Dimension and structure of the speech space

R. Togneri, BE, PhD
M.D. Alder, BSc, ARCS, PhD, MEngSc
Y. Attikiouzel, BSc, PhD, FIEE

Indexing terms: Pattern recognition, Speech space

Abstract: The paper presents evidence to support its claim that the space of trajectories of speech exists and may be approximated by a four-dimensional manifold which is nonlinearly embedded in both a space of LPC coefficients and also in a filter bank space. It also investigates the possibility that there are different dimensions for different phonetic categories, but finds no evidence to support this hypothesis. The dimension is of interest since it is the smallest number of independent parameters needed to specify speech.

1 Introduction

In an earlier paper [1], we considered the trajectories of utterances in both 12- and 16-dimensional spaces of (software simulated) filter bank values, and attempted to estimate the dimension of the subspace of points so formed. The dimension is, in principle, an interesting number to compute; it ought to be the minimum number of independent parameters needed to specify the production and recognition of speech, and it ought to tell us something of the effective number of degrees of freedom of vocal tracts. Phoneticians have been claiming that phoneme production can be described by a relatively simple model having a small number of parameters, yet engineers have tended to use FFTs and LPC models with a relatively large number of coefficients. Establishing the dimension therefore has both theoretical and practical significance.

What is not clear is that such a number can be estimated by examining the set of points obtained from discretised trajectories in a space of filter bank values (referred to below as the 'speech cluster'). If the subset of speech sounds were flatly embedded in the filter bank space, it would be possible to use the standard Karhunen–Loeve method, but K–L procedures suppose linearity and cannot be expected to give sensible results for nonlinear subspaces. In Reference 1 we described computational experiments on various data sets designed to validate our (nonlinear) method. Briefly, we used the Kohonen algorithm [3] to attract k-dimensional grids towards a set of points and to obtain a 'good fit' to the set. We refer the reader to this earlier paper for a more complete description of the background to the present work. Kohonen [3] gives pictures of one-grids trying to fill two-dimensional sets, and they confirm naive expectations that when the dimension of the grid is less than the dimension of the set, there will be a large amount of buckling associated with the grid. Conversely, one might naively expect that if the grid has higher dimension than the set, some relatively large amount of buckling of the grid will be necessary to obtain a 'good fit'. The mean absolute curvature of the converged grid should therefore be a minimum when the dimensions are the same. Experiments confirmed this for the case of known manifolds, and manifolds with noise, embedded in $\boldsymbol{R}^n$ for various n. We were able to provide evidence that the dimension is higher than two but less than 12. We also applied other dimension estimators and obtained values of between two and five, with an uncertain degree of confidence.

In this paper we are concerned with refining the bounds obtained in Reference 1. We also need to determine whether the dimension so estimated is really a property of the space and not an artefact of the measuring process. Finally, we attempt to determine whether different regions of the space corresponding to vowels, nasals and fricatives have a dimension which is different from that for the space as a whole. Considerations of the production process make this eminently plausible.

We use the same statistic as in Reference 1, a measure of the mean absolute curvature of the converged grid referred to as the 'crinkle statistic'. We consider a number of cases of different embeddings of grids; in particular we treat different numbers of simulated filters and we also examine the case where the space is a space of LPC coefficients. The embedding is apparently very different as Figs. 9 and 10 indicate. We are able to infer with modest confidence that overall the speech space is best approximated by a three- or four-dimensional grid, independent of the method of measuring it.

In the second part of the paper we are concerned with the issue of discriminable substructure in the speech cluster. It would not be surprising if the subcluster of fricatives had a different dimension from the subcluster of vowels, for example. We therefore labelled the speech from six sentences spoken by six speakers and divided the speech space into three regions: vowels (including diphthongs), nasals, and fricatives (voiced and unvoiced). By assigning different colours to the different categories, and projecting down into two dimensions, it is possible to get a 'view' of the speech space and the position of the different types of speech sound. Figs. 7–10 show the results. The projection down into two dimensions loses a good deal of information, and so we rotated the space prior to projection, by hand, looking for a 'good' orientation which would allow us to separate the different components by eye. This visual inspection of the data is useful in suggesting approaches and in verifying that the Kohonen process produces sensible results. Badly

Paper 8514I (E5), first received 19th September 1990 and in revised form 23rd September 1991

The authors are with the Centre for Intelligent Information Processing Systems, The University of Western Australia, Nedlands, Perth, Western Australia 6009, Australia

knotted two-dimensional grids, or grids that are far from the cluster, are not, we are happy to report, observed. On the other hand, the images strongly suggest a nonlinear embedding in all representations.

We attempted to assess the uniformity of the extent to which the grid fitted the space in several ways. The simplest was by taking the trajectory of labelled speech for an utterance, and measuring the distance of points of the trajectory to the grid. This gives a set of numbers which could then be partitioned into subsets corresponding to nasals, vowels and fricatives. We hoped that, if the dimension of fricatives or nasals was different from the dimension of vowels, this would manifest itself in the amount of variation in the mean distances for the three categories. Our last approach was to determine the crinkle statistic for each of the three regions of the speech cluster separately.

As will be plain from this introduction, a certain degree of modern mathematical background is required of the reader, but we endeavour to stress the conceptual rather than the formal. Reference 7 contains definitions of many of the terms we employ and pointers to the others.

2 Dimension of the speech cluster

The crinkle statistic was introduced in Reference 1 and is found by taking an $r \times r \times \cdots \times r$ (k factors) grid in some space of dimension bigger than k, embedded in some way. For example, it might be a two-dimensional square grid, stretched and rolled to parametrise a torus. Each grid point (referred to for reasons connected with the origins of the Kohonen process as a *neuron*) has 2^k nearest neighbours, paired off as originally opposite in direction; for example, in two dimensions we have north, south and east, west as opposite pairs. We simply measure the cosine squared of half the angles subtended at the grid point by opposite pairs of neighbours, and sum all the 2^{k-1} values for each grid point not on the boundary, to get the total 'crinkle' for the grid. We divide by the number of nonboundary grid points to obtain a mean for the grid. (See Reference 1 for the above definition stated algebraically.) If the embedding is flat (linear or affine), the statistic is zero; if the space itself is flatly embedded, the grid in general will not be, but it is found experimentally that the value of the statistic is still small. In our applications the grid has some position in a high-dimensional space after it has been attracted in to a set of points by the Kohonen process. It is to be hoped that the grid will have converged to the set of points and will give the best approximation possible for a grid of that dimension. Experiments with manifolds of known dimension supported this hope.

It is simple to construct sets of points in R^n which have been selected to lie on or near a prescribed embedded manifold. We used 2-, 3- and 4-tori as well as spheres and planes embedded in R^n in various ways. In all cases it was found that the mean crinkle statistic for a k-grid attracted to an r-manifold was a minimum when k equalled r. The result also holds when moderate amounts of higher-dimensional noise are added to the manifold points, where 'moderate' means that the mean perturbation is of the order of the distance between points. It also gave reasonable results when the attracting set was the Lorenz attractor, a strange attractor with nonintegral dimension.

We therefore concluded that both the Kohonen process and the crinkle statistic were behaving as expected, and that the application to speech data was defensible. We therefore took the data, consisting of approximately 16000 points, each representing a 25 ms frame from an FFT of speech from six speakers consisting of six phonetically balanced sentences. This represented more data and more speakers than were used in Reference 1. The results are given in Fig. 1.

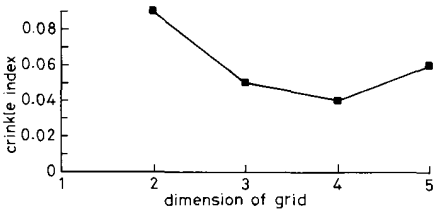

Fig. 1 *Dimension of enclosing space 12; filter bank space*

The values quoted are the means of five trials with different grids starting from different initial positions, and the standard deviation of the mean is about 0.005.

The (software simulated) filter bank was mel spaced; the number of units in two dimensions was 100 in a 10×10 array, in three dimensions it was 1000 in a $10 \times 10 \times 10$ array. In higher dimensions we used arrays of side eight in order to bring the computations within manageable limits.

When we repeated with 16 simulated filters we obtained the results shown in Fig. 2, where the increase is small but significant at the 0.1% level.

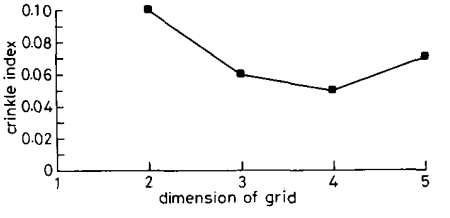

Fig. 2 *Dimension of enclosing space 16; filter bank space*

We also explored the space of 12 LPC coefficients obtained by Durbin's recursive algorithm (using the autocorrelation method [4]) applied to the same speech data. We used the euclidean rather than the Itakura metric on this space of LPC coefficients. The results are shown in Fig. 3. There were somewhat larger variances but the differences are still highly significant.

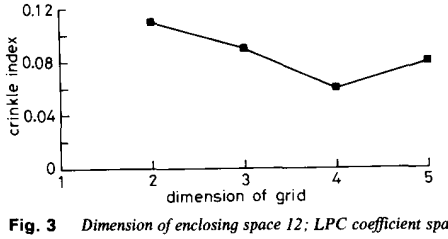

Fig. 3 *Dimension of enclosing space 12; LPC coefficient space*

Similarly, the results for a space of 16 LPC coefficients are shown in Fig. 4.

By comparison, the results for 12-dimensional random noise are shown in Fig. 5, where they are compared with the results of the dimension-12 filter bank speech space.

If instead of doing the experiment with the filter bank data we use a known (noisy) 4-manifold, a 4-torus in R^8, we obtain the results shown in Fig. 6 where again they are compared with a 12-dimensional filter bank speech space.

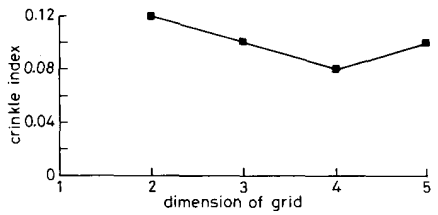

Fig. 4 *Dimension of enclosing space 16; LPC coefficient space*

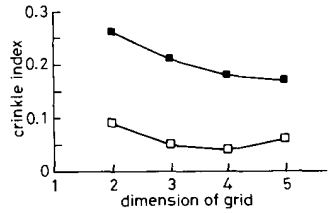

Fig. 5 *Random noise compared with filter bank space*
—■— random noise (dimension 12)
—□— filter bank space

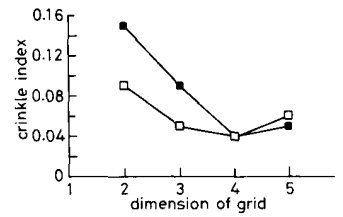

Fig. 6 *4-Torus in 8-space compared with filter bank in 12-space*
—■— noisy 4-torus in 8-space
—□— filter bank space

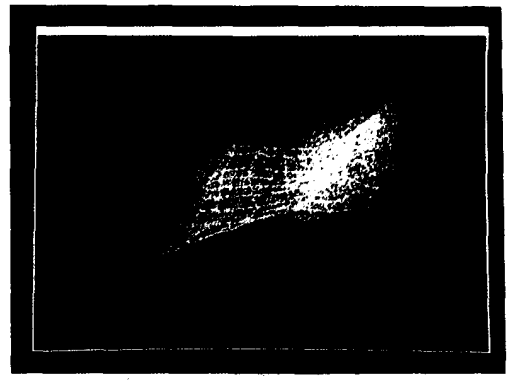

Fig. 7 *Two-dimensional projection from a 12-dimensional speech space of filter bank values*

Discrepancies between these results and those of Reference 1 are assigned to variations in the size of the grid (bigger grids magnify the mean absolute curvature when the dimensions of grid and data are disparate), and the fact that we have more diverse speech data.

It can be seen that in all cases of speech data and in the case of a known four-manifold with noise, there is a minimum in dimension four. This minimum is conspicuously absent (and the numbers are different) in the case of random data. There are good grounds for concluding that, overall, the speech cluster may be assigned a dimension of between three and four, and that this number is independent of the mode of measuring the speech data. This accords with the results obtained using the more conventional dimension estimators mentioned in Reference 1, and with anecdotal evidence obtained from synthesis [5].

Fig. 8 *Two-dimensional projection from a 12-dimensional speech space of LPC coefficients*

Figs. 7 and 8 show a projection from 12-dimensional spaces of a two-dimensional grid which has been attracted in to the cluster by the Kohonen process. Fig. 7 shows such a grid embedded in the simulated filter bank space, and Fig. 8 shows a similar embedding in a space of LPC coefficients. It is plain that the grid is indeed close to the cluster in each case and that the embeddings of the grids are nonlinear but not implausibly so.

3 Structure of the speech cluster

The speech data were labelled as described above and classified as vowel, nasal and fricative. Our first investigation into the substructure of the speech cluster consisted simply of colouring the different regions: blue for fricatives, red for nasals, and grey for vowels. Figs. 9 and 10 show projections from 12-space of the result for mel filter space and LPC space, respectively. It is possible to discriminate by eye two of the three regions from the perspective shown. The two-dimensional grid is also shown in its converged position.

The second stage consisted of tracking an utterance through the space and measuring its distance at each point from the grid. This entailed finding the nearest $k + 1$ grid points when the grid was of dimension k and projecting the trajectory point to the k simplex spanned by the $k + 1$ points. The distance between the original trajectory point and the projection of it on the simplex was then computed. Each trajectory point was classified as being either semi-vowel (S), vowel (V), diphthong (D), nasal (N), plosive (P), fricative (F), aspirate (A), or unidentifiable (L), according to the previously classified closest grid point.

It is worth contemplating the geometry in order to frame some expectations. Suppose a two-dimensional grid were fitted to two regions, one being two-dimensional and one region three-dimensional. Then it would be reasonable to expect the grid to be highly curved, i.e. to exhibit a higher crinkle index in the region

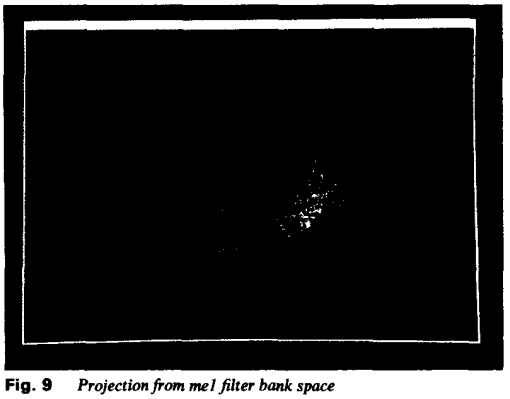

Fig. 9 *Projection from mel filter bank space*
blue points are fricatives, red are nasals, grey are vowels

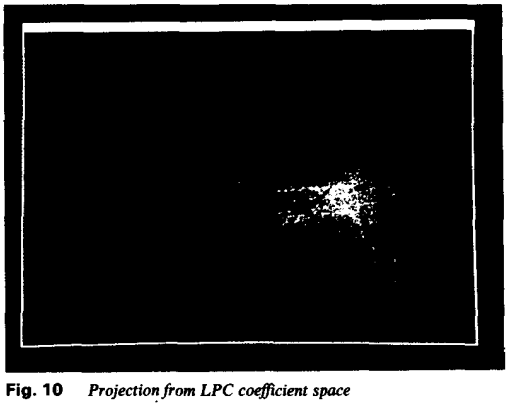

Fig. 10 *Projection from LPC coefficient space*
blue points are fricatives, red are nasals, grey are vowels

where the data are three-dimensional. One might expect that the mean distance from the grid would be higher in this region. So if the dimension of the cluster region is higher than the dimension of the grid, it is reasonable to expect that the mean distance of the trajectory from the grid will be larger than where they are the same. Now consider the case where the grid has higher dimension than the cluster, for example where the actual cluster lies along a curve and is being approximated by a two-dimensional grid. The mean distance of the trajectory from the closest point of the grid, if there are about as many points, will not be smaller and may be larger, particularly if the grid has not completely converged and still retains any two-dimensionality. We may tentatively conjecture that the mean distance will be a minimum when the two dimensions are equal. This conjecture can be tested by computational experiment on known data sets. In the light of this conjecture, the results in Fig. 11 are of interest, where the values are mean distances in the mel filter bank space.

One might expect that the distances themselves should give some measure of the relative goodness of fit; the obvious complication is the relative sizes of the sub-clusters, as the Kohenen process will put more grid points where there are more cluster points [6]. One might expect, *a priori*, that distances will be less for the cases where the numbers of points are large since that

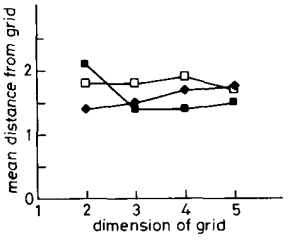

Fig. 11 *Vowels, nasals and fricatives*
—■— vowel (77 data points)
—□— nasal (14 data points)
—◆— fricative (70 data points)

would be where the grid points would have highest density.

The evidence is far from compelling but suggests that the vowels have dimension of around three to four, and is inconclusive about nasals and fricatives.

Finally, we computed the crinkle statistic for those parts of the grid associated with the separate phonemic categories. Each computation takes about a week on a Silicon Graphics IRIS workstation, so there is not a great deal of data, but the results are suggestive and are shown in Fig. 12.

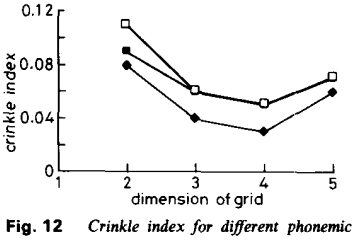

Fig. 12 *Crinkle index for different phonemic categories; filter bank space*
—■— vowel
—□— nasal
—◆— fricative

We have also explored the crinkle statistic in the case of the space of LPC coefficients. Preliminary results give extremely low values for the crinkle statistic in dimension two; this may be an anomally due to small amounts of data, or it may be telling us about the merits of LPC analysis.

4 Conclusions

There is some reason to believe that speech is, overall, inherently either three- or four-dimensional, that is to say, can be described by three or four independent parameters. The ordinary ways of measuring it by FFTs and LPC analysis do not yield an affine embedding of the speech space; the speech cluster can be viewed under various projections (as in Figs. 7–10) and can be seen to have significant nonlinearities. The conclusion that four numbers suffice is obtained by geometric methods of analysis derived from an application of the Kohenen process. What physical variables correspond to these dimensions is not clear, and is evidently a matter of considerable interest.

Attempts to find significant substructure in the space for regions corresponding to vowels, nasals and fricatives

have made it plain that the regions are relatively well delineated under suitable projection, but do not appear to be distinguishable so far as the dimension is concerned. Our methods here are crude, however, and are limited by our resources and the considerable amount of computation required.

5 References

1 ALDER, M.D., TOGNERI, R., and ATTIKIOUZEL, Y.: 'Dimension of the speech space', *IEE Proc. I, Commun. Speech & Vision*, 1991, **138**, (3), pp. 207–214

2 TATTERSALL, G., LYNFORD, P., and LINGGARD, R.: 'Neural arrays for speech recognition', *Brit. Telecom J.*, 1988, **6**, (2), pp. 140–163

3 KOHONEN, T.: 'Self-organisation and associative memory' (Springer Series in Information Science 8, 1988, Springer, Berlin)

4 RABINER, L., and SCHAFER, R.: 'Digital processing of speech signals' (Prentice Hall, Englewood Cliffs, NJ, 1978)

5 PARTHASARATHY, S., and COKER, C.: 'Phoneme parametrisation of speech using an articulatory model'. ICASSP-90

6 ALDER, M., TOGNERI, R., LAI, E., and ATTIKIOUZEL, Y.: 'Kohonen's algorithm for the numerical parametrisation of manifolds', *Patt. Recog. Lett.*, 1990, **11**, pp. 313–319

7 LIBERMANN, P., and MARLE, C.M.: 'Symplectic geometry and analytical mechanics' (Reidel, Dordrecht, 1987)