AN UNSUPERVISED ALGORITHM FOR THE EXTRACTION OF FORMANT-LIKE FEATURES FROM LPC-CEPSTRAL SPACE

Simon Hawkins, Iain Macleod, and Bruce Millar

Computer Sciences Laboratory, Research School of Information Sciences and Engineering,
Australian National University

ABSTRACT - This study develops an unsupervised algorithm for extracting the perceptual dimensions of vowel backness (a correlate of F2) and vowel height (a correlate of F1) from the LPC-cepstral representation of a speaker's vocalic system.

INTRODUCTION

There is evidence that monophthongal vowels are perceived in terms of multi-valued or continuous dimensions which relate to the perceived backness of the vowel, the perceived height of the vowel and the duration of the vowel. Although the duration of a vowel can be measured in a straightforward fashion, measurements of the perceived backness and height of a vowel are much more difficult to obtain. These dimensions can only be accurately measured by performing a perceptual experiment in which listeners are required to perceptually discriminate between vowel segments of equal length. Multi-dimensional scaling can then be applied to the perceptual confusions between vowel pairs made by listeners in order to construct a perceptual space that reflects the perceptual discriminability of vowel segments of equal duration. Vowels that are easily discriminated will lie far apart in this space whereas vowels that are easily confused will lie close together. The first perceptual dimension, along which vowels show the greatest variance, relates to the perceived backness of a vowel (Pols et al, 1969; Klein et al, 1970). The second dimension, along which vowels display the next most variance, is orthogonal to the first and relates to the perceived height of a vowel. The third perceptual dimension, which is orthogonal to the first two, is more difficult to interpret but it seems to provide a degree of discrimination between rounded and unrounded vowels, at least for the vowel systems of some languages (Pols et al, 1969; Klein et al, 1970).

Although an **accurate** measure of the perceived backness and height of a vowel can only be obtained empirically, **approximate** estimates of perceived vowel backness and height can be obtained using the formants of the vowel. The perceived backness of a vowel is related to the frequency difference between the first and second formants of the vowel as measured on the perceptually relevant bark scale (Ladefoged, 1982; Ladefoged and Maddieson, 1989). An approximate estimate of the perceived height of a vowel can be obtained from the frequency of the first formant (Ladefoged, 1982; Ladefoged and Maddieson, 1989). It makes no difference whether the frequency of the first formant is measured on the Hertz, bark, or mel scales because these scales are linearly related in the frequency range in which the first vowel formant is found (ie., 200 - 800 Hz).

However, the well-known problems associated with extracting reliable formant estimates make this method of deriving estimates of vowel backness and height rather unattractive. It would be valuable if we could develop a robust method for estimating the perceived backness and height of a vowel that was not empirically based and did not require the estimation of formants.

SUPERVISED EXTRACTION OF PERCEPTUAL DIMENSIONS FROM FILTERBANK SPACE

Klein et al (1970) developed a supervised method for extracting the dimensions of vowel backness and height from the physical representation of vowel spectra in terms of the log energy levels in a bank of eighteen 1/3-octave bandpass filters. Their work applied to the steady-state vowel targets of the 12 Dutch monophthongal vowels as pronounced by 50 male speakers in the context /h-vowel-t/. Their computations were based on one-third octave spectra of 100 ms segments taken from the vowel targets. Using the 18 frequency bands, the 12 x 50 vowel sounds created a cloud of 600 points in the 18-dimensional space. Principal Components Analysis (PCA) was used to rotate these eighteen axes such that the first new dimension explained as much as possible of the original variance, the second dimension explained as much as possible of the variance left unexplained by the first, and so on. In this way, a subspace was derived that explained the maximum total variance with a minimum number of dimensions. The results showed that the first two Principal Components (PCs) explained 61% of the total variance for the 600 vowel segments and that this percentage increased to 83% for the 12 average vowel points. This result suggests that a two-dimensional subspace (ie, a plane) will give a reasonably accurate approximation of the spectral difference among average vowel spectra.

Klein et al also calculated the perceptual dissimilarities between the same 600 vowel segments used in the analysis of vowel spectra dissimilarities. The same 50 speakers used in the vowel spectra study were required to listen to a series of vowel segments of 100 ms duration and to identify which of the 12 Dutch vowels they thought they had heard. Kruskal's multidimensional scaling technique was used to construct a spatial representation in which the distance between vowels corresponded to their perceptual discriminability. The only criterion in this technique is a monotonic relationship between interpoint distances and their corresponding dissimilarity indices. The extent to which this relation is violated for a particular number of dimensions is expressed by a "stress" percentage. Kruskal's technique, applied to the confusion matrix, produced a configuration in one dimension with a stress of 21.2% , a stress of 7.7% in two dimensions, a stress of 4.7% in three dimensions, and a stress of 2.3% in four dimensions. This suggests that the underlying perceptual configuration is at least two dimensional but may well be three or four dimensional. The first of these perceptual dimensions correlated highly with F2, suggesting that it was a measure of vowel backness. The second of the dimensions correlated highly with F1, suggesting that it was a measure of vowel height.

Having obtained a configuration of vowel points representing perceived vowel **timbre**, Klein et al rotated the four-dimensional principal component reduction of the sound spectrum measured in one-third octave bands in a trial-by-error fashion until it matched the four-dimensional perceptual configuration of the 12 vowels. After being rotated to congruence, there was a close correlation between the first three dimensions of the four-dimensional representation of vowel timbre and the first three dimensions of the four-dimensional physical representation. This demonstrated that an **appropriate rotation** of the four-dimensional **physical** representation of vowel spectra will generate an excellent first-order approximation of dissimilarity in the **perceived timbre** of a vowel.

We wished to compare the *shape of the surfaces* formed by the average Dutch vowels in physical space and in perceptual space. Our interest lay in the shape of the vowel surface in physical space *after it had been rotated to congruence* with a four-dimensional *perceptual* subspace. To visualise these surfaces, we considered only the first three dimensions of each space. This is acceptable because the fourth dimension in both the physical and perceptual space was found to contribute little to the total variance of the Dutch vowel system (5.8% and 6.2% respectively of the total variance of the 600 vowel points).

The physical subspace was derived from a graphical presentation of this space by Klein et al (1970, Fig. 4). The four-dimensional subspace was obtained by applying Kruskal's multidimensional scaling method to the matrix of dissimilarities between vowels pairs reported by Klein et al (1970, Table V). This was the same approach used by Klein et al (1970) to to derive the four-dimensional perceptual space that is neither graphed nor tabulated in this work.

We found high correlations between the first and second dimension of the physical representation and the first and second perceptual dimensions (r = 0.94413 and r=0.93174 respectively). There were also high correlations between the first and second physical dimensions and the second and first formants (r=0.96921 and r=0.85414 respectively). This finding suggests that the first and second dimensions of the three-dimensional physical configuration were providing an approximate measure of vowel backness and vowel height. The poor correlation between the third dimension of physical and perceptual space suggests that it is only possible to extract a **two-dimensional perceptual** representation from a **three-dimensional physical** representation.

TABLE 1. Correlation (N=12 average Dutch vowels) between the 3D perceptual space (D1, D2, D3), 3D physical space (X,Y,Z) and F1 and F2.

	PHYSICAL SPACE			PERCEPTUAL SPACE	
	X	Y	Z	D1	D2
D1	0.94413	-0.03627	-0.2796	-0.51359	0.89315
D2	-0.20101	0.93174	-0.2272	0.75029	-0.16901
D3	0.23396	0.324	0.2299		
F1	-0.53942	0.85414	0.10295		
F2	0.96921	-0.07230	0.03247		

UNSUPERVISED EXTRACTION OF PERCEPTUAL DIMENSIONS FROM CEPSTRAL SPACE

The principal components derived from the physical representation of vowel spectra are not laws of nature. These components will not necessarily be more theorectically meaningful than any other linear combination of log energy levels. Consequently, many researchers prefer to use components to reduce dimensionality by eliminating neglible variation, and then rotate once more in the smaller dimensional space to achieve some meaningful criterion (cf. Wilkinson, 1989, p77). It is also important to understand that the signs of loadings within components are arbitrary. Relative signs of loadings are artifactual and not theoretically significant. We can therefore rotate a factor 180 degrees by changing negative loadings to positive loadings and positive loadings to negative loadings (cf. Wilkinson, 1989).

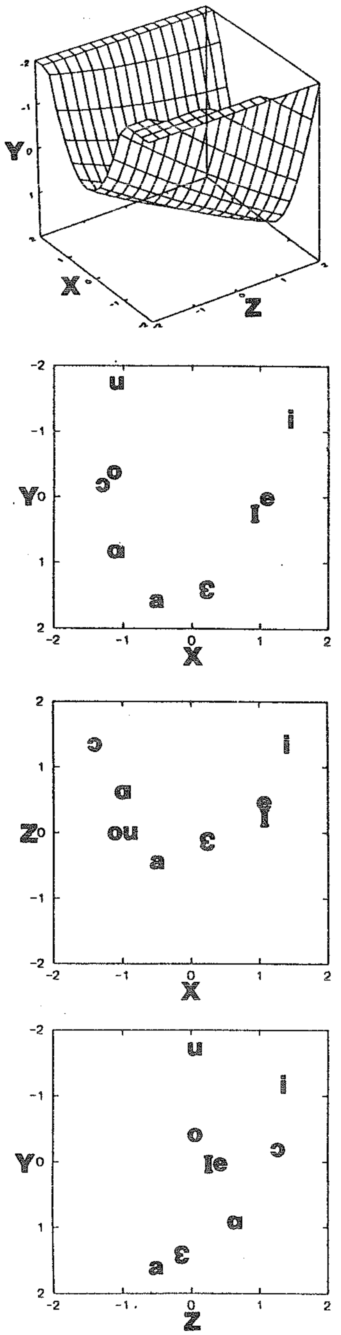

FIG.1 **PHYSICAL SPACE** (Klein et al) : Plot of the 12 Dutch vowels, excluding the rounded front vowels, in 3D physical space

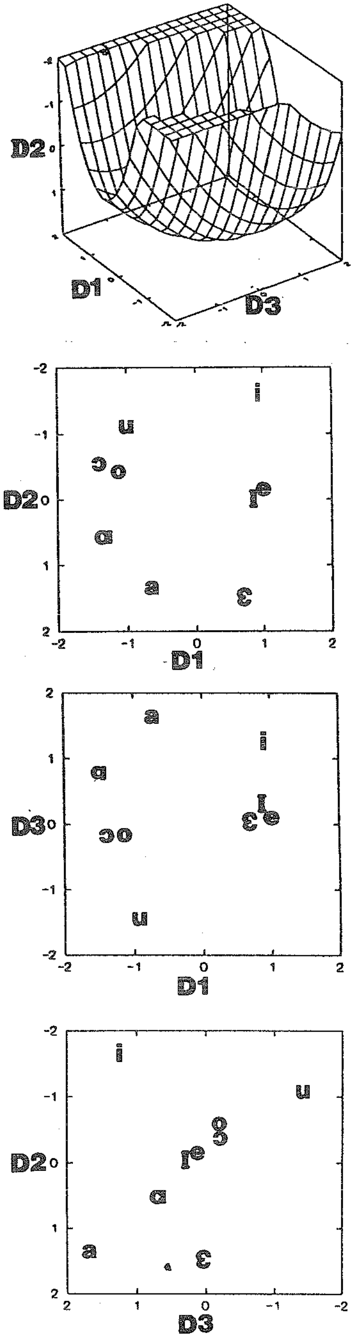

FIG.2 **PERCEPTUAL SPACE** (Klein et al): Plot of the 12 Dutch vowels, excluding the rounded front vowels, in 3D perceptual space

Our simplified version of the work by Klein et al (1970) demonstrates the *possibility* of rotating a three Principal Component representation of average vowel spectra such that the first two rotated Principal Components correlate strongly with the perceptual dimensions of vowel backness and vowel height. This was achieved by Klein et al using a trial-by-error approach to rotate the 4D physical representation of the12 average Dutch vowel spectra to congruence with a 4D perceptual representation. Careful observation of the *structure of the vowel surface in the first three dimesnions* of the physical subspace *after it had been rotated to congruence* (see Figure 1) suggests that the required rotation could be achieved in an *unsupervised* fashion. When 4D physical space had been rotated to congruence with 4D perceptual space, the Dutch vowel spectra in the first three dimensions of the physical space formed a surface that is well described by the four-term equation for a parabolic surface with tilt,

$$\hat{Y} = 2.00X^2 + 0.098X + 1.801 + 1.027Z \qquad \text{(Eqn. 1)}$$

where X, Y and Z are the first, second and third rotated principal components. This vowel surface is graphed in the first three dimensions (X-Y-Z) of the physical space in Figure 1 (and can be compared with the vowel surface in the first three dimensions of perceptual space (D1-D2-D3) in Figure 2). There was a correlation of 0.705 between estimates of Y derived using the equation and actual Y coordinates. The equation therefore represents the shape of the vowel surface reasonably well. This equation describes a parabolic surface with an axis of symmetry that lies perpendicular to the first rotated principal component (X). The parabolic surface is tilted such that the projection of vowels onto the plane formed by the second and third rotated principal components (the Y-Z plane) lie at an angle of about 45 degrees. In Equation (1) above, this 45 degree orientation of the vowel distribution in the Y-Z plane is captured by the relationship,

$$\hat{Y} = 1.027Z \qquad \text{(Eqn. 2)}$$

When the vowel surface in the physical representation of vowel spectra lies ***in this particular orientation***, there is a strong correlation between the first rotated principal component (X) and perceived vowel backness(D1) and between the second rotated principal component (Y) and perceived vowel height (D2). It may be possible to achieve these high correlations in an unsupervised fashion by rotating the parabolic vowel surface in the 3D physical representation to achieve *two criteria:*

- first, to maximise the parabolic shape of the vowel distribution in the Y-X plane, and
- second, to ensure that the vowel distribution in the Y-Z plane is oriented at 45 degrees.

DEVELOPMENT OF AN UNSUPERVISED ALGORITHM IN LPC-CEPSTRAL SPACE

These ideas were used to develop an unsupervised algorithm for extracting these two perceptual dimensions of vowel backness and vowel height from the LPC-cepstral representation of a speaker's vowel system. These dimensions correlate strongly with F2 and F1 respectively.

It is certainly possible to apply the **supervised** extraction technique developed by Klein et al using the *filterbank* representation of vowel spectra to the representation of vowel spectra in terms of *cepstral coefficients.* There is a linear relationship between the cepstral representation of a vowel system and its representation in terms of the log energy levels in a bank of filters. This means that it should be possible to find dimensions in LPC-cepstral space that correspond to the perceptual dimensions of vowel backness and vowel height that Klein et al extracted from the filterbank representation of a vowel system.

However, rather than using the supervised extraction technique of Klein et al, we developed an **unsupervised** technique that does not require a priori knowledge of the perceptual representation of a vocalic system. To develop this algorithm, we studied the structure of the vowel distribution in the first three dimensions of a four-dimension physical representation of the average Dutch vowel spectra This four-dimensional physical space had been rotated to congruence with the four-dimensional perceptual representation of the same vowels (see Figure 1). The Dutch vowels formed a parabolic surface in the subspace formed by the first three principal components of the filterbank representation. This parabolic surface had an axis of symmetry that lay at right angles to the first rotated principal component. The axis of symmetry of the surface was tilted at a 45 degree angle to the second principal component.

We can characterise the orientation of the vowel surface within this space by projecting vowels onto the Y-X and Y-Z facets of this space. When vowels were projected onto the Y-X facet formed by the first and second principal components, they formed a distribution that was shaped like a parabola. This parabola had an axis of symmetry that lay at right angles to the first principal component. When the vowels were projected onto the Y-Z facet formed by the second and third principal components, the vowels formed a wedge-shaped distribution that lay at 45 degrees to the Y axis.

A speaker's vowels form a parabolic surface in the space formed by the log energy levels in a bank of filters. They must form a surface of the same shape in *cepstral space* because the cepstral and filterbank representations are linearly related. In another paper , Hawkins et al (1994a) demonstrate that a speaker's vocalic distribution does in fact form a parabolic surface in the subspace formed by the first three principal components of the cepstral representation of a speaker's vocalic system.

To transform the first three principal components of LPC Cepstral space into perceptual dimensions, we must reorient the parabolic vocalic surface. *In particular,* we must rotate this three-dimensional subspace until the projection of vowels on the various facets of the space looks the same as it did in the experiment by Klein et al (1970) when physical space had been rotated to congruence with perceptual space (see Figure 1). The projection of vowels onto the Y-X facet formed by the first and second principal components should look like a parabola that has an axis of symmetry lying perpendicular to the first rotated principal component X. If we call the three principal components X, Y, and Z, then this vowel distribution is easily achieved by rotating the X-Y plane about the Z axis to maximise the correlation *r*

$$r(Y',\hat{Y}' = aX'^2 + bX' + c) \qquad \text{(Eqn. 3)}$$

The projection of vowels onto the Y-Z facet formed by the second and third principal components forms a wedge-shaped distribution that lies at approximately 45 degrees to the Y axis. This orientation can be achieved by rotating the Y-Z plane anticlockwise 45 degrees. In practice, a further 180 degree rotation is required to reverse the polarity of the principal component Y to match it to the polarity of the second dimension (D2) of perceptual space. When these two rotations have been performed, the parabolic vowel surface in principal component space should lie in an orientation that will maximise the correlation between the first principal component and the perceptual dimension of perceived vowel backness and the between the second principal component and and perceived vowel height.

TABLE 1. Correlations are between the PCs of LPC-cepstral space and F1 and F2 (1) prior to any rotation, (2) after rotation about the Z axis and (3) after a further rotation about the X axis. The rotation angle of the X-Y plane about the Z axis (in degrees) that maximised the correlation in Equation (3) is shown.

	% total variance			Unrotated		After rotation about Z			After rotation about X		
Sp No	PC1	PC2	PC3	r(X,F2)	r(Y,F1)	angle	r(X,F2)	r(Y,F1)	angle	r(X,F2)	r(Y,F1)
01	43.3	22.3	10.4	0.6832	0.4951	131	0.8579	0.7644	225	0.8579	0.9032
04	44.6	25.8	8.7	0.8271	0.5584	326	0.9488	0.8102	225	0.9488	0.8923
10	60.4	15.8	9.0	0.9287	.5874	315	0.9661	0.6654	225	0.9661	0.8894
11	50.1	20.7	8.0	0.9303	0.6644	329	0.9523	0.8480	225	0.9523	0.8994
12	48.7	20.1	8.3	0.8290	0.5357	217	0.8922	0.8060	225	0.8922	0.9159
13	46.4	21.7	8.5	0.7684	0.3798	323	0.8805	0.7797	225	0.8805	0.7684
19	41.1	23.0	12.1	0.7617	0.6414	317	0.8997	0.8272	225	0.8997	0.7501
20	46.5	25.8	7.3	0.5725	0.4335	131	0.8137	0.8506	225	0.8137	0.8259
24	49.5	18.1	9.4	0.7666	0.3601	226	0.8924	0.8684	225	0.8924	0.9248
30	42.3	27.7	7.8	0.8116	0.6411	209	0.8833	0.7785	225	0.8833	0.8786
31	39.5	35.9	9.4	0.3066	0.4886	277	0.8643	0.6086	225	0.8643	0.7275
34	45.2	22.7	9.3	0.7355	0.5323	131	0.9167	0.8403	225	0.9167	0.9027
35	47.8	16.3	13.1	0.8457	0.4561	320	0.9431	0.7676	225	0.9431	0.9400
36	43.5	22.1	11.8	0.7666	0.3601	11	0.9272	0.6649	225	0.9272	0.8849
M	46.4	22.0	9.5	0.75244	0.5096	234	0.9027	0.7771	225	0.9027	0.8645

PERFORMANCE OF THE UNSUPERVISED ALGORITHM

To test this unsupervised algorithm, we examined speech frames taken from the vocalic nuclei of the 19 Australian English vowels as uttered by 14 male speakers in /h-vowel-d/ context. The speech data is described in detail in Hawkins et al (1994a). For each of the 14 speakers, the first three principal components of the LPC-Cepstral representation of the speaker's 19 vocalic nuclei were calculated. The speaker's speech frames were then plotted as a cloud of points in the 3D space formed by the first three principal components of the cepstral representation (which were labelled X, Y and Z).

The Y-X plane was then rotated about the axis formed by the Z axis to maximise the correlation specified by Equation (3). After this rotation, F2 always correlated strongly with the first rotated principal component X' and F1 with the second rotated principal component Y'. The correlation between the X' and F2 increased from an average across the 14 speakers of 0.75244 prior to rotation to an average of 0.9027 after rotation. The correlation with our measure of vowel backness (ie. F2-F1 bark) was about the same at 0.9074.

The Y-Z plane was then rotated 225 degrees anticlockwise about the X axis. This rotation did not affect X'. After this rotation, F1 always correlated strongly with Y' but with some speakers, there was a slightly stronger correlation with the third rotated principal component Z'. Prior to any rotation, the correlation between Y' and F1 was only 0.5096 on average across speakers. This increased to 0.7771 after the first rotation and further increased to 0.8645 after the final rotation. These findings indicate that this unsupervised algorithm is capable of automatically rotating a speaker's vocalic surface within the space formed by the first three principal components of LPC-cepstral space in order to extract correlates of F1 and F2 (or perceived vowel height and vowel backness).

TABLE 2. Correlation of F1, F2 and vowel backness, as measured in Hertz, mel, and bark, with the first and second rotated PCs (X' and Y') for each of 14male speakers

							vowel backness		
	r (F1, Y')			r (F2, X ')			r (F2-F1 , X')		
Sp No	linear	Mel	Bark	linear	Mel	Bark	linear	Mel	Bark
	.9032	.8955	.8945	.8579	.8493	.8418	.8447	.8689	.8743
04	.8900	.8909	.8911	.9456	.9407	.9359	.9584	.9702	.9711
10	.8894	.8847	.8841	.9661	.9612	.9566	.9492	.9539	.9534
11	.8983	.8933	.8927	.9438	.9499	.9496	.8989	.9129	.9155
12	.9159	.9115	.9109	.8921	.8953	.8934	.8603	.8898	.8969
13	.7684	.7705	.7706	.8805	.8750	.8696	.9094	.9201	.9193
19	.7501	.7527	.7535	.8997	.9026	.8997	.8660	.8876	.8924
20	.8259	.8350	.8359	.8137	.7862	.7719	.8474	.8353	.8297
24	.9248	.9234	.9236	.8924	.8805	.8635	.9064	.9174	.9181
30	.8786	.8760	.8761	.8833	.8783	.8726	.8565	.8717	.8739
31	.7275	.7324	.7333	.8642	.8482	.8393	.9430	.9475	.9466
34	.9027	.9037	.90486	.9166	.9104	.9042	.8771	.8865	.8875
35	.9400	.9370	.9364	.9431	.9329	.9256	.9430	.9447	.9423
36	.8849	.8807	.8805	.9272	.9185	.9108	.8839	.8859	.8827
M	.8645	.8634	.8634	.9027	.8949	.8882	.8960	.9066	.9074

CONCLUSION

Hawkins et al (1994a) demonstrated that a speaker's vocalic nuclei lie on a parabolic surface in the 3D space formed by the first three principal components of LPC-cepstral space. This robust finding applied to all but one of the speakers examined. Even the exceptional speaker had a vowel surface which did not deviate greatly from this shape. In a second paper, Hawkins et al (1994b) demonstrated that the relation between LPC-cepstral space and formant space is approximately linear for the first and second formants but nonlinear for the third formant. This raised the possibility of finding a plane in a speaker's LPC cepstral space that approximately correlates with F1 and F2 but precludes the possibility of finding an F3 correlate. In the current paper, we have developed an unsupervised algorithm that is capable of extracting this plane from the LPC-cepstral representation of a speaker's vocalic system in an unsupervised fashion.

REFERENCES

Hawkins, S., Macleod, I. & Millar, B. (1994a). Modelling individual speaker characteristics by describing the distribution of a speaker's vowels in articulatory, cepstral, and formant space",
Proc. 5th International Conf. on Speech, Science, and Technology, Perth, to appear.

Hawkins, S. , Macleod, I. & Millar, B. (1994b). An ab initio analysis of relationships between cepstral and formant spaces, Proc. 5th International Conf. on Speech, Science and Technology, Perth, to appear.

Klein, W. , Plomp, R. & Pols, L.C.W. (1970). "Vowel spectra, vowel spaces, and vowel identification", J. Acoustical Society of America, 48, 999-1008.

Ladefoged, P. (1982). A course in phonetics. Orlando: Harcourt Brace Jovanovitch.

Ladefoged,P. & Maddieson, I. (1989). "Vowels of the world's languages", J. of Phonetics,18, 93-122.

Pols, L.C.W., and van der Kamp, L.J., Th., and Plomp, R. (1969). "Perceptual and physical space of vowel sounds", J. Acoust. Soc. Am., 46, 458-467.

Wilkinson, L. (1989) SYSTAT: The system for statistics. Evanston, IL: SYSTAT, Inc, 1989.