Frequency analysis of Dutch vowels from 50 male speakers

L. C. W. Pols, H. R. C. Tromp,* and R. Plomp
Institute for Perception, TNO, Soesterberg, The Netherlands
(Received 29 February 1972; revised 11 September 1972)

The frequencies and levels of the first three formants of 12 Dutch vowels were measured. The vowels were spoken by 50 male speakers in an h (vowel) t context. Statistical analysis of these formant variables confirmed that F_1 and F_2 are the most appropriate two distinctive parameters for describing the spectral differences among the vowel sounds. Maximum likelihood regions were computed and used to classify the vowels, and a score of 71.3% correct classification in the $\log F_1$-$\log F_2$ plane was obtained (87.3% if three related pairs are grouped together). These scores rose to 78.3% and 95.2%, respectively, when a simple speaker-dependent correction was applied. The scores are comparable with those obtained in an earlier study in which a principal-components analysis was applied to the 1/3-oct filter levels of the same vowel sounds [Klein, Plomp, and Pols, J. Acoust. Soc. Amer. 48, 999-1009 (1970)]. From the latter data a two-dimensional representation ("optimal plane") equivalent to the $\log F_1$-$\log F_2$ plane could be derived. The relative merits of the two approaches are discussed. For automatic speech recognition in particular, the dimensional analysis is much more attractive than the formant analysis because it is much simpler and can be carried out in real time.

Subject Classification: 9.3, 9.7.

INTRODUCTION

In a previous paper by Klein, Plomp, and Pols,[1] we presented a dimensional analysis of the frequency spectra of 12 Dutch vowels (|u|, |o|, |ɔ|, |ɑ|, |a|, |ɛ|, |e|, |ɪ|, |i|, |y|, |œ|, and |ø|), each pronounced by 50 male speakers. Diphthongs were excluded. A principal-components analysis of the sound-pressure levels (SPL) in 18 $\frac{1}{3}$-oct filter bands showed that the spectral differences among the 12 vowels could be represented satisfactorily in a four-dimensional factor space. The configuration of the average vowels in this *factor space* appeared to be highly correlated with the configuration of the average vowels in the F_1-F_2 *formant plane* and with their configuration in a four-dimensional *perceptual space* derived from confusion data.

The $\frac{1}{3}$-oct frequency analysis, applied in this earlier investigation, is not suited for accurate determination of the formant characteristics of individual speakers. Since we wanted to study the relative merits of the principal-components analysis and the more traditional formant frequency and level analysis more carefully, the formant data for each speaker were obtained with a narrow-band frequency analysis. We used the same 50×12 vowel sounds as in the previous, principal-components study. This paper presents the formant data and compares them with the principal-components representation. This comparison is presented in statistical measures such as the percentage of correct identifications obtained by applying an algorithm for recognizing the vowels based on a computation of the maximum-likelihood regions in the multidimensional representation.

I. FORMANT ANALYSIS

A. Method

The determination of the frequencies and levels of the first three formants of each of the 50×12 vowel segments consisted of the following successive steps:

(1) The word of the type h(vowel)t, recorded in a nonreverberant room (see previous paper[1]), was sampled via an 8-bit analog-to-digital converter at a rate of 20 kHz. The 8-bit samples were stored in the memory of a digital computer (DEC PDP-7, 8K memory).

(2) A number of these samples, comprising 10 periods of the initial, constant vowel waveform, were selected out and then generated, as a continuous periodic signal, with a digital-to-analog converter.

(3) This analog signal was fed to a wave analyzer (Hewlett-Packard, Model 302A) and a detailed frequency analysis was made over the range from 50 up to 5000 Hz, with a bandwidth of 7 Hz. The frequency was varied automatically (Hewlett–Packard Sweep Drive, model 297A) with a speed of 1000 Hz/min. The spectrum was recorded with the aid of a logarithmic converter, to register amplitude in decibels (Hewlett–Packard, Moseley Division, model 7560AM) and an $X-Y$ recorder (Hewlett–Packard, Moseley Division, model 7035B).

(4) From this recording, the frequencies and levels of the first three formants were determined by drawing the envelope of the spectrum by eye. Figure 1 gives an example of such a recording. In addition to the formants F_1, F_2, and F_3, the fundamental frequency F_0 of about 150 Hz and a large series of low-level harmonics of a fundamental of about 15 Hz can be seen. The latter series stems from the fact that the waveforms of the ten periods of the vowel segment are generally not completely identical. For most vowel segments, there were

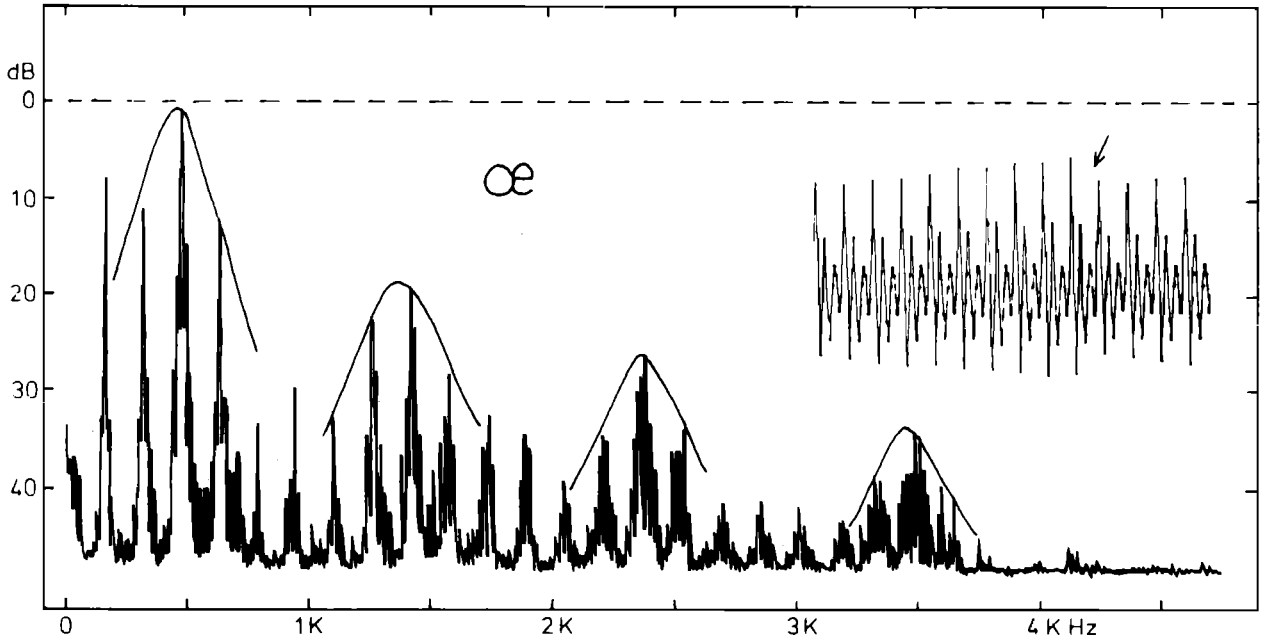

FIG. 1. Example of a recording representing the result of a narrow-band frequency analysis of the /œ/ of one of the 50 speakers. The spectral envelope of the formant regions is fitted by eye. The insert shows the repeated waveform of ten periods as used for the analysis, with the arrow indicating where the end and the beginning of the segment were connected.

no difficulties in locating in an objective way the formants in the frequency spectrum. We feel, however, that in a number of cases our *a priori* knowledge of where the formant should be located played a significant role in our decision. Also Fant[2] (pp. 66–67) mentions this problem.

B. Results

Table I and Fig. 2 present the means and standard deviations of the formant frequency and level data, pooled over the 50 speakers. The formant levels are given in decibels below the overall SPL of that particular vowel segment. For certain corresponding vowels the average values can be compared with data of Peterson and Barney[3] (33 male speakers) and with data of Fant[2] (seven male speakers). There is a satisfactory agreement in formant frequencies but not in the formant levels. Recently, Koopmans[4] determined F_1 and F_2 for Dutch vowels spoken by 10 males and 10 females in one-syllable words, using a method very different from ours: She measured period durations in the vowel waveforms. Despite this difference, her F_1 values, averaged over ten male speakers, are in excellent agreement with ours; most F_2 values are, however, about 10% higher than in Table I.

C. Information Content of the Formant Variables

The main purpose of deriving formant frequencies and formant levels is to characterize the various vowels

TABLE I. Average frequencies and levels, and their standard deviations, of the first three formants of 12 Dutch vowels pronounced by 50 male speakers. The formant levels are given in decibels below overall SPL.

Dutch vowel	IPA symbol	Formant frequency and standard deviation in Hz						Formant level and standard deviation in dB					
		F_1	σ_{F1}	F_2	σ_{F2}	F_3	σ_{F3}	L_1	σ_{L1}	L_2	σ_{L2}	L_3	σ_{L3}
1 h*oe*t	/u/	339	46	810	85	2323	211	5.2	2.4	18.2	4.5	41.2	5.1
2 h*oo*t	/o/	487	42	911	90	2481	224	5.7	3.1	13.1	3.7	35.6	4.4
3 h*o*t	/ɔ/	523	49	866	72	2692	189	6.1	2.5	13.9	4.6	34.3	5.0
4 h*a*t	/ɑ/	679	80	1051	89	2619	172	8.4	3.0	12.0	3.4	31.2	4.5
5 h*aa*t	/a/	795	95	1301	113	2565	199	8.2	2.3	13.8	3.4	28.7	4.9
6 h*e*t	/ɛ/	583	67	1725	164	2471	213	7.2	3.2	18.7	5.2	25.4	5.8
7 h*ee*t	/e/	407	52	2017	161	2553	171	5.0	2.4	21.0	5.8	23.3	5.3
8 h*i*t	/ɪ/	388	53	2003	180	2571	189	5.0	2.1	22.3	5.5	24.5	4.7
9 h*ei*t	/i/	294	38	2208	169	2766	203	5.5	3.4	25.0	6.2	27.9	5.9
10 h*uu*t	/y/	305	42	1730	152	2208	226	5.1	3.1	23.2	6.8	28.0	7.0
11 h*u*t	/œ/	438	48	1498	159	2354	201	4.8	2.4	20.5	5.3	28.2	5.8
12 h*eu*t	/ø/	443	46	1497	115	2260	140	5.1	2.5	20.7	4.9	27.9	5.3

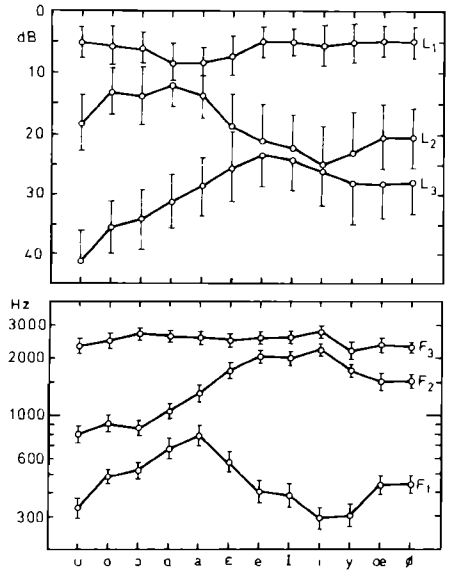

FIG. 2. Average frequencies and levels below overall SPL of the first three formants of 12 Dutch vowels pronounced by 50 male speakers. The vertical dashes indicate the standard deviations (in a few cases these dashes are given only in one direction to avoid overlapping).

by parameters (acoustic features) which take specific values for each vowel. The more different, in terms of standard deviations, the values of a particular variable (e.g., F_1) for two vowels are, the more appropriate this variable is to discriminate between these two vowels. Since the values for one vowel vary from subject to subject, the differences between vowels should be large compared to this variability within one vowel. Figure 2 shows that no single parameter has a different value for all vowels. This implies that the vowels cannot be described uniquely by only one variable (out of the group F_1, F_2, F_3, L_1, L_2, L_3). A combination of at least two is necessary. A visual inspection of the graphs makes it clear that L_1 and F_3 are not very appropriate as vowel descriptors because the variation of their average values, expressed in terms of their standard deviations, is small. Apparently, F_1 and F_2 are better because their standard deviations are much smaller than of L_2 and L_3.

This analysis of the formant data in search of the most characteristic variables can be done in a more quantitative way by computing for each variable how the total variance in the 50×12 individual data points is composed. Part of the total variance is the variance of the 12 average vowel points; this represents the difference between vowels. The remaining variance represents the spread of the individual data points around the 12 average values. In order to learn whether there is some systematic difference between speakers this remaining variance can be split up into two parts: the variance of the 50 average data points, representing the difference between speakers, and the residual variance, representing the random spread of the individual points.

TABLE II. Percentages of the total variance of each formant variable due to the different sources.

Source of variance	$\log F_1$	$\log F_2$	$\log F_3$	L_1	L_2	L_3
Vowels	85.8	93.6	42.6	17.0	40.3	46.5
Speakers	4.9	2.3	23.2	25.8	26.3	28.0
Residue	9.3	4.1	34.2	57.2	33.4	25.5

Table II gives the result of such a computation. The variances due to the different sources are expressed in percent of the total variance. In this calculation, $\log F_1$, $\log F_2$, and $\log F_3$ rather than F_1, F_2, and F_3 are used, since we prefer to use a logarithmic frequency scale, more in line with the hearing process than a linear frequency scale. (Unless otherwise stated, this will be the case for the rest of the article.) Table II shows that, of the six variables, $\log F_1$ and $\log F_2$ have the largest part of their variance "explained" by the vowels, the smallest part explained by the speakers, and also the smallest residuals. This confirms quantitatively the tradition of considering F_1 and F_2 as the most characteristic two acoustic features of vowels. The other four variables are much more speaker-dependent and their residual variances are also much larger than that of $\log F_1$ and $\log F_2$. Without any correction for the differences between speakers, the rank order of specificity of the six formant variables is $\log F_2$, $\log F_1$, L_3, $\log F_3$, L_2, L_1; with speaker-dependent correction, their rank order is $\log F_2$, $\log F_1$, L_3, L_2, $\log F_3$, L_1.

It is clear from Fig. 2 that the six formant variables are not independent. It appears that there is a good correlation between $\log F_1$ and L_2, and between $\log F_2$ and L_3. This correlation roughly follows Fant's[2] so-called low-pass filter rule (12 dB/oct). The upper-right part of Table III presents the various correlation coefficients in the average vowel data (each computation based on 12 pairs of numbers). The lower-left part of the table gives the correlation coefficients in the individual data (each computation based on 50×12 pairs of numbers). In addition to the correlation between $\log F_1$ and L_2, and $\log F_2$ and L_3 we see that also $\log F_1$ and L_1, $\log F_2$

TABLE III. Correlation matrix of the six formant variables. The upper-right part gives the correlation coefficients of the average data, the lower-left part the correlation coefficients of the individual data.

	$\log F_1$	$\log F_2$	$\log F_3$	L_1	L_2	L_3
$\log F_1$	1.000	−0.359	0.275	0.840	−0.806	0.032
$\log F_2$	−0.302	1.000	0.063	−0.278	0.796	−0.927
$\log F_3$	0.195	0.120	1.000	0.392	−0.241	−0.161
L_1	0.370	−0.090	0.116	1.000	−0.692	0.057
L_2	−0.533	0.512	−0.044	−0.042	1.000	−0.547
L_3	−0.021	−0.605	0.017	0.085	0.127	1.000

and L_2, and L_1 and L_2 are significantly correlated. L_1 and L_2 are only correlated in the average data.

D. Identification Score

We have concluded that the formant frequencies F_1 and F_2 are the most characteristic acoustic features of vowels. In Fig. 3, $\log F_2$ vs $\log F_1$ is plotted for all 50×12 spoken vowels. The better the 12 vowels are represented by separate clusters of 50 individual data points, the better each vowel is characterized by specific values of $\log F_1$ and $\log F_2$. We should like to have a quantitative measure of the degree to which the 12 clusters do overlap each other.

In the previous paper[1] such a quantitative measure was developed for the data resulting from a principal-components analysis of the vowel spectra. Described in simple terms, this procedure goes as follows: Assuming the points in each vowel cluster are distributed normally along the $\log F_1$ and $\log F_2$ scales, ellipses of equal probability (e.g., 1σ, 2σ, σ = standard deviation) around each average vowel point (+ symbols in Fig. 3) can be calculated. The points where the ellipses of equal probability of each pair of two neighboring clusters cross each other define likelihood regions for the various vowels. Each individual vowel point is

TABLE IV. Identification scores of the 50×12 individual vowel sounds as a function of the number of formant variables taken into account.

	Noncentered data		Centered data	
Variable	Non-grouped (%)	Grouped (%)	Non-grouped (%)	Grouped (%)
$\log F_2$	44.2	59.7	52.0	69.7
$+\log F_1$	71.3	87.3	78.3	95.2
$+\log F_3$	75.3	89.3	80.5	95.5
$+L_3$	78.0	90.5	82.8	96.0
$+L_2$	80.0	91.5	85.0	96.7
$+L_1$	79.5	91.8	85.2	97.0

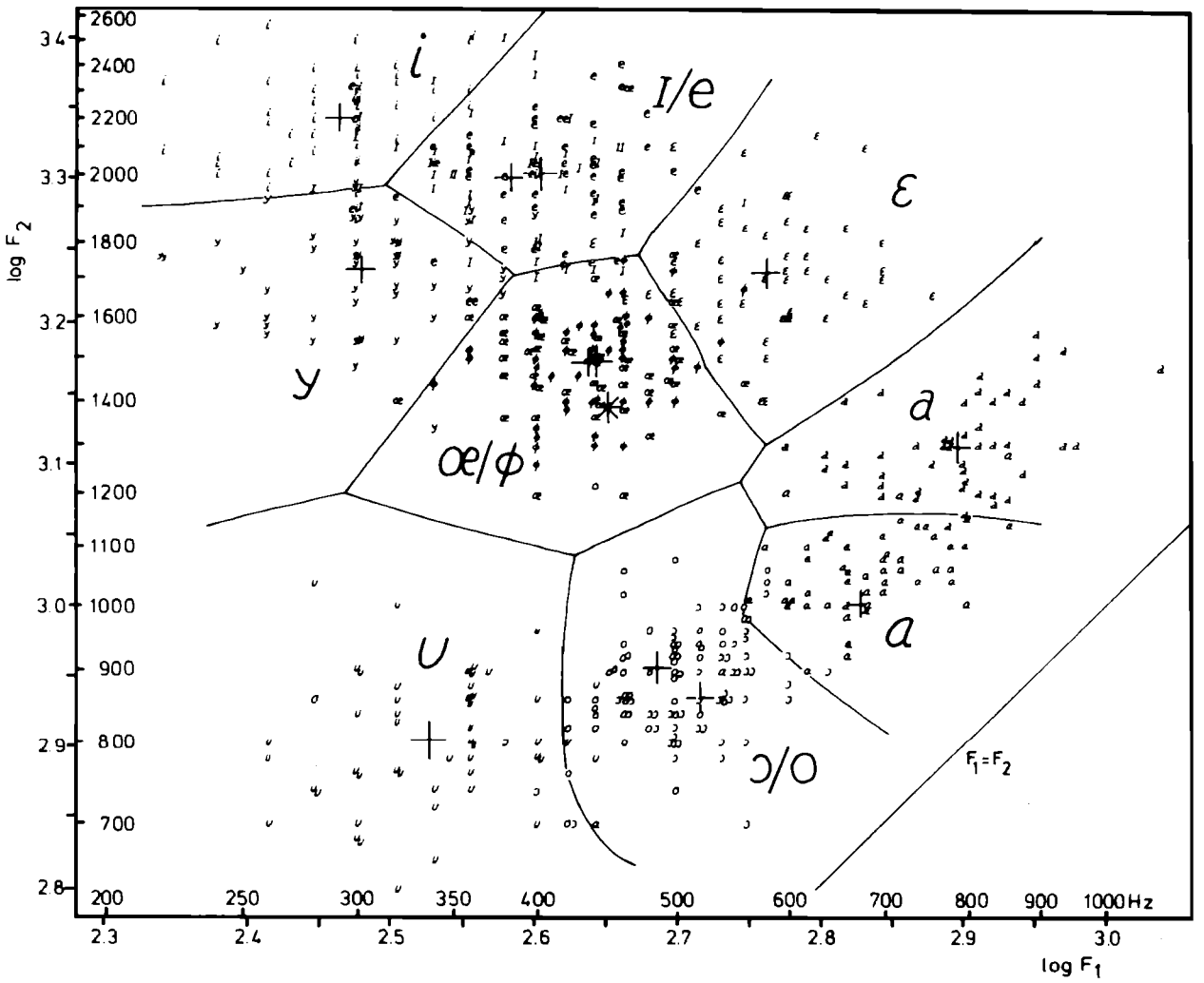

FIG. 3. $\log F_2$ vs $\log F_1$ of the 50×12 individual vowel sounds. The + symbols indicate the average values of the 12 vowels. The curves represent the boundaries of the maximum-likelihood regions of the *grouped* data (three pairs of vowels grouped together). In this graph, no speaker-dependent correction is applied.

considered to be correctly situated if it is located in the likelihood region of that particular vowel. In this way an indentification score can be determined representing the percentage of points, added over all vowels, situated in the correct likelihood regions.

In Fig. 3 the boundaries of nine maximum-likelihood regions are drawn; the regions for the vowel pairs /ɔ/–/o/, /ɪ/–/e/, and /œ/–/ø/ are combined. Figure 2 showed that the vowels within these three pairs have very similar average formant frequencies and levels. The main difference among them is duration. The identification score for these so-called *grouped* data is 87.3%. If all 12 vowels are considered separately (*nongrouped* data), this score drops to 71.3%.

The identification score can be used as a measure to investigate whether clustering is improved if more than two variables are taken into account. For example, L_3 might be considered as a third dimension, since Table II indicates that, next to $\log F_1$ and $\log F_2$, L_3 is the most characteristic parameter of the vowels. Computation has shown, however, that the identification score is improved more by adding $\log F_3$ than L_3 as a third dimension. This might be due to the fact that L_3 is highly correlated with $\log F_2$ (see Table III), so it does not provide much extra information, whereas $\log F_3$ is rather independent of $\log F_1$ and $\log F_2$.

In a similar way, a fourth dimension can be added, and so on. Table IV gives the identification scores for 1–6 dimensions if, step by step, that formant variable is added which contributes most to the identification score. The scores are computed both for the nongrouped and the grouped data.

The whole computation was repeated after applying a speaker normalization. As Table II showed, for each of the six formant variables, a part of the total variance was due to differences among speakers. One possible way of eliminating this source is by subtracting, for each speaker individually, his average value on each of the six variables. The results of the identification-score calculation for these *centered* data are also given in Table IV. We see that the identification scores for the centered data are consistently higher than for the *noncentered* data. In both cases the scores computed using only $\log F_1$ and $\log F_2$ are favorable compared with what is obtained when all six variables are used. This demonstrates again that most of the vowel information is covered by these formant frequencies. We should keep in mind, however, that the formants were determined by drawing the best-fitting envelopes of the frequency spectra. Probably, the identification scores would have been significantly lower if a more objective decision procedure were used, such as the recently published technique of Schafer and Rabiner.[5]

II. FORMANT ANALYSIS VERSUS PRINCIPAL-COMPONENTS ANALYSIS

As was mentioned in the Introduction, we applied the formant analysis to the same vowel segments which had previously been used in a principal-components analysis.[1] Before comparing the results of these two different approaches, the principal-components technique will be described and the main results given.

TABLE V. Identification scores of the 50×12 individual vowel sounds as a function of the number of factors taken into account. The factors were derived from the $\frac{1}{3}$-oct frequency spectra by applying a principal-components analysis. 2′ stands for the maximally discriminating plane.

	Noncentered data		Centered data	
Number of factors	Non-grouped (%)	Grouped (%)	Non-grouped (%)	Grouped (%)
1	36.8	51.0	44.3	60.2
2	62.8	78.2	70.0	88.0
2′			74.5	92.2
3	73.0	86.7	84.0	97.2
4	75.0	88.7	84.7	97.5
5	76.0	89.3	84.0	97.2
6	80.7	93.2	85.8	97.7

A. Method and Results of the Principal-Components Analysis

The method consisted of the following successive steps.

(1) The frequency spectra of 100-msec vowel segments were measured with a set of one-third octave bandpass filters ranging from 100 to 10 000 Hz. The outputs of the filters with center frequencies of 100, 125, and 160 Hz, as well as the outputs of the filters with center frequencies of 200 and 250 Hz were combined (energies were added), both in order to reduce the influence of differences in voice pitch and so that the bandwidths used would be comparable with the ear's critical bandwidth.

(2) In order to normalize the differences in the overall SPL of the vowels the output levels (in decibels) of the resulting 18 frequency bands were subtracted from the overall SPL of each individual vowel segment. The resulting numbers were considered as the coordinate values of 50×12 points in an 18-dimensional Euclidean space. A principal-components analysis was carried out on these data. As a result, new directions (factors) were obtained. The first "explains" most of the total variance of the points, the second most of the variance unexplained by the first one, etc. A speaker-dependent correction was also applied, it was identical to the one used for the formant data (Sec. I-D). This means that a translation of the 12 vowel points of each speaker was performed in such a way that the 50 centers of gravity (representing average vowel spectra) for the various speakers coincided. Several other speaker normalizations [1,6–9] were tried but none of them were more effective than the very simple translation procedure which we used here.

(3) Identification scores were computed for both the noncentered and the centered data, taking into account 1–6 factors, respectively.

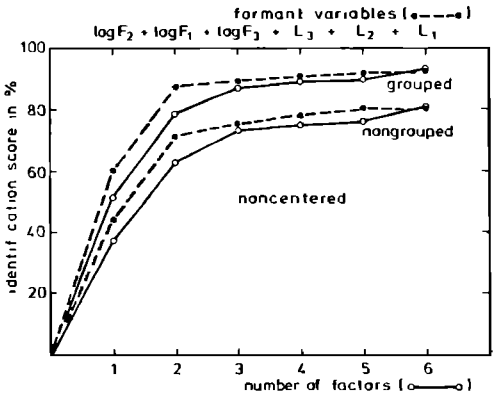

FIG. 4. Comparison of the identification scores of the 50×12 individual vowel sounds as a function of the number of formant variables (●——●) and of the number of factors (○——○) taken into account. The curves hold for the noncentered data, both nongrouped and grouped.

The identification scores found by this procedure are presented in Table V. The scores are also given for the grouped data (the three pairs of very similar vowels /ɔ/–/o/, /ɪ/–/e/, and /œ/–/ø/ are combined). The row marked by 2′ gives the identification scores, for the centered data, in that plane for which these scores are maximal. This maximally discriminating plane was found by rotating the I–II factor plane in small steps in the three-dimensional factor space, computing in each case the identification score for the data points projected on that plane. If a step resulted in an increased score, a subsequent step in the same direction was made, if not, then another rotation axis was tried. This iterative procedure produced the plane for which the vowels are maximally discriminated. Since the scores obtained are significantly higher than for the original plane (see Table V), we must conclude that principal-components analysis, although attractive and efficient for many applications, is not the most optimal technique for the reduction of this sort of data in identification experiments. For a further discussion of this question see Appendix A.

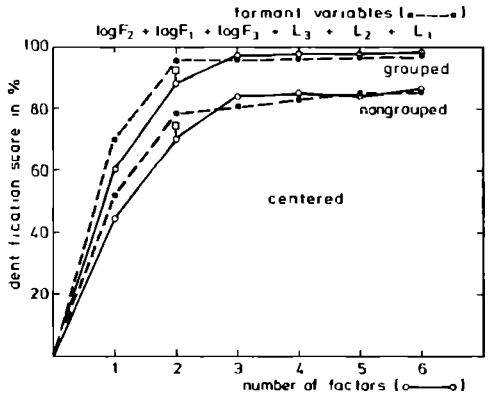

FIG. 5. As Fig. 4, but for the centered data (speaker-dependent correction). The square symbols represent the scores for the maximally discriminating plane.

B. Comparison of the Two Approaches

The information content of the formant variables and of the $\frac{1}{3}$-oct levels can be compared using identification score as a criterion. For this, the data of Tables IV and V are plotted in Figs. 4 and 5 for the noncentered and centered data, respectively. We see that, for the noncentered data, the formant variable analysis gives (for up to five dimensions) somewhat higher scores than the principal-components analysis. For the centered data, however, three factors give a better score than the three formant frequencies do. In this case, the scores with the maximally discriminating plane are only slightly less than with the $\log F_1 - \log F_2$ plane. Generally, we may conclude that a description of vowel sounds in terms of the formant variables and a description in which a principal-components analysis of the whole frequency spectrum is carried out, result in quite comparable identification scores.

For an evaluation of the principal-components approach it is also of interest to compare the configuration of vowel points in the $\log F_1 - \log F_2$ plane with their configuration in the plane that gives maximal discrimination, based on factor analysis of the $\frac{1}{3}$-oct SPLs. Consider first the data averaged over the 50 speakers. The two centered configurations of 12 points each can be matched only if first the total variance of the two configurations is equalized. Then the maximally discriminating plane is rotated over such an angle that it coincides as well as possible with the $\log F_1 - \log F_2$ plane. As a criterion for best coincidence, the method of least squares was used. In this case we minimized the sum of the squares of the distances between corresponding points in the two superimposed planes. The result of this matching procedure is presented in Fig. 6.

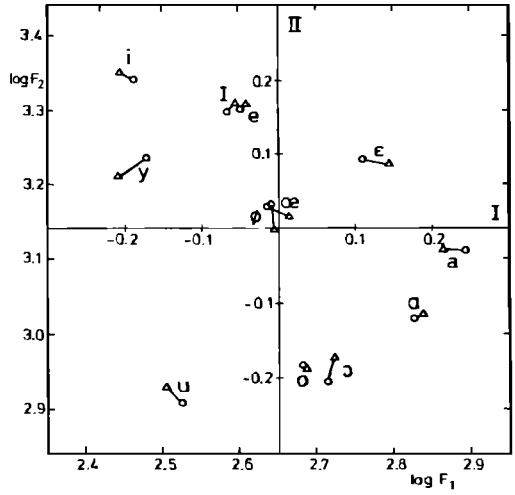

FIG. 6. Result of matching the configuration of the average centered vowel points in the maximally discriminating plane (△) to the configuration of average centered points in the $\log F_1$–$\log F_2$ plane (○). If to the coordinate values along dimensions I and II are added 2.652 and 3.141, respectively (being the overall averages of $\log F_1$ and $\log F_2$), these axes again represent $\log F_1$ and $\log F_2$ (outer scales).

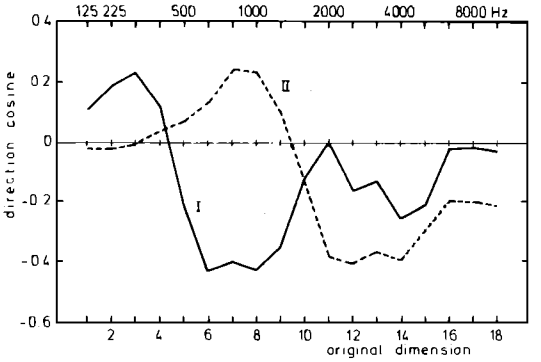

FIG. 7. Direction cosines between these axes in the maximally discriminating plane, related to best match with the $\log F_1$–$\log F_2$ plane, and the original 18 dimensions representing the levels below overall SPL in 18 $\frac{1}{3}$-oct bands.

There is an excellent agreement between the corresponding points indicating that, for the average vowels, the maximally discriminating plane is approximately equivalent to the $\log F_1 - \log F_2$ plane. This equivalence is illustrated by the very high correlations between the coordinate values (0.989 and 0.993 for the horizontal and vertical axes, respectively). The direction cosines (eigenvectors) between these axes and the original 18 dimensions are given in Fig. 7. We can see that these eigenvectors are appropriate to extract information from the frequency spectra comparable with F_1 and F_2: The eigenvectors I and II have their steepest slopes over the ranges 315–630 and 1000–2000 Hz, respectively, agreeing rather well with the ranges in which F_1 and F_2 vary (see Fig. 2). A steep slope of the direction-cosine curve at a certain frequency means that the filter band levels above and below that frequency are well distinguished.

The direction cosines presented in Fig. 7 make it possible to derive *directly* from the $\frac{1}{3}$-oct frequency spectrum a two-dimensional representation (called the *optimal* plane) in which the coordinates correlate highly with $\log F_1$ and $\log F_2$. Since this relation was computed for average vowel data, we should like to know also how well this derivation holds for the data points of each speaker individually. Therefore, the same eigenvector base, which was found to be optimal for the average data, was also applied to all centered individual data, using the same scale factor. As a measure of correspondence we used, for each individual, the distance between corresponding vowel points if both planes (optimal plane and $\log F_1 - \log F_2$ plane) were superimposed. Of these 600 distances 48% was smaller than 0.06, and 91% smaller than 0.12 (see Fig. 3 for an interpretation of these numbers). Though these distances are not large, they indicate that if we are only interested in F_1 and F_2 it will be better to measure them directly instead of computing them from $\frac{1}{3}$-oct-level data. The clusters of the various vowels in the $\log F_1 - \log F_2$ plane and in the optimal plane can be superimposed very well, but this does not hold for the corresponding data points individually.

III. DISCUSSION

In the previous sections, we compared two apparently different approaches to analyzing the spectra of vowel sounds. The first consisted of a narrow-band frequency analysis followed by a determination of the frequencies and levels of the lower three peaks in the envelope of the frequency spectrum, the formants. Thus we used the most prominent acoustical features to characterize the various vowels. The second approach consisted of a $\frac{1}{3}$-oct frequency analysis followed by a principal-components analysis. In order to compare the two approaches, we computed likelihood regions for the various vowels and, on their basis, identification scores. Both methods resulted in rather similar scores, particularly if more than two dimensions were used (Figs. 4 and 5). As far as higher identification scores were found using the formant data, we should remember that lower values would have been obtained had a more objective technique been used to determine the formant frequencies and formant levels. The three-dimensional data derived with principal-components analysis were used to find a maximally discriminating plane by optimizing the identification score. After an optimal rotation of this plane to the $\log F_1 - \log F_2$ plane (Fig. 6), the high correlation suggests that the two approaches are closely related. The optimal plane may be considered as an alternative for the formant description of the data having about the same "information content" in terms of recognition results. For the individual data points, however, the correlation between the optimal plane and $\log F_1 - \log F_2$ plane is lower. This means that the position of an individual vowel point in the optimal plane cannot be considered as an accurate substitute for the formant frequencies.

Having found this alternative for the formant description of the data, we may consider the question if this or the formants themself might be preferred as a description of the spectra of vowel sounds. In our opinion, no general answer to this question can be given. Preference depends upon the goal one sets for studies of vowel sounds. At least four different interests can be distinguished.

(1) If one is interested in the relation between vowel spectra and vowel *production*, one would like to describe the vowel sounds in terms that are related to parameters of the vocal tract. Though it is reasonable to suppose that, just as the factor representation compared so well with the formant representation, a similar comparison of the factor representation with articulatory features (like tongue height and tongue advancement) would be successful,[10] it is obvious that the formant description has advantages. The formant variables certainly give a more direct insight in the physical properties of the vocal tract than the factors do.

(2) If one is interested in the relation between vowel spectra and vowel *perception*, the situation is different. There is no *a priori* reason why, in this case, the spectral differences between vowels should be described by distinctive features related to their production. It was reported earlier[1,11] that the vowel representation in terms of principal components derived from $\frac{1}{3}$-oct frequency analysis corresponds quite well with perceptual representations derived from confusion and scaling experiments. Generally, the spectral differences between complex tones of equal loudness and pitch are highly correlated with their perceptual differences, irrespective of whether the spectra are characterized as formants or not.[12] The presence of the formants in vowel spectra does not imply that they are perceived in a specific way by some sort of "formant extractors" in the auditory system. Whether this is the case or not has to be decided by psychoacoustical experiments. At the moment, no decision can be made.

(3) In studying the *dynamic structure* of speech, the factor analysis approach has some clear advantages. The coordinate values along the axes in the reduced dimensional representation can be computed in a very short time. Thus it is easy to follow sound transitions rather precisely as changing trajectories in that space (for instance, every 10 or 15 msec a sample can be taken). Since the approach is not restricted to vowel sounds but can be applied to any sound, also complete words and sentences can be represented and studied in this way.

(4) The advantage of dimensional analysis over formant analysis is most obvious in *automatic speech recognition* systems. As we have seen, both approaches give comparable identification scores for vowels. Since it is rather difficult to develop a fast algorithm to extract formant frequencies,[5] the alternative technique is very attractive, because it can be applied in real time and because it is both objective and simple. In this case as well, it is a great advantage that the approach is not restricted only to vowel sounds. The technique has been successfully used already by one of the authors[13] in on-line speech analysis and real-time word recognition.

IV. CONCLUSIONS

- Statistical analysis of formant frequencies and formant levels of 12 Dutch vowels confirms that F_1 and F_2 are the most appropriate two distinctive parameters for describing the spectral differences among the vowel sounds.
- By means of a principal-components analysis of $\frac{1}{3}$-oct power levels, identification scores are obtained which are comparable with scores based upon formant analysis.
- For the data averaged over the speakers, the optimal plane derived from the $\frac{1}{3}$-oct data is equivalent with the $\log F_1 - \log F_2$ plane suggesting that the two approaches may be closely related.
- Principal-components analysis is preferred above formant analysis in automatic speech recognition because it is much faster and simpler and is also applicable to nonvowel sounds.

ACKNOWLEDGMENT

The authors wish to thank D. J. P. van Nierop, who did part of the computation in the final stage of the research.

APPENDIX A: DATA REDUCTION TECHNIQUES

For a multidimensional set of data, as we have with our 18-dimensional $\frac{1}{3}$-oct vowel spectra, there are various possible techniques to diminish the number of dimensions in order to reduce the amount of data. An important requirement is that as much of the original information as is possible should be preserved by this reduction. Some of the possible procedures are:

(1) *Analysis of variance.* This ranking technique looks to the ratio (F ratio) between the within-class variance and the between-class variance per dimension. A large ratio for a certain dimension means that the variation between the different vowels is large relative to the variation within a vowel for different speakers. Then one can single out the dimensions with the highest F ratios. This dimension reduction (feature evaluation) is used by Pruzansky[14] and Das and Mohn[15] in speaker identification and verification experiments. The main disadvantage of this technique is that the interactions among dimensions are neglected.

(2) *Principal-components analysis.* This technique[16] successively calculates new directions, being linear combinations of all original dimensions, which explain as much of the residual variance as is possible. Here, dependency between the original dimensions is taken into account. However, the variance between vowels is not optimized relative to the within-vowels variance. This may result in a subspace in which the variance is maximal, but in which identification is not optimal for that number of dimensions.

(3) *Discriminant analysis.* This technique is a combination of (1) and (2); it maximizes the between-class differences relative to the within-class differences, in a reduced number of dimensions, being linear combinations of the original dimensions. The simple case of classifying two clusters of data is fundamental to this analysis.[17] However, in our data, we do not have two but 12 clusters, moreover with unequal within-class variances. Mohn[18] describes a modified discriminant analysis which is applicable also to this type of data. For our data, the identification scores for the centered data in two, three, and four dimensions, computed from the modified discriminant analysis, are 89.3%, 93.0%, and 99.0%. These scores are in general somewhat higher than in Table V, despite the fact that for this discriminant analysis the different within-class covariance

matrices had to be pooled to one average within-class covariance matrix.

(4) *Maximally discriminating plane.* As discussed under (2), a principal-components analysis gives a subspace with a maximal amount of explained variance but not necessarily with maximal identification. So, if one is primarily interested in a subspace which gives highest identification scores, then it is better to make this the prime criterion. As far as we know, no algorithm exists to find this space directly, so we have done it with an iterative procedure. Computational limitations did not make it possible for us to determine iteratively a more than two-dimensional maximally discriminating space in a more than three-dimensional subspace. So we started with the three-dimensional factor space and rotated the I–II factor plane in small steps in that space. After each small rotation the percentage correct score was determined for the data projected on this plane. If the correct score became larger, one more rotation step in the same direction was done, otherwise another direction was chosen. In this iterative way, the best plane within the three-dimensional factor space was found; it made angles of 32° and 12° with the I–II factor plane. We call this the maximally discriminating plane; the correct score in this plane for the centered data is 92.2%. This score is significantly higher than in the I–II factor plane (88.0%).

Which type of data reduction technique one should prefer depends on the type of data, the computational limitations, and the final goal of the research project. If one has very many parameters,[15] the analysis of variance seems most appropriate. If one wants to use a general data reduction technique and is not primarily interested in optimal recognition results, then the principal-components analysis seems to be a good technique.[1] The discriminant analysis seems to give the best recognition results but is rather complicated and in many situations not directly applicable.[19] The fourth iterative technique is very time consuming and limited in its use since no general algorithm has yet been developed.

*Present address: Postal and Telecommunications Services, Leidschendam, The Netherlands.

[1]W. Klein, R. Plomp, and L. C. W. Pols, "Vowel Spectra, Vowel Spaces, and Vowel Identification," J. Acoust. Soc. Am. **48**, 999–1009 (1970).

[2]G. Fant, "Acoustic Analysis and Synthesis of Speech with Applications to Swedish," Ericsson Tech. **1**, 3–108 (1959).

[3]G. E. Peterson and H. L. Barney, "Control Methods Used in a Study of the Vowels," J. Acoust. Soc. Am. **24**, 175–184 (1952).

[4]F. J. Koopmans, "Vergelijkend fonetisch klinkeronderzoek," Inst. Phonetic Sci. U. Amsterdam Publ. No. 32 (1971).

[5]R. W. Schafer and L. R. Rabiner, "System for Automatic Formant Analysis of Voiced Speech," J. Acoust. Soc. Am. **47**, 634–648 (1970).

[6]J. F. Boehm and R. D. Wright, "Speaker Normalization for Automatic Word Recognition," J. Acoust. Soc. Am. **49**, 133(A) (1971).

[7]L. J. Gerstman, "Classification of Self-Normalized Vowels," IEEE Trans. Audio Electroacoust. **AU-16**, 78–80 (1968).

[8]B. M. Lobanov, "Classification of Russian Vowels Spoken by Different Speakers," J. Acoust. Soc. Am. **49**, 606–608(L) (1971).

[9]G. Fant, "A Note on Vocal Tract Size Factors and Non-Uniform *F*-Pattern Scaling," STL-QPSR **4**, 22–30 (1966).

[10]S. Singh and D. R. Woods, "Perceptual Structure of 12 American English Vowels," J. Acoust. Soc. Am. **49**, 1861–1866 (1971).

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

[12]R. Plomp, "Timbre as a Multidimensional Attribute of Complex Tones," in *Frequency Analysis and Periodicity Detection in Hearing*, edited by R. Plomp and G. F. Smoorenburg (A. W. Sijthoff, Leiden, The Netherlands, 1970), pp. 397–411.

[13]L. C. W. Pols, "Real-Time Recognition of Spoken Words," IEEE Trans. Comput. **C-20**, 972–978 (1971).

[14]S. Pruzansky, "Talker-Recognition Procedure Based on Analysis of Variance," J. Acoust. Soc. Am. **36**, 2041–2047 (1964).

[15]S. K. Das and W. S. Mohn, "A Scheme for Speech Processing in Automatic Speaker Verification," IEEE Trans. Audio Electroacoust. **AU-19**, 32–43 (1970).

[16]H. H. Harman, *Modern Factor Analysis* (U. of Chicago Press, Chicago, 1967).

[17]R. B. Cattell, Ed., *Handbook of Multivariate Experimental Psychology* (Rand McNally, Chicago, 1966).

[18]W. S. Mohn, "Statistical Feature Evaluation in Speaker Indentifiction," Ph.D. dissertation, North Carolina State U., Rayleigh, N. C. (1969).

[19]P. D. Bricker, R. Gnanadesikan, M. V. Mathews, S. Pruzansky, P. A. Tukey, K. W. Wachter, and J. L. Warner, "Statistical Techniques for Talker Identification," Bell Syst. Tech. J. **50**, 1427–1454 (1971).