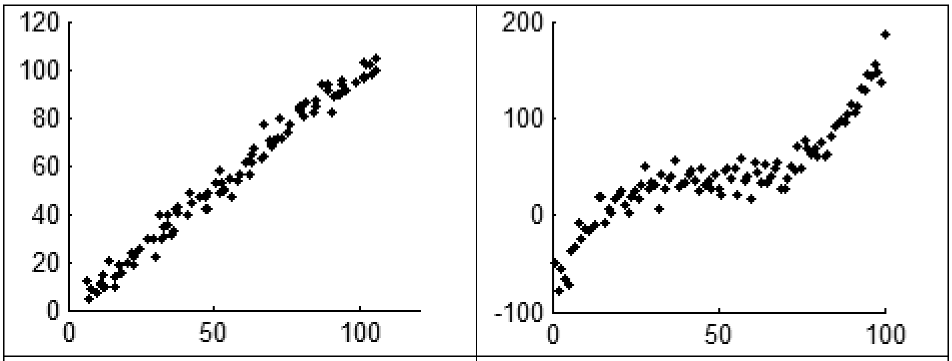
A function which does not satisfy these two properties is nonlinear, and so is a data matrix in which the relationship between two or more of its columns is nonlinear.

2. Nonlinearity Detection

It is not in general obvious whether a given data matrix contains nonlinearity, and the only way to find out if it does is to test for it. In practice, data abstracted from observation is likely to contain at least some noise, and it is consequently unlikely that strictly linear relationships between variables will be found. Instead, one is looking for degrees of deviation from linearity. Three ways of doing this are presented. Two of them are well-established, and the third is a proposal based on graph distance measurement.

2.1 Graphical Identification of Nonlinearity

The graphical method is based on pairwise scatter-plotting of variables and subsequent visual identification of deviation from linearity. In figure 1a, for example, the essentially linear relationship of variables *v1* and *v2* is visually clear despite the scatter, and the nonlinear relationship in figure 2b equally so.



a: Linear relationship of *v1* and *v2*

b: Nonlinear relationship of *v1* and *v2*

Figure 1: Scatter plots of linear and nonlinear bivariate data

Looking for nonlinearity in this way involves plotting of all possible distinct pairings of data variables and visual identification of any nonlinearity. This can be a fairly onerous but generally not insuperable undertaking where the number of variables is large. A more serious problem is that visual interpretation of scatter plots is subjective, and where the shape of the relationship between variables is not as unambiguous as those in figure 1 different observers are likely to draw different conclusions. For example, is the relationship in figure 2 linear with substantial noise, or nonlinear?

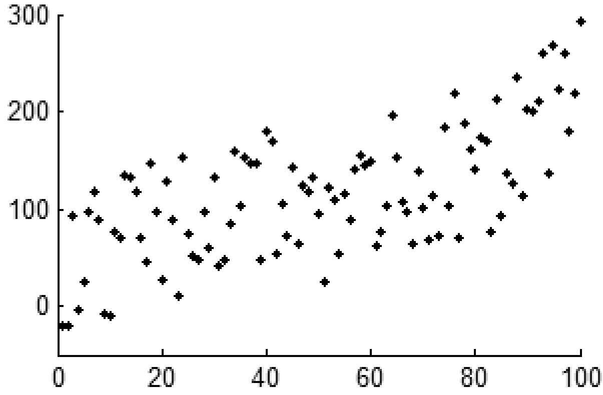


Figure 2: Possibly noisy linear, possibly nonlinear bivariate data

2.2 Identification of Nonlinearity Based on Regression

To be fully useful, graphically-based identification of nonlinearity needs to be supplemented by quantitative measures of the degree of nonlinearity. Regression analysis provides this [6, 7]. Regression attempts to model the relationship between one or more independent variables and a dependent variable whose values are hypothesized to be causally determined by the independent one(s), by finding a mathematical function which best fits the data distribution. Because the aim is simply to decide whether given data is linear or nonlinear rather than to find the optimal mathematical fit for it, the discussion confines itself to parametric regression.

The first step in parametric regression is to select a mathematical model that relates the values of the dependent variable *y* to those of the independent variable *x*. A linear model proposes a linear relationship of the general form

$$y = ax + b \tag{1}$$

where *a* and *b* are scalar constants representing the slope of the line and the intercept of the line with the y-axis respectively; *a* and *b* are unknown and are to be determined. This is done by finding values for *a* and *b* such that the sum of squared residuals, that is, distances from the line of best fit to the dependent-variable values on the *y*-axis, is minimized. The line determined by the values for *a* and *b* is the best linear fit for the hypothesized relationship between *x* and *y*. A nonlinear model proposes a nonlinear relationship between *x* and *y*. Numerous nonlinear models are available. Frequently used ones in regression are polynomials with the general form

$$y = a_n x^n + a_{n-1} x^{n-1} \ldots + a_2 x^2 + a_1 x + a_0 \tag{2}$$

where the $a_n..a_0$ are constants and *n* is the order of the polynomial; where *n* = 1 the polynomial is first-order, where *n* = 2 it as second-order and so on, though traditionally orders 1, 2, and 3 are called ‘linear’, ‘quadratic’, and ‘cubic’ respectively. As with linear regression, nonlinear regression finds the line of best fit by calculating the coefficients $a_n..a_0$ which minimize the sum of squared residuals between the line and the *y* values.

Using regression to identify nonlinearity in data would appear simply to be a matter of comparing the goodness of fit of the linear model with that of whatever nonlinear model has been chosen: the data is linear if a straight line provides as good a fit as any other mathematical function [11], and nonlinear if the nonlinear model is a significantly better fit than the linear one [7]. In figure 3, for example, the cubic model looks like it fits the data best, the quadratic less well, and the linear least well; based on visual inspection, one would say that this data is nonlinear.

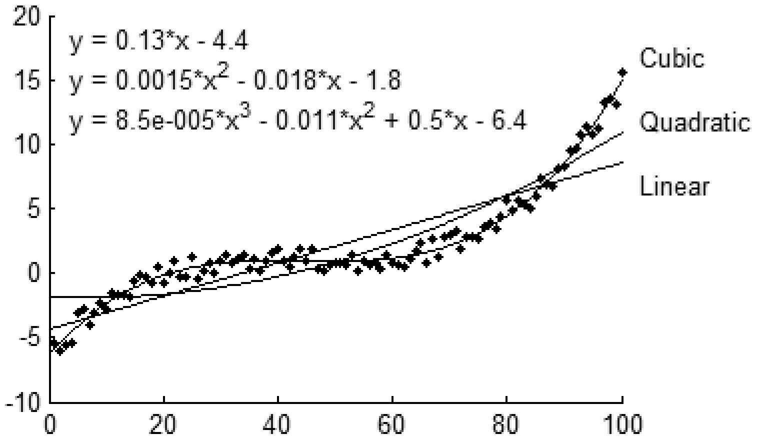


Figure 3: Linear, quadratic and cubic polynomials with curves of best fit

Such direct visual interpretation can be corroborated by residual analysis and various goodness-of-fit statistics like the runs test, summed square of errors (SSE), root mean squared error (RMSE), and R^2 [8 - 12]. These statistics all look reasonable but have an underlying problem. For a given family of models such as polynomials, the model with more parameters typically fits the data better than one with fewer; the more parameters the more convoluted the line of best fit can be and thus the closer it can get to the data values, thereby reducing SSE and affecting RMSE and R^2. Use of the foregoing statistics for identification of nonlinearity therefore implies that the best model is always the one which comes closest to the data points. Where the relationship between variables is perfectly linear this is not a problem because increasing the number of parameters will not affect the statistics: the linear model is optimal. But, as already noted, empirical data typically contains noise, and that is where the problem lies. Given data that is not perfectly linear and a model for it with $n > 2$ parameters, there are two possible interpretations. On the one hand, it may be that the model is fitting noise and thereby obscuring a relationship between the variables which is better captured by a model with fewer than *n* parameters. On the other, it may be that the nonlinearity is not noise but a genuine reflection of the nonlinear relationship between those aspects of the domain which the data describes, and that the model with *n* parameters is the preferred one. Which interpretation is correct? Knowledge of the likelihood and scale of noise in the domain can help in deciding, but this is supplemented by an extensive range of model selection methods [13, ch.5]. Two of the more frequently used methods are the extra sum-of-squares F-test and Akaike's information criterion [14; 7, ch.22].

2.3 Identification of Nonlinearity Based on Graph Distance

An alternative to regression proposed here is to make the ratio of mean nonlinear to mean linear distances among points on the data manifold the basis for nonlinearity identification. This is motivated by the observation that the shape of a manifold represents the real-world interrelationship of objects described by variables, and curvature in the manifold represents the nonlinear aspect of that interrelationship. Linear metrics ignore the nonlinearity and will therefore always be smaller than nonlinear ones; a disparity between nonlinear and linear measures consequently indicates nonlinearity, and their ratio indicates the degree of disparity.

Given a set X, a metric [15; 16] is a function d:X * X > R if, for all **x**,**y**,**z** ϵ X, the following properties hold:

- d(**x**,**y**) >= 0, that is, the distance between any two vectors is non-negative.
- d(**x**,**y**) = 0 if and only if **x** = **y**, that is, the distance from a vector to itself is 0, and for vectors which are not identical is greater than 0.
- d(**x**,**y**) = d(**y**,**x**), that is, distances are symmetrical.
- d(**x**,**z**) <= d(**x**,**y**) + d(**y**,**z**), that is, the distance between any two vectors is always less than or equal to the distance between them and a third vector.

A metric space M(V,*d*) is a vector space V on which a metric *d* is defined, which returns the distance between any two points in the space.

Numerous distance metrics exist [16]. For present purposes these are divided into two types: linear metrics, where the distance between two points on a manifold is the length of the straight line joining the points without reference to the shape of the manifold, and nonlinear metrics where the distance is the length of the shortest line joining them along the surface of the manifold, which need not be flat. Where the manifold is flat, linear and nonlinear measures are identical. Where it is curved, however, linear and nonlinear measurements can differ to varying degrees depending on the nature of the curvature, as shown in figure 4.

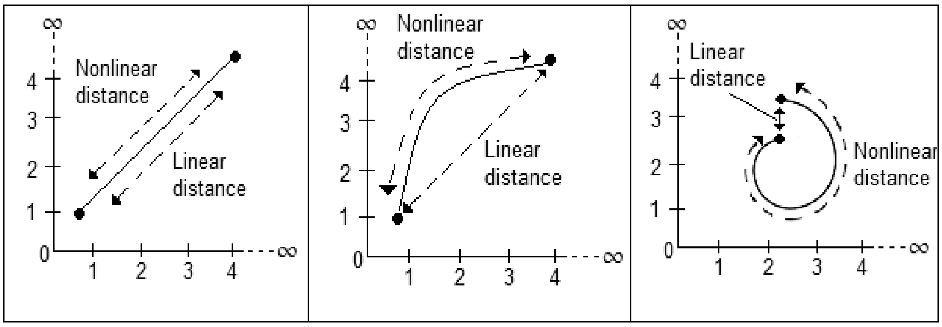


Figure 4: Linear and nonlinear distance on flat and curved manifolds

Euclidean distance is here used for linear measurement and geodesic distance for nonlinear. Euclidean distance is well known and commonly used [16]; geodesic distance requires a little explanation. Etymologically, the word ‘geodesy’ comes from Greek *geodaisia*, ‘division of the earth’; geodesic distance is the shortest distance between any two points on the Earth measured along its curved surface. Mathematically, geodesic distance is a generalization of linear to nonlinear distance measurement in a space: the geodesic distance $g(\mathbf{x},\mathbf{y})$ is the shortest distance between two points x and y on a manifold measured along its possibly-curved surface [16, ch.6]. What follows develops a method for approximating geodesic distance on manifolds using graph distance. Figure 5 shows a small nonlinear matrix M and the associated scatterplot.

	V1	**V2**
A	0.27	0.20
B	0.32	0.50
C	0.40	0.70
D	0.50	0.80
E	0.60	0.70
F	0.68	0.50
G	0.73	0.20

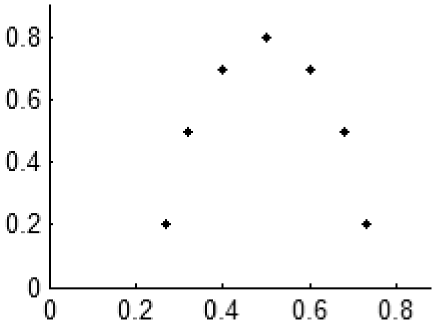


Figure 5: Nonlinear data matrix and corresponding scatterplot

For a data matrix M with m rows and n columns, a Euclidean distance matrix D is an m x m matrix each of whose values $D_{i,j}$ (for i, j = 1..m) is the Euclidean distance from row i to j of M in n-dimensional space. Figure 6 shows D for M in figure 5.

	A	**B**	**C**	**D**	**E**	**F**
A	0	.30	.52	.64	.60	.51
B	.30	0	.22	.35	.34	.36
C	.52	.22	0	.14	.20	.34
D	.64	.35	.14	0	.14	.35
E	.60	.34	.20	.14	0	.22
F	.51	.36	.34	.35	.22	0
G	.46	.51	.60	.64	.52	.30

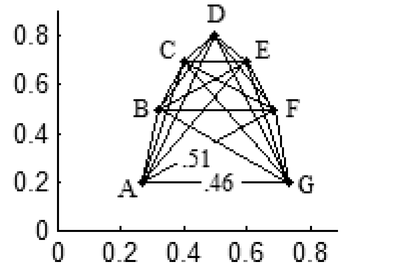


Figure 6: Euclidean distance matrix for the data in figure 5 and interpretation of the manifold as a connected graph with Euclidean distances as arc labels.

M is interpretable as a connected graph G each of whose arcs from *i* to *j* is labelled with the Euclidean distance between G_i and G_j, as shown in figure 6; the distance between node A and node B, for example, is given in the table as 0.30, between A and G as 0.46, and so on; only two arcs are labelled to avoid clutter.

A spanning tree for G is an acyclic subgraph of G which contains all the nodes in G and some subset of the arcs of G [17]. A *minimum* spanning tree of G, as its name indicates, is a spanning tree which contains the minimum number of arcs required to connect all the nodes in G, or, if the arcs have weights, the smallest sum of weights. The minimum spanning tree for G in figure 6 is shown in figure 7, with the arcs comprising the tree A>B>C>D>E>F>G emboldened.

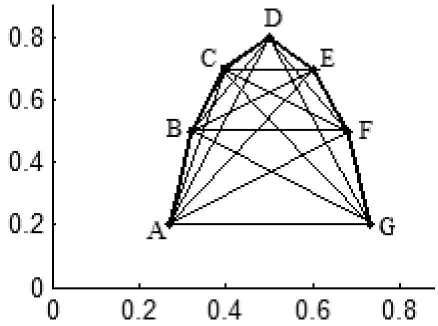


Figure 7: Minimum spanning tree for the graph in figure 6

A minimum spanning tree can be used to approximate the geodesic distances using the Euclidean distances because the distance between any two nodes is guaranteed to be minimal; in figure 7, from A to B the geodesic and Euclidean distances are identical, but from A to C the geodesic is AB + BC rather than the Euclidean AC, and so on. Figure 8 shows a table constructed in this way together with the corresponding Euclidean one.

	A	B	C	D	E	F	G
A	0	.30	.52	.64	.60	.51	.46
B	.30	0	.22	.35	.34	.36	.51
C	.52	.22	0	.14	.20	.34	.60
D	.64	.35	.14	0	.14	.35	.64
E	.60	.34	.20	.14	0	.22	.52
F	.51	.36	.34	.35	.22	0	.30
G	.46	.51	.60	.64	.52	.30	0

a. Euclidean distance matrix for M.

- Sum of distances: 16.52
- Mean distance: 0.34
- Distance A-G: 0.46

	A	B	C	D	E	F	G
A	0	.30	.52	.66	.80	1.0	1.3
B	.30	0	.22	.36	.50	.72	1.0
C	.52	.22	0	.14	.28	.50	.80
D	.66	.36	.14	0	.14	.36	.66
E	.80	.50	.28	.14	0	.22	.52
F	1.0	.72	.50	.36	.22	0	.30
G	1.3	1.0	.80	.66	.52	.30	0

b. Graph distance matrix for M.

- Sum of distances: 22.60
- Mean distance: 0.46
- Distance A-G: 1.3

Figure 8: Euclidean and geodesic distance matrices for the data in figure 14a

The sum of distances and mean distance for the Euclidean matrix are both substantially less than for the graph, and the graph distance between A and G is almost three times larger than the Euclidean, which figure 7 confirms visually.

The ratio of mean graph to mean Euclidean distance between all pairs of nodes in a graph gives a measure of the amount of nonlinearity in a data manifold. If the manifold is linear then the two means are identical and the ratio is 1; any nonlinearity makes the mean of graph distances greater than the Euclidean mean, and the ratio is greater than 1 in proportion of the degree of nonlinearity. Figure 9 is an example based on the Swiss roll data extensively used in discussions of nonlinearity, and shows the path of the shortest graph distance from A to B.

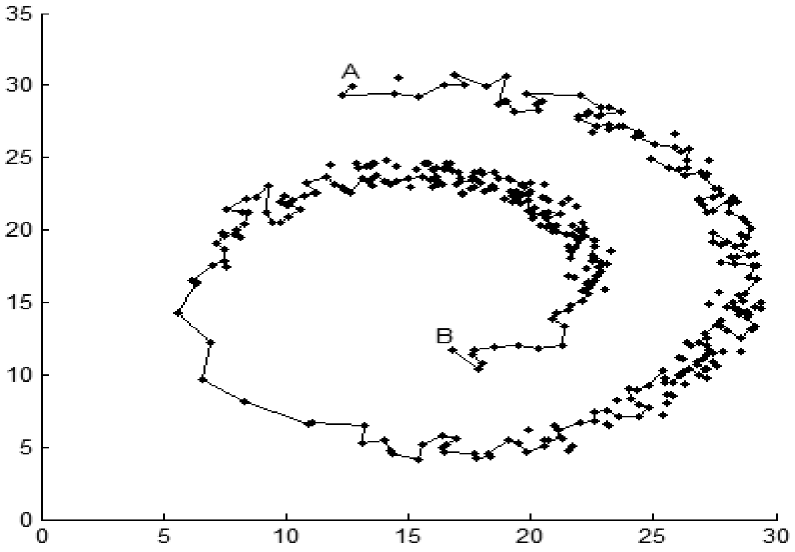


Figure 9: Euclidean and graph distances on a nonlinear data manifold

The ratio of mean graph to mean Euclidean distance in figure 9 is 3.7, and the ratio of graph to Euclidean distance from A to B is 6.6, that is, almost seven times as far.

3. Case Study

This final section presents a case study to show that substantial nonlinearity does in fact occur in a particular speech corpus.

3.1 Corpus Data

The *Diachronic Electronic Corpus of Tyneside English* (DECTE) [3] includes phonetic transcriptions of 63 audio recordings, and the data for what follows is abstracted from these. Each speaker was represented by a 156-element vector each element of which represents a different phonetic segment in the DECTE transcription scheme, and the value at any given element is the frequency with which the speaker uses the associated segment in his or her interview. The set of speaker vectors was assembled into a matrix M in which the rows *i* (for *i* = 1..63) represent the speakers, the columns *j* (for *j* = 1..156) represent the phonetic segments, and the value at $M_{i,j}$ is the number of times speaker *i* uses segment *j*. M was normalized using mean document length [18] to remove the effect of variation in interview length.

3.2 Identification of Nonlinearity in M

Using graphical and regression-based methods, no strictly or even approximately linear relationships between pairs of variables were found in M. In a few cases the relationships looked random, but most showed a discernible pattern; the segment pair *ɔ:* and *a:* is representative and is used as the basis for discussion in what follows.

3.2.1 Graphical Identification of Nonlinearity

A scatter plot of *ɔ:* on the horizontal axis and *a:* on the vertical in figure 10 shows a visually clear nonlinear relationship.

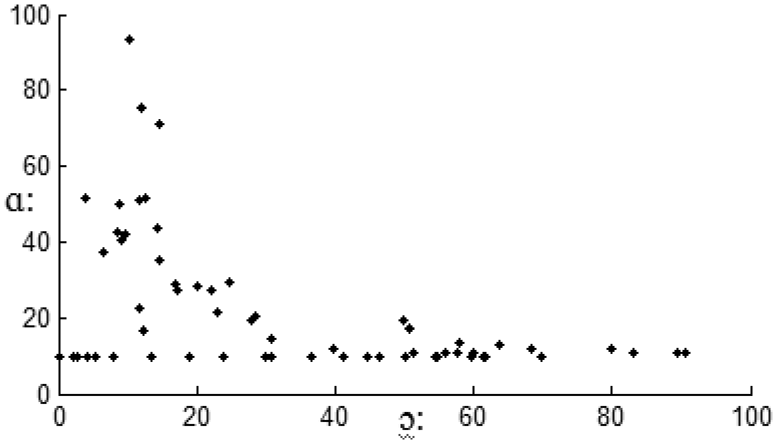


Figure 10: Scatter plot of column values in M representing the phonetic segments *ɔ:* and *a:*

3.2.2 Regression-based Identification of Nonlinearity

Using *ɔ:* as the independent variable and *ɑ:* as the dependent, a selection of polynomials was used to model the relationship. These are shown in figure 11.

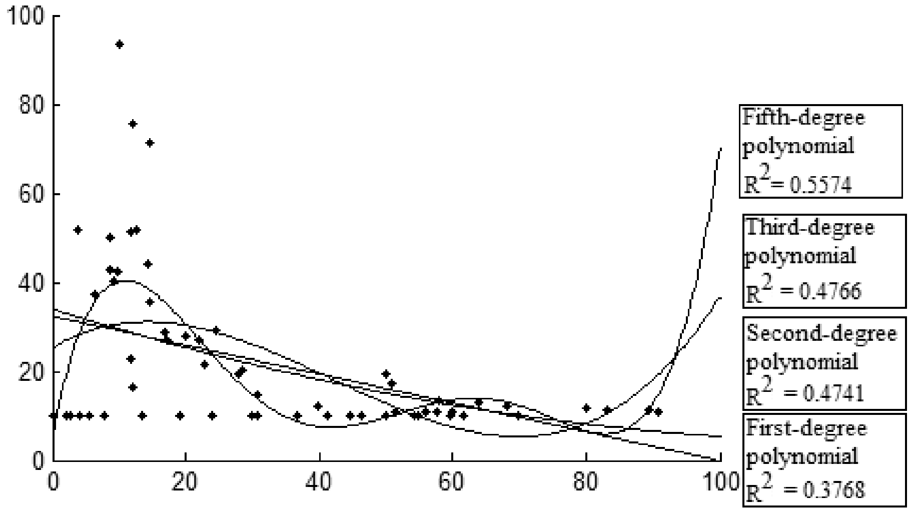


Figure 11: Polynomial regression models of the ɔ: / ɑ: relationship

Visually, the linear model appears to fit least well and the 5th-degree polynomial best, as expected, and this is confirmed by runs tests, residual plots, and the goodness of fit statistics in table 1.

Table 1: Goodness of fit statistics for figure 11

	SSE	**RMSE**	**R^2**
Degree 1	12420	15.03	0.3768
Degree 2	10480	13.93	0.4741
Degree 3	10390	14.00	0.4786
Degree 5	8821	13.15	0.5574

The extra sum-of-squares F-test and AIC test further support the indications so far: that the first-order model is worst, that second-order is better than third, but that the fifth-order model is preferred.

3.2.3 Graph Distance-based Identification of Nonlinearity

The Euclidean 63 x 63 distance matrix E was calculated for M, the minimum spanning tree for E was found, and the graph distance matrix G was derived by tree traversal, all as described in the foregoing discussion. The rows of both E and G were then linearized into vectors of length 63 x 63 = 3969, sorted, and co-plotted to get a representation of the relationship between linear and graph distances in the two matrices. This is shown in figure 12.

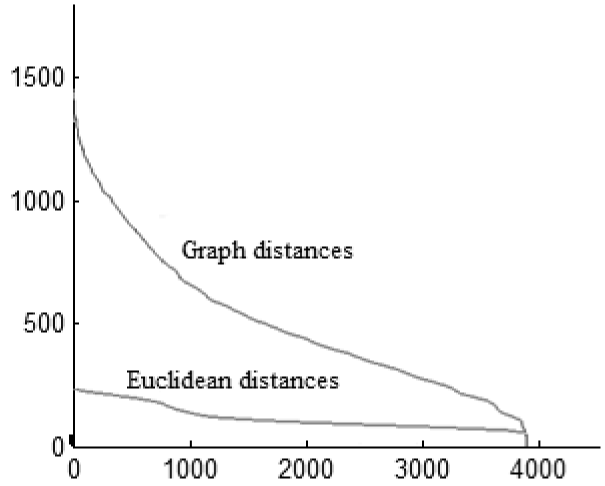


Figure 12: Comparison of Euclidean and graph distances for M

The graph distances between and among the speakers in M are consistently larger than the Euclidean ones over the entire range. This is summarized in the ratio *mean(G) / mean(E)* of mean distances, which is 3.89. On these indicators, M can be said to contain a substantial amount of nonlinearity.

Conclusion

This discussion set out to show how nonlinearity can be detected in data derived from linguistic corpora using established graphical and regression-based methods and proposing a method based on approximation of geodesic distance measurement with graph distance. These methods were applied to frequency data abstracted from phonetic transcriptions of speech from DECTE, a dialect corpus, and all the methods agreed that substantial nonlinearity was present. DECTE is typical of the many digital electronic language corpora that have appeared in recent years [19,20], and it is reasonable to suspect that nonlinearity will be present in data abstracted from these as well. Where, therefore, measurement of distance among data objects is a factor in analysis, as it is, for example, in cluster analysis, the data should first be screened for nonlinearity and the selection of analytical method should be guided by the result.

References

1. Strogatz, S.: Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry and Engineering. Perseus Books, New York (2000)
2. Moisl, H.: Data nonlinearity in exploratory multivariate analysis of language corpora. In: Nerbonne, J., Ellison, M., Kondrak, G. (eds.) Computing and Historical Phonology. Proceedings of the Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology, pp. 93--100. Association for Computational Linguistics (2007)
3. Corrigan, K., Moisl, H., Buchstaller, I.: The Diachronic Electronic Corpus of Tyneside English. http://research.ncl.ac.uk/decte/index.htm (2012)

4. Bertuglia, C., Vaio, F.: Nonlinearity, Chaos, and Complexity: The Dynamics of Natural and Social Systems. Oxford University Press, Oxford (2005)
5. Lay, D.: Linear Algebra and its Applications, 4th ed. Pearson, New York (2010)
6. Seber, G., Wild, C.: Nonlinear Regression. Wiley-Interscience, Hoboken, NJ (2003)
7. Motulsky, H., Christopoulos, A.: Fitting Models to Data using Linear and Nonlinear Regression. Oxford, Oxford University Press (2004)
8. Mark, H., Workman, J.: Linearity in calibration: the importance of nonlinearity. Spectroscopy 20(1), (2005)
9. Mark, H., Workman, J.: Linearity in calibration: the Durbin-Watson statistic. Spectroscopy 20(3) (2005)
10. Mark, H., Workman, J.: Linearity in calibration: other tests for nonlinearity. Spectroscopy 20(4) (2005)
11. Mark, H., Workman, J. Linearity in calibration: how to test for non-linearity. Spectroscopy 20(9) (2005)
12. Mark, H., Workman, J.: Linearity in calibration: quantifying nonlinearity. Spectroscopy 20(12) (2005)
13. Izenman, A.: Modern Multivariate Statistical Techniques. Springer, Berlin (2008)
14. Burnham, K., Anderson, D.: Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach, 2nd ed. Springer, Berlin (2002)
15. O Searcoid, M.: Metric Spaces. Springer, Berlin (2006)
16. Deza, M., Deza, E.: Encyclopedia of Distances. Springer, Berlin (2009)
17. Gross, J., Yellen, J.: Graph Theory and its Applications, 2nd ed. Chapman & Hall, London (2006)
18. Moisl, H.: Variable scaling in cluster analysis of linguistic data. Corpus Linguistics and Linguistic Theory 6, 75--103 (2010)
19. Beal J., Corrigan K., Moisl H.: Creating and Digitizing Language Corpora, Volume 1: Synchronic Databases. Palgrave Macmillan, Basingstoke (2007)
20. Beal J., Corrigan K., Moisl H.: Creating and Digitizing Language Corpora, Volume 2: Diachronic Databases. Palgrave Macmillan, Basingstoke (2007)