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

Citation for published version (APA):
Perceptual and Physical Space of Vowel Sounds

L. C. W. I'OLS, 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), 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 3-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). 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) and sound-pressure levels (L) of the formants. In this respect, the work of Peterson and Barney is of prime importance. They tried to relate vowel qualities with formant patterns.

SOMLs can be considered as an indication of the number of dimensions required to describe the differences between 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). 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 (Hendal and Hughes). Cohen et al. 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 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, Pickering found the same three features to be the most important ones, and to a lesser degree, relative intensity. Hanson 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 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.

---

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 psychosociology are:

(1) Short-term recall. 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. 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). 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. In that article, attention will also be given to methodological issues.

(3) Semantic scaling. In this procedure, the subject has to associate presented stimuli with a set of bipolar adjectival scales (semantic differential; Osgood). 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 impressions. Furthermore, the preselection of component scales restricts the final solution of the analysis.

(4) Direct scaling by ratio estimation. 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, 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. 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. 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. 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

---

L. J. Th. van der Kamp and L. C. W. Pols, "Perceptual Analysis from Confusions among Vowels" (to be published).
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\(^7\) 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 distances in distances between \(k-j\) and \(i-k\). The only assumption made by Kruskal\(^8\), who worked out the ideas of Shepard\(^9\), is that of a monotonic inverse relationship between interpoint distances and similarity indices.

The characteristic 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\(^3\)).

The multidimensional-scaling computer program, originally described by Kruskal\(^4\) 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\(^5\)), 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\(^6\)). 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\(^7\), "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

---


constant vowel parts of 11 CVC words of the type h(vowel)t, spoken by one of the authors. The vowels used were [ce], [a], [u], [y], [i], [a], [e], [o], Dutch vowels u, o, a, oe, ut, bie, 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 separately. 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). 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. Resampling each signal in such a way as to get 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. Resampling each signal in such a way as to get 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. Resampling each signal in such a way as to get 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.

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

<table>
<thead>
<tr>
<th></th>
<th>[a]</th>
<th>[a]</th>
<th>[a]</th>
<th>[a]</th>
<th>[a]</th>
<th>[a]</th>
<th>[a]</th>
<th>[a]</th>
<th>[a]</th>
<th>[a]</th>
</tr>
</thead>
<tbody>
<tr>
<td>[a]</td>
<td>168</td>
<td>140</td>
<td>115</td>
<td>145</td>
<td>100</td>
<td>108</td>
<td>250</td>
<td>180</td>
<td>145</td>
<td>201</td>
</tr>
<tr>
<td>[a]</td>
<td>162</td>
<td>152</td>
<td>116</td>
<td>51</td>
<td>99</td>
<td>144</td>
<td>230</td>
<td>138</td>
<td>130</td>
<td>130</td>
</tr>
<tr>
<td>[a]</td>
<td>99</td>
<td>87</td>
<td>78</td>
<td>205</td>
<td>144</td>
<td>195</td>
<td>209</td>
<td>130</td>
<td>130</td>
<td>130</td>
</tr>
<tr>
<td>[u]</td>
<td>208</td>
<td>78</td>
<td>66</td>
<td>94</td>
<td>144</td>
<td>144</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
</tr>
<tr>
<td>[u]</td>
<td>155</td>
<td>68</td>
<td>134</td>
<td>106</td>
<td>118</td>
<td>84</td>
<td>84</td>
<td>84</td>
<td>84</td>
<td>84</td>
</tr>
<tr>
<td>[u]</td>
<td>70</td>
<td>115</td>
<td>57</td>
<td>91</td>
<td>104</td>
<td>104</td>
<td>104</td>
<td>104</td>
<td>104</td>
<td>104</td>
</tr>
<tr>
<td>[u]</td>
<td>122</td>
<td>144</td>
<td>187</td>
<td>127</td>
<td>127</td>
<td>127</td>
<td>127</td>
<td>127</td>
<td>127</td>
<td>127</td>
</tr>
<tr>
<td>[e]</td>
<td>156</td>
<td>120</td>
<td>225</td>
<td>156</td>
<td>156</td>
<td>156</td>
<td>156</td>
<td>156</td>
<td>156</td>
<td>156</td>
</tr>
<tr>
<td>[e]</td>
<td>163</td>
<td>135</td>
<td>163</td>
<td>135</td>
<td>135</td>
<td>135</td>
<td>135</td>
<td>135</td>
<td>135</td>
<td>135</td>
</tr>
</tbody>
</table>

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, Knops), 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


32 L. Knops (personal communication), Catholic Univ., Dep. Psychol., Leuven, Belgium.
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 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). Alternative approaches that would preserve interindividual differences are suggested by Tucker and Messick and by McGee. Research in this line is still in progress. To get an estimate of the homogeneity of the similarity ratings of the subjects, the (15X55) 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 sinusoids were analyzed with 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.), 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). The sound-pressure levels in decibels in the 18 frequency bands constitute an (11X18) 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

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). Alternative approaches that would preserve interindividual differences are suggested by Tucker and Messick and by McGee. Research in this line is still in progress. To get an estimate of the homogeneity of the similarity ratings of the subjects, the (15X55) 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 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.), 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). The sound-pressure levels in decibels in the 18 frequency bands constitute an (11X18) 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

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). Alternative approaches that would preserve interindividual differences are suggested by Tucker and Messick and by McGee. Research in this line is still in progress. To get an estimate of the homogeneity of the similarity ratings of the subjects, the (15X55) 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 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.), 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). The sound-pressure levels in decibels in the 18 frequency bands constitute an (11X18) 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

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). Alternative approaches that would preserve interindividual differences are suggested by Tucker and Messick and by McGee. Research in this line is still in progress. To get an estimate of the homogeneity of the similarity ratings of the subjects, the (15X55) 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 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.), 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). The sound-pressure levels in decibels in the 18 frequency bands constitute an (11X18) 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

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) was performed with the following results. The first factor (new dimension) 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 94.1% (see Fig. 3). It is clear that the increase in variance accounted for by taking a more than three-dimensional solution is necessarily pronounced. The opinions about the 11th signal, which was not recognizable as the appropriate vowels, differed.

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

<table>
<thead>
<tr>
<th>Signal</th>
<th>$F_1$</th>
<th>$F_2$</th>
<th>$F_3$</th>
<th>$L_1$</th>
<th>$L_2$</th>
<th>$L_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>[a]</td>
<td>500</td>
<td>1450</td>
<td>2150</td>
<td>37.5</td>
<td>22</td>
<td>15</td>
</tr>
<tr>
<td>[e]</td>
<td>500</td>
<td>750</td>
<td>2750</td>
<td>33.5</td>
<td>29.5</td>
<td>9</td>
</tr>
<tr>
<td>[o]</td>
<td>720</td>
<td>930</td>
<td>2830</td>
<td>34</td>
<td>31</td>
<td>9.5</td>
</tr>
<tr>
<td>[o]</td>
<td>250</td>
<td>620</td>
<td>2050</td>
<td>35</td>
<td>33</td>
<td>5.5</td>
</tr>
<tr>
<td>[u]</td>
<td>250</td>
<td>1600</td>
<td>2720</td>
<td>37.5</td>
<td>26</td>
<td>14</td>
</tr>
<tr>
<td>[i]</td>
<td>250</td>
<td>2100</td>
<td>3100</td>
<td>35.5</td>
<td>23</td>
<td>21</td>
</tr>
<tr>
<td>[a]</td>
<td>990</td>
<td>1450</td>
<td>2550</td>
<td>34</td>
<td>26</td>
<td>4.5</td>
</tr>
<tr>
<td>[e]</td>
<td>495</td>
<td>1520</td>
<td>2220</td>
<td>35.5</td>
<td>20</td>
<td>18.5</td>
</tr>
<tr>
<td>[o]</td>
<td>550</td>
<td>920</td>
<td>2595</td>
<td>33</td>
<td>26</td>
<td>2.5</td>
</tr>
<tr>
<td>[e]</td>
<td>740</td>
<td>1800</td>
<td>2660</td>
<td>35</td>
<td>22</td>
<td>10.5</td>
</tr>
<tr>
<td>[e]</td>
<td>495</td>
<td>2100</td>
<td>3300</td>
<td>36</td>
<td>19.5</td>
<td>4</td>
</tr>
</tbody>
</table>

The line spectra of the signals were also computed, with a method described by Ralston and Wilf. 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 $F_3$ 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 $(11 	imes 6)$ 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 1-Oct analysis can be related with the formant frequencies and levels. For that, we used the canonical-matching procedure that is originally described by Cliff 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...
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 \( F_2 \). \( L_3 \) 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 V** Correlation coefficients between the projections of the points on the vectors corresponding maximally to \( F_1 \) and \( L_1 \) 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).

<table>
<thead>
<tr>
<th>( F_1 )</th>
<th>( F_2 )</th>
<th>( F_3 )</th>
<th>( L_1 )</th>
<th>( L_2 )</th>
<th>( L_3 )</th>
<th>Mult. corr. coeff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>( F_1 )</td>
<td>( -0.1781 )</td>
<td>( -0.1656 )</td>
<td>( 0.9472 )</td>
<td>( 0.9412 )</td>
<td>( 0.9533 )</td>
<td>++ ( p &lt; 0.01 )</td>
</tr>
<tr>
<td>( F_2 )</td>
<td>( -0.8472 )</td>
<td>( -0.8692 )</td>
<td>( 0.8959 )</td>
<td>( 0.9556 )</td>
<td>++ ( p &lt; 0.01 )</td>
<td></td>
</tr>
<tr>
<td>( L_1 )</td>
<td>( -0.8568 )</td>
<td>( 0.9860 )</td>
<td>+ ( p &lt; 0.05 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( L_2 )</td>
<td>( -0.7440 )</td>
<td>( 0.8394 )</td>
<td>++ ( p &lt; 0.01 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( L_3 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>++ ( p &lt; 0.01 )</td>
</tr>
</tbody>
</table>

---

**Table VI** Correlation coefficients between the projections of the points on the vectors corresponding maximally to \( F_1 \) and \( L_1 \) in the six-dimensional physical space. The multiple correlation coefficients are given in the last column with the significance levels.

<table>
<thead>
<tr>
<th>( F_1 )</th>
<th>( F_2 )</th>
<th>( F_3 )</th>
<th>( L_1 )</th>
<th>( L_2 )</th>
<th>( L_3 )</th>
<th>Mult. corr. coeff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>( F_1 )</td>
<td>( -0.0149 )</td>
<td>( 0.6931 )</td>
<td>( 0.5201 )</td>
<td>( -0.0538 )</td>
<td>( -0.5144 )</td>
<td>0.983 ++</td>
</tr>
<tr>
<td>( F_2 )</td>
<td>( 0.6417 )</td>
<td>( 0.5620 )</td>
<td>( 0.7879 )</td>
<td>( 0.9626 )</td>
<td>( 0.984 ++ )</td>
<td></td>
</tr>
<tr>
<td>( L_1 )</td>
<td>( -0.1296 )</td>
<td>( -0.1487 )</td>
<td>( 0.5518 )</td>
<td>( 0.579 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( L_2 )</td>
<td>( -0.5478 )</td>
<td>( 0.6135 )</td>
<td>( 0.920 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( L_3 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.821 n.s.</td>
</tr>
</tbody>
</table>

---


Ref. 42, p. 92.

---

The Journal of the Acoustical Society of America 465
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 role. 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\(^4\)).

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 \( \Delta f \) 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 \( L_s \) 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 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 0.713, 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.

---


---

Table VII. Correlation coefficients between the projections of the stimulus points on the vectors corresponding maximally to \( F_1 \) and \( L_s \) in the three-dimensional perceptual space. The multiple correlation coefficients are given in the last column with the significance levels.

<table>
<thead>
<tr>
<th>( F_1 )</th>
<th>( F_2 )</th>
<th>( F_3 )</th>
<th>( L_1 )</th>
<th>( L_2 )</th>
<th>( L_3 )</th>
<th>Mult. corr. corr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>( F_1 )</td>
<td>-0.3506</td>
<td>0.5099</td>
<td>-0.7975</td>
<td>-0.0145</td>
<td>-0.6479</td>
<td>0.972 + + +</td>
</tr>
<tr>
<td>( F_2 )</td>
<td>0.2216</td>
<td>0.6785</td>
<td>-0.5663</td>
<td>-0.0595</td>
<td>0.899 + + +</td>
<td></td>
</tr>
<tr>
<td>( F_3 )</td>
<td>-0.5663</td>
<td>0.9441</td>
<td>0.742 + + +</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( L_1 )</td>
<td>-0.5663</td>
<td>0.9441</td>
<td>0.742 + + +</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( L_2 )</td>
<td>-0.5085</td>
<td>0.899</td>
<td>0.718 + + +</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( L_3 )</td>
<td>-0.5805</td>
<td>0.899</td>
<td>0.718 + + +</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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 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.), 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.