Affine Transformations of the Speech Space

Mylène Pijpers and Michael D. Alder
Centre for Intelligent Information Processing Systems
Department of Electrical and Electronic Engineering
Department of Mathematics
The University of Western Australia
Nedlands W.A. 6009, AUSTRALIA

Abstract

The papers *Speaker Normalization of static and dynamic vowel spectral features* (J.A.S.A **90**, July 1991 pp 67-75) and *Minimum Mean-Square Error Transformations of Categorical Data to Target Positions* (IEEE Trans Sig.Proc,**40** Jan 1992, pp13-23) by Zahorian and Jagharghi describe an algorithm for transforming the space of speech sounds so as to improve the accuracy of classification. Classification was accomplished by both back-propagation neural nets and by a Bayesian Maximum Likelihood method on the model of each vowel class being specified by a gaussian distribution. The transformation was an affine transformation obtained by choosing ideal 'target' points for each cluster in a second space and minimising the mean square distance of the points in the speech space from the appropriate target. The speech space itself was a space of cepstral coefficients obtained from a Discrete Cosine Transform.

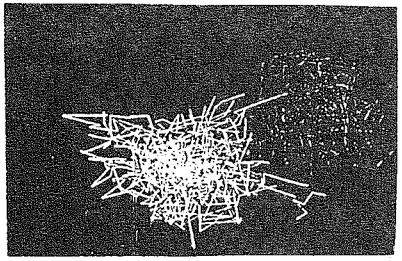
These findings are remarkable, indeed almost unbelievable. The reason is that both the maximum likelihood classification on the gaussian model, and the Neural Net classifier are essentially affine invariant. In the case where the transformation is invertible, this is clearly the case. When the transformation has non-trivial kernel, it may happen that the classification gets worse, but it cannot get better.

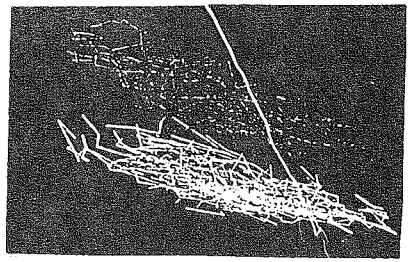
A back-propagation neural net in effect classifies by dividing the space into regions by means of hyperplanes. The gaussian model does so by means of quadratic forms, with quadratic discrimination hypersurfaces. Projecting a hyperplane by any non-zero affine map which is onto the target space will usually give another hyperplane in the target space, and if the second separates points, so will the first. Conversely, if there is a solution in the target space, it can be pulled back to a solution in the domain space. It is not hard to show that similar considerations apply to the case where we use quadratic hypersurfaces.

In this paper, we attempt to account for the results of Zahorian and Jagharghi by investigating vowel data. We describe a simple projection algorithm which may be applied to high dimensional data to give a view on a computer screen of the data and of transformations of it.

1 Introduction

It is well known that the conventional back-propagation algorithm for a three layered neural net performs classification by deciding into which region of a space of measurements a particular observation falls, and that the partition of the space into regions is essentially by means of hyperplanes, each unit in the hidden layer defining one such hyperplane. So if we use a neural net to classify vowel sounds, then distinguishing between /a/ and /u/ (represented in what follows as AA and UX, the conventional Carnegie Mellon nomenclature) then whether we choose to use k Cepstral coefficients or m filterbank values to characterise each utterance, we have a cluster of points representing the AA sound in $\mathcal{R}^k$, $\mathcal{R}^m$ respectively, and a different cluster of points representing the UX sounds. In *Figure 1* and *Figure 2* we show the result of projecting from a twelve dimensional space of simulated filter bank values the results of a number of utterances of each of these sounds: the data was extracted from the TIMIT data base and shows each vowel sound in a different colour, reproduced here as different grey scales. The figures show two different projections obtained by rotating the containing space of dimension 12 between projections.





Figures 1. and 2.
Two different views of 12 dimensional points projected to a 2 dimensional space. The frames of an utterance are connected by lines, showing the trajectory of that vowel. The UX vowel data are printed in grey, the AA vowel data in white.

As might be expected, the clusters form two roughly multivariate gaussian distributions, with some small degree of overlap. Similar clusters are obtained for each distinct vowel although the degree of overlap and the locations of the centres in the space may vary widely. If instead of using filterbank values we were to use cepstral coefficients, we have found in earlier work [5], [6] that a non-linear transformation is applied to the representation space. It remains however true that the vowels still form approximately gaussian clusters, although diphthongs or glides can form rather different shaped clusters.

If we wish to separate the clusters, we could do so by assuming that the clusters are indeed approximately gaussian and computing the covariance matrix for each cluster, giving a quadratic discrimination hypersurface if we use a Bayesian classification scheme based on likelihood ratio. Or we could apply a number of methods to simply place a hyperplane

between the two clusters. A three layer neural net with one unit in the hidden layer accomplishes this.

Multiple classification of many vowel classes may be reduced to pairwise binary classification, and it is therefore unnecessary to consider the problem of distinguishing between more than two vowel utterances.

Although the quadratic discriminant hypersurface would seem to be a better approach, in the case when the covariance matrices for the two clusters are the same, the quadratic reduces to a hyperplane, and in general, it is only in extremal cases that the quadratic hypersurface is sensibly non-affine in the region between the two clusters. It is common therefore to rely on affine methods; these are well known and robust, and easily implemented with the classical neural nets. The solutions obtained by the perceptron convergence algorithm, to which the back-propagation algorithm reduces in the case of a single unit in the hidden layer and a binary output, are optimal when the sets are indeed linearly separable, that is to say, if there is any solution, the algorithm will find one. If the two clusters overlap however, all that can be said is that the *mean* position of the solution hyperplane is optimal.

The papers [1] and [2] make the claim that an affine transformation of the vowel space can improve the effectiveness of classification of a neural net. Now this is a very surprising result and on the face of things impossible. If two clusters are linearly separable, then there is by definition a hyperplane which separates them. Any affine transformation which is 1-1 can only take this separating hyperplane to another such separating hyperplane. If an affine transformation is not 1-1, then it has a non-trivial kernel, by elementary linear algebra. The separating hyperplane has an orthogonal complement which determines a unique subspace of dimension 1, and if this subspace is in the kernel of the linear part of the transformation, the separating hyperplane is 'killed' by the transformation. Separation of the image of the two clusters by the transformation by some new hyperplane in the image space may or may not be possible. The MSECT algorithm described in [1] and [2] requires the user to assign to each cluster some suitable 'target centre' in $\mathcal{R}^m$ for any choice of m. It then finds the affine transformation, unique if the target space has lower dimension than the domain space, which minimises the sum of squared distances of each point from its intended target. The declared intention is to use it as a pre-processor for a neural net or other such classification system.

2 Mathematical Issues: Paradox Found.

If there is a separating hyperplane U in the image space, and if

$$T : \mathcal{R}^n \longrightarrow \mathcal{R}^m$$

is the affine map, then $T^{-1}U$ must have dimension $n-1$ since we may easily construct a projection p from $\mathcal{R}^m$ to $\mathcal{R}$ which has U as kernel and has rank 1. If we now consider the composite $p \circ T$ we see that it has rank 1, since T is onto, and kernel $T^{-1}U$, and the result follows from the rank nullity theorem. Moreover it is immediate that if the map p takes one cluster of points to positive values and the other to negative values, i.e. U separates the clusters, then so does $p \circ T$ i.e. $T^{-1}U$ separates the clusters too.

If the clusters are not linearly separable, but some subset is, then the same argument applies to the subset and moreover to the largest such subset. Hence any solution to a separation by neural nets in some second space reached by an affine transformation onto the second space from a starting space, gives immediately a solution in the starting space. So how can any such transformation improve matters?

3 Experimental testing: Paradox confirmed.

In order to investigate the claim that it mysteriously does, we extracted from the TIMIT data base a set of pure vowels, labelled AA (/a/), AE (/æ/), AH (/ʌ/), AO (/o/), EH (/ɛ/), IH (/I/), IY (/iʸ/) and UX (/u/). These were transformed into trajectories in $\mathcal{R}^{12}$ by taking a simulated filter bank obtained by binning the results of an FFT into 12 mel spaced bins as described in [3], [4]. In order to confirm the appropriateness of the analysis, we investigated the vowel data visually by projection methods: random rotations of the vowel space were performed while a projection to the screen of the computer was imposed on the result. This allows the user to see the clusters corresponding to the trajectories. We chose different colours for each vowel sound.

We then trained a neural net with one unit in the hidden layer to separate out the two clusters. This was repeated with different pairs of vowels. Results are shown in table 1. , with two repetitions of the training and classification process.

	Original Data	Original Data	Transformed Data	Transformed Data
Vowels	Percentage correct on training data	Percentage correct on test data	Percentage correct on training data	Percentage correct on test data
AA UX	99.50 %	99.13 %	100.00 %	99.68 %
	100.00 %	98.97 %	100.00 %	99.68 %
AA AE	87.00 %	84.49 %	90.50 %	92.98 %
	87.00 %	84.28 %	94.50 %	92.30 %
EH IH	82.50 %	74.93 %	81.50 %	79.86 %
	86.50 %	73.04 %	83.00 %	83.04 %
AH EH	83.00 %	70.55 %	86.00 %	82.21 %
	85.50 %	80.58 %	87.50 %	80.95 %
AA IH	99.00 %	98.00 %	99.00 %	98.87 %
	99.50 %	96.61 %	99.50 %	98.70 %

Table 1. Separation of 2 vowel data clusters by a NN. The original data are in 12 dimensions, the transformed data in 2 dimensions.

Next we applied the MSECT algorithm as described in [1] and [2] to the original 12 dimensional data. We did this several ways, one by choosing 'natural' points in a two and three dimensional space for the clusters to get sent to. Some of these results are shown in table 2. finally, we settled on doing the process for two vowels at a time into the real line, sending one vowel to 1 and the other to -1. The same neural net was then applied to the one

dimensional data in order to classify the points representing vowel trajectories, now spread out along the real line. The results are shown in table 1.

It may be seen that the results of Zahorian and Jagharghi are confirmed. Applying MSECT to the vowel data can increase the percentage of points correctly classified. It is true that the dimension reduction carries a cost and that it is high when we reduce the 8 clusters to two or three dimensions. On the other hand, when we simply take two clusters, two vowels, and try to separate them with a hyperplane, it is more effective to project onto a line and then separate with a single point, than it is to separate directly in the domain space. The neural net ought to be finding the hyperplane which gets sent to that single point by the MSECT algorithm. But it generally doesn't.

Data Set	Percentage correct on testdata	Percentage correct on training data
Original Data (12D)	48.73 %	50.75 %
	40.19 %	51.88 %
Transformed to 2D	46.47 %	47.50 %
	54.62 %	52.00 %
Transformed to 3D	44.43 %	45.12 %
	57.07 %	53.88 %

Table 2. Test results of classification of 8 vowel data clusters projected from 12 dimensions to 2D and 3D.

4 Paradox tentatively explained.

We have a number of explanations for the observed phenomena.

It is reasonable to suppose that the clusters are reasonably well approximated by gaussian distributions in dimension 12. The trajectories are highly autocorrelated, as can be seen by eye in *Figures 1 and 2* but ignoring this consideration, the points appear to be rasonably well described by gaussian clusters, and we shall make that assumption in what follows.

First, it is not hard to see that the number of moves taken in order to get convergence in the case of a linearly separable data set will generally increase with the dimension. The same can be expected when the sets are two gaussians with some degree of overlap. It may be therefore that simply running the Neural Net program for longer will decrease the discrepancy between the 12 dimensional and the 1 dimensional solutions. Experiments show that this generally appears to be the case.(See table 3.) Second, the existence of local minima when the data is not linearly separable is easily seen to be a minor problem in the case where the dimension is low. Cases where the neural net failed to yield any reasonable classification at all occur in dimension 12. They do not occur in dimension 1. Nor could they; it is easy to see that the solution hyperplane (a point in dimension 1) must oscillate between the largest and smallest of the actual locations of the minimum error locations, with some overlap which depends on the step size.

Number of iterations	*AA AE* original	*AA AE* transformed	*IY UX* original	*IY UX* transformed
20 000	84.96 %	92.63 %	79.30 %	82.22 %
	84.16 %	91.53 %	75.52 %	81.36 %
40 000	88.70 %	91.15 %	76.66 %	81.54 %
	88.23 %	92.67 %	77.57 %	81.43 %
60 000	87.74 %	92.86 %	67.34 %	81.72 %
	89.87 %	92.23 %	79.17 %	81.80 %
100 000	89.49 %	92.82 %	68.91 %	82.33 %
	88.98 %	92.39 %	78.82 %	82.28 %

Table 3. The mean percentages correct on testing for different amounts of training for 2 vowel pairs.

5 References

[1] Zahorian S.A. and Jagharghi A.J. *Speaker normalization of static and dynamic vowel spectral features* J.Acoust.Soc.Am.**90**(1), July 1991 pp67-75.

[2] Zahorian S.A. and Jagharghi A.J. *Minimum Mean-Square Error Transformations of Categorical Data to Target Positions* IEEE Trans. Sig.Proc. Vol 40, No. 1 January 1992 pp 13-23.

[3] Zhang, Y., deSilva, C. J. S., Togneri, R., Alder, M. D. and Attikiouzel, Y. (1992) *A Multi-HMM Isolated Word Recognizer*, Proceedings SST-'92 pp ??

[4] Zhang, Y., deSilva, C. J. S., Attikiouzel, Y. and Alder, M. D. (1992) *A HMM/EM Speaker-Independent Isolated Word Recognizer*, accepted for publication by *The Journal of Electrical and Electronic Engineers, Australia,*.

[5] Alder, M., Togneri R., and Attikiouzel J. *Dimension of the Speech Space*, IEE Proceedings-I, Vol. 138, No.3, June 1991 pp.207-214.

[6] Alder, M., Togneri R., and Attikiouzel Y. *Dimension and Structure of the Speech Space*, IEE Comm, Speech and Vision, April 1992 pp. 123-127.