Vowel spectra, vowel spaces and vowel identification

Klein, W.; Plomp, R.; Pols, L.C.W.

Published in:
The Journal of the Acoustical Society of America

DOI:
10.1121/1.1912239

Received 15 April 1970

9.6, 9.5

Vowel Spectra, Vowel Spaces, and Vowel Identification

W. KLEIN,* R. PLOMP, AND L. C. W. POLS

Institute for Perception RVO-TNO, Soesterberg, The Netherlands

Twelve Dutch vowels, each pronounced by 50 male speakers, were analyzed in 18 filter bands comparable in bandwidth with the ear's critical band. By considering the sound levels (in decibels) in these filter bands as dimensions, with a principal-component analysis the 18 dimensions per sound were reduced to four factors which together explain 75% of the total variance. The configuration of the average vowels in the factor space appeared to be highly correlated with their configuration in the F_1–F_2 formant plane. After matching to maximal congruence, correlation coefficients along corresponding axes were 0.997 and 0.979. Machine vowel identification, based upon the position of the individual vowels in the four-dimensional factor space, resulted (after three pairs of related vowels were grouped together) in 98% correct identifications if correction was applied for personal timbre of the speakers' voices. Ten listeners, to whom the 600 vowels were presented as 100-msec segments, gave 86% correct responses in identifying the intended vowels. The confusions between the vowel types were basis for a multidimensional scaling (Kruskal) to construct a perceptual configuration of the vowels. In four dimensions the solution showed 2.3% stress. Perceptual configuration and factor configuration, maximally matched, had correlation coefficients along corresponding axes of 0.997, 0.995, 0.907, and 0.794, respectively.

INTRODUCTION

The differences between vowel spectra are usually described in terms of the formant frequencies F_1 and F_2. In a previous paper[1] a more general approach, consisting of a multivariate analysis of the sound levels in $\frac{1}{3}$-oct frequency bands, was introduced. It appeared that the 15 Dutch vowels investigated could be represented by a configuration of 15 points in a four-dimensional space in which the distance between any two points is a measure of the spectrum difference between the corresponding two vowels. This configuration will be called *factor configuration*, as the four dimensions can be regarded as the principal factors accounting for the vowel differences. More recently, it was shown[2] that such a factor configuration is in excellent agreement with the configuration of the same sounds in a perceptual space derived from listening experiments (triadic comparisons).

The present paper, based on data from 50 male speakers, is an extension of the first one, in which only 10 speakers were employed. This extension appeared to be desirable for the following reasons: (1) to determine the average points and the spread of the individual points as representative of the Dutch vowels, pronounced by male speakers; (2) to provide data in order to test techniques for machine identification of spoken vowels based on their position in the factor space. As the position of a vowel in the factor space can be determined easily, even in running speech, such a technique would have great advantage to any technique based on formants.

Attention will be paid also to the correlation between the factor space and the formant plane of F_1 vs F_2, to the relation between the identification of the individual vowels by human observers and by the machine, and to the effect of bandwidth.

I. VOWEL SPECTRA

The vowel spectra were determined in a way rather similar to the one described earlier.[1] Summarizing this technique, it consisted of the following successive steps:

(1) Each subject pronounced in a nonreverberant room 12 vowels in the context /*h*(vowel)*t*/. The first and second columns of Table I represent these words as written in Dutch and the vowels in phonetic symbols adopted from the IPA,[3] respectively. The 50 speakers were young male adults with a pronunciation representative of correct Dutch.

(2) From each of the recorded words a 100-msec segment out of the beginning of the vowel was singled out by a relay-controlled gate. The gate opened at the moment that the over-all sound-pressure level (SPL)

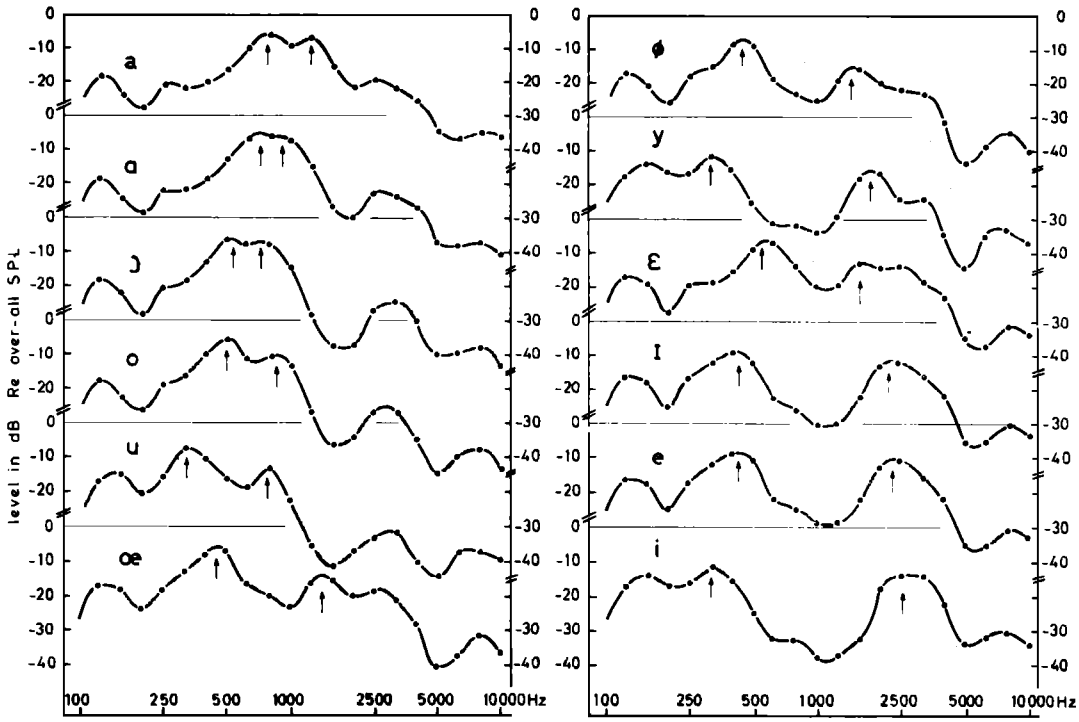

FIG. 1. Average $\frac{1}{3}$-oct spectra of the 12 Dutch vowels used.

passed a fixed value; by means of an oscilloscope, it was checked visually that the obtained 100-msec segment was within the constant vowel part.

(3) The frequency spectra of the vowel segments were measured with a set of $\frac{1}{3}$-oct bandpass filters (Brüel & Kjær spectrometer, model 2112) and recorded by a level recorder (Brüel & Kjær, model 2304). The 21 $\frac{1}{3}$-oct bands ranged from 100 to 10 000 Hz. The outputs of the band filters with center frequencies of 100, 125, and 160 Hz were added energetically (by computation) and were represented by one number; the same was done for the band filters with center frequencies of 200 and 250 Hz. The aim of these combinations was to reduce the influence of differences in voice pitch on the low-frequency data and to apply bandwidths comparable with the ear's critical bandwidths. In this way, the total number of frequency bands was reduced to 18.

(4) In order to correct for differences in the over-all SPL of the vowels, the output levels of the 18 bands were subtracted from the over-all SPL of that particular vowel. In this way, for every individual vowel a series of 18 numbers was obtained representing for the 18 filter bands the sound level in decibels below over-all SPL. As 12 vowels and 50 speakers were involved, 600 series of 18 numbers became available as a basis for further calculations. The reason why decibel values were used in the calculations of, for instance, averages and variances is that this logarithmic measure is a fair approximation of how the hearing organ evaluates sound-pressure differences (doubling in loudness agrees over a large range with 9 to 10 dB difference in SPL).

TABLE I. Symbols and formant frequencies of the 12 Dutch vowels used.

	12 Dutch vowels	Symbols after IPA[a]	Average F_1 and F_2 in Hz		F_1 and F_2 in Hz after Meinsma	
			F_{1a}	F_{2a}	F_{1b}	F_{2b}
1	h*aa*t	/a/	790	1250	730	1350
2	h*a*t	/ɑ/	710	900	700	1300
3	h*o*t	/ɔ/	530	720	380	750
4	h*oo*t	/o/	500	820	410	700
5	h*oe*t	/u/	320	750	300	700
6	h*u*t	/œ/	460	1400	410	1800
7	h*eu*t	/ø/	440	1500	390	1800
8	h*uu*t	/y/	300	1800	300	2000
9	h*e*t	/ɛ/	560	1600	630	1950
10	h*i*t	/ɪ/	420	2200	410	2500
11	h*ee*t	/e/	430	2300	400	2600
12	h*ie*t	/i/	300	2500	300	3000

[a] See Ref. 3.

Figure 1 represents the $\frac{1}{3}$-oct frequency spectra of the vowels, averaged over the 50 speakers. Although these band filters are too broad to show the formant frequencies accurately for each individual speaker, these frequencies can be derived rather well from the average spectra (arrows). Only in the case of the vowel /a/ is there some difficulty in distinguishing the formants. The average formant frequencies thus determined are reproduced as F_{1a} and F_{2a} in Table I. They agree satisfactorily with the formant frequencies, marked as F_{1b} and F_{2b} in the same table, adopted from Meinsma (see Cohen *et al.*[4]).

The frequency spectra were averaged also over the 12 vowels instead of over the 50 speakers. In this case, an average spectrum was obtained for each speaker. It appeared that these spectra differ significantly from subject to subject, indicating a personal "touch" that should be taken into account (see Bordone-Sacerdote and Sacerdote[5]).

II. PHYSICAL VOWEL SPACE

A. Calculation of the Main Factors

As a result of the vowel-spectra measurements, we obtained in the previous section 600 (12 vowels, 50 speakers) series of 18 numbers representing the relative sound levels in the various frequency bands. By using these numbers as coordinates, each series can be plotted as a point in an 18-dimensional Euclidean space. Therefore, the 600 particular vowel spectra can be represented by a cloud of 600 points in that space. As the frequency spectra of the same vowel pronounced by different speakers are rather similar, the cloud will consist of 12 clusters of 50 points each.

We can get some insight into the way in which the 600 points spread by computing how the total variance of the cloud (equal to the sum of squares of distances of points from their "center of gravity" divided by number of points) is composed. If we substitute each vowel cluster by its center of gravity, the variance of the resulting 12 points, equal to 60% of the total variance, represents the differences between vowels. The remaining 40% stand for the variance within the 12 clusters of 50 individual vowel points. A further analysis showed that the variance of the 50 centers of gravity for the 12 vowel points of each speaker accounts for 17% of the total variance. This percentage represents the differences between speakers. So, if we translate the set of vowel points for each speaker in such a way that the centers of gravity coincide for all speakers, the percentage of total variance of the points within the 12 clusters is reduced to 23%.

Now, the question is: Do we actually need 18 dimensions to describe the differences between the 600 vowel spectra and, if not, how can we derive a subspace with fewer dimensions still fitting the 600 vowel points? This question was solved by applying the technique of principal-component analysis[6] to the cloud of 600 points in 18 dimensions.

The solid curve in Fig. 2 gives the variance of the 600 points along each of the 18 axes. The variance is expressed in the percentage of the total variance of the cloud "explained" by each dimension. (Pythagoras' theorem implies that the total variance is equal to the sum of the variances along any set of 18 orthogonal axes.) We see that no single dimension explains more than 10% of the total variance. This does not mean, however, that no specific direction explaining more than 10% can be found. We should like to rotate the original

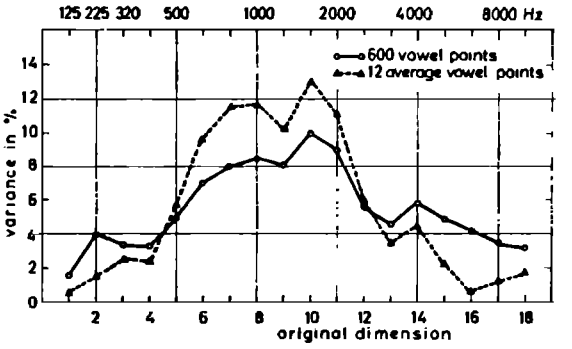

FIG. 2. Percentage of the total variance explained by the 18 original dimensions.

axes in such a way that the first *new* dimension will explain as much as possible of the total variance, the second *new* dimension as much as possible of the variance left unexplained by the first, and so on. This is just what the technique of principal-component analysis does. We call the new dimensions, being linear combinations of the original ones, the *factors*; the coordinates of a vowel point along these factors the *factor scores*; and the subspace found in this way the *factor space*. The factors can be determined by computing the eigenvectors of the covariance matrix of the 600 points. The eigenvector with the largest eigenvalue defines the direction of the first factor, the corresponding eigenvalue is the variance explained by that factor, and so on.

Figure 3 shows the results of the principal-component analysis (solid curve). The first four factors, I–IV, explain 33.7%, 27.2%, 8.7%, and 5.8% of the total variance, respectively. These four factors together leave 24.6% of the variance unexplained. Whether more factors are necessary to describe the spectral differences between vowel sounds can be studied by repeating the whole procedure for the average vowel spectra. The dashed curve in Fig. 2 represents, for the 12 points, the

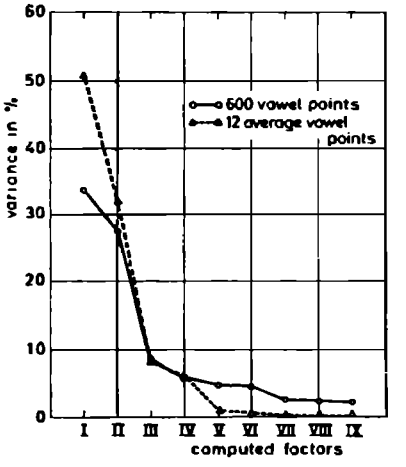

FIG. 3. Percentage of the total variance explained by the first nine factors.

TABLE II. Direction cosines of the factors I to IV with respect to the original 18 dimensions.

	1 125 Hz	2 225	3 315	4 400	5 500	6 630	7 800	8 1000	9 1250	10 1600	11 2000	12 2500	13 3150	14 4000	15 5000	16 6300	17 8000	18 10000
I	0.082	0.135	0.134	0.049	−0.183	−0.369	−0.448	−0.433	−0.177	0.194	0.382	0.231	0.199	0.144	0.084	0.118	0.151	0.157
II	−0.063	−0.144	−0.173	−0.087	0.115	0.237	0.146	0.226	0.469	0.518	0.325	0.267	0.173	0.254	0.164	0.025	0.078	0.104
III	−0.050	−0.094	−0.078	−0.100	0.068	0.155	0.132	0.072	0.305	−0.496	−0.114	0.263	0.320	0.418	0.393	0.198	0.124	0.117
IV	0.017	−0.123	0.173	0.460	0.736	0.192	−0.078	−0.150	−0.173	0.017	0.088	0.136	0.055	0.033	−0.202	−0.180	0.017	−0.053

percentage of variance explained by each of the 18 original dimensions. The result of a principal-component analysis for these points is reproduced by the dashed curve in Fig. 3. The first four factors explain in this case 51.0%, 32.1%, 8.1%, and 6.6% of the variance, respectively. Only 2.2% is left unexplained. This strongly suggests that, also for the 600 points, only four factors are necessary if we are interested exclusively in differences between the vowels; the 24.6% variance unexplained must be mainly due to individual spread and to differences between the speakers.

B. Factor Space

Computation of the factor scores for every vowel point along the first four factors results in a four-dimensional cloud, again with 12 clusters corresponding to the vowels. In Fig. 4 the projections of the average vowel points on the I–II, I–III, and I–IV planes are plotted. In the I–II plane, which is the most important one, the points form the well-known vowel triangle with /a/,/u/, and /i/at the angular points.

The 12 average points are each the center of gravity of a cluster of 50 individual points. Assuming for each vowel normal distribution of the points in all directions, we can represent these clusters by four-dimensional 1-σ ellipsoids (σ is the standard deviation). The directions of the axes of each ellipsoid are the eigenvectors of the covariance matrix of each set of 50 individual points, and the lengths of these axes are equal to the square roots of the eigenvalues along the axes. The eigenvectors and eigenvalues were computed for the covariance matrix of each cluster separately; these calculations should be clearly distinguished from the principal-component analysis resulting in the eigenvectors and eigenvalues of the covariance matrix of the cloud of 600 points as a whole. The projections of the 1-σ ellipsoids on the I–II, I–III, and I–IV planes are also drawn in Fig. 4. One should realize that, theoretically, in two dimensions the 1-σ ellipse includes only 39% of the individual points. Actually, it appears to be about 45% in our case.

Besides the fact that some vowels have smaller ellipsoids than others, it is striking that the longest axes of all ellipsoids tend to have the same direction. There are indications that this orientation is related to the differences in the average frequency spectra of the speakers, already referred to in Sec. I. The average frequency spectrum for each speaker is represented by the center of gravity of the 12 vowel points of that speaker. The dashed 1-σ ellipses in Fig. 4 give the spread of these centers of gravity. Their orientation is similar to the orientation of the vowel ellipses.

This speaker bias was eliminated by such a translation of each personal set of 12 points in the original 18-dimensional space that the 50 centers of gravity came to coincide in the origin of the coordinate system. Once again, a principal-component analysis was carried out.

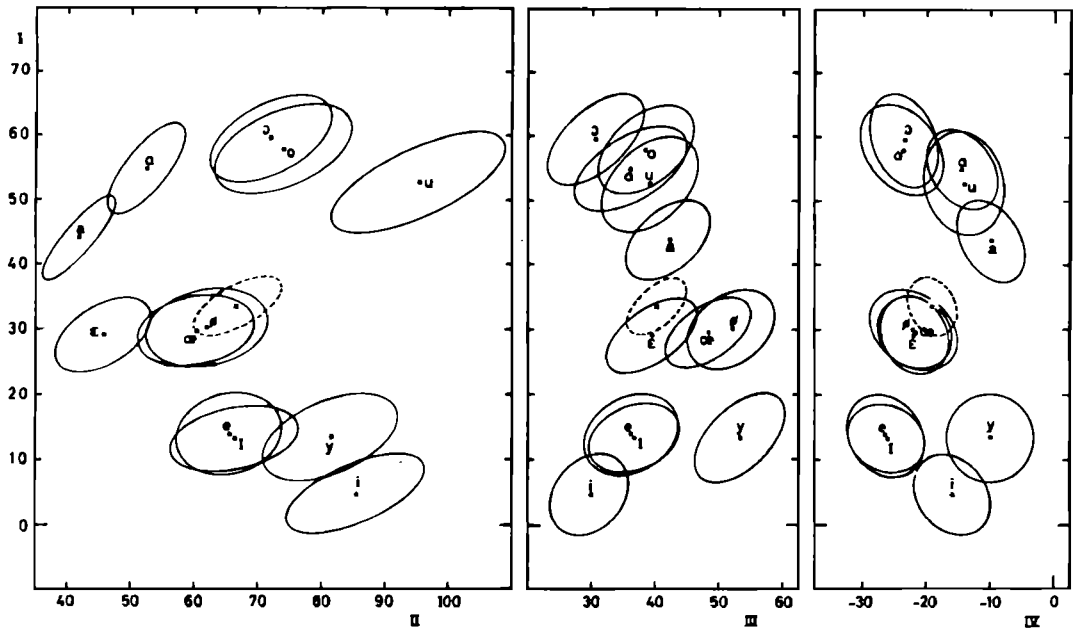

FIG. 4. Average vowel positions and 1-σ ellipses in the I–II, I–III, and I–IV planes (original data).

This time it was based on the "corrected" 600 vowel points. Nearly the same four principal factors as before were found, in this case explaining 39.0%, 27.7%, 8.5%, and 6.2% of the total variance, respectively, together 81.4%. Table II gives the cosines of the angles between the factors I–IV and the original dimensions. Computation of the corrected factor scores along the four factors gives average vowel points and 1-σ ellipses as are drawn in Fig. 5. As was expected, the ellipses are smaller than those in Fig. 4. This means that the described speaker-dependent correction improves the separation of the vowel clusters. This will be important for the vowel-identification procedure to be described in Sec. III.

Rather than interpreting the longest axes of the 1-σ ellipsoids in Fig. 4 to be parallel, one could interpret them as pointing to the origin. Furthermore, the lengths of these axes for vowels close to the origin tend to be somewhat smaller than for more distant vowels. These facts do not favor a correction just by translation of the personal sets of 12 points, but would suggest a correction

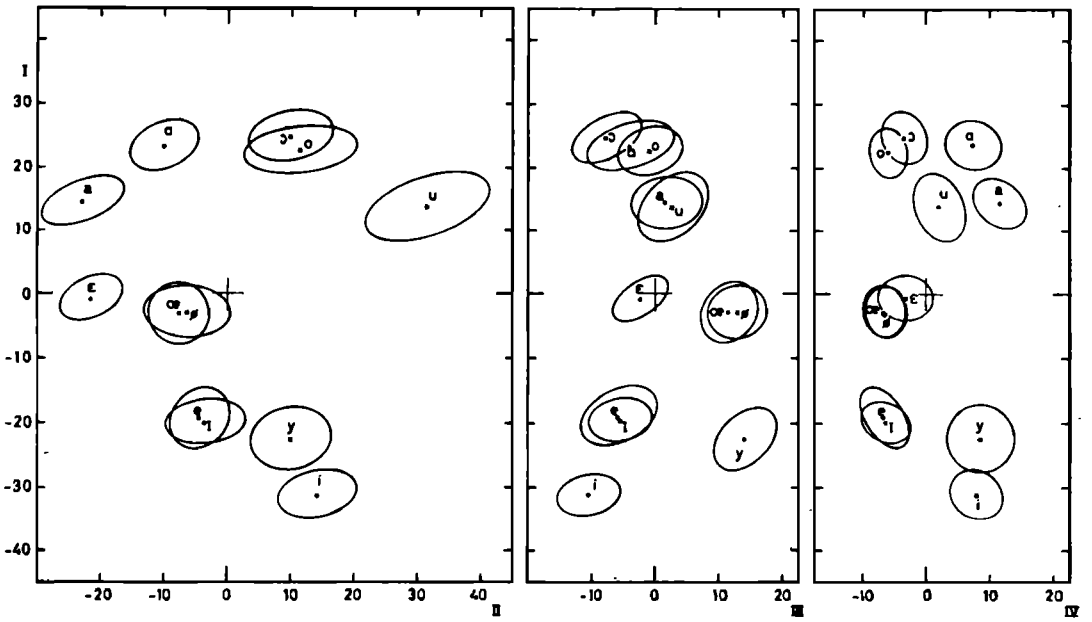

FIG. 5. Average vowel positions and 1-σ ellipses in the I–II, I–III, and I–IV planes (after speaker-dependent correction by translation).

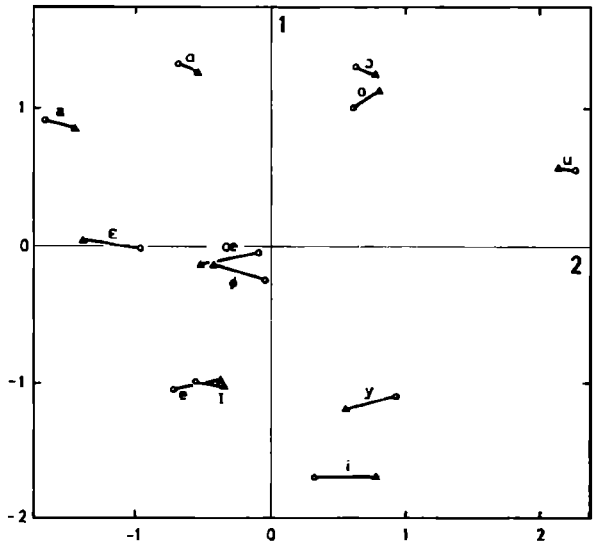

FIG. 6. F_1–F_2 configuration (○) and I–II configuration (Δ) after rotation to optimal congruence.

by the following two steps: (1) multiplication of each set with respect to the origin in such a way that all centers of gravity get the same distance from the origin, and (2) rotation of each set, also with respect to the origin, to make the centers of gravity coincide. Speaker-dependent correction of the data in this way also results in smaller 1-σ ellipsoids of the vowels in the factor space than without correction. However, as will be shown in Sec. III, this does not lead to a better machine vowel identification than the easier way of correction by translation.

The consistent orientation of the 1-σ ellipsoids even after translation per speaker (see Fig. 5) is partly due to the correction for the over-all SPL. As the over-all SPL is mainly determined by the highest maximum in the spectrum, we may expect, within 50 individual vowels, a small spread in the sound levels in a filter band of which the average level is near this maximum. On the other hand, the sound levels in a filter band of which the average level is far below the maximum will have a larger spread. This implies that all vowel ellipsoids in the 18-dimensional space will have largest spread in the direction towards the origin. This is not changed by the rotation to the factor space.

The factors I, II, III, and IV together appear to enclose the same subspace as the four factors found in the earlier experiment with 10 speakers and 15 vowels.[1] The individual factors, however, are not completely identical, but are rotated within the four-dimensional subspace. This will be caused primarily by the different sets of vowels.

C. Factor Space Versus Formant Plane

The configuration of the 12 average vowels in the factor space, especially in the I–II plane, bears much resemblance to the configuration of the same vowels in the formant plane with F_1 and F_2 plotted logarithmically along the axes. To investigate this resemblance mathematically, we used the so-called canonical-matching procedure.[2,7] Before the actual matching, both configurations are normalized to make their variances in all directions equal to unity. Matching then consists of rotating both normalized configurations individually in such a way that the corresponding new coordinates of the points of the two configurations show maximal correlation. The canonical-correlation coefficients indicate how well the configurations can be matched.

Matching of the F_{1a}–F_{2a} formant configuration, derived from the average vowel spectra of Fig. 1, and the I–II factor configuration of the average vowels of Fig. 5 results in canonical-correlation coefficients 0.997 and 0.945. Matching the same formant configuration with the four-dimensional I–II–III–IV configuration gives coefficients 0.997 and 0.979. The difference between these pairs of excellent correlation coefficients is so small that we concluded that the I–II plane is a fair approximation of the plane that correlates best with the formant plane. Although it seems that the improvement by matching with the four-dimensional instead of the two-dimensional configuration is not significant, more elaborate investigation on the significance of the canonical-correlation coefficients will be necessary. In Fig. 6 the matched F_{1a}–F_{2a} and I–II configurations are reproduced. As the shape of both configurations is changed by the normalization, no calibration of the axes is possible.

III. MACHINE VOWEL IDENTIFICATION

For many people, a main criterion in evaluating an alternative vowel-description technique will be whether it is successful in developing a vowel-identification apparatus or algorithm. An algorithm on which a vowel-identification apparatus can be based is described below. Conclusions about the sufficiency of three or four dimensions and about the usefulness of a speaker-dependent correction will be drawn from the identification scores of the algorithm in the specific cases.

A. Algorithm

As a first-order approximation, it is possible to base a vowel-identification algorithm on the Euclidean distances, in the factor space, of an unknown vowel point to all the average vowel points. This procedure was used in the previous investigation,[1] in which the limited number of speakers (10) did not allow a more elaborate approach. With 50 speakers, we have a much better insight into the orientation and size differences of the 1-σ ellipsoids, so that we can take these differences into account.

Let $\mathbf{m}_i$ be the vector pointing to the average position of the ith vowel. This vector is defined by the n coordinates of the average point in n dimensions.

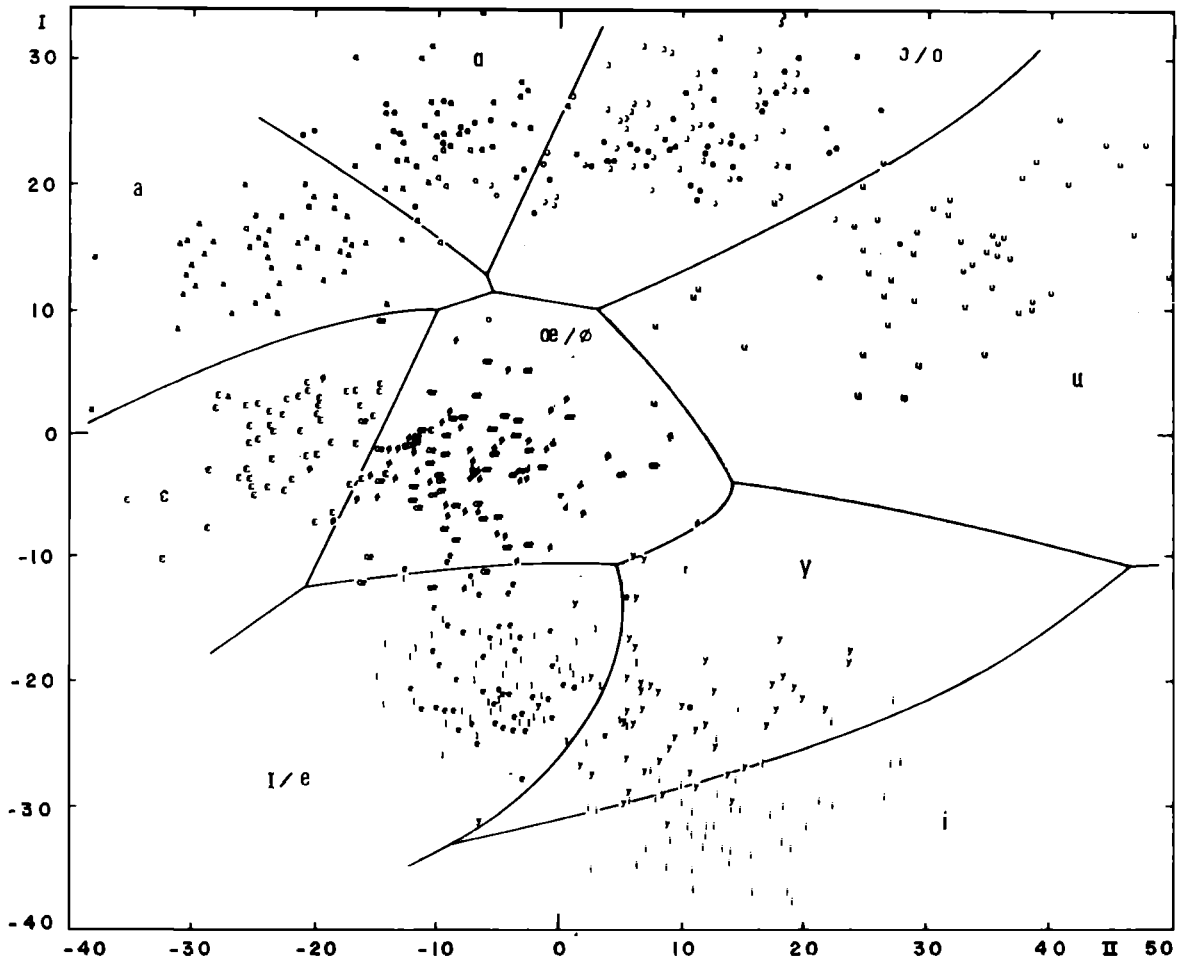

FIG. 7. Data points in the I–II plane after speaker-dependent correction by translation. Identification areas in two dimensions.

Secondly, let S_i be the $n \times n$ covariance matrix of the 50 individual points clustering around the average point. Assuming a normal distribution of the individual vowel points, this matrix is representative of the spread of the points around the average point and determines the 1-σ ellipsoid of the ith vowel. For any point in the n-dimensional space, defined by its vector $\mathbf{x}$, one can compute the probability that a specimen of the ith vowel will be found at that point. A measure for this probability is the multidimensional density function[8]

$$f_i(\mathbf{x}) = (2\pi)^{-n/2} \cdot |S_i|^{-\frac{1}{2}} \times \exp[-\tfrac{1}{2}(\mathbf{x}-\mathbf{m}_i)^T \cdot S_i^{-1} \cdot (\mathbf{x}-\mathbf{m}_i)],$$

in which $(\mathbf{x}-\mathbf{m}_i)^T$ stands for the transposed of vector $\mathbf{x}-\mathbf{m}_i$. There will be an area in the space within which the probability of finding a specific vowel is larger than the probabilities of finding the other vowels. The space can be divided into as many of such "identification areas" (maximum likelihood regions) as there are vowels to discriminate. The boundaries of these areas are multidimensional quadratic "surfaces."

As an illustration, in Fig. 7 the I–II plane of the factor space is plotted with all 600 vowel points, corrected by translation for the personal bias of the speakers. The plane is divided by quadratic curves into nine identification areas. The vowels /o/, /ø/, and /e/ are regarded to be identical to /ɔ/, /œ/, and /I/, respectively. This makes sense especially for the first 100 msec of these vowels, since phoneticians claim that the long vowels /o/, /ø/, and /e/ in Dutch tend rather to diphthongs, in which the first part is equal to /ɔ/, /œ/, and /I/, respectively.[4] Although the same is claimed for /a/–/ɑ/, our data do not support this view (see, for instance, the clearly different average positions of /a/ and /ɑ/ in Fig. 5). Unless explicitly stated otherwise, the three pairs of vowels mentioned above are treated as three single vowels throughout the following exposition.

B. Identification Score

One can check in Fig. 7 that 528 of the 600 points lie within the correct areas. Thus, the identification score of a vowel-identification apparatus designed after this

TABLE III. Identification scores, before and after speaker-dependent correction, with use of one, two, three, four, or six factors.

Number of factors used	1	2	3	4	6
	Identification score in %				
Original data	51.0	78.2	86.7	88.7	93.2
Data corrected by translation	60.2	88.0	97.2	97.5	97.7
Data corrected by multiplication + rotation	52.5	88.8	94.2		

algorithm would be 88% in two dimensions. Visualization of the points and the identification areas in the case of including more than two factors is scarcely possible. However, the check whether the points will be correctly identified can be done directly by determining for each individual vowel point whether the vowel that has the largest probability to be found at that point is the same vowel as was intended by the speaker. We computed the identification scores of the imaginary apparatus in the case of using one, two, three, four, or six factors, respectively, both for the original data and for data corrected for the personal bias of the speakers. Table III and Fig. 8 give the results.

Correction for personal bias improves the identification score appreciably. A correction by multiplication and rotation, however, has no better results than a correction by translation. Therefore, we prefer the latter as it is easiest to accomplish. Apparently at least three factors are necessary to obtain almost maximal identification; with more factors, only a slight further improvement is obtained. This procedure gives better scores than the procedure based on distances, mentioned earlier. With three factors and after correction by translation, this last procedure had a score of 92%, to be compared with the score of 97% in the case of the procedure based on probabilities.

It appeared that the procedure based on probabilities has been applied also by Welch and Wimpress[9] on the formant-based data of Peterson and Barney.[10] Using

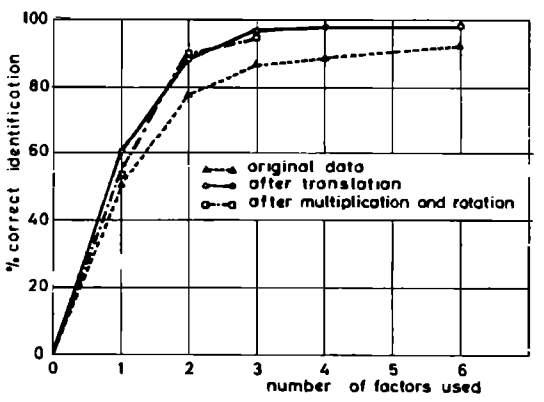

FIG. 8. Identification score as a function of the number of factors used.

only F_1 and F_2, they could identify correctly 87% of the vowels. By adding F_3, they obtained an identification score of about 92%. The best result, about 94%, was obtained by using six dimensions: F_1, F_2, F_3, fundamental F_0, and the levels L_1 and L_2 of the first two formants. We should be cautious to compare these percentages with those of Table III (for the original data) because (1) the nine Dutch and the ten American vowels used differ greatly, and (2) Welch and Wimpress used only words that were unanimously recognized by human listeners. Within these restrictions, however, we may regard these percentages to affirm the view that the factor space bears at least the same information as the formant space.

C. Reduction of Number of Band Filters

The choice of the original 18 frequency bands from 100 Hz up to 10 000 Hz in resolving the frequency spectrum was not arbitrary. The high identification scores obtained suggest that the 18 numbers include all basic information. One could ask, however, whether the result would have been just as good if we had chosen wider bandpass filters, or a smaller total frequency range, or both. Figure 2 suggests that the frequency bands below 500 Hz and above 4000 Hz do not contain much information. Therefore, we repeated the complete procedure of speaker-dependent correction, principal-component analysis and identification on the data of the ten $\frac{1}{3}$-oct bands from 500 Hz up to 4000 Hz. The resulting identification score, in the case of using three factors, was 95% now, only a bit less than the 97% obtained when starting with all 18 bands.

A second step was to combine in pairs the sound levels of the ten $\frac{1}{3}$-oct bands from 500 Hz up to 4000 Hz energetically, resulting in five $\frac{2}{3}$-oct bands. These data, reduced to three factors, gave an identification score of 94%. Even when we started from three original numbers per vowel spectrum, giving the sound levels in two $\frac{3}{3}$-oct bands and one $\frac{4}{3}$-oct band in the frequency region from 500 Hz up to 4000 Hz, we obtained 88% correct identifications. We may conclude that, although the $\frac{1}{3}$-oct band analysis gives highest identification scores, even a drastic reduction of the number of filter bands has amazingly little influence.

IV. HUMAN VOWEL IDENTIFICATION; PERCEPTUAL VOWEL SPACE

It would be quite interesting to compare the score obtainable with vowel-identification equipment with the score of listeners. Therefore, a listening test was performed. The same 100-msec segments of the 600 vowels used for analysis were presented by earphone to ten listeners; none of them had been members of the group of 50 speakers. The vowel segments were presented in random order, one every 2 sec, with a 5-sec

TABLE IV. Matrix of the confusions $c(i,j)$ made by 10 listeners in identifying 12 vowels of 50 speakers.

	Responses											
	/a/	/ɑ/	/ɔ/	/o/	/u/	/œ/	/ø/	/y/	/ɛ/	/I/	/e/	/i/
Stimuli												
/a/	297	187	—	—	—	6	3	—	7	—	—	—
/ɑ/	38	383	65	14	—	—	—	—	—	—	—	—
/ɔ/	—	12	388	99	—	1	—	—	—	—	—	—
/o/	1	14	130	340	6	7	1	—	1	—	—	—
/u/	—	—	4	23	471	1	—	1	—	—	—	—
/œ/	—	1	1	2	2	378	103	6	6	1	—	—
/ø/	—	—	2	4	—	137	336	12	5	2	1	1
/y/	—	—	—	1	—	16	13	460	—	—	—	10
/ɛ/	—	—	2	—	1	43	16	—	422	12	4	—
/I/	—	—	—	—	—	42	27	21	12	335	42	20
/e/	—	—	—	—	—	5	41	19	16	158	229	32
/i/	—	—	—	—	—	—	—	48	—	15	7	430

pause after every 12 presentations, in three sessions of 17 subsets of 12 each. The samples of a 51st speaker were included to obtain similar sessions. The subjects had to write down the vowels they thought to be spoken. They knew which 12 vowels were involved and were forced to make a choice anyhow. Only the 6000 responses referring to the original 50 speakers were used in the following calculations.

The responses were cumulated in a confusion matrix (see Table IV). Of these responses 74% are correct; if confusions /ɔ/–/o/, /œ/–/ø/, and /I/–/e/ are neglected, this score is 86%. For nine American vowels, presented in 300-msec segments, 74% correct identifications were found by Fairbanks and Grubb.[11] [Presentation of complete /*h*(vowel)*t*/ words gives much higher scores.[10]] The score of 86% is equal to the machine vowel-identification score for the uncorrected data in the case of using three factors (see Fig. 8). This seems to make sense. Within the 100 msec of a segment, the listener will not be able to get accustomed to a speaker's voice and cannot take his personal touch into account.

From the confusion matrix of Table IV, a perceptual configuration of the stimuli can be found with Kruskal's multidimensional-scaling technique.[12,13,2] A problem is how to deal with the asymmetry of the confusion matrix. A discussion on several methods of solving this problem is given by van der Kamp and Pols.[14] One method is the construction of two configurations, a stimulus configuration and a response configuration. We have no idea, however, what the interpretation of two such configurations in vowel perception could be. Of the other methods, one symmetrizes the matrix by correcting for a supposed response bias,[15] and others symmetrize by some averaging process. Especially in our case, where we presented 100-msec segments of vowels, we would not be surprised if listeners were biased to respond short vowels more often than long ones. To investigate this possible response bias, we computed from the original matrix all possible 2×2 submatrices, each belonging to one pair of stimuli and the same pair of responses. This computation can be made if we assume that the ratio of the four relevant entries does not depend on the possible presence of other stimuli in the set (Clarke's constant-ratio rule[16]). None of the submatrices found showed appreciable response bias.

As response bias appeared to be negligible, there is no objection against symmetrizing the confusion matrix by means of an averaging method. These methods use some sort of an averaging process to derive the similarity element $s(i,j)=s(j,i)$ from the four confusion elements $c(i,j)$, $c(j,i)$, $c(i,i)$, and $c(j,j)$ of the confusion matrix. There is, however, more information present in the confusion matrix about the similarity of stimuli i and j than is represented by just the four mentioned confusion elements. The more i and j are similar in perception, the more their response distribution over the total set of response categories will also be similar. This degree of similarity can be expressed by the number of times that i and j have resulted in the same response, summated over all response categories. So, in Table IV, the stimuli /a/ and /ɑ/ have 38 /a/ responses and 187 /ɑ/ responses in common, resulting in a similarity index of 225; the stimuli /a/ and /ɔ/ have 12 /ɑ/ responses and 1 /œ/ response in common, resulting in a similarity index of 13; and so on (Table V). This symmetrizing method, worked out by our associate T. Houtgast, has the additional advantage of reducing the number of empty cells. It can be denoted by the formula

$$s(i,j)=s(j,i)=\tfrac{1}{2}\sum_{k=1}^{12}[c(i,k)+c(j,k)-|c(i,k)-c(j,k)|].$$

$\sum|c(i,k)-c(j,k)|$ represents the dissimilarity of stimuli i and j. Subtraction from the constant $\sum[c(i,k)+c(j,k)]$ gives the similarity element of i and j.

Kruskal's technique, applied to the obtained similarity matrix of Table V, results in a configuration in one dimension with 21.2% stress, in two dimensions

TABLE V. Matrix of similarity indices $s(i, j)$, derived from confusions of Table IV.

	/a/	/ɑ/	/ɔ/	/o/	/u/	/œ/	/ø/	/y/	/ɛ/	/ɪ/	/e/
/ɑ/	225										
/ɔ/	13	91									
/o/	23	94	243								
/u/	1	18	28	34							
/œ/	16	4	5	15	7						
/ø/	14	6	7	15	8	255					
/y/	10	1	2	9	3	36	43				
/ɛ/	16	2	3	12	4	68	69	29			
/ɪ/	17	1	2	10	2	83	90	60	86		
/e/	15	0	1	7	2	59	70	47	53	283	
/i/	0	0	0	0	1	7	16	58	16	63	73

with 7.7% stress, in three dimensions with 4.7% stress, or in four dimensions with 2.3% stress. According to criteria developed by Wagenaar and Padmos,[17] this suggests that the underlying perceptual configuration is at least two-dimensional but may well have a higher dimensionality. Canonical matching of the obtained four-dimensional perceptual configuration with the four-dimensional factor configuration of Fig. 5 gives canonical-correlation coefficients 0.997, 0.995, 0.907, and 0.794, respectively. Up to the third dimension, the correlation between the two configurations is excellent. Both matched configurations are represented in these three dimensions in Fig. 9. Also the correlation in the fourth dimension is good, which supports the view that, although the identification score is not much improved by including a fourth factor, this last factor does contain relevant information.

V. CONCLUSIONS

The $\frac{1}{3}$-oct-band spectral information present in a vowel sound can be condensed in scores along four factors as described in this paper. No more than these four factors are necessary for describing all occurring vowel utterances. The 12 average vowels form a configuration in the factor space that, as far as the first two factors are concerned, is almost identical to the F_1–F_2 configuration of the same vowels. So we killed two birds with one stone: we proved that the F_1–F_2 plane of vowels is the plane that includes the most possible information and, secondly, we showed that there is a much easier way to obtain this same information.

Apart from its easier determination, the $\frac{1}{3}$-oct-band spectral approach combined with the principal-component analysis has more advantages: (1) one can easily use more factors to obtain even more information; (2) one can apply the method to sounds with no apparent maxima in the spectrum; and (3) the method is in tune with our knowledge of the ear's analyzing power. The four-dimensional factor configuration of the average vowels can be matched in an excellent way with the four-dimensional perceptual configuration, which supports the view that the presented approach of the vowel spectra is a useful model of what a human listener unconsciously does.

Individual utterances of a vowel form a cluster around the average vowel position in the factor space, with unequal extent in different directions. This is partly due to the individual touch of every speaker. The identification scores for corrected and uncorrected data show clearly that for optimal identification this individual touch should be taken into account. The identification scores as a function of the number of factors used show that at least three factors are necessary for identification. The fourth factor scarcely improves the score. The necessity of speaker-dependent correction (as well as the necessity of using three dimensions for good identification) agrees with the experience of Gerstman.[18]

We are aware that with the foregoing analysis the basic data are not exhausted. Further computations will be of interest to learn more about the correlation between the factor configuration and the formant configuration for each speaker individually, whether an Euclidean space is the most appropriate one or whether some other technique may be preferred to the principal-component analysis. New experiments are in progress in which female speakers are employed, so that the results for male and female speakers can be compared.

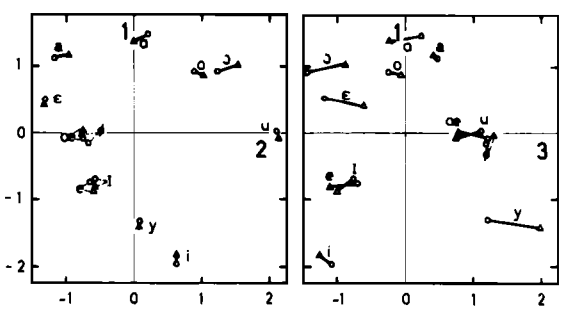

FIG. 9. Four-dimensional factor configuration (Δ) and four-dimensional perceptual configuration (○) after rotation to optimal congruence. Only the three most correlating dimensions are represented.

* Present address: Akoestisch Adviesbureau Ir. V.M.A. Peutz N. V., Nijmegen, The Netherlands.

[1] R. Plomp, L. C. W. Pols, and J. P. van de Geer, "Dimensional Analysis of Vowel Spectra," J. Acoust. Soc. Amer. 41, 707–712 (1967).

[2] L. C. W. Pols, L. J. Th. van der Kamp, and R. Plomp, "Perceptual and Physical Space of Vowel Sounds," J. Acoust. Soc. Amer. 46, 458–467 (1969).

[3] International Phonetic Association, *The Principles of the International Phonetic Association* (Department of Phonetics, University College, London, 1967).

[4] A. Cohen, C. J. Ebeling, K. Fokkema, and A. G. F. van Holk, *Fonologie van het Nederlands en het Fries* (Martinus Nijhoff, 's-Gravenhage, 1961), 2nd ed.

[5] C. Bordone-Sacerdote and G. G. Sacerdote, "Some Spectral Properties of Individual Voices," Acustica 21, 199–210 (1969).

[6] P. Horst, *Factor Analysis of Data Matrices* (Holt, Rinehart and Winston, New York, 1965).

[7] N. Cliff, "Orthogonal Rotation to Congruence," Psychometrika **31**, 33–42 (1966).
[8] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis* (Wiley, New York, 1958).
[9] P. D. Welch and R. S. Wimpress, "Two Multivariate Statistical Computer Programs and their Application to the Vowel Recognition Problem," J. Acoust. Soc. Amer. **33**, 426–434 (1961).
[10] G. E. Peterson and H. L. Barney, "Control Methods Used in a Study of the Vowels," J. Acoust. Soc. Amer. **24**, 175–184 (1952).
[11] G. Fairbanks and P. Grubb, "A Psychophysical Investigation of Vowel Formants," J. Speech and Hearing Res. **4**, 203–219 (1961).
[12] J. B. Kruskal, "Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis," Psychometrika **29**, 1–27 (1964).
[13] J. B. Kruskal, "Nonmetric Multidimensional Scaling: A Numerical Method," Psychometrika **29**, 115–129 (1964).
[14] L. J. Th. van der Kamp and L. C. W. Pols, "Perceptual Analysis from Confusions among Vowels" (to be published).
[15] W. A. Wagenaar, "Application of Luce's Choice Axiom to Form-Discrimination," Ned. Tijdschr. Psychol. **23**, 96–108 (1968).
[16] F. R. Clarke, "Constant Ratio Rule for Confusing Matrices in Speech Communication," J. Acoust. Soc. Amer. **29**, 715–720 (1957).
[17] W. A. Wagenaar and P. Padmos, "Quantitative Interpretation of Stress in Kruskal's Multidimensional Scaling Technique," J. Math. Statist. Psychol. (to be published).
[18] L. J. Gerstman, "Classification of Self-Normalized Vowels," IEEE Trans. Audio Electroacoust. **AU-16**, 78–80 (1968).