Mapping Phonetic Variation in the *Newcastle Electronic Corpus of Tyneside English*

Hermann Moisl

September 21, 2011

The *Newcastle Electronic Corpus of Tyneside English* (NECTE) is a sample of dialect speech from Tyneside in North-East England (Corrigan *et al* 2006; Allen *et al.* 2007). Jones-Sargent (1983), Moisl and Jones (2005), and Moisl, Maguire and Allen (2006) used cluster analysis to show that the speakers who constitute the earlier of the two chronological strata in the corpus fall into distinct groups defined by relative frequency of usage of phonetic segments, and Moisl and Maguire (2008) went on to identify the main phonetic determinants of that grouping by comparing cluster centroids. The present discussion develops these findings by constructing a map which comprehensively describes the pattern of phonetic variation across the NECTE speakers, and, in combination with the earlier studies just cited, is intended as a contribution to a methodology for corpus-based mathematical and statistical study of language variation.

The discussion is in two main parts: the first part briefly describes NECTE, the second constructs the phonetic variation map.

1 The *Newcastle Electronic Corpus of Tyneside English*

NECTE is a corpus of dialect speech from Tyneside in North-East England, shown as the boxed area in Figure 1.

It is based on two pre-existing corpora of audio-recorded speech, one of them gathered in the late 1960s by the Tyneside Linguistic Survey (TLS) (Strang 1968; Pellowe *et al.* 1972), and the other between 1991 and 1994 by the *Phonological Variation and Change in Contemporary Spoken English* (PVC) project (Milroy *et al.* 1994). This discussion, like the earlier ones cited in the Introduction, deals with the TLS component of NECTE only, which is henceforth referred to as NECTE/TLS.

NECTE/TLS includes phonetic transcriptions of each of 64 recordings which the TLS team produced with the aim of determining whether systematic phonetic variation among Tyneside speakers of the period could be significantly correlated with variation in their social characteristics. To this end the TLS developed a methodology which was radical at the time and remains so today: in contrast to the then-universal and still-dominant theory driven approach, where social and linguistic factors are selected by the analyst on the basis of some combination of an independently-specified theoretical framework, existing

Figure 1: The Tyneside area of North-East England

case studies, and personal experience of the domain of enquiry, the TLS proposed a fundamentally empirical approach in which salient factors are extracted from the data itself and then serve as the basis for model construction.

To realize this research aim using its empirical methodology, the TLS had to compare the audio interviews it had collected at the phonetic level of representation. This required that the analog speech signal be discretized into phonetic segment sequences, or, in other words, to be phonetically transcribed. Two levels of transcription were produced, a highly detailed narrow one designated 'State', and a superordinate 'Putative Diasystemic Variables' (PDV) level which collapsed some of the finer distinctions transcribed at the 'State' level. Like the earlier studies cited in the Introduction we shall be dealing with the PDV level, which defines 156 different phonetic segments; details of the TLS transcription scheme are available in Jones-Sargent (1983) and Corrigan *et al.* (2006).

2 Construction of the phonetic variation map

In mathematics, a map is a function. Colloquially, a map is a graphical representation of the relationships between and among objects of interest, such as the relative distances and connectedness of real-world towns and cities. The latter usage is intended here: the map constructed in this part of the discussion is a graphical representation of how the variabilities of the phonetic segments defined by the TLS transcription scheme relate to one another.

2.1 Data

The data used in this discussion is identical to that used in Moisl, Maguire and Allen (2006) and Moisl and Maguire (2008). It was created as the basis for answering the following research question:

> Is there nonrandom phonetic variation among speakers in the Tyneside speech community as represented by NECTE/TLS and, if so, what are the primary determinants of that variation?

The objects of interest were and are the 64 NECTE/TLS speakers, and the variables describing the speakers are the 146 of the 156 phonetic segments defined by the TLS transcription scheme that are actually used by speakers in their interviews. Each speaker S_i (for $i = 1..64$) is described by a 146-element vector v each of whose elements v_j (for $j = 1..146$) represents a different segment, and the value at v_j is the frequency with which S_i uses segment j, as in Table1.

Table 1: Vector representation of a NECTE/TLS speaker's phonetic usage

Vector index	1	2	3	...	146
Phonetic segment	g	i	t	...	x
Frequency	31	28	123	...	0

The set of 64 speaker vectors was assembled into a matrix M in which the 64 rows represent the 64 speakers, the 146 columns represent the 146 PDV variables, and the value at $M_{i,j}$ is the number of times speaker i uses the phonetic segment j. A fragment of this 64 x 146 matrix M is shown in Table 2.

Table 2: Fragment of the NECTE/TLS data matrix M

Vector index	1	2	3	...	146
Phonetic segment	g	i	t	...	x
S_1	31	28	123	...	0
S_2	22	8	124	...	0
...	...	...	...	...	...
S_{64}	19	3	73	...	0

Because the transcriptions differ substantially in length, M was normalized by mean document length (Robertson and Spärck-Jones 1994; Spärck-Jones *et al.* 2000) to eliminate the distorting effect of that variation on subsequent analysis. This normalization involved transformation of the row vectors of M in relation to the average length of the 64 transcriptions.

$$M_i = M_i(\frac{\mu}{length(T_i)}) \tag{1}$$

where M_i is the matrix row representing the frequency profile of transcription T_i, $length(T_i)$ is the total number of phonetic segments in T_i, and μ is the mean number of segments across all 64 transcriptions:

$$\mu = \sum\nolimits_{i=1..m} \frac{length(T_i)}{m} \tag{2}$$

The values in each row vector M_i are multiplied by the ratio of the mean number of phonetic segments per transcription across the whole corpus to the number of segments in transcription T_i. The longer the document the numerically smaller the ratio, and vice versa; the effect is to decrease the values in the vectors that represent long transcriptions, and increase them in vectors that represent short ones, relative to average transcription length.

2.2 Map construction

The object is to construct a map that describes how the variabilities of the NECTE/TLS phonetic segments relate to another. Principal component analysis (PCA) is used for this purpose in what follows. A brief outline of PCA is first given, and it is then used in map construction.

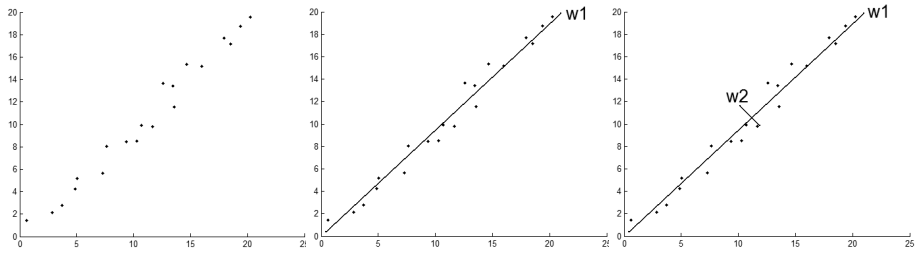
2.2.1 PCA outline

Data are an interpretation of some domain of study. Such an interpretation is a description of objects in the domain in terms of variables, where a variable is a symbol, that is, a physical entity to which a meaning is assigned by human interpreters. Each variable represents an aspect of the domain considered to be relevant in answering the research question, and the set of selected variables constitutes the template in terms of which the domain is interpreted. Selection of variables appropriate to the research question is, therefore, crucial in scientific research.

Which variables are appropriate in any given case? The fundamental principle is that the ones chosen must represent all and only those aspects of the domain which are relevant to the research question. In general, this is an unattainable ideal. Any domain can be described by an essentially arbitrary number of finite sets of variables; selection of one particular set can only be done on the basis of personal knowledge of the domain and of the body of scientific theory associated with it, tempered by personal discretion. In other words, there is no algorithm for choosing an optimally relevant set of variables. Where, as in the present case, the research question involves variability, a variable can be suboptimal in two ways. On the one hand, some variables are more useful than others in describing the domain: the values taken by a variable like, say, 'income' are likely to vary much more across a random sample of a human population than, say, the values for 'number of limbs', and as such 'income' is much more informative. On the other, two or more variables might by redundant, that is, correlated, to greater or lesser degrees because they overlap in terms of the information they represent: 'age', 'height', and 'weight' are correlated in that, for example, adults generally weigh more than children. Any given data matrix might contain variables that are suboptimal in one or both these ways.

Given an n-dimensional data matrix containing some significant degree of suboptimality, where n is the number of columns representing variables, PCA is a method for expressing most of the total variability across the values in all n columns using a smaller number $k < n$ of uncorrelated variables, thereby eliminating any uninformative variables and redundancy in the original matrix. These new variables are found using the shape of the manifold in the original n-dimensional space. Figure 2 shows how this is done using a geometric interpretation of a fictitious data matrix, where $n = 2$ for ease of exposition.

The horizontal and vertical axes represent two variables v_1 and v_2, the objects described by these variables are represented as vectors in the two-dimensional space, and the set of vectors constitutes a manifold.



(a) Correlated variables (b) W1: Main variability (c) W2: Remaining variability

Figure 2: Main directions of variability in a two-dimensional manifold

Visual inspection of Figure 2a shows that v_1 and v_2 are strongly correlated, and the main direction of variability in the manifold can be visually identified; the least-squares line of best fit drawn through the manifold in that direction, as in Figure 2b, is the first new variable w_1: it captures most of the variability in the manifold, and its length is the amount of variability that it captures. A second line of best fit is now drawn orthogonally to the first (Figure 2c) to capture the remaining variability: this is the second new variable w_2, and its length is again the amount of variability it represents. We now have two new uncorrelated variables in addition to the two original ones. What is gained? Note the disparity in the lengths of w_1 and w_2: it's clear that w_1 captures almost all the variability in the manifold and w_2 very little, and one might conclude that w_2 can simply be omitted with minimal loss of information. Doing so reduces the dimensionality of the original data from 2 to 1. This extends to any dimensionality. It might, for example, be found that most of the variability in, say, a 24-dimensional matrix with substantial redundancy can be expressed in, again say, 3 or 4 uncorrelated variables.

Details of how Pca finds uncorrelated variables are available in most non-elementary statistics textbooks such as (Tabachnik and Fidell 2005: ch. 13); the standard reference work is Joliffe (2002), and Shlens (2009) is an excellent online introduction.

2.2.2 Application of Pca to map construction

The variables in the Necte/Tls matrix M are suboptimal in both of the above senses. On the one hand, nearly half of them are almost entirely uninformative in terms of their variability. This can be shown by calculating the variance of each of the 146 columns, sorting the variances in descending order of magnitude, and plotting the result. The plot is shown in Figure 3.

The variables from about the 80^{th} to the 146^{th} contribute effectively nothing to the total variability in M. And, on the other hand, there is substantial redundancy among variables in M. The correlation matrix for the column vectors of M was calculated and examined for significant correlations, where the threshold value for significant' is taken to be the de facto standard absolute value $+/-0.3$ (Tabachnik and Fidell 2005: ch.13). Discounting the 146 perfect

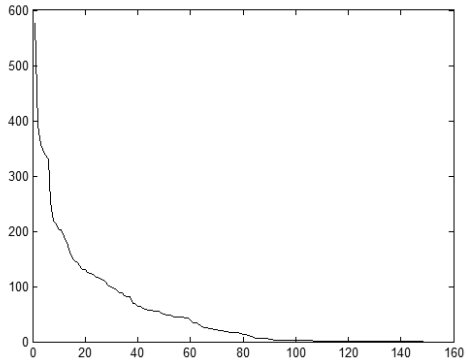


Figure 3: Sorted variances of the columns of M

correlations for variables with themselves on the main diagonal of the matrix, 888 significant positive and negative correlations were found out of a total 9198, which is 9.7%; the distribution of significant correlations, sorted in descending order of magnitude, is shown in Figure 4.

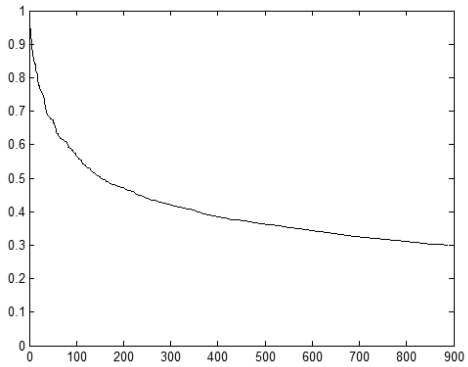


Figure 4: Absolute correlations ≥ 0.3 in the correlation matrix for M

PCA was applied to M to eliminate the very low variance variables and redundancy among the others. The available software implementations of PCA typically generate a range of outputs; two are important for present purposes:

1. *The eigenvector matrix EVECT(M)*

   For any n-dimensional data matrix D, PCA calculates an $n \times n$ matrix EVECT:

   - The n rows of EVECT represent the n variables of D.
   - The n columns of EVECT represent the n new variables; these are the principal components of D, also known as the eigenvectors after which EVECT is named. The columns of EVECT are sorted in descending order of how much of the variability in D each one represents. The first column c_1 represents the direction of greatest variability in D; in Figure 2c this is w_1 The second column c_2 represents the second greatest direction of variability in D; this is w_2 in Figure 2c. And so on to n.

- The values at EVECT_{ij} (for $i, j = 1 \ldots n$) are the coefficents of the linear combinations of the original variables of D that generate the principal components. These values are also known as loadings, and can be thought of as a measure of the contribution which each of the original variables makes to each of the new ones.

2. *The eigenvalue vector EVAL(M)*

   For any n-dimensional data matrix D, PCA also calculates an n-element vector EVAL whose values quantify the amount of variability which each successive eigenvector in EVECT represents. The first value in EVAL corresponds to the first eigenvector in EVECT and shows the amount of variability in D that that eigenvector represents; this is the length of w_1 in Figure 2c. The second value in EVAL corresponds to the second eigenvector in EVECT and shows how much of the remaining variability in D that that eigenvector represents; this is the length of w_2 in Figure 2c. And again so on to n. By examining the values in EVAL, one can decide the point at which the amount of variability captured by successive eigenvectors of EVECT becomes negligible and then simply delete all the eigenvectors beyond that point from EVECT, thereby reducing the dimensionality. If, for example, the k^{th} alue in EVAL were selected as the threshold, then the eigenvectors $1 \ldots k$ in EVECT would be retained, and those from $k + 1$ to n deleted.

The eigenvector matrix EVECT(M) and the eigenvalue vector EVAL(M) for M are both far too large to be shown in their entirety, so only small fragments are given in Tables 3 and 4 by way of example.

Table 3: Fragment of the eigenvector matrix EVECT(M) for M

		v_1	v_2	v_3	...	v_{146}
1.	ɔ: goat	-0.130	-0.634	-0.079	...	-0.001
2.	eI mine	0.424	-0.082	-0.229	...	0.003
3.	n nice	0.200	0.066	0.550	...	0.003
...	...	...	...	...	...	...
146.	ɐ bird	0.000	0.000	0.000	...	0.9985

Table 4: Fragment of the eigenvalue vector $EVAL(M)$ for M

v_1	v_2	v_3	...	v_{146}
1853.6	1271.5	644.56	...	0.0

How many eigenvectors should be retained in any given case? There is no right answer; selection of a threshold is a matter of judgement by the researcher in relation to the research question. Various criteria have been proposed (Tabachnik and Fidell 2005: ch. 13; Joliffe 2002: ch. 6). A widely used one is the scree plot, that is, a plot of the values in the main diagonal of the eigenvalue matrix. The scree plot for EVAL(M) is shown in Figure 5.

It is clear from Figure 5 that the amount of variability in M captured after the 40^{th} or so eigenvector is negligible, and as such the first 40 eigenvectors should be retained as the uncorrelated variables to replace the original 146

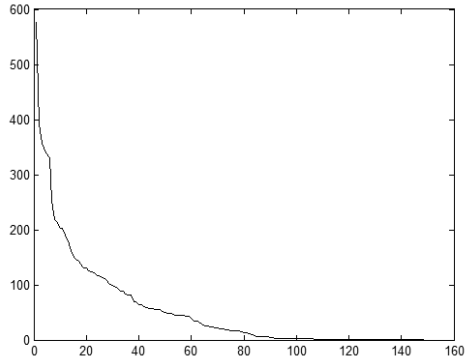


Figure 5: Screen plot of the values on the main diagonal of $EVAL(M)$

variables of M. For expository purposes, however, only the first 10 eigenvectors are retained, which captures 71% of the variance in M. The result is an 146×10 eigenvector matrix $\text{EVECT}(M)_{10}$ in which the 146 rows represent the original variables of M, the 10 columns represent the new variables, and the value at $\text{EVECT}(M)_{10,ij}$ (for $i = 1 \ldots 146, j = 1 \ldots 10$) is the loading of original variable i on the new variable j. A fragment is shown in Table 5.

Table 5: Fragment of $\text{EVECT}(M)_{10}$

		v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9	v_{10}
1.	ɔː goat	-0.130	-0.634	-0.079	0.164	-0.343	-0.174	0.080	0.018	0.092	0.043
2.	eI mine	0.424	-0.082	-0.229	-0.037	-0.136	0.031	-0.145	0.038	0.090	0.115
3.	n nice	0.200	0.066	0.550	0.243	-0.121	0.185	0.239	-0.132	-0.010	0.076
...	...	...	...	...	...	...	...	...	...	...	...
146.	ɐ bird	0.000	0.000	0.000	0.000	0.001	0.000	-0.001	0.002	0.000	0.000

Variable v_1 ɔː in Table 5 has a strong negative loading on v_2, a significant negative one on v_5, and weak loadings on the remaining variables; variable v_2 eI has a significant positive loading on v_1 and weak loadings on the remainder; and so on.

$\text{EVECT}(M)_{10}$ is the basis for the NECTE/TLS phonetic variability map. Geometrically, the column vectors of $\text{EVECT}(M)_{10}$ constitute an orthogonal basis for a 10-dimensional vector space in which

- each basis vector represents one of the 10 main directions of variability in M,
- the values in each of the row vectors are the coordinates of one of the original 146 phonetic segment variables in the 10-dimensional space, and
- the set of 146 row vectors constitutes a manifold representing the variability of these phonetic segment variables in the space.

The relatively high dimensionality of the space prevents the manifold being directly visualized via two or three dimensional plotting, but hierarchical cluster analysis provides an indirect visualization by representing the similarity relations of the vectors in the manifold as a cluster tree. The row vectors of $\text{EVECT}(M)_{10}$ were hierarchically cluster analyzed using the squared Euclidean distance and the complete link clustering algorithm; other clustering

algorithms were tried, with similar results. The tree in Figure 6 only includes the 80 highest-variance variables in M because, as noted earlier, the remainder are negligible in terms of the amount of total variability they contribute, and including them would both clutter the tree and make it too large to display.

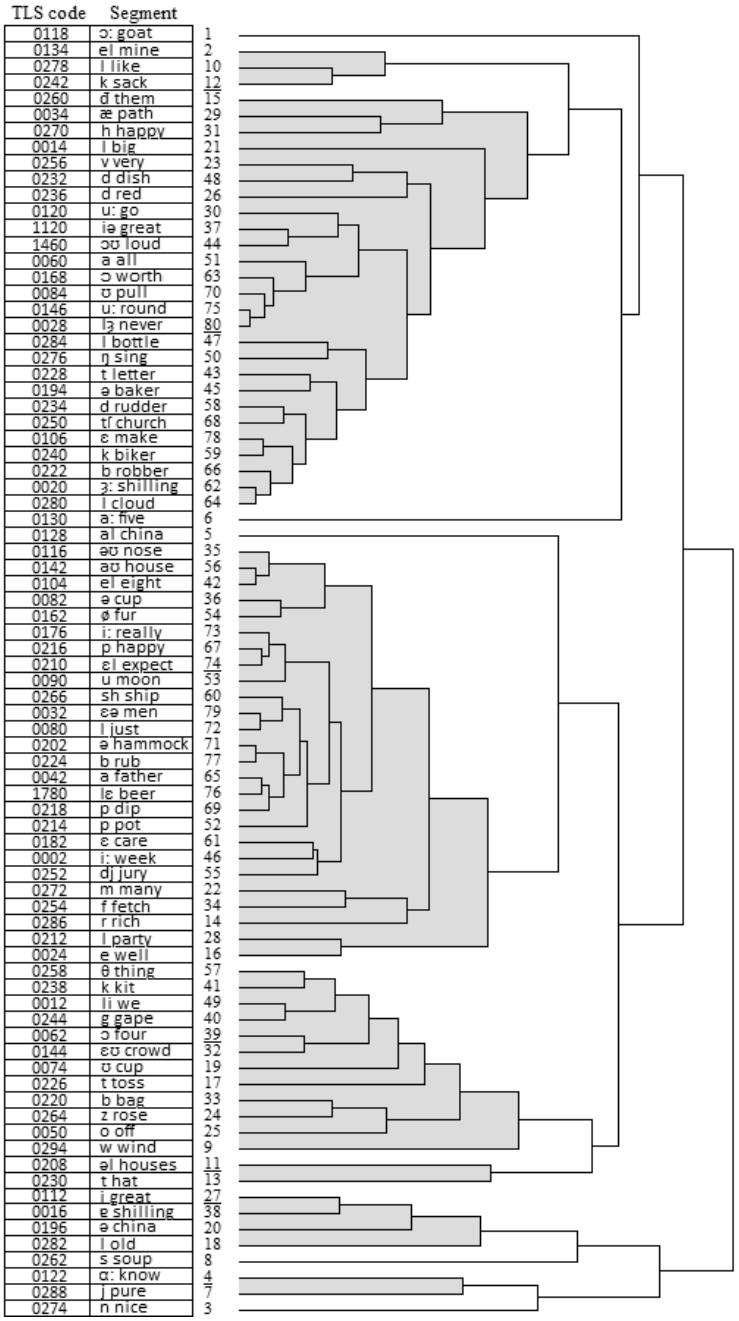


Figure 6: Cluster analysis of the first 80 rows of EVECT$(M)_{10}$

Column 1 of the labels at the leaves of the tree lists the codes of the phonetic

segments in the TLS transcription scheme for convenience of cross-referencing with the code list in Jones-Sargent (1983) and at the NECTE website (Corrigan *et al.* 2006), and column 2 lists the phonetic symbols themselves together with an example of each.

The cluster tree in Figure 6 is the map which this discussion sets out to create. The variability of each of the 80 highest-variance phonetic segments in relation to that of all the others can be seen by direct inspection, and the map thereby provides a comprehensive basis for understanding of phonetic variation in the NECTE/TLS speaker community. Most fundamental to that understanding is the observation, based on the well defined cluster structure, that the pattern of phonetic variation in the community was strongly non-random.

3 Conclusion

This discussion sets out to develop existing work by Jones-Sargent (1983), Moisl and Jones (2005), and Moisl, Maguire and Allen (2006) on phonetic variation in Tyneside English as represented by the *Newcastle Electronic Corpus of Tyneside English* by constructing a map that comprehensively describes the pattern of phonetic variation across the NECTE/TLS speakers. This map is a hierarchical cluster tree which represents the variability of the phonetic segments in the NECTE transcription scheme in a vector space whose orthogonal axes represent the main directions of variability in the NECTE data.