Perceptual and physical space of vowel sounds

Pols, L.C.W.; van der Kamp, L.J.T.; Plomp, R.

Published in:
The Journal of the Acoustical Society of America

DOI:
10.1121/1.1911711

9.5, 9.3

Received 16 January 1969

Perceptual and Physical Space of Vowel Sounds

L. C. W. Pols, L. J. Th. van der Kamp,* and R. Plomp

Institute for Perception RVO–TNO, Soesterberg, The Netherlands

Experiments were carried out to investigate the correlation between the perceptual and physical space of 11 vowel sounds. The signals were single periods out of the constant vowel part of normally spoken words of the type h(vowel)t, generated continuously by computer. Pitch, loudness, onset, and duration were equalized. These signals were presented to 15 subjects in a triadic-comparison procedure, resulting in a cumulative similarity matrix. Multidimensional scaling (Kruskal) of this matrix resulted in a three-dimensional perceptual space with 1.6% stress. The signals were also analyzed physically with ⅓-oct band filters. Principal-components analysis of the decibel values per frequency band indicated that three dimensions accounted for 81.7% of the total variance. Matching the perceptual and the physical configurations to maximal congruence yielded an excellent result with correlation coefficients of 0.992, 0.971, and 0.742 along the corresponding dimensions. The formant frequencies and levels were correlated also with both configurations.

INTRODUCTION

The relation between the perceptual differences in vowel sounds and the differences in their articulatory and physical properties has received more and more attention in recent years.

Our inability to order the various vowel sounds along a single perceptual scale means that a complex attribute is involved. The complexity of the attribute can be met by a psychological (perceptual) space in which each stimulus is represented by a point. The dimensionality of the space and the positions of the points can be determined by multidimensional scaling.

An articulatory description of the vowel sounds can be given in terms of tongue-hump position (front–back), and degree of constriction (high–low).[1] Apart from some references, this approach is left out of consideration in this paper.

The multidimensional character of vowels is also apparent from a physical analysis of the sounds by means of narrow- or wide-band filtering. From the resulting frequency spectra, specific data can be extracted, for example, the frequencies (F_i) and sound-pressure levels (L_i) of the formants. In this respect, the work of Peterson and Barney[2] is of prime importance. They tried to relate vowel qualities with formant patterns. Plotting 10 vowels, spoken by 76 persons, in the F_1-F_2 plane showed an overlap for some vowel areas, even when consideration was limited to those vowels which were unanimously correctly classified by a group of listeners. The addition of the frequency of the third formant as an extra dimension did not greatly reduce this overlapping. Axis transformations applied to the three-dimensional representation of the data of Peterson and Barney simplified the boundaries, which is of value for developing a computer recognition logic, but did not reduce the overlapping (Foulkes[3]). Welch and Wimpress[4] used the same data to show that it may be possible, by applying multivariate statistical techniques, to subdivide the space in an optimal way with respect to recognition. An economical decision tree can then be extracted. About 13% errors remained when only the F_1-F_2 information was used. With F_3 added, the error rate reduced to 9%, and if the fundamental frequency and two formant levels are taken into account, still, an error rate of about 6% was present.

The spread of the vowel areas in the formant space is partly due to interindividual differences. When the vowels are spoken a number of times by the same person, then the corresponding vowel points in the two-formant plane result in strictly bounded areas with no overlap

* Psychological Institute, University of Leyden.

[1] J. L. Flanagan, *Speech Analysis, Synthesis and Perception* (Springer Verlag, Berlin, 1965), p. 16.

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

[3] J. D. Foulkes, "Computer Identification of Vowel Types," J. Acoust. Soc. Amer. 33, 7–11 (1961).

[4] 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).

(Potter and Steinberg[5]; Fant[6]). In this respect, it would be of great interest if a speaker-dependent correction could be found. Particular experiments suggest that a specific "reference space" is built up in the course of speech perception, a space which depends on the incoming speaker-dependent information (Ladefoged and Broadbent[7]). Gerstman,[8] adopting the data of Peterson and Barney,[2] introduced a speaker normalization based on only three vowels, resulting in a two-dimensional representation in which nearly all vowels were distinguishable (97.5%). Actually, he used a third dimension to distinguish the vowel [ɝ] from the rest of the vowels. This had been suggested previously by Potter and Steinberg.[5]

Another approach to studying the characteristics of vowel spectra was introduced by Plomp, Pols and van de Geer.[9] A dimensional analysis was carried out on frequency spectra, determined with $\frac{1}{3}$-oct band filters, of 15 Dutch vowels spoken by 10 subjects. Of the total variance, 84.1% could be "explained" by four factors. Using the shortest distance between the individual points and the mean vowel positions in this four dimensional space as a criterion, 90% correct identifications resulted. This value reduced to 85% when three instead of four dimensions were used. Comparable data reduction techniques, involving the use of dimensional analysis, have recently been described by Boehm and Wright[10] and Li *et al.*[11]

From these and other measurements (e.g., Fant[6]), we may conclude that at least three dimensions are necessary to describe the vowel sounds physically. It may be expected that the perceptual space is related to the physical space, since at least some of the physical dimensions must correspond to the way in which subjects discriminate between stimuli (Wilson and Saporta[12]). The minimal number of physical dimensions required to describe the differences between vowel sounds can be considered as an indication of the number of perceptual dimensions required. Thus, one can also expect that a perceptual specification of vowel sounds must comprise about three dimensions.

In specifying vowel sounds in terms of distinctive features, one needs four to describe them well: acute/grave, flat/plain, compact/diffuse, and tense/lax (Hemdal and Hughes[13]). Cohen *et al.*[14] found that, apart from specific formant bandwidths, at least the three factors F_1, F_2, and duration had to be combined in an optimal way for maximal recognition of synthetic vowels. From a confusion experiment with low-pass-filtered vowels, Miller[15] concluded that the same three features are necessary to specify every sound on a binary scale. By studying perceptual confusions among 12 vowels masked with noise, Pickett[16] found the same three features to be the most important ones, and to a lesser degree, relative intensity. Hanson[17] found, in scaling experiments with Swedish vowels, three perceptual dimensions. Two of these dimensions were related to F_1 and F_2, as well as to the distinctive features acute/grave and diffuse/compact. The meaning of the third dimension was less clear; he called it the perceptual contrast factor. Mohr and Wang[18] determined similarity matrices for vowels by aid of a paired-comparison procedure and coupled the rank order of the similarity indices with physiological features (high, mid, labial, palatal, nasal). Some of these features had significant effects on the similarity scores.

From this brief survey of the literature we can conclude that vowel sounds have a multidimensional character. At least three factors can be derived which must be related to acoustical, articulatory, and linguistic features. In most of the articles mentioned, this relation is indicated in an arbitrary way, such as by comparing the rank order of the stimuli along a perceptual dimension with one or another feature. By optimal rotation of the configuration and mathematical matching techniques the relation between perceptual and physical dimensions can be examined more thoroughly. We decided to study this relation by applying the most advanced (in our opinion) techniques for perceptual and physical analyses and data processing.

---

[5] R. K. Potter and J. C. Steinberg, "Toward the Specification of Speech," J. Acoust. Soc. Amer. 22, 807–820 (1950).

[6] C. G. M. Fant, "Acoustical Analysis and Synthesis of Speech with Application to Swedish," Ericsson Tech. 1, 1–108 (1959).

[7] P. Ladefoged and D. E. Broadbent, "Information Conveyed by Vowels," J. Acoust. Soc. Amer. 29, 98–104 (1957).

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

[9] 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).

[10] J. F. Boehm and R. D. Wright, "Dimensional Analysis and Display of Speech Spectra," J. Acoust. Soc. Amer. 44, 386(A) (1968).

[11] K. P. Li, A. S. House, and P. W. Hughes, "Vowel Classification using a Dispersion Analysis Method," J. Acoust. Soc. Amer. 44, 390(A) (1968).

[12] K. Wilson and S. Saporta, "Linguistic Organization," in *Psycholinguistics. A Survey of the Theory and Research Problems*, C. E. Osgood and T. A. Sebeok, Eds. (Indiana University Press, Bloomington, Ind., 1965), pp. 77–83.

[13] J. F. Hemdal and G. W. Hughes, "A Feature-Based Computer Recognition Program for the Modelling of Vowel Perception," in *Proceedings of the Symposium on Models for the Perception of Speech and Visual Form* (MIT Press, Cambridge, Mass., 1967), pp. 440–452.

[14] A. Cohen, I. H. Slis and J. 't Hart, "On Tolerance and Intolerance in Vowel Perception," Phonetica 16, 65–70 (1967).

[15] G. A. Miller, "The Perception of Speech," in *For R. Jakobson; Essays on the Occasion of his Sixtieth Birthday*, M. Halle *et al.*, Eds. (Mouton and Company, 's-Gravenhage, The Netherlands 1956), pp. 353–359.

[16] J. M. Pickett, "Perception of Vowels Heard in Noises of Various Spectra," J. Acoust. Soc. Amer. 29, 613–620 (1957).

[17] G. Hanson, "Dimensions in Speech Sound Perception. An Experimental Study of Vowel Perception," Ericsson Tech. 23, 3–175 (1967).

[18] B. Mohr and W. S. I. Wang, "Perceptual Distances and the Specification of Phonological Features," Phonetica 18, 31–45 (1968).

I. PERCEPTUAL ANALYSIS

A. Introduction

The aim of a perceptual analysis is to determine a psychological stimulus space on the base of observations concerning the relative similarity of the stimuli. Some of the methods used in the field of psychoacoustics are:

(1) Short-term recall.[19] This is an interesting technique in which the subjects are asked to repeat, successively, a number of presented stimuli, followed by a recall of the total set. On the basis of the errors made, an error matrix can be determined, which provides information about the coding mechanism. This coding mechanism includes, however, the memory function, in which we are not presently interested. A further disadvantage of this technique is that it is unsuitable for stimuli that cannot be easily denominated.

(2) Scaling based on perceptual confusion.[15,20] With undistorted signals, perceptual confusions will be rare. Therefore, some sort of distortion has to be introduced to prevent too many empty cells in the error matrix. This can be a serious disadvantage. The mathematical techniques required to deduce a perceptual space from a confusion matrix are still in development. Special difficulties are related to asymmetry and response bias in such matrices. Usually the information in an error matrix is partly used by looking only to the trend of the confusions (e.g., Pickett[16]). We, in fact, also used the method of perceptual confusion to determine a perceptual space. Because of the variety of possible techniques for handling the confusion data, these results will not be included in this paper but will be published separately.[21] In that article, attention will also be given to methodological issues.

(3) Semantic scaling.[22,23] In this procedure, the subject has to associate presented stimuli with a set of bipolar adjectival scales (semantic differential; Osgood[24]). The subject's task is to indicate for each stimulus on, for instance, a seven-point scale, which of the polar terms, and to what extent, applies to the stimulus. The main drawback of this technique is that the subjects are forced to judge the stimuli in terms of prescribed bipolar scales or verbal categories. Such categories may well be different from his auditory impression. Furthermore, the preselection of component scales restricts the final solution of the analysis.

(4) Direct scaling by ratio estimation.[17] In direct scaling, the magnitude of similarity or dissimilarity between pairs of stimuli is judged on a numerical or graphical scale, whether or not the pair is in relation to a standard stimulus pair. Our experience is that untrained subjects often find it difficult to make consistent judgments, resulting in a large spread in their responses. Hanson,[17] using both direct (ratio estimation) and indirect (triadic comparison) scaling techniques for different numbers of vowels in the stimulus sets, found no essentially different results. The published individual results of the direct ratio estimations, however, suggest large interindividual differences. He made no attempt to study the specific interindividual differences by using, for example, a technique proposed by Tucker and Messick.[25] Besides the fact that human observers consider it easier to provide information at an ordinal level than at a ratio level, an objection of quite another type can be made against the direct use of ratio-judgment results in multidimensional scaling techniques (i.e., transforming the observed (dis)similarities into scalar products and then factor analyzing these products). The objection concerns the strong assumptions to be made to justify the application of factor-analytic techniques.

(5) Scaling based on triadic comparison.[26] In this method, an extended form of paired comparison, one has to decide, for each possible subset of three stimuli, which pair is most similar and which pair is least similar, without further indicating the degree of similarity. Moreover, the subjects are not obliged to make their judgments in relation to specific categories. Subjects consider this decision task to be rather simple, and hardly any instructions need be given. This technique thus has some essential advantages over other scaling methods.

B. Method

On the basis of the foregoing considerations, we decided to collect our similarity data by the method of triadic comparison. From the single decisions of the subject, a similarity matrix is built up in the following way. The subject, presented with a given triad, has to select the pairs of stimuli that are, in his opinion, most similar and most dissimilar. Now the three pairs of triads can be ordered with respect to similarity. The most similar pair receives two points; the intermediate pair, one point; and the least similar pair, no point. These scores, cumulated for all triads, result in a similarity matrix in which every cell contains the

[19] W. A. Wickelgren, "Distinctive Features and Errors in Short Term Memory for English Vowels," J. Acoust. Soc. Amer. 38, 583 588 (1965).

[20] W. E. Castle, "The Effect of Narrow Band Filtering on the Perception of Certain English Vowels," *Janua Linguarum, Series Practica 13* (Mouton and Company, 's-Gravenhage, The Netherlands, 1964).

[21] L. J. Th. van der Kamp and L. C. W. Pols, "Perceptual Analysis from Confusions among Vowels" (to be published).

[22] L. N. Solomon, "Semantic Approach to the Perception of Complex Sounds," J. Acoust. Soc. Amer. 30, 421–427 (1958).

[23] J. P. van de Geer, W. J. M. Levelt, and R. Plomp, "The Connotation of Musical Intervals," Acta Psychol. 20, 308–319 (1962).

[24] C. E. Osgood, "The Nature and Measurement of Meaning," Psychol. Bull. 47, 197–237 (1952).

[25] L. R. Tucker and S. Messick, "An Individual Difference Model for Multidimensional Scaling," Psychometrika 28, 333–368 (1963).

[26] W. J. M. Levelt, J. P. van de Geer, and R. Plomp, "Triadic Comparison of Musical Intervals," Brit. J. Math. Stat. Psychol. 19, 163–179 (1966).

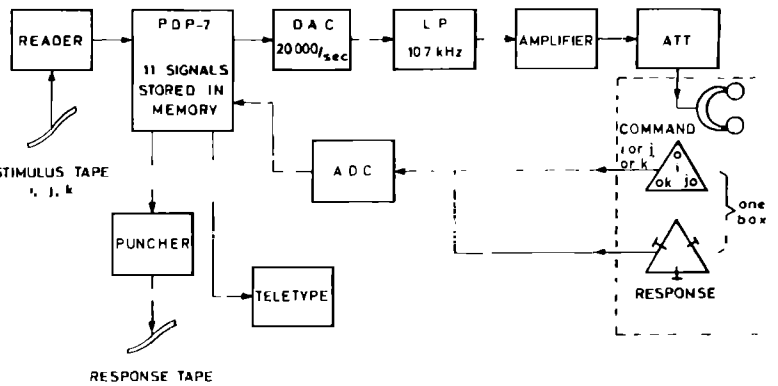

FIG. 1. Block diagram of the experimental setup.

number of times a pair is judged more similar than the other pairs. A large value, therefore, means a very similar pair; in other words, a perceptually short distance, and a small value means a highly dissimilar pair.

The similarity judgments of the subjects have to be transformed into distances in a perceptual space. The nature of the relation between similarity indices and interpoint distances is more or less arbitrary. Torgerson[27] used the law of comparative judgment to relate the proportion of times that Stimulus k is judged closer to Stimulus i than to j, to the corresponding differences in distances between k–j and i–k. The only assumption made by Kruskal,[28] who worked out the ideas of Shepard,[29] is that of a monotonic inverse relationship between interpoint distances and similarity indices. The character of this relation is not further restricted. This assumption means only that, if the similarity index of one pair is smaller than that of another one, the interpoint distance of the first pair in the multidimensional representation must be larger than the distance of the latter pair. Therefore, the absolute values are not important, only the rank order. Other possibilities of analyzing our data would have been those proposed by Guttman and Lingoes. The programs required for such analysis, however, were not yet available; furthermore, there are indications that in regard to the final configurations the different methods give similar results (e.g., Lingoes[30]).

The multidimensional-scaling computer program, originally described by Kruskal[31] and adapted by us for our computer, starts with an arbitrary configuration in a number of dimensions chosen beforehand, calculates the distances between the points in this configuration, and makes a scatter diagram with the similarity indices along one axis and the calculated distances along the other. Then, a monotonic regression of distances upon similarity indices is performed, and the residual variance, after suitable normalization, is used as a quantitative measure of "goodness of fit," called *stress*. In fact, the stress is the square root of a residual sum of squares, which can be expressed as a percentage. By changing the configuration in an iterative way by the method of steepest descent, a configuration with a minimal value for the stress can be obtained. The configuration with minimal stress gives those coordinates of the points, in the desired number of dimensions, whose rank order of distances fits best with the rank order of the similarities. An interpretation of the amount of stress is a matter of intuition and experience (Roskam[32]), and an objective criterion for evaluating the stress cannot be given. It depends on the kind of data and the number of dimensions. A comparison with the distribution of stress percentages found by analysis of random data can be an indication for deciding whether a stress value for a given configuration is significant or not (Wagenaar and Padmos[33]). Neither is there an objective way of determining the number of dimensions concealed in a given similarity matrix. The usual criteria in the Kruskal technique are: (1) looking for "elbows" in the curve that represents minimal stress as a function of the number of dimensions and (2) the interpretability of the coordinates. According to Kruskal,[28] "it is reasonable to choose a value of the dimensionality which makes the stress acceptably small, and for which further increase in dimensionality does not significantly reduce stress."

C. Experimental Setup

As stimuli, 11 vowel-like sounds were used. These signals were derived by taking one period out of the

[27] W. S. Torgerson, *Theory and Methods of Scaling* (John Wiley & Sons, Inc., New York, 1958).

[28] J. B. Kruskal, "Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis," Psychometrika **29**, 1–27 (1964).

[29] R. N. Shepard, "The Analysis of Proximities: Multidimengional Scaling with an Unknown Distance Function," Psychometrika **27**, 125–140, 219–246 (1962).

[30] J. C. Lingoes, "Recent Computational Advances in Nonmetric Methodology for the Behavioral Sciences," in *Proceedings of the International Symposium: Mathematical and Computational Methods in the Social Sciences* (International Computation Centre, Rome, 1967).

[31] J. B. Kruskal, "Nonmetric Multidimensional Scaling: A Numerical Method," Psychometrika **29**, 115–129 (1964).

[32] E. E. Ch. I. Roskam, "Metric Analysis of Ordinal Data in Psychology," thesis, University of Leyden (1968).

[33] W. A. Wagenaar and P. Padmos, "The Significance of a Stress Percentage Obtained with Kruskal's Multidimensional Scaling Technique," Rep. No. IZF 1968-22, Institute for Perception RVO-TNO, Soesterberg, The Netherlands (1968).

TABLE I. Cumulative similarity matrix of 11 vowel-like sounds (15 subjects).

	[œ]	[ɔ]	[ɑ]	[u]	[y]	[i]	[a]	[ɸ]	[o]	[ɛ]	[e]
[œ]	···	168	140	115	145	100	108	250	180	145	201
[ɔ]		···	162	152	116	51	99	144	230	158	130
[ɑ]			···	99	87	78	205	144	195	209	130
[u]				···	208	78	66	94	144	141	85
[y]					···	155	68	134	106	118	84
[i]						···	70	115	57	91	104
[a]							···	122	141	187	127
[ɸ]								···	156	150	225
[o]									···	163	135
[ɛ]										···	160
[e]											···

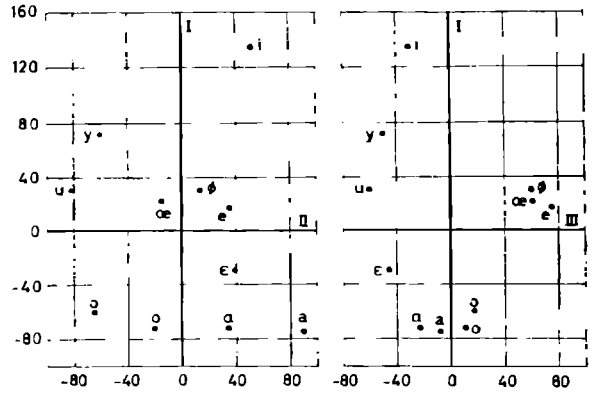

FIG. 2. Projections of the 11 stimulus points on two perpendicular planes of the three-dimensional perceptual space.

constant vowel parts of 11 CVC words of the type h(vowel)t, spoken by one of the authors. The vowels used were [œ], [ɔ], [ɑ], [u], [y], [i], [a], [ɸ], [o], [ɛ], and [e] (IPA[34]), which are in written language the Dutch vowels u, o, a, oe, uu, ie, aa, eu, oo, e, and ee, respectively. In order to make it possible to cut one period out of the spoken words, each word was sampled (eight bit) via an analog-to-digital converter (ADC) with a rate of 20 000 samples per second, and all samples were stored in the memory of a digital computer (DEC PDP-7). By means of a simple machine-language program it was possible to generate every wanted part of the stored word via a digital-to-analog converter (DAC). A low-pass filter with a cutoff frequency of 10.7 kHz and a slope of −42 dB per oct filtered out the 20-kHz sample frequency. In this way, one specific period, of about 8 msec, taken from the spoken word, could be repeated continuously.

Since we wanted to reduce the number of physical parameters of the sounds as much as possible, the signals were modified in such a way that only the information present in the frequency spectrum was varied. The fundamental frequency of the signals was equalized by resampling each signal in such a way that we got the same number of samples (162) for all vowel periods. This resulted in a fundamental frequency of 123.5 Hz for the vowels. The first sample of the vowel periods was on, or in the neighborhood of, the zero line, thus minimizing the onset transients as far as possible. By repetition of one period a fixed number of times, the duration of all signals was made exactly the same (405 msec). Five subjects matched the loudness of all stimulus pairs in order to determine the mean loudness deviations. By correction of the amplitudes, the loudness levels of all 11 stimuli were made equal. Applying the above described procedure, we obtained stimuli that sounded like sustained vowels. The essential information necessary to generate these stimuli now only consists of one-period samples. Eleven of these sampled periods constituted the whole listening material; this material could easily be stored in the memory of the computer and generated at request by asking for the number of the signal. The stimuli were presented binaurally (Beyer headphones DT-48) at a sensation level of 50 dB, in a quiet room.

Application of the computer as a stimulus generator has the great advantage that the stimuli are momentarily available in any wanted order. The use of tape recorders, as done by others (Hanson,[17] Knops[35]), involves the restriction that the stimuli per triad have to be presented after each other in a fixed order.

In our setup, the whole triadic experiment is controlled by the computer (see Fig. 1). A paper tape is read in, on which the numbers of the signals for all triadic combinations are available in a random order, with the constraint that no two successive triads have any pair of stimuli in common. If the triad with the stimuli i, j, and k has to be compared, each of these stimuli is generated by request of the listener. For that purpose the subject pushes one of three stimulus buttons located at the vertices of an equilateral triangle. By operating the three buttons, he can listen in any order to the three different stimuli (maximal duration, 405 msec; the subject can, however, switch to another stimulus within this 405 msec). When he has decided which pair is, in his opinion, most similar, he pushes the response button positioned between the two stimulus buttons corresponding to the two stimuli. He does the same for the most dissimilar pair. His responses are automatically recorded with a teletypewriter and punched out on a response paper tape. Immediately thereafter, the code for the next triad is read in and the subject can compare the stimuli of that triad. In this way, one needs about 1 h to judge all 165 triads that are possible with 11 signals (11.10.9/3.2.1). During the experiment, the presence of an experimenter is not necessary. The similarity judgments of the subjects are gathered in a similarity matrix. This matrix is the input

[34] International Phonetic Association, *The Principle of the International Phonetic Association* (Department of Phonetics, University College, London, W. C. 1, 1967).

[35] L. Knops (personal communication), Catholic Univ., Dep. Psychol., Leuven, Belgium.

TABLE II. Minimal stress percentages of 15 subjects in three and four dimensions.

Subject	Minimal stress in 4 dim.	3 dim.
1	3.2	4.0
2	3.4	4.1
3	2.4	4.2
4	1.8	4.4
5	1.3	4.1
6	0.9	2.9
7	4.2	5.4
8	3.0	3.4
9	4.3	7.4
10	1.6	3.9
11	6.6	7.9
12	5.5	8.8
13	5.0	8.3
14	5.7	10.5
15	5.4	6.9
Cumulative	0.5	1.6

for the Kruskal multidimensional-scaling program which determines the spatial configuration with interpoint distances that best fit the similarity indices.

D. Results

Fifteen subjects (four female), all with normal hearing and between 20 and 30 years old, participated in the triadic comparison. Each subject judged 165 triads presented in a random sequence, yielding as data 15 similarity matrices. By summation, one cumulative matrix was determined, which is presented in Table I. With the Kruskal multidimensional-scaling program, a three-dimensional configuration with a minimal stress value of 1.6% was found. With a four-dimensional configuration, the minimal stress value was 0.5% and in two dimensions, 8.2%. On the basis of the criteria mentioned in Sec. I-B, the three-dimensional configuration was chosen for further analysis. In Fig. 2, the positions of the points in this configuration are given as projections on two perpendicular planes. The orientation of the coordinate axes is not mathematically

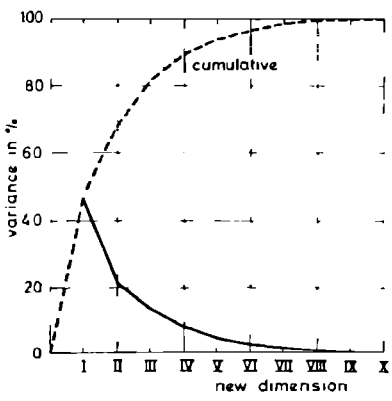

FIG. 3. Percentages of the total variance explained by the computed new dimensions.

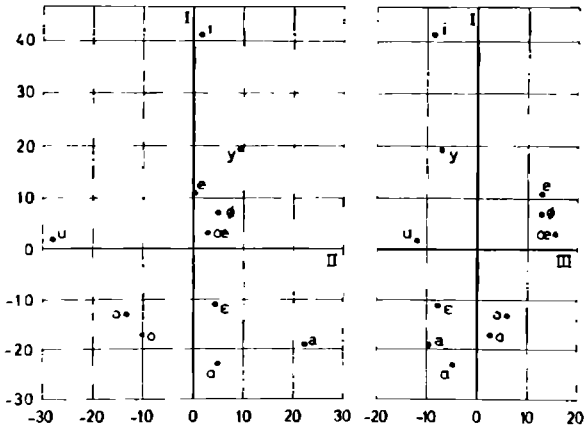

FIG. 4. Positions of the 11 signals in the physical I–II and I–III planes.

unique; therefore, suitable rotation of this configuration is permitted.

Also, the individual similarity matrices were analyzed. The minimal stress values in three and four dimensions are gathered in Table II. We were interested in the question of whether interindividual differences in the perceptual structure would merge, after suitable rotation, in some compromise position. Such an analysis, however, would destroy individual differences (McGee[36]). Alternative approaches that would preserve interindividual differences are suggested by Tucker and Messick[25] and by McGee.[36] Research in this line is still in progress. To get an estimate of the homogeneity of the similarity ratings of the subjects, the (15×55) matrix, with the similarity indices per subject on the rows, was analyzed according to a theorem by Eckart and Young (see Ref. 25). As a result, it was concluded tentatively that one factor accounted for the differences in similarity ratings, i.e. that the similarity judgments of the subjects were homogeneous.

II. PHYSICAL ANALYSIS

In order to determine the physical space of the used signals, the continuously generated sounds were analyzed with $\frac{1}{3}$-oct band filters (Brüel & Kjær spectrometer 2112); below, we also discuss some other analysis techniques. As pointed out in an earlier article (Plomp *et al.*[9]), this bandwidth was chosen because it agrees rather well over a large frequency range with the critical bandwidth of the ear's analyzing mechanism (Plomp and Mimpen[37]). The sound-pressure levels in decibels in the 18 frequency bands constitute an (11×18) data matrix. In terms of a geometrical model, we can say that the sound spectra of the 11 vowels result in a set of 11 points in an 18-dimensional space. These

[36] V. E. McGee, "Multidimensional Scaling on N Sets of Similarity Measures: A Nonmetric Individual Difference Approach," Multivariate Behavioural Res. 4, 233–248 (1968).

[37] R. Plomp and A. M. Mimpen, "The Ear as a Frequency Analyzer. II," J. Acoust. Soc. Amer. 43, 764–767 (1968).

TABLE III. Formant frequencies in hertz, and levels in decibels, for the 11 used vowel-like signals.

	F_1	F_2	F_3	L_1	L_2	L_3
[œ]	500	1450	2150	37.5	22	15
[ɔ]	500	750	2750	33.5	29.5	9
[ɑ]	720	950	2850	34	31	9.5
[u]	250	620	2050	35	33	5.5
[y]	250	1600	2720	37.5	26	14
[i]	250	2100	3100	35.5	23	21
[a]	990	1450	2550	34	26	4.5
[ø]	495	1520	2220	35.5	20	18.5
[o]	550	920	2595	33	26	2.5
[ɛ]	740	1800	2600	35	22	10.5
[e]	495	2100	3300	36	19.5	4

points can always be described in a 10-dimensional space.

It is of interest to determine the minimal number of dimensions required to describe the data without loss of too much information. The maximal variance in any of the original dimensions was only 15%. Principal-components analysis (Horst[38]; Harman[39]) was performed with the following results. The first factor (new dimension 1) explained 46.6% of the total variance; the second one, 21.4%; and the third one, 13.6%. This means that, in three dimensions, 81.7% of the total variance could be explained. With four dimensions, the figure rose to 89.6%, and with five dimensions, to 94.0% (see Fig. 3). It is clear that the increase in variance accounted for by taking a more than three-dimensional solution is relatively small. So, for further analysis and for comparison with psychological and other physical structures, mainly the three-dimensional solution was chosen.

A representation of the points in the I–II and I–III plane of the physical space is given in Fig. 4, with the center of gravity of the set of points at the origin. This configuration is somewhat different from the one found in earlier experiments (Plomp *et al.*[9]). Partly, the difference can be explained by an interchange of Dimensions I and II. Moreover, the signals used do not necessarily represent the average Dutch vowel sounds by which name they are described in this article, owing to the use of only one period out of vowels spoken by only one person. This does not mean, however, that the signals were not recognizable as the appropriate vowels. Despite the modifications that were carried out on the sounds, and despite their isolated presentation, 10 of the 11 signals were practically unanimously denominated by 15 subjects as the vowels which were originally pronounced. The opinions about the 11th signal, [e], differed.

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

[39] H. H. Harman, *Modern Factor Analysis* (The University of Chicago Press, Chicago, 1967).

TABLE IV. Correlation coefficients between formant frequencies and levels for the 11 vowel-like sounds.

	F_1	L_1	F_2	L_2	F_3	L_3
F_1	...	−0.4677	−0.0282	−0.0605	−0.4088	0.0055
L_1		...	0.5338	−0.4545	−0.0907	0.5287
F_2			...	−0.8353	0.4271	0.5124
L_2				...	−0.2229	−0.3702
F_3					...	−0.0545
L_3						...

The line spectra of the signals were also computed, with a method described by Ralston and Wilf.[40] From the structure of these spectra, we determined the formant frequencies and levels. The formant levels were defined as the decibel values of the formant peaks relative to an arbitrary zero level. This information is summarized in Table III. The linear regression between these variables was determined in order to get an idea about the interdependency of these variables. The correlation coefficients are given in Table IV. From these coefficients, we may conclude that, for this group of signals, F_1 and F_2 are independent, and F_2 and L_2 are highly correlated. We can demonstrate the dependency between the different factors by a principal-components analysis of an (11×6) data matrix, consisting of the numbers given in Table III, but then normalized per dimension (equal variance along the axes). Three new factors already explain 86.3% of the total variance.

It is, of course, most interesting to find out whether the physical dimensions extracted from the results of the $\frac{1}{3}$-oct analysis can be related with the formant frequencies and levels. For that, we used the canonical-matching procedure[26] that is originally described by Cliff[41] under the name "orthogonal rotation to congruence (Case 1)." This procedure makes it possible to determine an optimal relationship between two sets of variables, e.g., the physical versus the formant configuration. In order to derive a maximal congruence, both configurations are transformed orthogonally in such a way that the covariance between projections of the points of both configurations on corresponding axes is maximal. This means also that the sum of the squares of the distances between corresponding points is minimized. One way to express the degree of correspondence is in terms of correlation coefficients computed for corresponding orthogonal axes. As far as we know, no significance tests for such correlations exist. The coefficients resulting out of a matching of the F_1–F_2 plane with the three-dimensional physical configuration (81.7% explained variance) are 0.974 and 0.816. Matching with the six-dimensional physical space (96.6% variance) raises these coefficients to 0.985 and 0.981. Projections of the points on the superimposed

[40] A. Ralston and H. S. Wilf, *Mathematical Methods for Digital Computers* (John Wiley & Sons, Inc., New York, 1958).

[41] N. Cliff, "Orthogonal Rotation to Congruence," Psychometrika 31, 33–42 (1966).

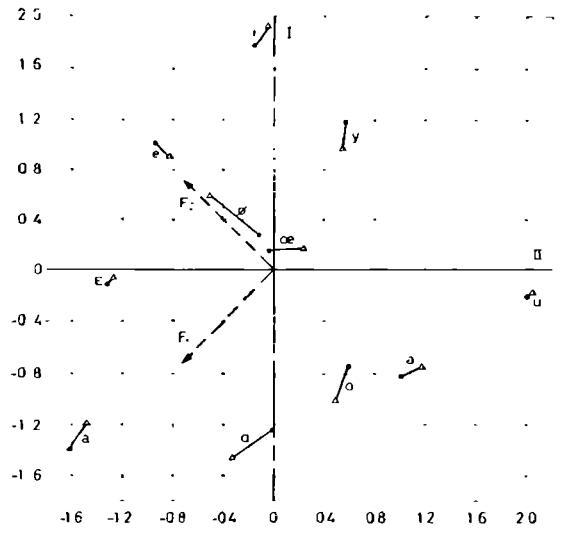

FIG. 5. Positions of the points when the F_1-F_2 configuration (●) is matched maximally with the six-dimensional physical configuration (△). The original orientations of the F_1 and F_2 axes are also given.

planes are given in Fig. 5. It is apparent that there is a large correspondence between both configurations, which brings us to the conclusion that the F_1 and F_2 information is almost completely present in our multidimensional physical representation, despite the fact that that one is derived from a broad-band ($\frac{1}{3}$-oct) analysis. The correspondence with the two dimensional physical space (68.0% variance) appears to be less good, thus indicating that if one wants to describe the spectral information of vowel sounds as positions of points in a plane, the F_1-F_2 plane is not the most proper one.

In order to provide the relation between all formant frequencies and levels and the physical space, we used the multiple-correlation technique.[42,43] This implies that, in a multidimensional space (e.g., physical), direction is determined which correlates maximally with a reference vector, being one of the outside variables (e.g., F_1). The multiple correlation coefficient defines the measure of correlation. This procedure is carried out for all formant frequencies and levels individually. The concerning multiple correlation coefficients are given in the last column of Table V for the three dimensional physical space, and in Table VI for the six-dimensional physical space. The p values indicating the level of statistical significance of these correlations[44] are also given. The results show that it is possible to find, in the three-dimensional physical space, directions which are highly correlated to F_1, F_2, and L_2. For F_1 and F_2, we already knew this from the canonical matching. The correlations and the significance levels were still improved in the six-dimensional physical

[42] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis* (John Wiley & Sons, Inc., New York, 1958).

[43] J. P. van de Geer, *Inleiding in de Multivariate Analyse* (Van Loghum Slaterus, Arnhem, The Netherlands, 1967).

[44] Ref. 42, p. 92.

TABLE V. Correlation coefficients between the projections of the points on the vectors corresponding maximally to F_i and L_i in the three-dimensional physical space. The multiple correlation coefficients are given in the last column (significance: ++ $p=0.01$, + $p=0.05$, n.s. = not significant).

	F_1	F_2	F_3	L_1	L_2	L_3	Mult. corr. coeff.
F_1	...	−0.1781	−0.1656	0.5272	0.0642	−0.5386	0.938 ++
F_2		...	0.9736	0.9112	0.8353	0.9251	0.860 +
F_3			...	0.8432	−0.6902	0.8959	0.344 n.s.
L_1				...	−0.8268	0.9850	0.740 n.s.
L_2					...	−0.7440	0.839 +
L_3						...	0.712 n.s.

space. No directions were found which correlated significantly with F_3, L_1, and L_3. The next question is if these "images" of the outside variables are independent, or perhaps more or less associated. An appropriate measure for that is the correlation between the projections of the points on the vectors corresponding maximally to the outside variables. These correlation coefficients are given in Tables V and VI. It is quite clear that there are, at least for this group of stimuli, only two independent factors, being F_1 and F_2. L_2 is negatively correlated to F_2. These relations already existed in the original formant frequency and level data (Table IV). It is, however, interesting that they can be found back in the same way in our multidimensional physical representation.

III. RELATION BETWEEN PHYSICAL AND PERCEPTUAL SPACE

As already mentioned in the Introduction, our main interest was in the relation between physical and perceptual space.

Applying the earlier-described canonical-matching procedure, the three-dimensional perceptual configuration (1.6% stress), computed from the cumulative results of the triadic experiment, was matched with the three-dimensional physical configuration (81.7% explained variance). The correlation coefficients for the three optimal dimensions were 0.992, 0.971, and 0.742, which means an excellent matching, at least in two dimensions. The lower value for the third dimension is mainly due to the position of only one vowel [y] (see Fig. 6). It seems reasonable to suppose that this sound is

TABLE VI Correlation coefficients between the projections of the points on the vectors corresponding maximally to F_i and L_i in the six-dimensional physical space. The multiple correlation coefficients are given in the last column with the significance levels.

	F_1	F_2	F_3	L_1	L_2	L_3	Mult. corr. coeff.
F_1	...	−0.0319	0.0901	0.5201	−0.0538	−0.5134	0.983 ++
F_2		...	0.6417	0.5620	−0.8709	0.6216	0.984 ++
F_3			...	−0.1296	−0.4487	0.5513	0.679 n.s.
L_1				...	−0.5478	0.6135	0.920 n.s.
L_2					...	−0.4577	0.974 +
L_3						...	0.821 n.s.

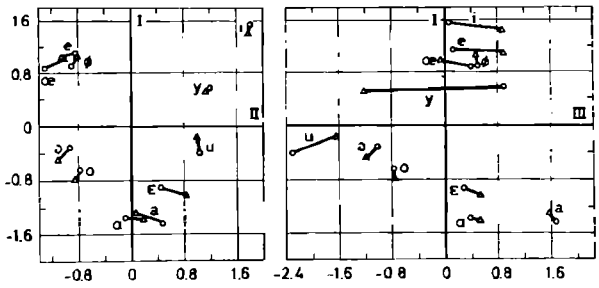

FIG. 6. Positions of the points in the optimal I–II and I–III planes when the three-dimensional physical configuration (○) is matched with the three-dimensional perceptual configuration (△).

specific and that a specific dimension is needed to describe this vowel adequately. Supporting this explanation is the fact that matching the perceptual space with the six-dimensional physical space indeed gives better results (0.999, 0.987 and 0.974), (see Fig. 7.)

From this remarkable correspondence, it can be concluded that the subjects used for their perceptual judgments information comparable with that present in the physical representation of these signals. The perceptual differences between the stimuli, to be considered as timbre differences, appear to be qualified by their differences in frequency spectra. Since these signals were analyzed with $\frac{1}{3}$-oct filters, comparable in bandwidth to the critical bands of the hearing organ, we may suppose that also in vowel detection the critical bandwidth plays an important rôle. The results show that it is not necessary to determine the spectra with narrow-band filters, but that $\frac{1}{3}$-oct filtering is sufficient. This makes it possible also for all kinds of other periodic signals, for instance those of musical instruments, to relate the multidimensional perceptual-attribute timbre of these sounds to the frequency spectra determined by $\frac{1}{3}$-oct analysis. This approach is worked out further in our institute (Plomp and Steeneken[45]).

In order to evaluate the merits of the $\frac{1}{3}$-oct filtering, the signals used were also analyzed with other filter systems, both with constant Δf and constant $\Delta f/f$. In no case could a better correlation with the results of the perceptual analysis be achieved than was obtained with the $\frac{1}{3}$-oct filters.

The data give us, also, the possibility of relating the found perceptual dimensions with the formant frequencies and levels of the sounds. The results of these multiple correlations are presented in the last column of Table VII. The correlation coefficients for F_1, F_2, and L_2 are high, but only the factors related to F_1 and F_2 are independent, as can be seen from the correlations between the projections on the vectors corresponding maximally to the outside variables (see Table VII). We repeated this analysis for the perceptual

[45] R. Plomp and H. J. M. Steeneken, "Effect of Phase on the Timbre of Complex Tones," J. Acoust. Soc. Amer. 46, 409–421 (1969).

TABLE VII. Correlation coefficients between the projections of the stimulus points on the vectors corresponding maximally to F_i and L_i in the three-dimensional perceptual space. The multiple correlation coefficients are given in the last column with the significance levels.

	F_1	F_2	F_3	L_1	L_2	L_3	Mult. corr. coeff.
F_1	...	−0.0566	0.3699	−0.7975	−0.0148	−0.6479	0.972 + −
F_2		...	0.8957	0.6268	−0.8492	0.7922	0.883 + +
F_3			...	0.2176	−0.7128	0.4674	0.411 n.s.
L_1				...	−0.5663	0.9441	0.742 n.s.
L_2					...	−0.5895	0.839 +
L_3						...	0.718 n.s.

space of each individual. These results showed a great resemblance to those given in Table VII, indicating that the subjects agreed closely in their way of judging the signals.

In general, the results support the idea that the first and second formant frequencies are the most important factors in vowel perception. A description of the third dimension is hard to give.

IV. DISCUSSION

The most remarkable result of the above-described experiments is the fact that such an excellent correspondence could be achieved between the physical data and the perceptual data derived from the judgments of the subjects. Since most of the subjects did not even realize that the stimuli were taken from speech sounds, we may assume that they did not use linguistic information in their judgments. In their opinion, they were presented with complex synthetic signals, and they based their decisions on physical cues present in the signals. The effect of familiarity with the Dutch vowels may be considered as negligible, as judged from the results obtained by using as subjects two foreign visitors who were so kind as to participate in the experiment. One of them, a Welshman, obtained a three-dimensional perceptual space with 5.0% stress, which could be matched very well with the three-dimensional physical space (correlation coefficients 0.975, 0.946, and 0.785, respectively). The other, a native Japanese, obtained a three-dimensional perceptual space with 6.3% stress, and correlation coefficients of 0.972, 0.826, and 0.173, respectively. These results are comparable with the individual results of our 15 Dutch subjects.

Our proposed dimensional analysis of spectra, based on a $\frac{1}{3}$-oct frequency analysis, delivers three or four well-defined factors. These factors are sufficient information to come to a fairly high recognition rate of vowel sounds.[9] The factors are not only obtained in a correct statistical way, but they are also in good agreement with the results of a perceptual evaluation of the sounds by observers. Moreover, they are in accordance with parameters such as formant frequencies and distinctive features.

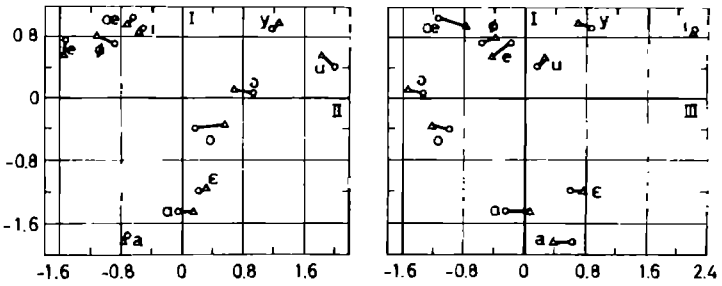

FIG. 7. Positions of the points in the optimal I–II and I–III planes when the six-dimensional physical configuration (○) is matched with the three-dimensional perceptual configuration (△).

So, in a speech-recognition device, the information necessary to recognize vowel sounds might be based on three or four parameters, these being weighted sums of the outputs of a set of $\frac{1}{3}$-oct filters, after logarithmic detection. By defining specific regions per vowel sound, or by determining the shortest distance to the mean vowel positions (Plomp *et al.*[9]), a vowel-recognition procedure can be carried out. Preliminary measurements already point out that the nasals and the liquids can also be described fairly well in the "vowel space." For the plosives and fricatives, a new set of factors must be introduced. Further study along this line is in progress.

ACKNOWLEDGMENTS

The authors wish to thank Dr. J. P. van de Geer for his valuable contributions to the statistical-analysis procedures and L. W. M. Spiekman for developing the program used in the on-line triadic-comparison experiment.