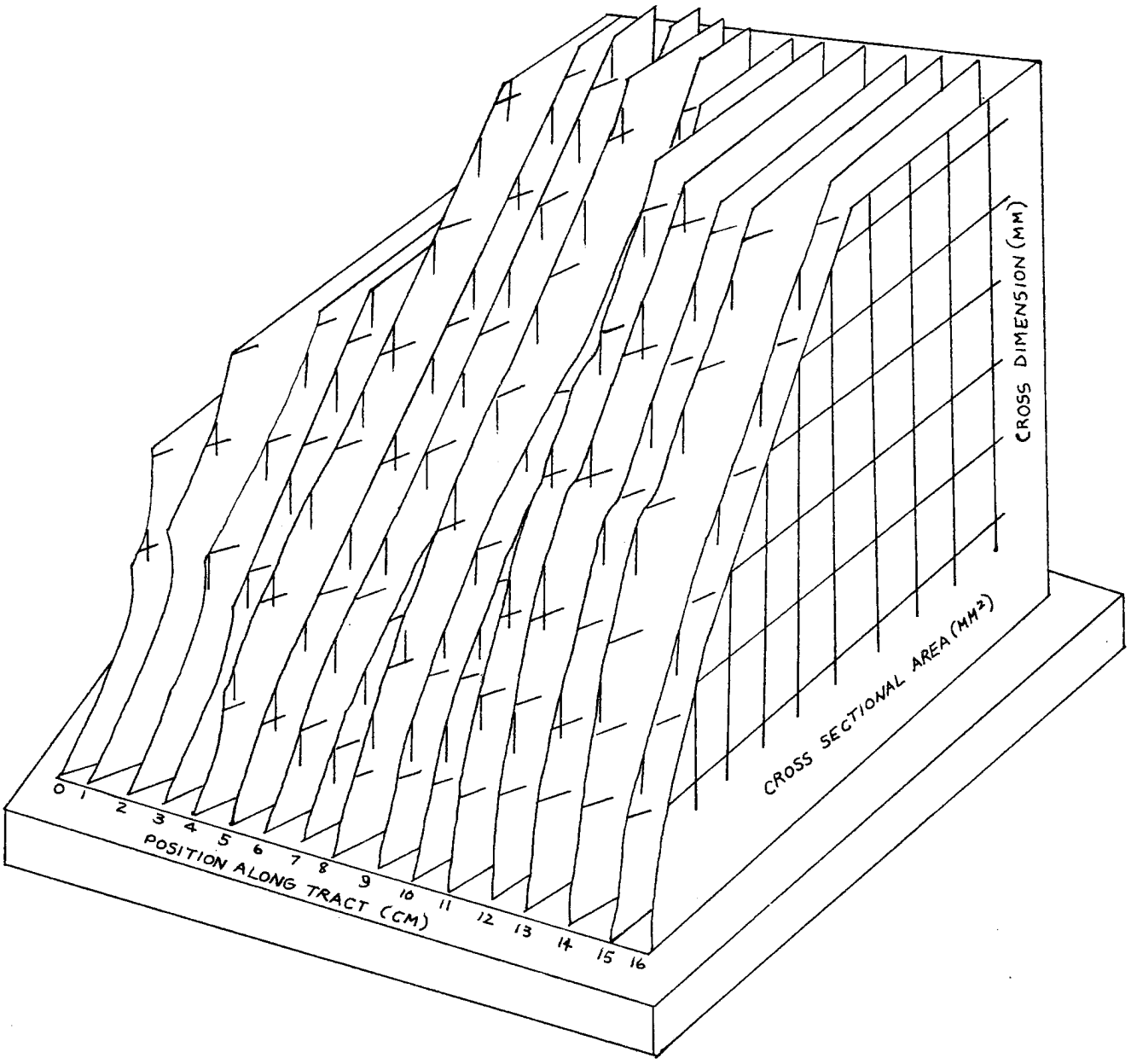
# WORKING PAPERS IN PHONETICS




NO. 19

UCLA

JUNE 1971

*UCLA Working Papers in Phonetics 19*

June, 1971

## The UCLA Phonetics Laboratory Group

### Research

Victoria A. Fromkin
Richard Harshman
Leon Jacobson
Stephen Krashen
Peter Ladefoged
John Laver
Mona Lindau
George Papçun
Lloyd Rice
Dale Terbeek
Robin Thelwall

### Technical and Secretarial

Larry Grant
Julie Haaker
Willie Martin
Renee Wellin
Jeanne Yamane

As on previous occasions, the material which is presented here is simply a record for our own use, a report as required by the funding agencies, and a preliminary account of work in progress.

Funds for the UCLA Phonetics Laboratory are provided through:

USPHS grant NB 04595
ONR contract NR 049-226
NSF grant GS 2741
NSF grant GS 2859
and the UCLA Department of Linguistics

Correspondence concerning this series should be addressed to:

Professor Peter Ladefoged
Phonetics Laboratory
UCLA
Los Angeles, California 90024

*Recent Publications*

Victoria A. Fromkin (1970), The concept of "naturalness" in a universal phonetic theory, *Glossa* 4:29-45.

Victoria A. Fromkin (1971), The non-anomalous nature of anomalous utterances, *Language* 47:27-51.

Victoria A. Fromkin (1971), In defence of systematic phonemics, *Journal of Linguistics* 7:75-83.

Peter Ladefoged (1971), The limits of phonology, *Form and substance: papers presented to Eli Fischer-Jorgensen*, Copenhagen: Akademisk Forlag, 47-56.

Peter Ladefoged (1971), *Elements of acoustic phonetics*, Chicago: University of Chicago Press (paperback edition; revised).

Clive Criper and Peter Ladefoged (1971), Linguistic complexity in Uganda, *Language use and social change* (ed. W. H. Whiteley), London: Oxford University Press, 145-159.

[Reprints of most of these papers are available from the Phonetics Laboratory, 2225 Humanities Building, UCLA, Los Angeles, California, 90024.]

# *Direct Measurement of the Vocal Tract*

Peter Ladefoged, J.F.K. Anthony*, and Cordell Riley

[Revised version of a paper presented at the 80th meeting of the Acoustical Society of America]

Models of the speech production process often involve generating the waveform that would be produced by a given shape of the vocal tract. Typically these models approximate the actual shape of the vocal tract by considering it to be equivalent to a fixed number of connected cylinders, all with the same length (perhaps 1 cm), but each having a variable cross-sectional area. For example, the vocal tract shape represented by the traditional phonetic diagram of the mid-sagittal section as seen in a lateral x-ray (Figure 1a) is considered to be equivalent to the set of cylinders shown in Figure 1(b).

There are two major sets of problems in trying to convert mid-sagittal diagrams of the positions of the vocal organs into equivalent area functions. In the first place, it is difficult to get reliable data on the shape of the vocal tract *in vivo*. Secondly, there is a non-unique relationship between data in the form of mid-sagittal sections and the equivalent area functions.

A considerable amount of data has now been collected on the vocal tract shape of one subject. Most of the information is limited to his pronunciation of four vowels, the cardinal vowels [e] and [o], and the English (RP) vowels /ə/ as in *her* and /ɑ/ as in *far*. These vowels were chosen because they are monophthongs which the subject can repeat with very little variation over long periods of time; and they constitute widely different qualities but do not include very close vowels such as [i] and [u] which have constrictions that make it difficult to use some of the measurement techniques to be described below.

Much of the data on these vowels came from lateral x-rays, which enabled us to determine the position of the tongue both along the mid-line and along the high points at the side. The tracings were made from very clear still x-rays, which had been further enhanced by logetronic processing. These

---

* Presently at University of Edinburgh, Scotland.

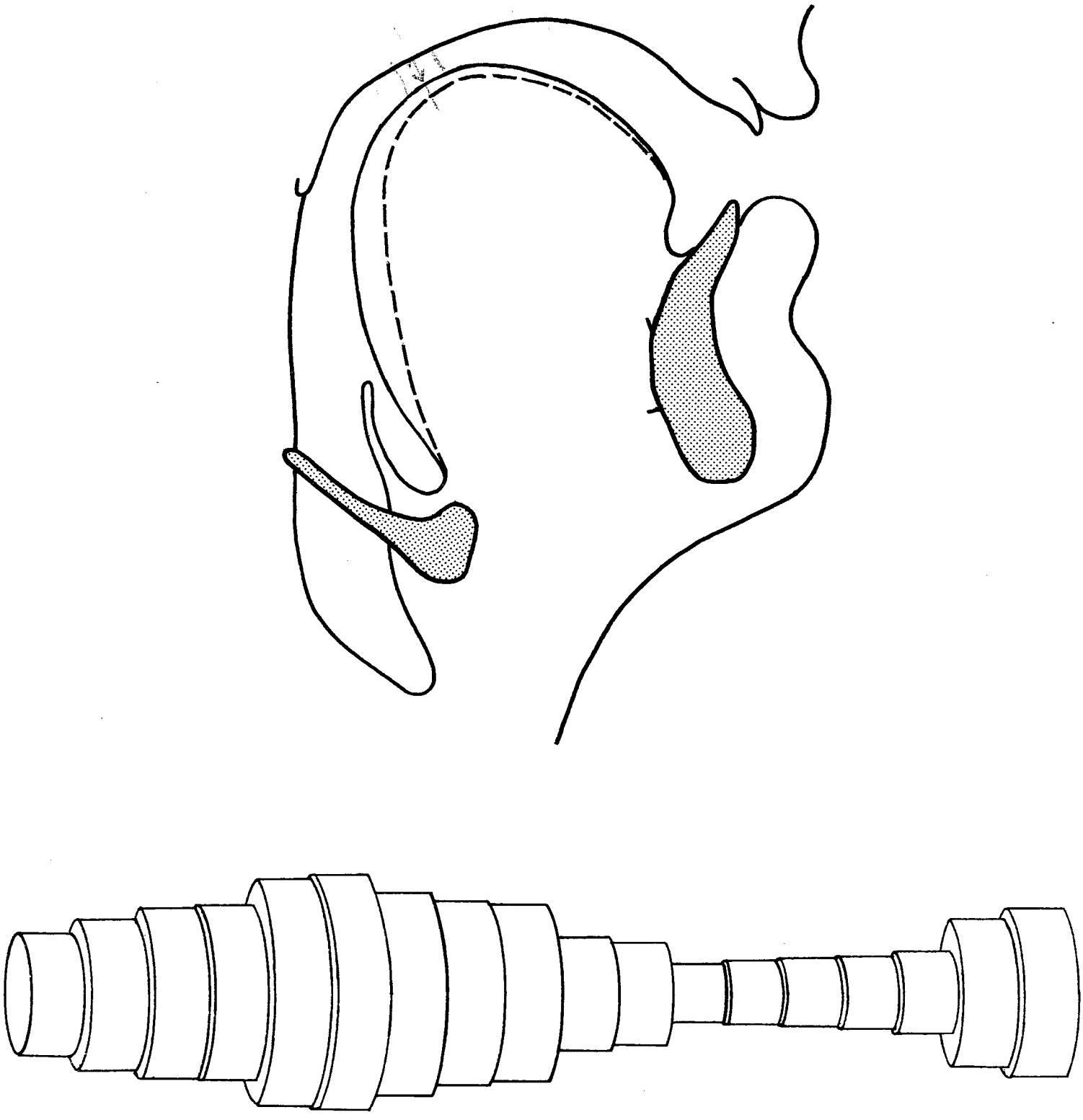


Figure 1 (a). Diagram based on a lateral x-ray showing the position of the vocal organs in the vowel /e/. Where there are two lines the dashed line indicates the position of the mid-line of the tongue, and the solid line the raised portion at the sides; (b) A set of connected cylinders with equivalent cross-sectional areas.

pictures showed a great deal more about the position of the center and sides of the tongue than can be seen on the single frames of cineradiology pictures of the kind that are nowadays most often used in research on speech. Motion pictures are the only valid technique for observing many aspects of speech; but high quality stills are far preferable for studying the shape of the vocal tract during the pronunciation of isolated vowels. It might be objected that steady state vowels are not typical of those in normal speech. This is true, but irrelevant to the present set of experiments. All that matters in determining the relation between the sagittal section diagram and the equivalent area function is that we have a range of possible vocalic positions of the vocal tract; these positions need not be the same as the vowels actually used in speech.

We also used transverse laminagrams showing a cut in the plane of the first and second molars; but these records were not satisfactory, since they were too blurred to give more than an indication of tongue shape. More information was gained from palatograms (Anthony 1954), which gave an accurate record of the positions of the sides of the tongue, and from records of the lips and the front of the tongue obtained by pouring dental impression material over the subject and into his mouth while he was lying down saying these vowels. The precise shape of the subject's lower and upper teeth and hard palate were also obtained from records made with dental impression material.

None of these investigations gave us much information about the shape of the pharynx in different vowels. But by using a slightly different technique we were able to record the position of the back wall of the pharynx and the sides of the pharyngal cavities while the subject was in a position similar to that involved in saying the vowel /ɐ/. The subject was inverted and dental impression material was inserted to the level of the arytenoid cartilages. This procedure was carried out in several stages, using a special impression material (Hydro-Cast; Kay See Dental Mfg. Co., 124 E. Missouri Ave., Kansas City, Mo, 64106.) This material has the property that even after it has been allowed to set in making a first impression of part of a cavity, it is still possible to add additional unsolidified material to it. Consequently an impression can be gradually extended until it fills the entire cavity. The material takes about 45 seconds to set, so the subject had to hold his breath for slightly longer in the final stage of making an impression of the pharynx in the region of the arytenoid cartilages.

We are confident that the impression accurately reflects the position of the back wall and sides of the pharyngal cavity. But the subject found it difficult to hold a steady position of the tongue, and there was some distortion of the impression that was made of the anterior part of the cavity. However, given an accurate record of the posterior part of the cavity, we are reasonably sure of the total pharyngal dimensions

reported, since we also have extensive visual observations of the pharynx (Anthony 1964), as well as the high quality x-ray data showing the mid-line of the tongue, and the superior surfaces at the sides in all four of the vowels /e, ɑ, o, ə/.

In calculating the vocal tract dimensions in these vowels we have also taken into account data reported by other investigators, particularly those showing the relation between the movements of the walls of the pharynx and the position of the tongue (Minifie et al. 1970). In addition we have incorporated our own recent observations on the possible movements of the tongue in the pharynx in cadavers which are less than 24 hours old.

Using all the sources of data discussed above, we constructed a model in dental stone of the left half of the subject's vocal tract in the position for saying /ə/. The mid-line on this model is as shown in Figure 2. The model was filled with an alginate impression material which was allowed to set, removed, and cut into 18 sections along the lines shown in the figure. These lines were arbitrarily arranged for this subject the criteria being that they should be approximately perpendicular to the midline and also approximately equidistant for all possible tongue shapes. The cut sections were used in drawing the outlines shown in Figure 3. Similar outlines were drawn using the data recorded for the other vowels. (The data for /e/ were used in preparing Figure 1.) By interpolation and extrapolation we then calculated a table showing the cross-sectional area for each section (excluding the lip section) for each possible mm opening of the tongue, up to a limit of 32 mm. Figure 4 shows these relations in graphical form. The dashed lines for the upper values for sections 16 (just behind the front teeth) and 17 (across the front teeth) indicate that these values are probably not valid since at these openings the effective outer end of the vocal tract is behind these sections. The horizontal lines that occur in the middle of sections 12, 13, 14 and 15 represent the moments at which increasing the aperture lowers the tongue below the upper teeth, so that there is a sudden increase in the size of the cross-sectional area.

Several problems have been disregarded in the presentation given so far. In the first place the vocal tract varies in length. The lips may be protruded, or the larynx may be lowered. These possibilities are not taken into account in our data. This may be a serious problem when trying to use the data for estimating the cross-sectional areas corresponding to other lateral x-rays. A given distance (or a given percentage of the distance) along the mid-line might on different occasions be in different parts of the vocal tract, and hence correspond to different cross-sectional areas. In addition, it should be noted that the mid-line of the vocal tract will be longer in a high vowel such as /i/ or /u/. In our model this is taken into account by our coordinate system, which makes the appropriate sections out of 9 through 15 appreciably further apart when the tongue approaches of roof of the mouth. In this

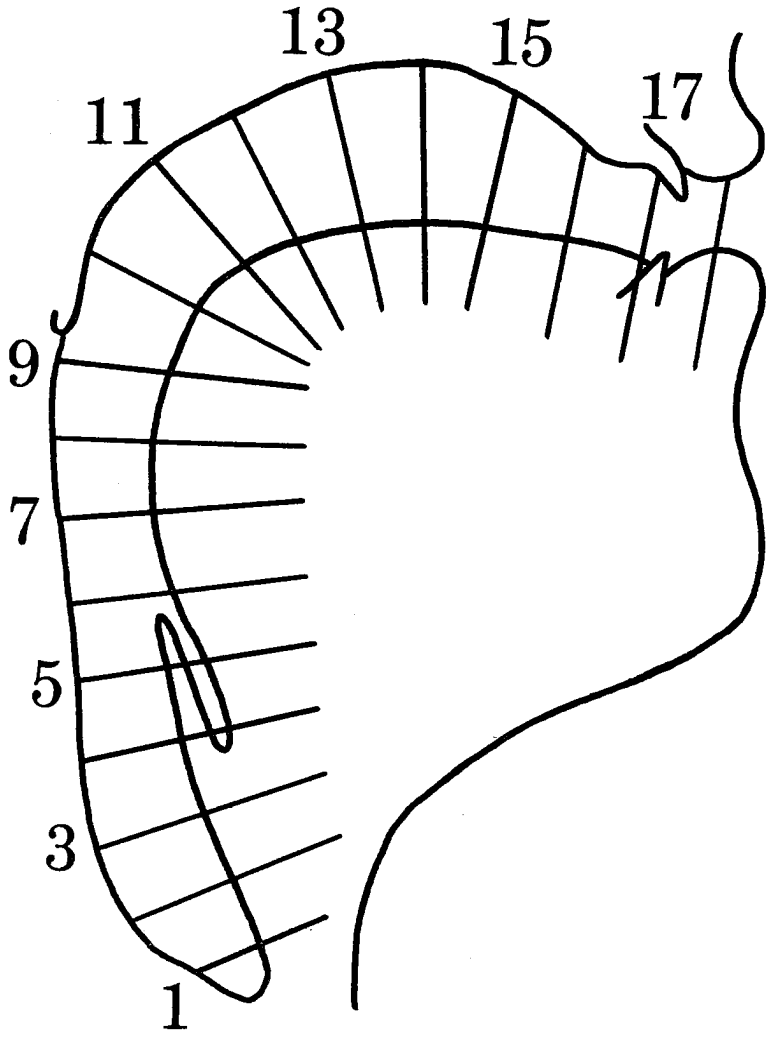



Figure 2. Mid-sagittal diagram of the position of the vocal organs in the vowel /ə/, with 18 arbitrary reference lines drawn so as to be approximately perpendicular to the mid-line of the cavity, and also approximately equidistant for all possible tongue shapes.

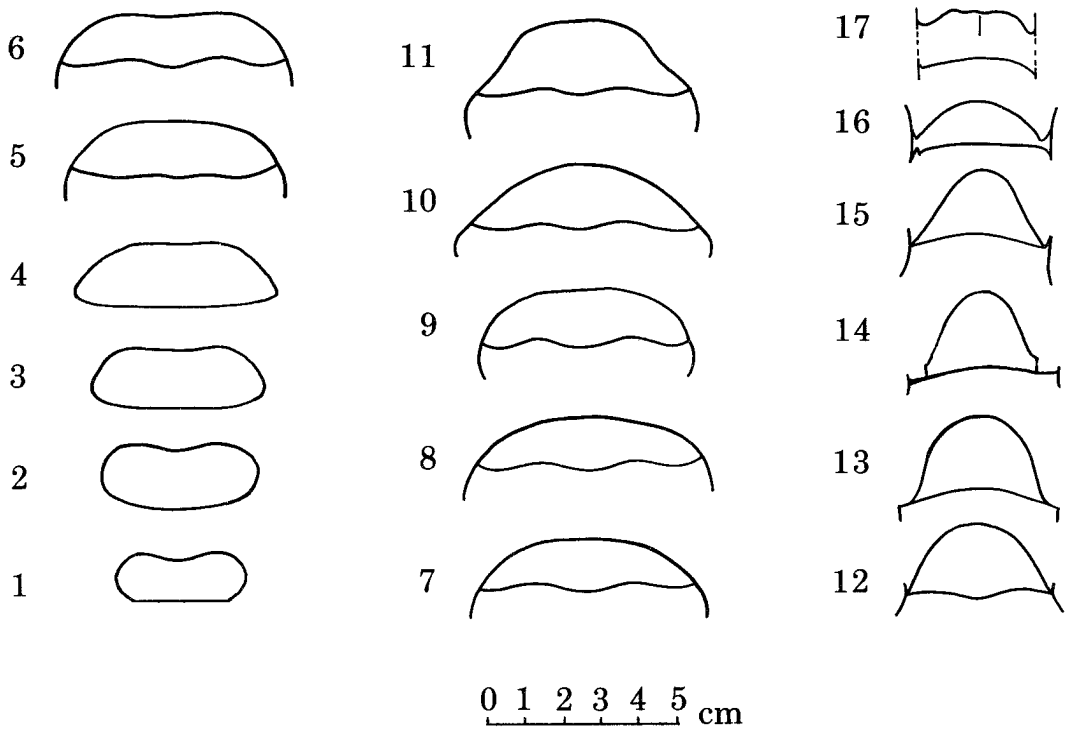



Figure 3. Coronal sections showing the cross-sectional areas of the vocal tract for each of the sections (excluding the lip section) indicated in Figure 2.

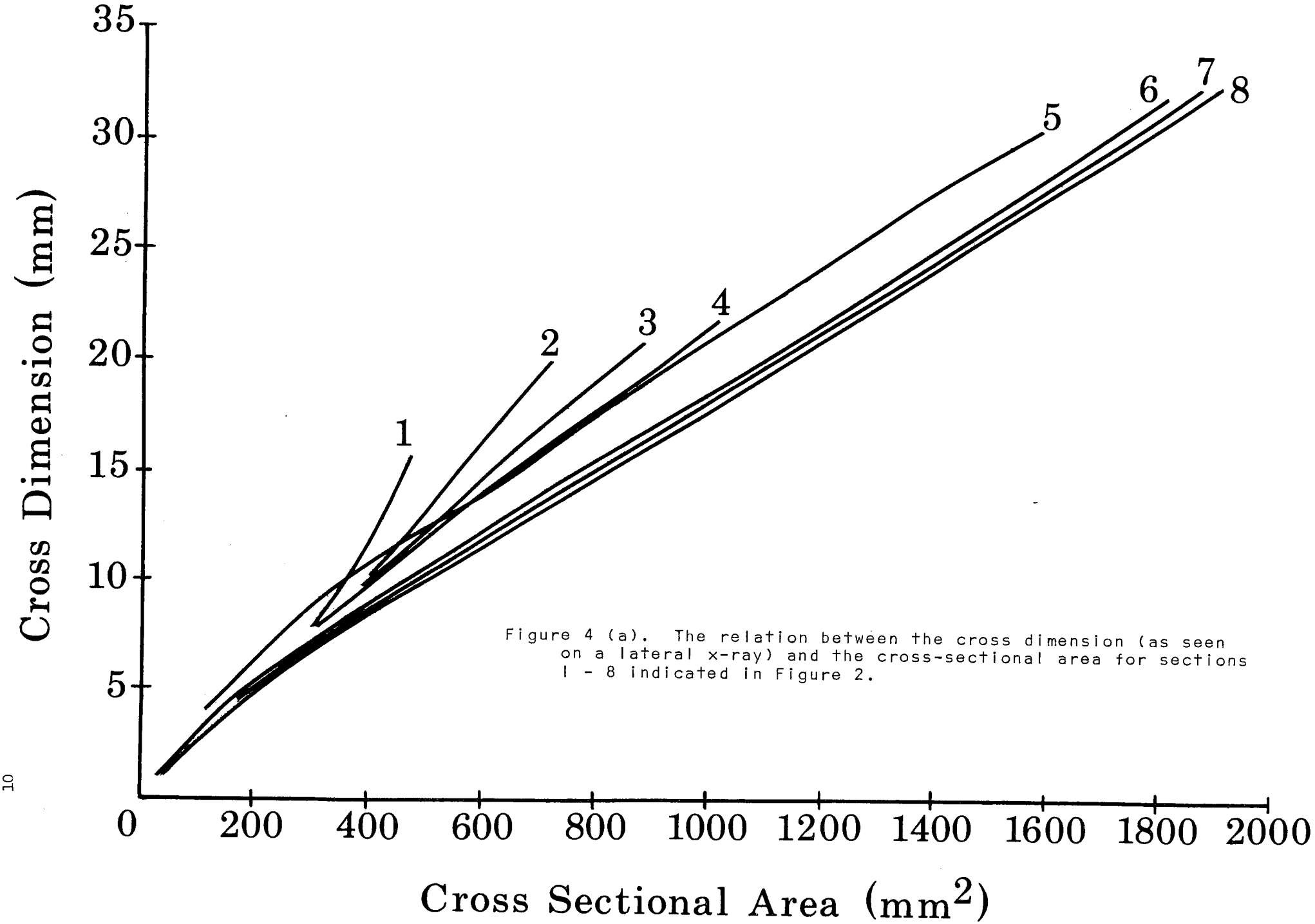



Figure 4 (a). The relation between the cross dimension (as seen on a lateral x-ray) and the cross-sectional area for sections 1 - 8 indicated in Figure 2.

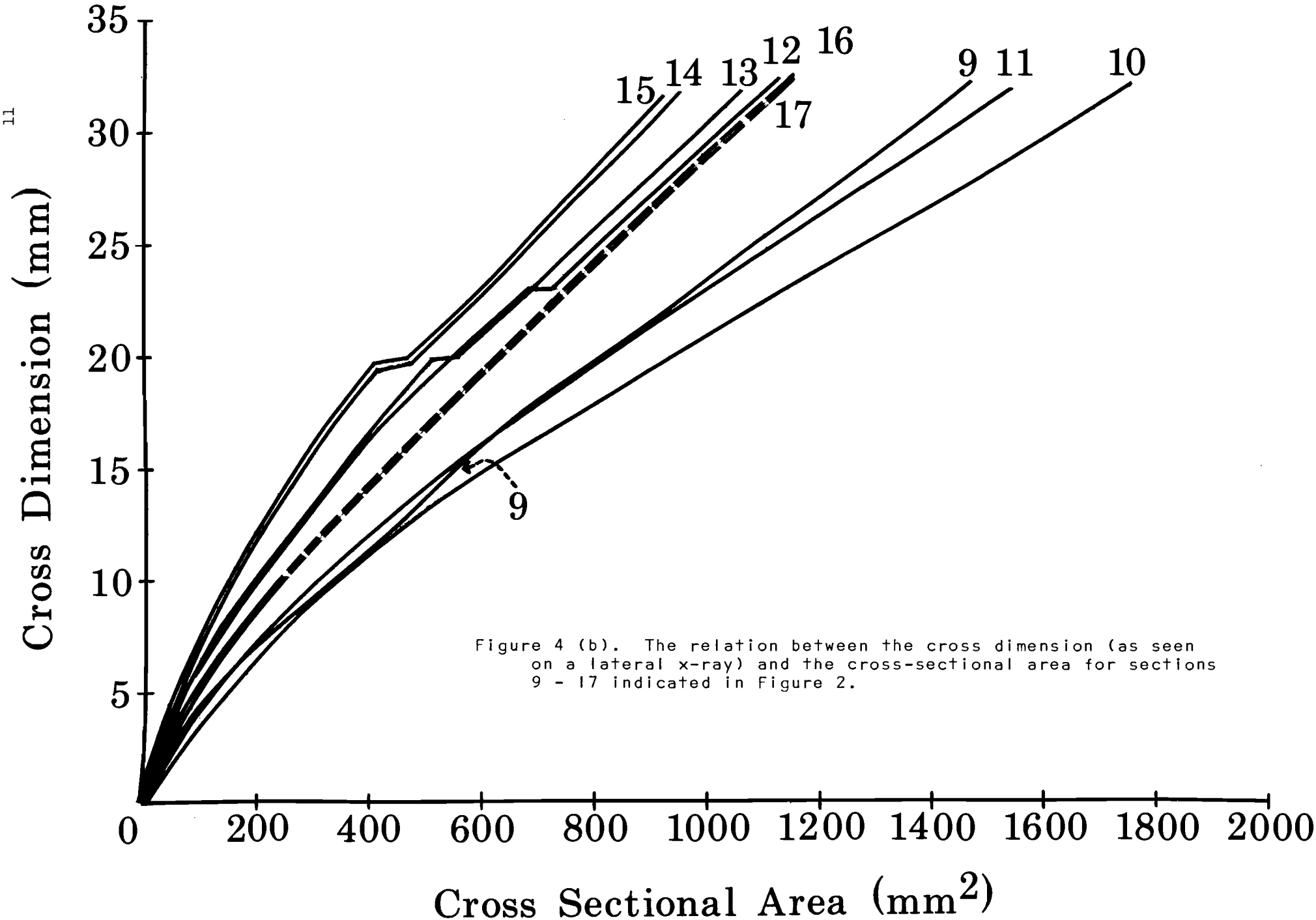



Figure 4 (b). The relation between the cross dimension (as seen on a lateral x-ray) and the cross-sectional area for sections 9 - 17 indicated in Figure 2.

way we can ensure that a given section always refers to a given place in the vocal tract.

Finally, it must be continually emphasized that our data apply only to vowels, and that certain assumptions had to be made concerning the shape of the tongue. In other articulations several of the sections could be either domed, with the center of the tongue being raised towards to roof of the mouth, or hollowed, with the center being depressed. A given value of the mid-sagittal cross-dimension would correspond to a smaller value of the cross-dimension when the tongue was in a hollowed position with the sides raised than when it was in a domed position with the sides further from the roof of the mouth than the center.

It is quite easy to measure the degree of hollowing of that part of the tongue which is in the oral cavity by comparing lateral x-rays which show the position of the center of the tongue with palatograms which give an accurate record of the contact between the lateral margins of the tongue and the upper teeth or hard palate. Thus (as can be seen from the data given in Ladefoged 1957), the shape of section 14(between the first and second molars) in /s/ as in *saw* is probably as shown on the left in Figure 5; whereas the shape of this section in /ʃ/ as in *shaw* is probably as on the right in Figure 5. The curvature of the tongue may not be exactly as shown, but the extent of the hollowing or doming must be as indicated since the position of the center of the tongue is known from x-rays in which a marker had been placed along the mid-line, and the sides of the tongue are known (from palatography) to touch the hard palate at the levels shown. Differences in curvature of this kind must be taken into account when our data are used in specifying consonants.

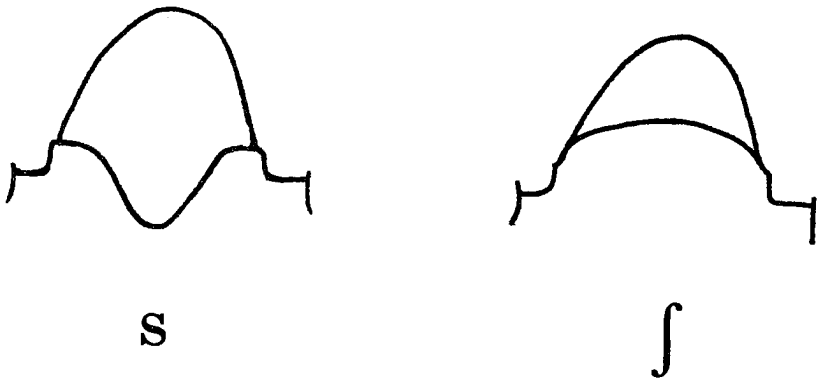



Figure 5. Estimates of the shape of section 14 (between the first and second molars) in /s/ as in *saw* and /ʃ/ as in *Shaw*. The positions of the sides of the tongue are accurately known from palatography, and the positions of the center of the tongue from lateral x-rays.

References

Anthony, J.F.K. (1954), "A new method of investigating the tongue positions of consonants," *Science Technologists' Bulletin* Oct-Nov., 2-5.

Anthony, J.F.K. (1964), "Replica of the vocal tract," *Working Papers in Phonetics No. 1*, 10-14.

Ladefoged, Peter (1957), "Use of palatography," *J. Speech and Hearing Disorders* 22.5, 764-774.

Minifie, Fred.D., Hixon, T.J., Kelsey, C.A., and Woodhouse, R.J. (1970), "Lateral pharyngeal wall movement during speech production," *J. Speech and Hearing Research* 13.3, 584-594.

# *PARAFAC: An "Explanatory" Factor Analysis Procedure*

Richard Harshman

[Paper presented at the 81st meeting of the Acoustical Society of America]

Factor analysis has long been used by scientists seeking to discover what was really going on underneath the surface of their data. Such scientists are looking for real empirical influences or for fundamental explanatory dimensions which would lead to useful predictions.

It is not easy, with traditional factor analysis, to discover such "explanatory" factors. This is because the classical factor model is not sufficiently constrained by the data, and therefore does not provide a unique solution. If there is no unique solution, then one cannot choose which of the possible alternative solutions in fact corresponds to the real pattern of underlying influences for a given data set.

Various selection principles have been suggested, the most common being "simple structure". But none of these principles provide convincing bases for choice except in certain special circumstances or with selected types of data. Therefore, scientists seeking "explanatory" factors have had to devise supplementary experiments. The results of these extra tests would be used to provide empirical grounds for selecting among different hypotheses suggested by the alternative sets of possible factors.

In contrast, the PARAFAC procedure approaches this problem by expanding the factor model itself. This expanded model is used to analyze repeated measures in such a way that it provides a unique solution for which no rotation is possible (so long as the data was adequate and appropriate). In a sense, PARAFAC is a technique for performing factor analysis which incorporates, *within the factor model*, certain basic tests for determining the explanatory factors.

The name PARAFAC comes from PARA for parallel and FAC for factors. The model looks for parallel sets of factors in parallel data sets. The basic idea comes from Cattell (1944). In his approach, two traditional factor analyses are performed on two related sets of data. The true explanatory factors would then be those sets of factors which would be the same in both analyses. But if both sets of data have the same factors acting to the same degree, then the second set is not different from the first and all systems of possible factors in one set will have a matching system in the other data set. It is important, therefore, that the degree of influence of the factors relative to one another be different in the two data sets. Then, Cattell believed, there would be only

one system of possible factors in either set which would have a similar corresponding system in the other. In this case, the pattern of effects of a given factor will be the same in both data sets, but the sizes will only be proportional, all the factor loadings of a given factor being stepped up or down proportionately as we go from one set to the other. The amount of stepping up or stepping down should be different for each factor.

PARAFAC generalizes this proportional profiles approach into a form of three-way or three-mode factor analysis which provides factor loadings not only for measures, and for objects being measured, but also estimates the relative influence of the factors across the intervals of a third mode such as different occasions or conditions of repeated measurement. By simultaneously solving for factors across all occasions, and requiring proportional loadings from one occasion to the next, a unique solution is discovered.

Because the uniqueness of the solution is obtained with the simple assumption of proportional factor loadings, a strong argument is made for the explanatory significance of the solution. Notice, for example, that no assumption is made about orthogonal or uncorrelated factors, or about "simplicity" of the solution. The method will yield whatever pattern of factor relationships is determined by the data, be it oblique or orthogonal.

It is important to note that not all types of three-mode data are appropriate for analysis by PARAFAC. Such analysis is appropriate only in cases where the relative influence of factors can be expected to proportionally increase or decrease *for all measures* from one occasion or condition to the next.* The conditions of data adequacy and results of inadequate data are explored in detail in *Working Papers in Phonetics 16*.

An important feature of PARAFAC type analysis is that it provides an additional guide to determining the correct number of factors to extract from a given set of data. With traditional factor analysis, one successively extracts more and more factors, noticing at each step the relative degree of improvement of prediction of the data. Hopefully a point is reached where, suddenly, no more significant improvement is accomplished by extracting more factors. Often, however, no sudden shift occurs and a subjective judgment must be made as to what is and what is not a real or significant improvement in fit.

---

* But see generalization PARAFAC2 in next issue of *Working Papers in Phonetics*.

With PARAFAC, all factors are extracted simultaneously. One tries two one-factor solutions, then two two-factor solutions, etc. Here the traditional examination of improvements in fit are still just as useful, but uniqueness of the solution provides an additional criterion for the true number of factors. The solution will be unique (given adequate data) whenever the number of factors being extracted is less than or equal to the true number of factors, and will be non-unique whenever too many factors are extracted.

Another benefit of the unique solutions obtained with PARAFAC is the possibility of identifying certain types of factor interaction and non-linear effects of factors by examining relationships between the loadings of extracted unique factors. An example of this kind is discussed in Terbeek and Harshman (1971).

The INDSCAL program developed for metric multidimensional scaling by Carroll and Chang (1970) is a special case of analysis by the proportional profiles model. The INDSCAL model is in fact equivalent to an analysis of scalar products matrices by PARAFAC or by what Carroll calls canonical decomposition of N-way tables.

Other papers illustrate both the factor analytic type of application of PARAFAC (Lindau et al. 1971; Ladefoged et al. 1971) and the multidimensional scaling application (Terbeek and Harshman 1971).

References

Carroll, J.D. and Chang, J.J. (1970), "Analysis of individual differences in multidimensional scaling via an N-way generalization of 'Eckart-Young' decomposition," *Psychometrika* 35, 283-319.

Cattell, R.B. (1944), "'Parallel Proportional Profiles' and other principles for determining the choice of factors by rotation," *Psychometrika* 9, 267-283.

Ladefoged, Peter, DeClerk, J., and Harshman, R. (1971), "Factor analyses of tongue shapes," *Working Papers in Phonetics No. 19*, 39-46.

Lindau, Mona, Harshman, R., and Ladefoged, P. (1971), "Factor analyses of formant frequencies," *Working Papers in Phonetics No. 19*, 17-25.

Terbeek, Dale and Harshman, R. (1971), "Cross-language differences in the perception of natural vowel sounds," *Working Papers in Phonetics No. 19*, 26-38.

*Factor Analyses of Formant Frequencies of Vowels*

Mona Lindau, Richard Harshman, and Peter Ladefoged

[Paper presented at the 81st meeting of the Acoustical Society of America]

This is a report of a study in progress where we want to sort out what parameters underlie the systematic variation in the acoustic speech signal which produces different phonetic vowel qualities. Phoneticians have had difficulty in discovering exactly what physical properties of the speech signal distinguish vowels; we do not know the precise acoustic basis of vowel quality. But it does not necessarily follow that the only interesting aspects of vowel quality are perceptual and "psychological". Clearly there must be specific properties of the acoustic signal which a listener uses in identifying vowel quality; and, in contrast with some modern phonologists, we expect that knowledge of the dimensions underlying these properties will lead to explanations of the form of some natural classes of vowels.

Acoustically vowel qualities may be defined in terms of formants. However, the whole spectrum of a spoken vowel contains information on more than the phonetic quality. Part of the spectrum is determined by the individual speaker's voice quality, and it is still not known to what extent each of these kinds of information contribute to the whole spectrum.

We carried out factor analyses of acoustic properties of vowels as uttered by several speakers using the PARAFAC procedure (Harshman 1970). Our aims were to sort out the phonetic quality from the spectrum, and also to discover acoustic dimensions in terms of which an objective definition and a physical explanation could be given to phonetic vowel quality. The method seemed justified, as our assumption was that the sets of vowels that are possessed by different speakers are related. We used formant frequencies as the input data, since Fant (1956) has shown that the whole spectrum envelope is predictable from the formant frequencies (if one assumes constant bandwidth and glottal source spectrum).

In the first set of analyses the data consisted of formant frequencies that had been converted into pitch values of mels for the eight primary cardinal vowels, as uttered by eleven phoneticians. The data were taken from Ladefoged (1967). The input data thus consisted of a 4 by 8 by 11 matrix of $F_0$, $F_1$, $F_2$, $F_3$, by vowel by speaker. The analysis of this data was previously reported in Harshman (1970), but is reviewed here because it provided the motivation for the subsequent analyses.

## Results

The results of the PARAFAC analysis showed that there were at least three factors underlying the systematic variation of the cardinal vowels across speakers. The three factor solution seemed to fit the data fairly well with approximately a 5% error across the whole data-set.

When the factor loadings for each vowel are plotted against each other in three dimensions and a box constructed around the vowel points, the vowels occupy a space as shown in Figure 1. The factor axes have been removed from the figure. There is a striking similarity between these objectively extracted dimensions and the dimensions of vowel quality that traditionally have been used by phoneticians to specify phonetic vowel quality. It is interesting to compare Figure 1 to Figure 2 which is a diagram by Ladefoged (1967, p. 140) of vowel qualities as points in a space of the three traditional dimensions that are referred to as "vowel height", "front-backness", and "lip rounding". These dimensions are deduced from theories of vowels that are based on *subjective* auditory and physiological impressions.

A preliminary interpretation of the extracted factors (Harshman 1970) hypothesized a one-to-one correspondence between factors and traditional dimensions. Factor 1 in the extracted box corresponds to "vowel height" in Ladefoged's diagram. The other dimensions in Figure 1 also look similar to the other two dimensions in Ladefoged's diagram. Factor 2 was tentatively identified as "front-backness" and factor 3 as "lip rounding". But it is of course very difficult to do so with any amount of certainty as the dimensions covary in our data: front vowels are unrounded, and back vowels are rounded. Factor 3 could even represent something other than "rounding", and "rounding" might not be a dimension at all.

In connection with these difficulties it is worth noting that for a theoretical "ideal speaker" English vowels or cardinal vowels can be plotted equally well in two acoustic dimensions, the first formant against the second formant. Three formants are not necessary to disambiguate these vowels (as spoken by an "ideal" speaker). Given the first two formant frequencies it is possible to assign a vowel a place on the dimension of "vowel height", and equally on the "front-back" dimension. But even with three formant frequencies it is not possible to tell front rounded and front unrounded apart.

In order to test these preliminary hypotheses and to disambiguate the two factors tentatively identified as "front-backness" and "rounding", we selected a new data set which contained vowels in which "backness" was independent of "rounding" because there were front rounded vowels. The new data set consisted of formant frequencies converted into pitch values

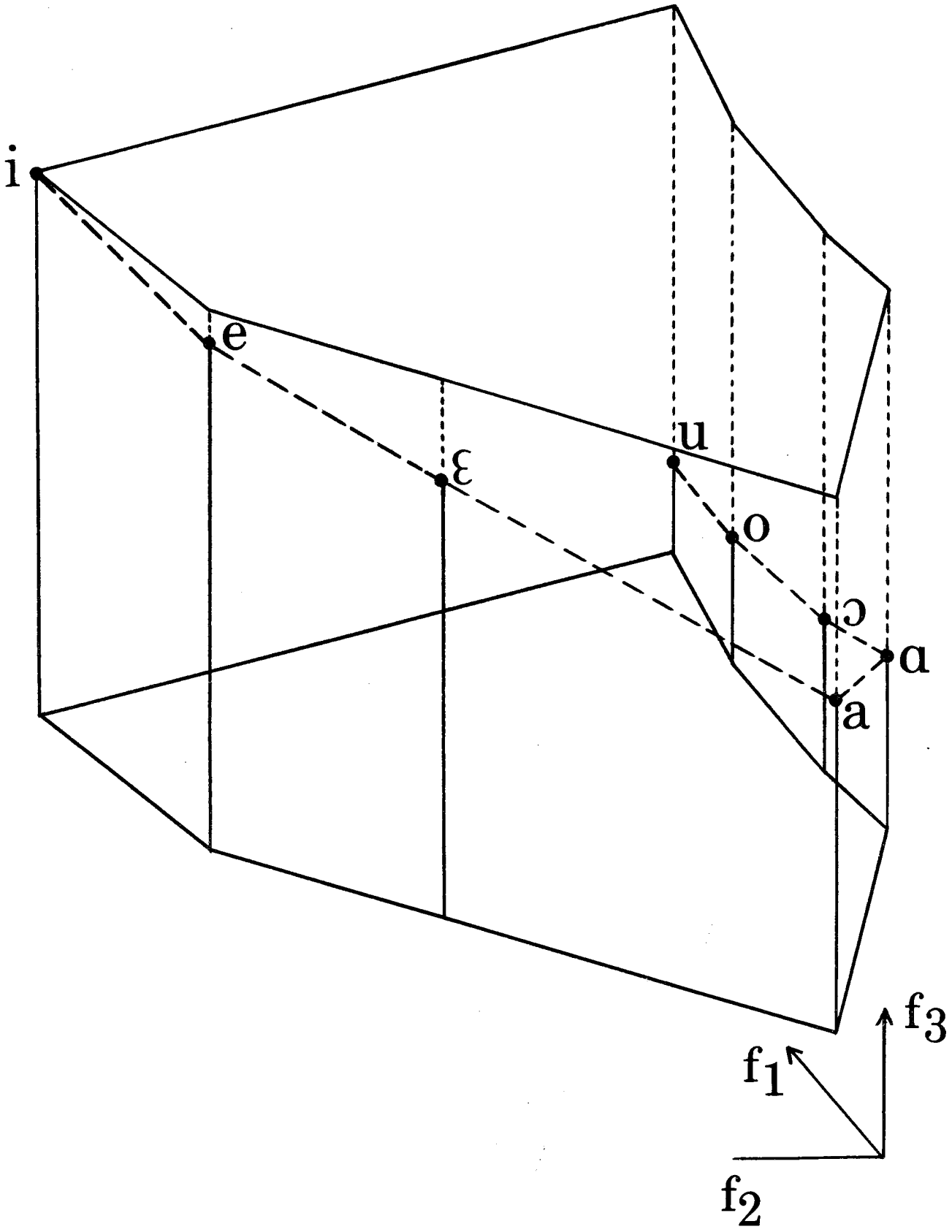



Figure 1. The cardinal vowels plotted on the three factors extracted from the analysis. The factor axes are removed and a box constructed around the vowel points.

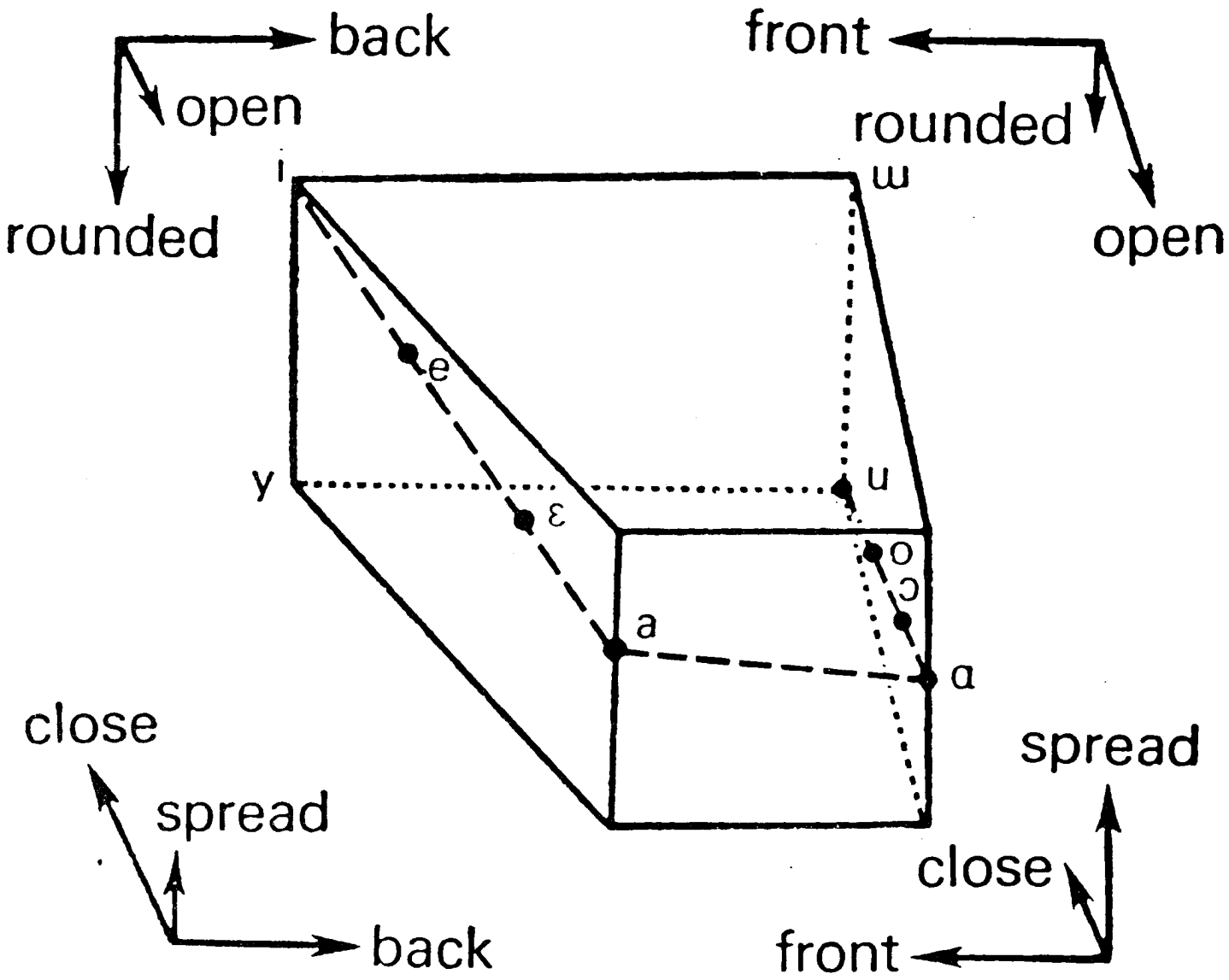



Figure 2. A three-dimensional vowel diagram with traditional dimensions for categorizing vowels.

in mels of a number of Swedish vowels.* We used two subsets for the PARAFAC analysis. One consisted of the eight vowels [i:, e:, u:, o:, ɑ:, y:, ø:, ʉ:], as uttered by seven speakers, the other of the seven vowels [i:, e:, u:, o:, y:, ø:, ɑ:] as uttered by eleven speakers. None of the vowels of any of the speakers departed from Standard Swedish pronunciation according to judgments from a panel of control listeners (Fant, Henningsson, Stålhammar 1970, p. 28). A subsidiary reason for selecting Swedish was that we were interested in finding out what point in an acoustic space the notoriously hard-to-define [ʉ:]-vowel would occupy.

These data sets were analyzed with and without the fourth formant, and with and without the fundamental. The supposition that the fourth formant does not contribute much to phonetic vowel quality tended to be supported by our statistical evidence and emphasis was subsequently placed on the analysis of three formant data. In these analyses there appeared to be a source of bias distorting the proportional relationships in the data. We attempted to eliminate this bias by hypothesizing that variations in the data were due to proportional deviations from a basic vowel, or articulatory basis, of each speaker rather than proportional deviations from zero frequency. We further assumed that, for the purpose of our analysis, this basic vowel was constituted by taking

$$F_{1_{basic}} = \frac{1}{2}\left(F_{1_{highest}} + F_{1_{lowest}}\right),$$

and similarly for $F_2$ and $F_3$. Each speaker's basic vowel was computed in this way and then subtracted from the data-set as follows:

$$\text{new } F_{1_{i\,j}} = \left(F_{1_{i\,j}} - F_{1_{i\,basic}}\right) \text{ for person } i, \text{ and vowel } j,$$

$$\text{new } F_{2_{i\,j}} = \left(F_{2_{i\,j}} - F_{2_{i\,basic}}\right) \quad \text{etc.}$$

Once these constants were deleted it was expected that the data-set would contain only the systematic variations between vowels across speakers. The validity of this "basic vowel" model has not yet been fully determined, but the resulting increase in the goodness of fit to the data looks encouraging. Most factor relationships are similar in either case. The PARAFAC results from the subtracted data-set were unique for up to three factors. Although analysis has not been completed, the patterns of error and uniqueness with different data sets and different numbers of factors

---

* We are grateful to Prof. G. Fant who provided us with the original data that was used in the article by Fant, Henningsson, and Stålhammar (1970). The conversion of the frequency values into mels used a program written by Dale Terbeek, to whom we are also grateful.

so far suggest three factors just as for the cardinal vowels (see Figure 1). The loadings of each factor in this three factor space were plotted against each other as before and the resulting diagram is shown in Figure 3. The vowels [ɛ] and [a] were not included in the data input, but they are part of the whole set of Swedish vowels. It did not seem unreasonable to extrapolate their positions along the line of [i:] and [e:] in the same way as they occur in the results from the cardinal vowels. The solution shown in Figure 3 is for the subtracted data-set. Because of the unfinished state of our analysis it is simply taken as interesting and representative of the type of results obtained. The rotation of the Swedish box is somewhat different from that of the cardinal vowel box, but there is a clear similarity between the two solutions in the acoustic vowel space.

If we were to speculate on the basis of these representative but incomplete results, many interesting hypotheses could be generated:

(1) The traditional dimension referred to as "vowel height" is represented by factor 1, along which the vowels lie from the back to the front of the box. Factor 1 could be interpreted as the relation of the first formant to the rest of the spectrum, that is as a feature that is related to the Jakobsonian feature diffuse/compact. Diffuse vowels have a relative dominance of formants that are non-centrally located, like [i:] and [u:] at the back of the box. Compact vowels have a relative dominance of a centrally located formant, as has [ɑ:] at the front of the box.

It is interesting to note that the factor of "vowel height" was one that turned up very persistently in almost every single analysis. It thus seems to be the most clearly identified physical dimension and might perhaps be the most "basic" property of vowels as has been suggested by Lindblom and Sundberg (1969).

(2) Factor 3 separated the vowels into two categories that correspond to the traditional "front" and "back" vowels. Also this dimension could be interpreted as a Jakobsonian feature: gravity. In grave vowels the lower part of the spectrum dominates, and the second formant is closer to the first than to the third formant as is the case for the vowels to the right in the box. In non-grave, or acute, vowels the upper part of the spectrum dominates, and the second formant is closer to the third than to the first formant which is progressively more the case the further to the left in the box of Figure 3. However, a much better interpretation of this factor is that it is a measure of the distance between the first and the second formant.

(3) The remaining factor, number 2, moves along the vertical dimension in Figure 3. It is very similar to factor 3. It orders

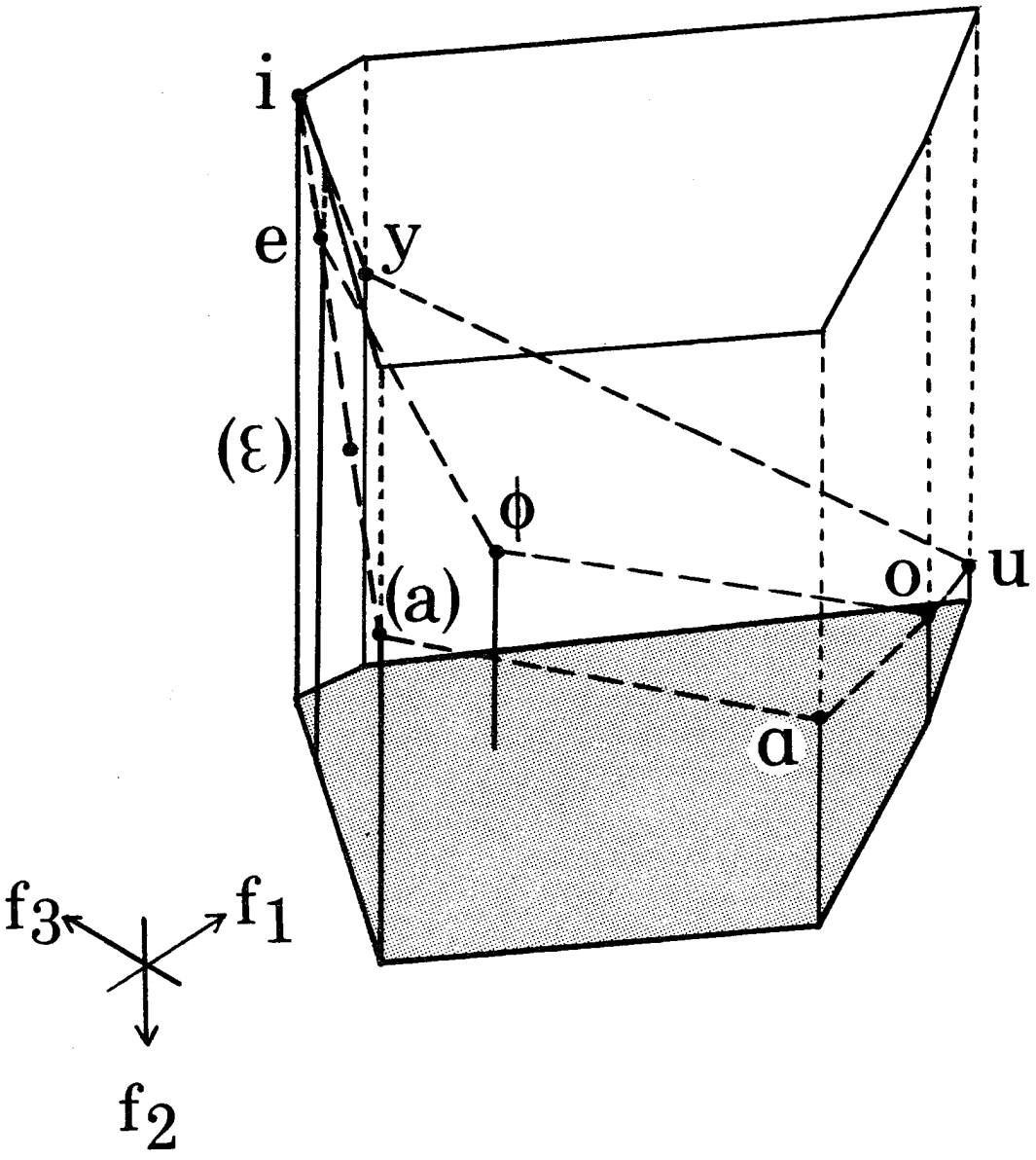



Figure 3. The Swedish vowels plotted on the three factor axes in the same way as in Figure 1.

the vowels in the same way as factor 3 does with [i:, e:, y:] at one end of the dimension and [o:, u:] at the opposite end. The difference between the two factors is that factor 3 separates the vowels [ø:] and [ɑ:] fairly widely, while they are extremely close together along factor 2. We find factor 2 hard to interpret.

However, it is important to note that including the front rounded vowels did *not* disambiguate the two ambiguous dimensions, as had been hoped. There is still no factor that puts a vowel in the bottom corner opposite to [i:]. Whatever factor 2 represents it does not seem to be "rounding", and the traditional category of "rounding" does not constitute a separate acoustic dimension.

The Swedish [ɯ:]-vowel was not included in the data-set from which Figure 3 was drawn. In other similar solutions still being evaluated it occupies a place to the right of [i:] and [e:], right between [y:] and [ø:]. Note in this context that the factor space between [y:] and [ø:] is relatively large, about double the distance between [i:] and [e:], so there is plenty of space between [y:] and [ø:] where another vowel could fit in without approaching any other vowel too closely.

Even tnough this is a progress report on incomplete results one might still speculate as to how phonological processes affecting vowels operate along these objectively extracted dimensions. Consider the vowel assimilation rule of i-umlaut. One could hypothesize that it operates in the domain of factor 3, since i-umlaut consists of a separation of formant one and formant two in the affected vowel. When umlaut occurs the vowel moves along factor 3, its value on the other two factors remaining relatively constant. This predicts that [u] will move to [y] or [i], [o] to [ø] or [e], and [ɑ] to [ɛ] (or [æ]), and that if all stages occur, [u] will reach [y] before [i], and [o] will reach [ø] before [e]. A stronger version of the latter hypothesis that is not contradicted by any data so far known to use would be that when an i-umlaut process causes [u] to change to [i], it has to do so by way of an intermediate stage of [y] (and [o] > [e] by an intermediate stage of [ø]).

In Swedish the i-umlaut process did not go beyond [u] > [y], [o] > [ø] and [ɑ] > [ɛ], but there are Norwegian dialects where it went all the way for some words, e.g. miolk > mjølk > mjelk, melk (Naess 1965). Historically, English contains many examples of the whole process, e.g. mūs > mȳs > mīs; dōmjan > dōeman > dēm. Until factor 3 had been established as a physical reality phonologists could not explain why the first stage of the process affected the "front-back" parameter and not "rounding", so that the stages were u > y > i, and not u > ɯ > i.

References

Fant, G. (1956), "On the predictability of formant levels and spectrum envelopes from formant frequencies," in *For Roman Jakobson* (M. Halle, H. Lunt, and H. MacLean, eds.), The Hague: Mouton.

Fant, G., Henningsson, G., and Stålhammar, U. (1969), "Formant frequencies of Swedish vowels," *STL-QPSR* 4, 26-31.

Harshman, R. (1970), "Foundations of the PARAFAC procedure: Models and conditions for an 'explanatory' multi-modal factor analysis," *Working Papers in Phonetics No. 16*, 84 pp.

Ladefoged, P. (1967), "The nature of vowel quality," in *Three Areas of Experimental Phonetics*, London: Oxford University Press.

Lindblom, B. and Sundberg, J. (1969), "A quantitative theory of cardinal vowels and the teaching of pronunciation," *STL-QPSR* 2-3, 19-25.

Naess, O. (1965), *Norsk Grammatik*, Oslo.

# *Cross-Language Differences in the Perception Natural Vowel Sounds*

Dale Terbeek and Richard Harshman

[Revised version of a paper presented at the 81st meeting of the Acoustical Society of America]

## I. Introduction

Recent studies of multidimensional scaling of vowel perception (Pols, van der Kamp, and Plomp 1969; Hanson 1967; and Singh and Woods 1970) have dealt with two basic questions:

(1) How many perceptual dimensions are there?

(2) How do these dimensions correlate with known characteristics of vowel sounds?

Hanson's extensive experiments dealing with Swedish suggest that the vowel perception space consists of three dimensions. One of these is related mostly to tongue height, or in acoustic terms, the position of the first formant; a second dimension is related mostly to tongue advancement, which involves the position of the second formant. The third dimension of Hanson's solution, which he labels a "perceptual contrast factor", has no clear relation to any acoustic or articulatory parameter. Pols et al., in an investigation of Dutch vowels, present data similar to Hanson's. That is, the perceptual space seems to be three-dimensional, and a plane can be found which reflects the two parameters of tongue height and tongue advancement. The third dimension, however, is difficult to interpret. Singh and Woods showed that two dimensions may be sufficient for the perception of American vowels if the vowel [ɔ] is not included in the analysis; if it is, three dimensions seem necessary to adequately reflect the perceived differences between the sounds.

Although there is still work to be done in solving puzzles posed by the above research, there are clear indications as to the general number of factors and what some of them might relate to in the physical world. But two additional questions have not as yet been considered. The first of these is:

(3) To what extent is one's vowel perception related to one's native language?

In each of the works mentioned above, listeners were asked to respond to sounds acoustically similar to vowels in their native languages. The main purpose of the present research was to compare the responses of speakers of different native languages to a common set of stimuli, in an attempt to answer question (3). In other words, we are investigating the possibility of language-specific answers to questions (1) and (2). It was decided to use natural stimuli, in order to insure as much as possible that subjects were aware of the stimuli as speech signals and not merely as complex tones. In the works mentioned above, it is questionable whether listeners were engaged in a speech perception mode, since the stimuli used were all steady-state sounds on the order of 400 milliseconds. The stimuli were not speechlike except in short-term spectral characteristics.

## II. Experimental design

Linguistically naive speakers of three different native languages -- English, Thai, and German -- were asked to perform triadic comparisons (Levelt, et al. 1966) on a set of common stimuli. It must be noted that all the Thai and German subjects spoke English, and several of the English subjects were familiar with a foreign language. It is not known to what extent the conclusions drawn in this paper would apply to monolingual speakers of the test languages.

Twelve stimuli were chosen so that each listener would hear some vowels similar to those in his native language, and some unfamiliar sounds. Figure 1 shows the stimulus vowels as points in the space defined by formants 1, 2, and 3. In addition to the vowels [i, e, æ, ɑ, o, u] which were similar to vowels found in all 3 languages tested, there were two front rounded vowels [y, œ], two unrounded vowels [ɨ, ʌ], and two retroflex vowels [ɚ] and [ɑ˞]. Figure 2 gives the approximate native/non-native status for each stimulus vowel for each of the languages used. The [ɨ] vowel, listed as native to Thai, does not correspond perfectly to the high back unrounded vowel present in Thai, but Thai subjects did in fact identify the [ɨ] stimulus as their [ɯ] vowel in informal questioning. Comments volunteered by the subjects indicated that the [ɑ˞], a low back vowel with some retroflexion, was perceived as different from the [ɑ] only in voice quality, not in phonetic quality.

The stimuli were recorded with stress and falling intonation in the context [bəb_] by a trained phonetician. Several tokens were recorded of each of the twelve different stimulus vowels in this context. The twelve tokens to be presented to the subjects were selected from this set on the basis of similarity to the other vowels with respect to duration, intonation pattern, and absolute pitch, and freedom from diphthongization. The selected stimuli were then low-pass filtered at 3500 Hz, sampled with 8-bit resolution at 11.3 kHz and stored on digital magnetic tape by a PDP-12 computer. The

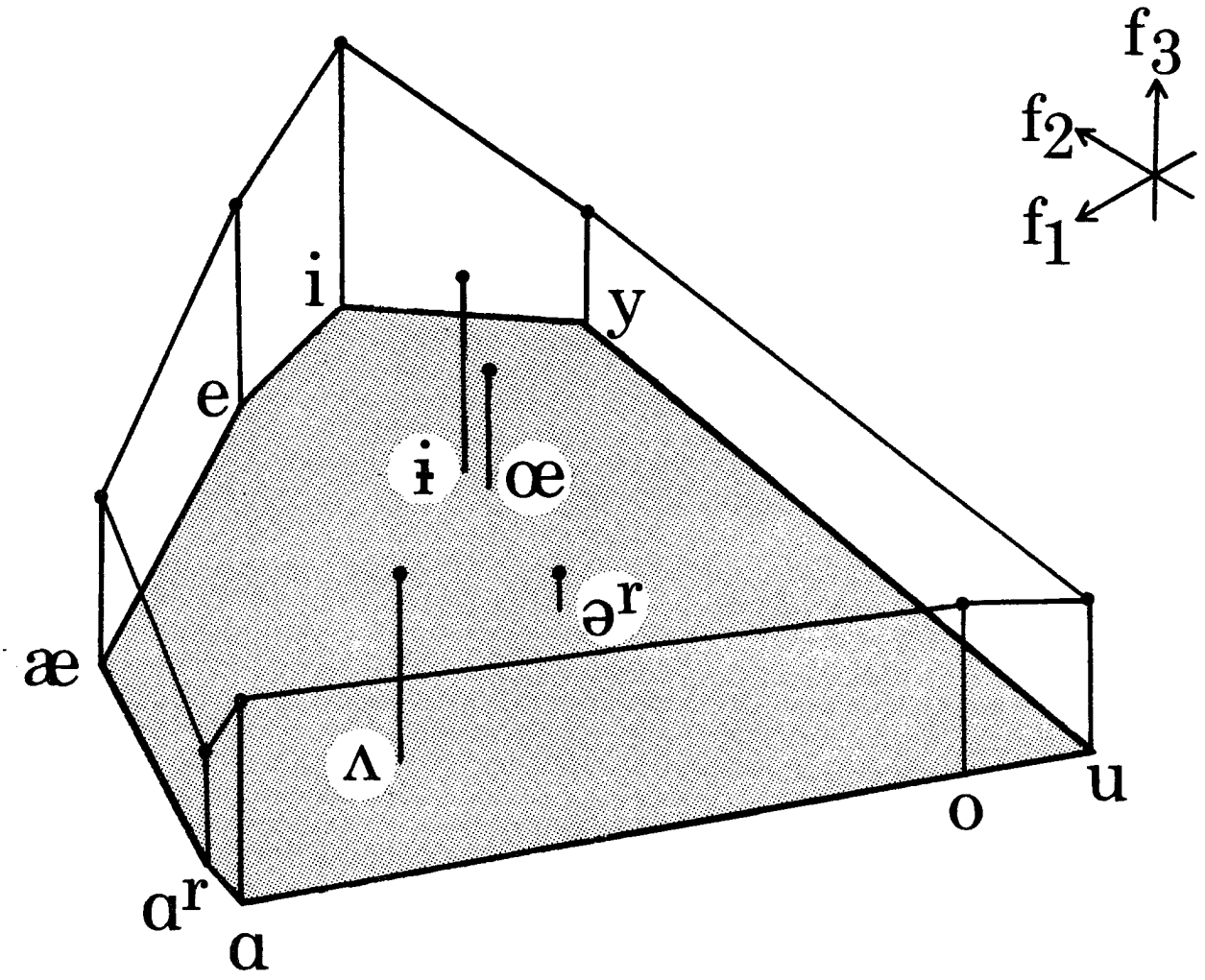



Figure 1. Stimulus space defined by formants 1, 2, and 3.

| | i | e | æ | ɑ | o | u | y | œ | ʌ | ɨ | ɐ | ɑ$^{r}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| German | X | X | X | X | X | X | X | X | | | | |
| English | X | X | X | X | X | X | | | X | | X | |
| Thai | X | X | X | X | X | X | | | | X | | |

Figure 2. Approximate native-nonnative status of stimulus vowels in each language tested. X indicates native status.

| | ɑ$^{r}$ | ɨ | y | œ | ɐ | æ | ʌ | u | o | ɑ | e | i |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ɑ$^{r}$ | 0 | | | | | | | | | | | |
| ɨ | 12 | 0 | | | | | | | | | | |
| y | 15 | 4 | 0 | | | | | | | | | |
| œ | 9 | 1 | 0 | 0 | | | | | | | | |
| ɐ | 10 | 6 | 10 | 6 | 0 | | | | | | | |
| æ | 5 | 9 | 15 | 15 | 14 | 0 | | | | | | |
| ʌ | 4 | 7 | 8 | 6 | 8 | 7 | 0 | | | | | |
| u | 14 | 5 | 2 | 5 | 14 | 16 | 9 | 0 | | | | |
| o | 16 | 9 | 4 | 9 | 17 | 11 | 13 | 7 | 0 | | | |
| ɑ | 0 | 8 | 13 | 10 | 14 | 5 | 2 | 12 | 16 | 0 | | |
| e | 14 | 9 | 13 | 7 | 5 | 12 | 10 | 14 | 16 | 13 | 0 | |
| i | 19 | 8 | 11 | 10 | 11 | 13 | 15 | 16 | 19 | 17 | 6 | 0 |

Figure 3. Sample response matrix computed from similarity and difference judgments.

filtering and sampling produced no noticeable degradation of the signals. The same computer presented the stimuli via loudspeaker by means of a set of remote pushbuttons operated by the listener.

Listeners were presented with all possible triples (= triads) of the stimuli in a pseudo-randomized order, so that each triad was completely different from the preceding one. (For 12 stimuli, the number of triads = 12!/(3!9!) = 220.) They were allowed to listen to each member of a triad as many times as they wished and in any order. The task was to make two judgments for each triad: which two vowels seemed most similar and which two seemed most different. These judgments were recorded by means of two additional pushbuttons connected remotely to the PDP-12, which stored the responses on digital tape for later analysis. Due to the large number of stimuli, subjects took short breaks every 15 minutes and spent no more than 1 to 1 1/2 hours per session. Three to four hours were generally sufficient to complete the required judgments. The first session for each subject was begun with 20 practice triads to provide familiarity with the stimuli and the task; subsequent sessions began with 10 practice triads.

Vowel-by-vowel matrices representing perceived distances between stimuli were computed from the lists of similarity and difference judgments compiled for each subject, according to the method described by Pols, van der Kamp and Plomp (1969). A judgment of "most different" for a given pair adds a score of 2 to the relevant cell in the matrix, a judgment of "most similar" adds a score of 0, and 1 point is added to the cell representing the remaining pair in a triad. A sample matrix is given in Figure 3. The diagonal cells are zero by definition; no triad with 2 or more identical stimuli was given. Since there were 12 items, each possible pair of stimuli occurred 10 times. Every cell, then, will contain some value between 0 and 20 inclusive.

## III. Analysis and results

The heart of the analysis procedure is PARAFAC (Harshman 1970), a multi-mode factor analysis method which replaces ad hoc factor rotation criteria with a principled algorithm for deriving explanatory rather than merely descriptive factors. In this multi-dimensional scaling application, solutions are expressed in terms of vowel loadings (loadings = values on a given dimension), which describe the group space, and person loadings, which indicate the relative size or importance of each dimension for each person.

The first step in the analysis was to apply Cooper's best-fit technique to the response matrices (Cooper 1970). This technique optimizes the fit in a given number of dimensions by optimizing the value of a constant to be added to each non-diagonal cell in the response matrix. That is, the proportionalities between entries in the matrix is adjusted

while leaving the additive relationships intact. Optimum additive constants for 1 through 5 dimensional solutions were obtained, yielding 5 absolute distance matrices for each subject. The absolute distance matrices were converted to scalar products of vectors according to Torgerson (1958) and, grouped according to language, were analyzed in 2 through 5 dimensions by the PARAFAC procedure. At this point it was concluded on several grounds that 2 German subjects and 2 English subjects were not responding in a way consistent with their respective groups, and were excluded from further analyses. PARAFAC was then used to extract 1 through 5 dimensions for each language group, with the number of persons reduced to 5 German, 6 Thai, and 6 English.

In order to determine the actual number of dimensions used in the perception of the 3 languages, it is useful to discuss Figure 4, a plot of fit against dimensions extracted. Fit refers to the amount of variation present in the data which has been accounted for by the analysis. Notice that this measure of "fit," devised by Cooper (op. cit.), does not involve the standard concept of percentage of variation accounted for *by each factor*. PARAFAC does not retain factors from one dimensionality to the next; rather, it derives a new set of factors for each analysis.

The number of dimensions constituting the data is usually determined by locating sharp bends in the fit curve. This method would suggest that English subjects responded in 2 dimensions, and that the Germans responded perhaps in three. There is no sudden bend in the Thai curve, and a guess of this type is difficult to make. PARAFAC, however, provides a supplementary criterion as to the number of dimensions, in that there will be a unique solution for all dimensionalities up to and including the correct one. (The uniqueness and non-uniqueness of results reported here are based on two different starting configurations of random loadings for each analysis.) Filled circles in Figure 4 represent unique solutions, and empty circles, non-unique ones. This criterion supports the conclusion that English listeners used 2 dimensions, and German listeners, 3. The uniqueness criterion further suggests that Thai listeners responded in 3 dimensions.

Of course, no matter how many dimensions are found to be "real" for each language group, they must be interpretable. After all, a good deal is known about the acoustic and articulatory structure of vowel sounds, and if the perceptual dimensions can not be related to any known characteristics of the stimuli, it is questionable whether the analysis has rendered any useful information at all.

In discussing the interpretation of the resulting perceptual spaces, we will refer to the perspective drawings in Figures 5, 6, and 7 to illustrate the overall configurations for the three language spaces.

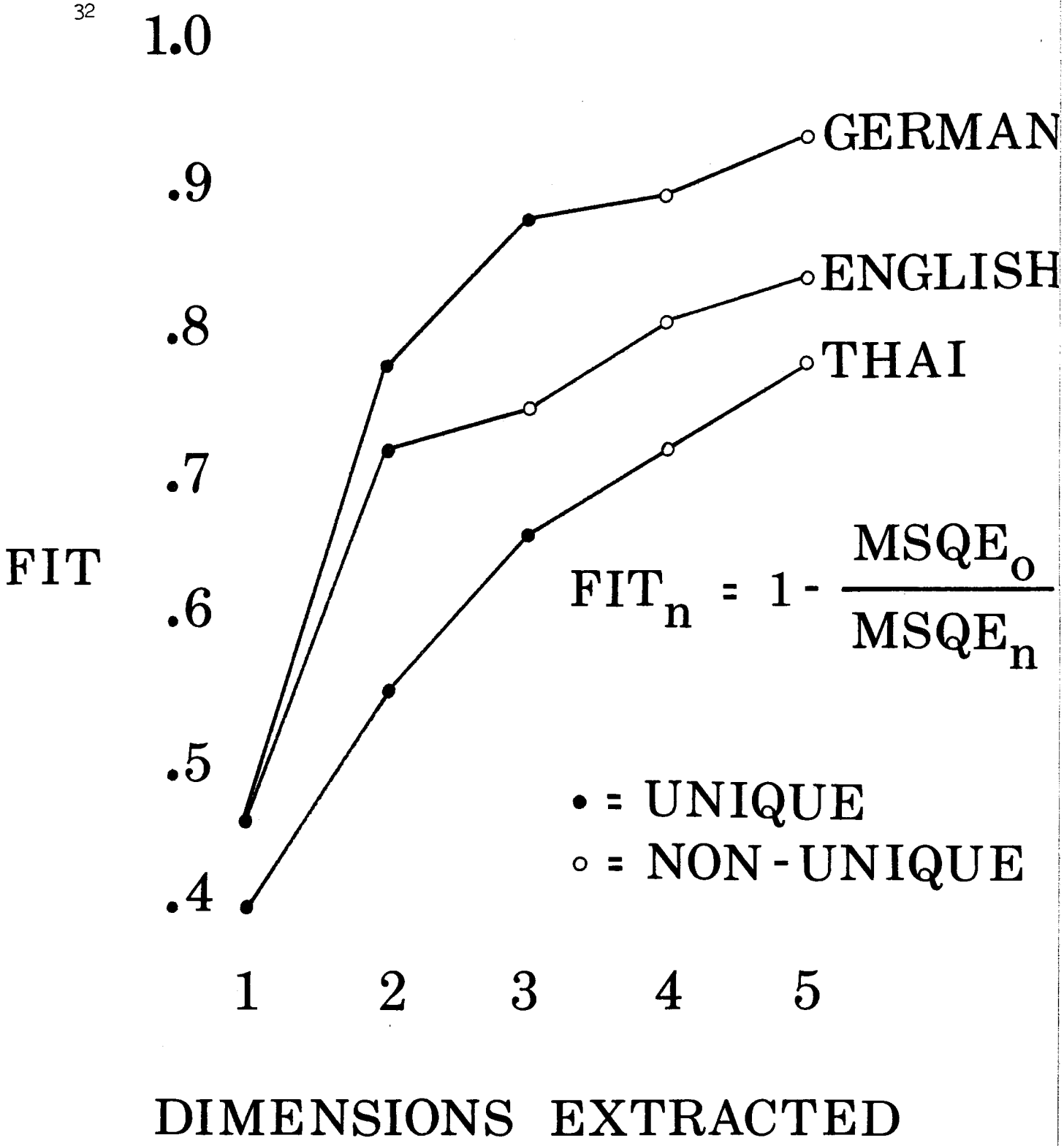



Figure 4. Goodness of fit against dimensions extracted for each language group.

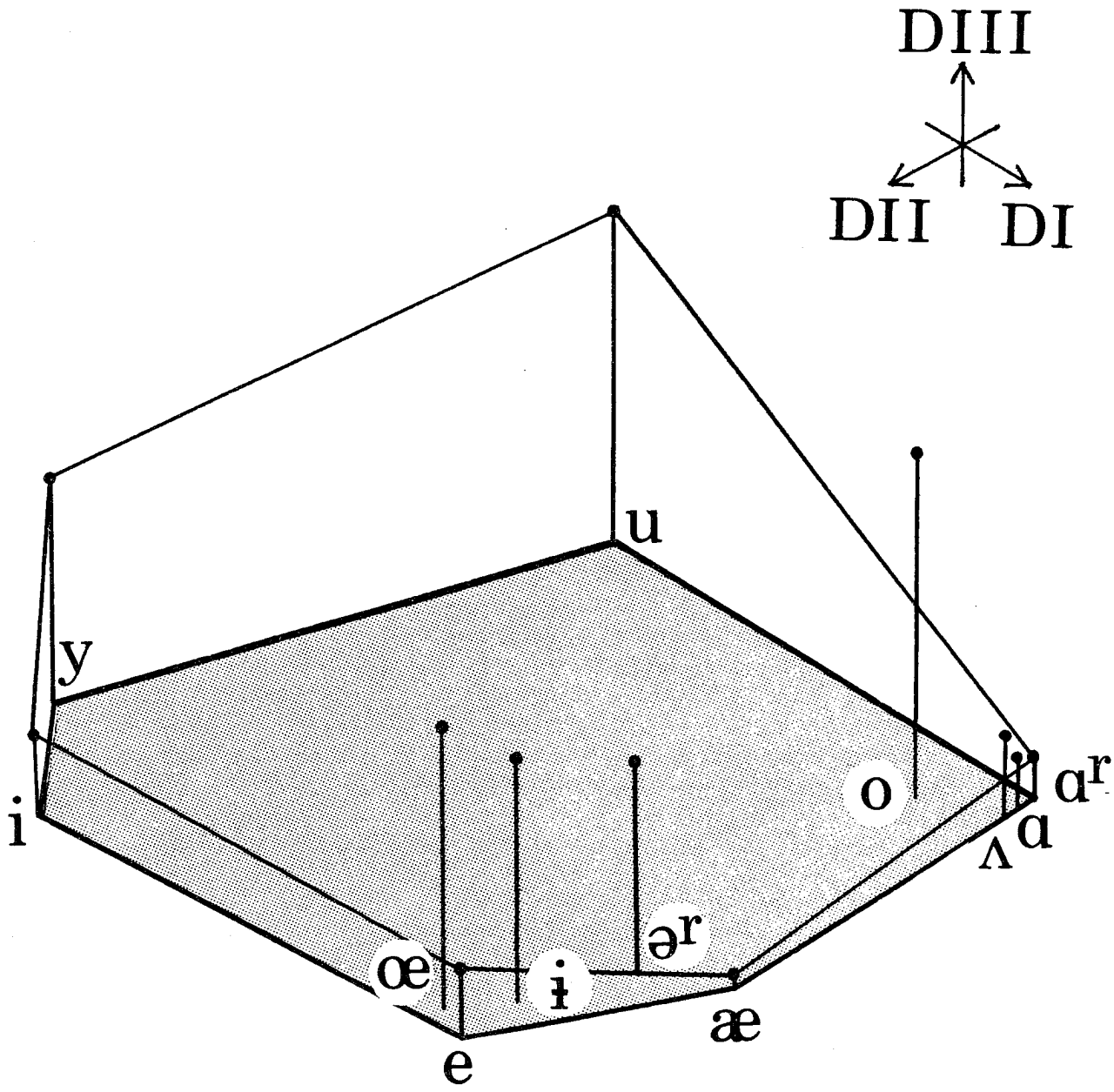



Figure 5. Three-dimensional perceptual vowel space for German.

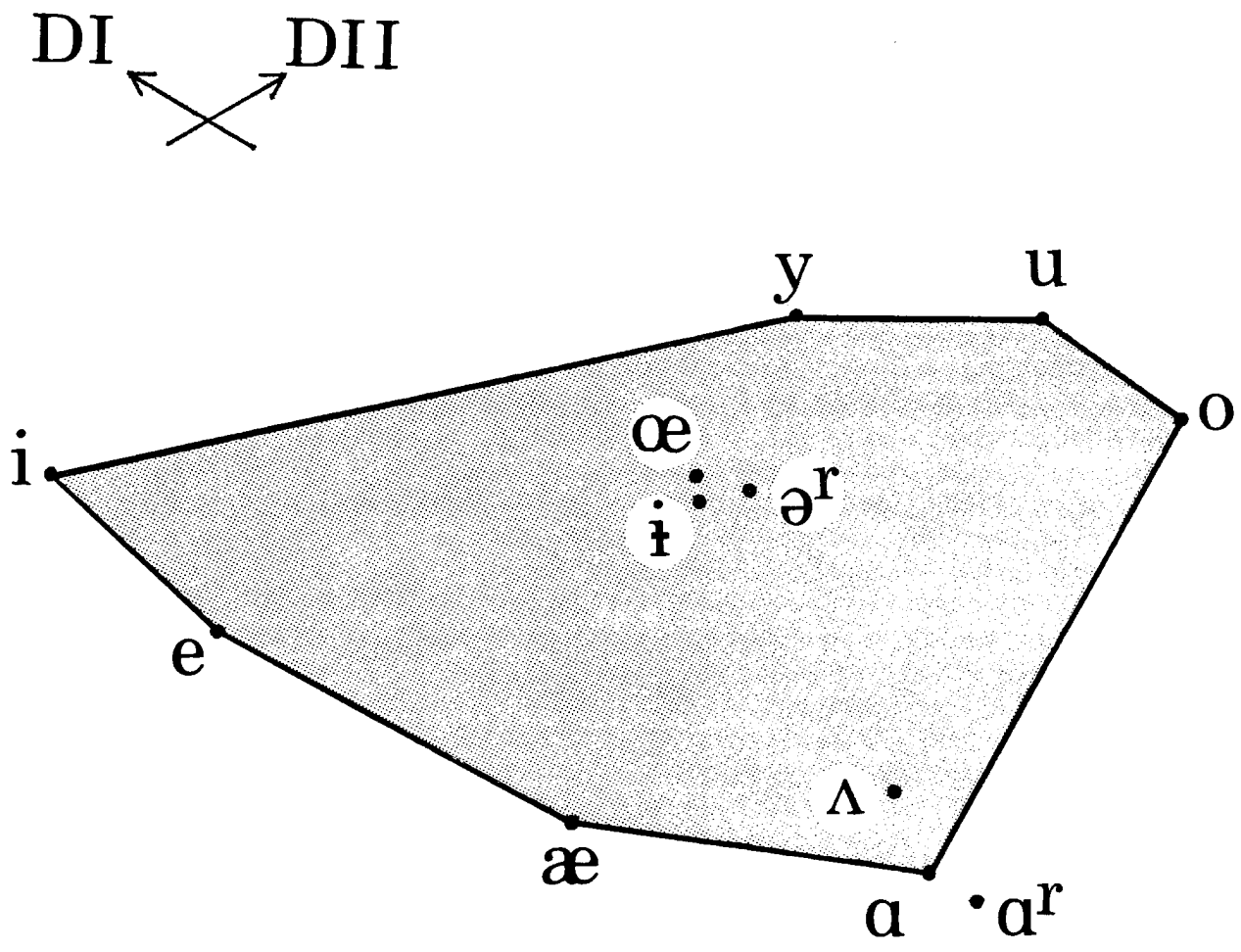



Figure 6. Two-dimensional perceptual vowel space for English.

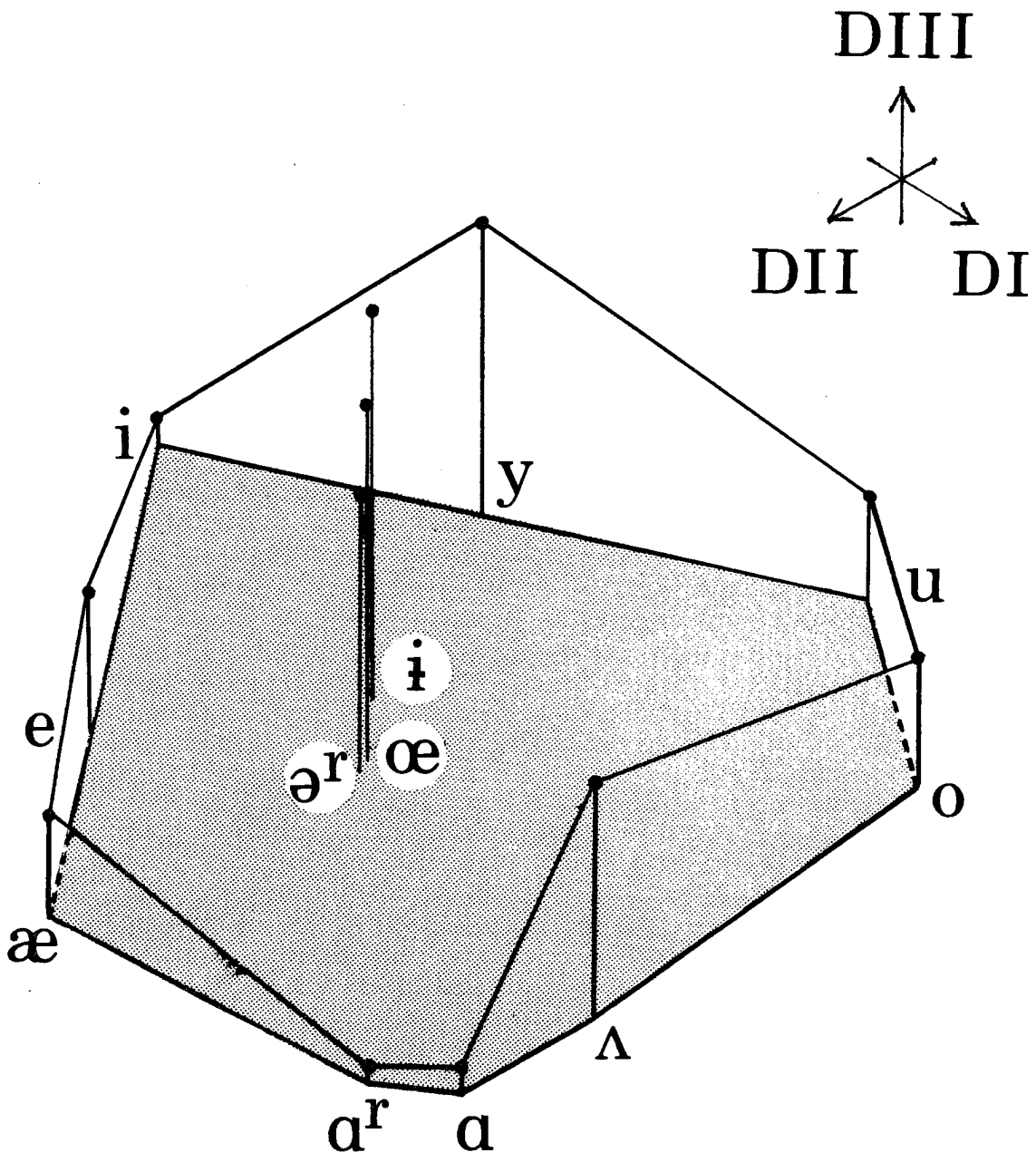



Figure 7. Three-dimensional perceptual vowel space for Thai.

We turn first to the German results, ignoring for a moment the vowels not native to German (the vowels [ɨ, ɞ, ʌ, ɑ˞]). Dimension DI in Figure 5 sets off the high vowels [i, y] and [u] from all the others. This distinction is related to the acoustic parameter of first formant height, although there is no large break between the first formant values of [i, u], and [y] as opposed to the other 9 vowels. DII separates back vowels from front vowels, and correlates well with the difference between the first and second formants. Finally, DIII distinguishes rounded from non-rounded vowels. There is no straightforward acoustic measure which reflects this distinction well; subjects may have been using an articulatory parameter -- lip rounding -- in making their judgments.

A striking characteristic of the German space is the tendency of vowels to cluster at the corners of an imaginary cube. That is, they cluster into 2 groups on each dimension, rather than spread out over more of a continuum. In fact, for the native vowels, the perceptual dimensions, when regarded as binary scales, conform exactly to the distinctive features [±high], [±round], and [±back]. It seems that the German subjects were responding less directly to acoustic properties of the stimuli, and more to phonological properties of the stimuli as related to their native language.

The non-native vowels conform well to these feature descriptions, with some discrepancy in the rounding dimension DIII for the vowels [ɨ] and [ɞ]. But these two vowels were informally labelled by the subjects as versions of their front rounded [œ] vowel. If they were basing their judgments on phonological distinctions present in German, it follows that [ɨ], [ɞ], and [œ] would be placed near one another in the perceptual space. For this to occur, however, implies the existence of some perceived acoustic similarity between these sounds. Perhaps a more general description of DIII should be attempted, since the separation of [i, e, æ, ɑ, ɑ˞] and [ʌ] from the other vowels occurs in the Thai and English spaces as well. Such an attempt will not be considered here except to note that some concept of "relative high-frequency de-emphasis" may be required to encompass the acoustic characteristics of [u, o, y, œ, ɨ] and [ɞ].

Moving on to the English results, we consider Figure 6. Although the configuration of vowels in the plane makes sense, the dimensions extracted by PARAFAC are not particularly satisfying. Dimension DII separates [i, e, æ, ɑ, ɑ˞, ʌ] from the other vowels, somewhat like dimension DIII of the German space, but further distinguishes [o, u, y] from [ɨ, ɞ, œ], the latter 3 vowels being placed near the center of the DI-DII plane. As with the German dimension DIII, it is difficult to find a clear acoustic or articulatory correlate for this dimension. DII also reflects to some extent the front-back distinction. Dimension DI obviously has something to do with vowel height, as seen in the series [æ, e, i] and the

relationship between [u], [o], and [ɑ], but the labelling of vowels as high, mid, or low is not perfect on the basis of DI alone. The pair [ɣ, e], for example, have approximately the same DI value, but are not of the same linguistic vowel height, nor do they have similar first formant values. The same comment holds for the pair [o, æ].

Notice that this 2-dimensional result for English seems at odds with Singh and Wood's conclusion that [ɚ] adds an extra dimension to the space. It may be that with a larger number of subjects we too would have found 3 dimensions. But a comparison of the results for the individual English listeners shows that the placement of [ɚ] is inconsistent from one person to the next, whereas the configuration of all other vowels, including non-native vowels is much the same from person to person. This suggests that the perception and identification of [ɚ] is not part of the vowel perception space, and that the variation in placement of [ɚ] was interpreted as noise in the data.

The Thai space (Figure 7), although a 3-dimensional one, is more similar to the English space than to the German. The DI-DII plane in the Thai solution shows much the same configuration as the English DI-DII plane, except that it seems to be rotated somewhat in comparison. Dimension DIII, though, is puzzling, in that it bears no relation to any linguistic or acoustic or articulatory measures at all. No plausible parameter groups [i, ɑ, u, o, æ, e] together towards one end of a scale opposed to [ɨ, œ, ɣ, ɚ, ʌ] at the other end. We might, however, look at DIII as a familiarity scale of vowel quality. Given the somewhat counter-to-fact assumption that the [ɨ] vowel was perceived as an unfamiliar sound, DIII can be seen to separate native Thai [i, e, æ, ɑ, o, u] from the remaining 6 vowels. But even if we generously accept this interpretation, it still remains to be explained why Thai subjects ranked vowels on a foreignness dimension and English and German subjects did not.

In comparing the results from the three language groups, it appears to be the case that there is no universal perceptual space. Speakers of different native languages respond to the same stimuli in different ways. Neither the number of dimensions nor the interpretation of the dimensions is consistent across all three languages tested.

It can hardly be overstated that our results are highly tentative, and have brought new questions to light. Among these are (1) Is there any clear interpretation for the third Thai dimension DIII? (2) Can more satisfying correlates be found for the English perceptual plane and the Thai plane DI-DII? (3) Why is the German space so different from the Thai and the English results?

Work in progress is contributing to answering these questions. An appeal to non-linearity of dimensions is promising for question 1; it seems that the third factor is an artifact of the linear model used by

PARAFAC, and that Thais in fact, use 2 dimensions, much like the English subjects. As to question 2, we have strong evidence that the DI-DII plane for Thai and English can be redescribed in terms of 2 oblique axes, which correlate with the first formant and the difference between the first two formants.

The third question is being investigated by including Swedish and Turkish speakers as subjects. Since Hanson's earlier results with Swedish are remarkably like our results for Thai and English, we predict that Swedish speakers will respond to our triadic comparison task more like the Thai and English subjects than the Germans. If this is the case, the mere presence of front rounded vowels is insufficient to explain the German subjects' departure from what we might with some boldness call the Thai-English-Swedish perceptual space.

The hypothesis we are investigating further is that there may be phonological relationships between sounds in a language which affect perceived distances between the sounds. The psychologically real umlaut rule in German, which transforms $\begin{bmatrix} +\text{back} \\ \alpha\text{round} \\ \alpha\text{high} \\ \alpha\text{low} \end{bmatrix}$ vowels to $\begin{bmatrix} -\text{back} \\ \alpha\text{round} \\ \alpha\text{high} \\ \alpha\text{low} \end{bmatrix}$ vowels, may be such a relationship. Turkish is of interest here since its highly symmetrical vowel system and vowel harmony constraints may likewise influence native speakers' perceptual judgments.

References

Cooper, Lee (1970), *Metric Multidimensional Scaling and the Concept of Preference*, Western Management Science Institute, UCLA.

Hanson, G. (1967), "Dimensions in speech sound perception: An experimental study of vowel perception," *Ericsson Technics* 23, 3-175.

Harshman, Richard (1970), "Foundations of the PARAFAC procedure: Models and conditions for an 'explanatory' multi-modal factor analysis," *Working Papers in Phonetics No. 16*, 84 pp.

Levelt, W.J.M., van der Geer, J.P., and Plomp, R. (1966), "Triadic comparison of musical intervals," *Brit. J. Math. Stat. Psychol.* 19, 163-179.

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

Singh, S. and Woods, G. (1970), "Multidimensional scaling of 12 American English vowels," abstract of paper read to the Acoustical Society of America, April, 1970.

Torgerson, W.S. (1958), *Theory and Methods of Scaling*, New York: Wiley and Sons.

# *Factor Analyses of Tongue Shapes*

Peter Ladefoged, Joseph L. DeClerk*, and Richard Harshman

[Paper presented at the 81st meeting of the Acoustical Society of America]

This is a progress report about a new technique which looks as if it might give us very significant information about the way the tongue is used in making vowel sounds. Our object is to discover which factors underlie the shapes of the tongue that occur in vowels. We hope that we will be able to characterize vowels in an explanatory way, rather than in terms of arbitrary tongue shapes. Our first hypothesis, which is apparently true only to a very limited extent, is that the factors would directly indicate the *causes* of the different shapes; that is to say, they would reflect the pulls of the muscles and the other forces which affect the shape of the tongue.

The data for our analysis were obtained by cineradiology. Six subjects were photographed while saying (among other things) ten sentences each of the form "say *h_d* again". The vowels in the frame were /i ɪ e ɛ æ a ɔ o ɷ u/. Spectrograms (as in Figure 1) were made of all 60 sentences (60 sentences since there were six speakers each saying ten vowels). The spectrograms included at the top a digitally coded frame count which had been recorded at the time the photographs had been taken. This enabled us to mark each spectrogram at an appropriate point in the middle of the vowel in a word, and to be able to locate the corresponding frame in the film.

We then made tracings (as in Figure 2) of each of these 60 frames. We decided that we could characterize the tongue in terms of six equally spaced points starting from the root of the tongue which was usually fairly well defined on the original x-rays. Accordingly, we measured along the tongue on each of the 60 frames, and marked these points. We also marked the lower edge of the upper incisors, and an arbitrary point on the floor of the nasal cavity. These points, shown as A and B on the figure, were used in lining up each of each subject's utterances. The relations between each subject and each of the other subjects were arbitrarily arranged so that it appeared by eye that there was the maximum similarity over all subjects. At this stage one subject appeared to be very different from all the others and he was discarded.

The (x,y) coordinates of each point on each tracing were digitally recorded on a PDP-12 computer, using a GRAF-PEN reader. The origin of the coordinate system was completely arbitrary, and of little relevance,

---

* Joseph L. DeClerk is at the U.S. Army Electronics Laboratory, Fort Monmouth, New Jersey.

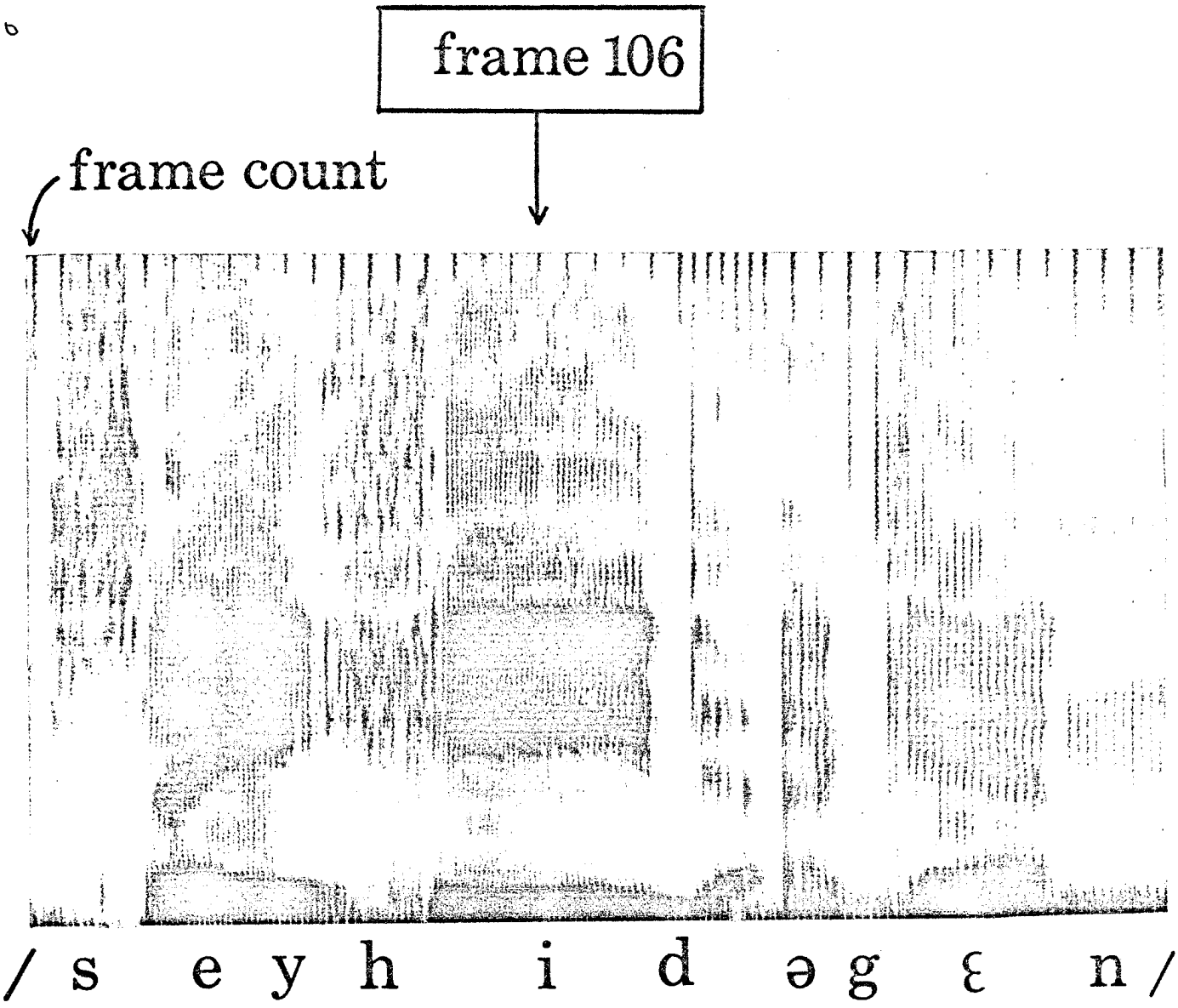



Figure 1. Spectrogram of the phrase "Say heed again" as said by subject 1, showing the digitally coded frame count, and the frame that was marked as being in the middle of the word *heed*.

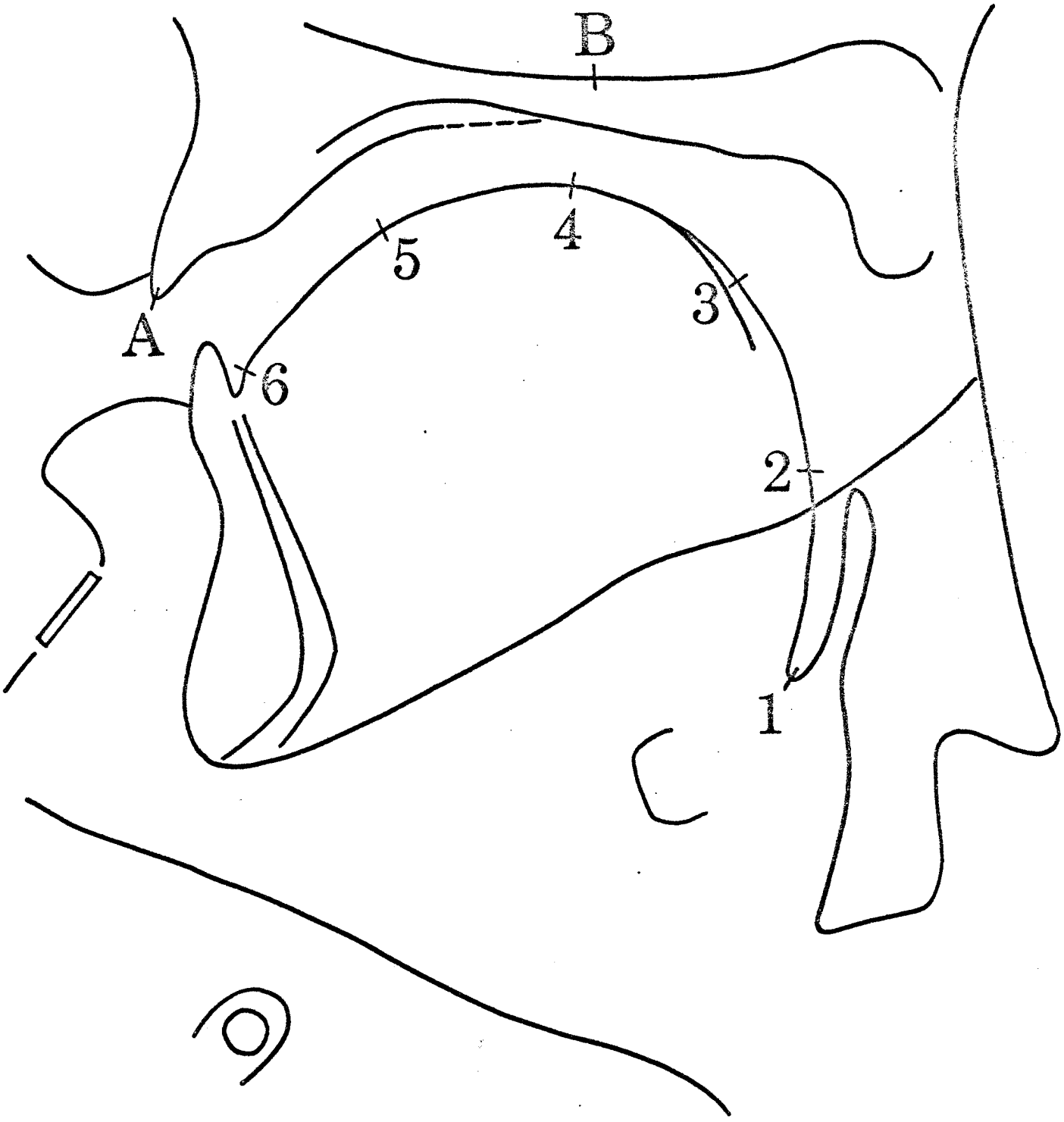



Figure 2. A tracing from the x-ray of subject 4 in the middle of the vowel in *hid*. The numbers designate six equally spaced points starting from the root of the tongue. The letters indicate fixed points used in lining up each of the subject's tracings.

since the first step in the analysis was to calculate for each subject a neutral tongue shape, which was taken to be the mean of his ten peripheral vowels. All subsequent analyses were then concerned with tongue displacements relative to this neutral shape. It should be noted that the neutral shape of the tongue was determined for use in these analyses only. We do not mean to imply that this is in any sense *the* neutral shape of the tongue.

The forces affecting the shape of the tongue are extremely complex. A very simplified diagram of three of the major influences is shown in Figure 3. The first is the degree of jaw opening, an influence which was somewhat neglected in previous descriptions, but which has been recently, and rightly emphasized by Lindblom and Sundberg (1971) who have pointed out that the major difference in the position of the tongue in vowels such as /ɪ ɛ æ/ is simply that the jaw is lowered. Next there is the action of the principle muscle of the tongue, the genioglossus, which pulls the root of the tongue forwards, and thus causes the front of the tongue to be pushed upwards. Then there is the upward and backward pulling action of the styloglossus. In addition, there are several other muscles, such as the glossopharyngeus, which pulls backwards, the hyoglossus pulling downwards, the mylohyoideus, which lifts the whole body of the tongue, and the longitudinal muscles which bunch the tongue lengthwise. None of these actions are shown on this very simplified diagram as we have not been able to isolate the contribution of any of them to the shapes of the tongue found in our x-rays.

We looked for the factors underlying tongue shapes by using the PARAFAC procedure for factor analysis (Harshman 1971). Our input data matrix consisted of twelve numbers representing the relative (x,y) coordinates of the six points, of each of ten vowels of each of five speakers. We have conducted a large number of analyses of this data set and of various subsets of it. Figure 4 shows one factor that occurs. This factor indicates that as the points near the root of the tongue move forward, the points at the front of the tongue move up. The length of the arrows indicates the degree of influence (the loading) of the factor on each point. This factor is obviously very similar to the action of the genioglossus muscle as shown in Figure 3. But on the whole, at this stage, we find that if the analysis is constrained so that the shape of the tongue is to be described in terms of, say, three factors, then these factors do not bear a simple relationship to the pulls of the muscles, or to jaw opening.

For example, another factor we found in the same analysis is shown in Figure 5. This factor indicates that the points on the front of the tongue tend to move forward together; and this movement has very little effect on points near the root of the tongue. There is no single muscle which could cause these movements.

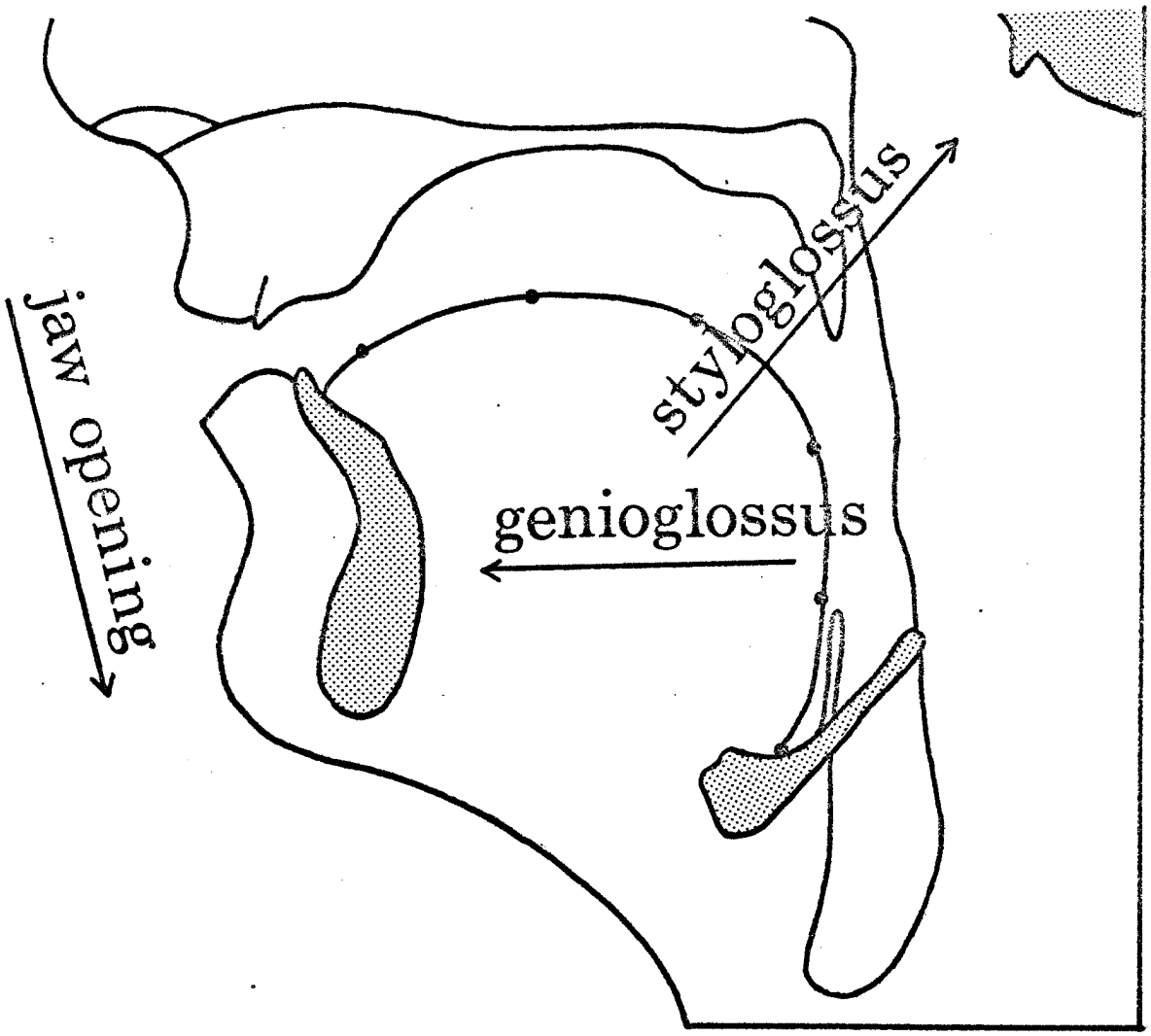



Figure 3. A simplified diagram of three of the major influences on the shape of the tongue.

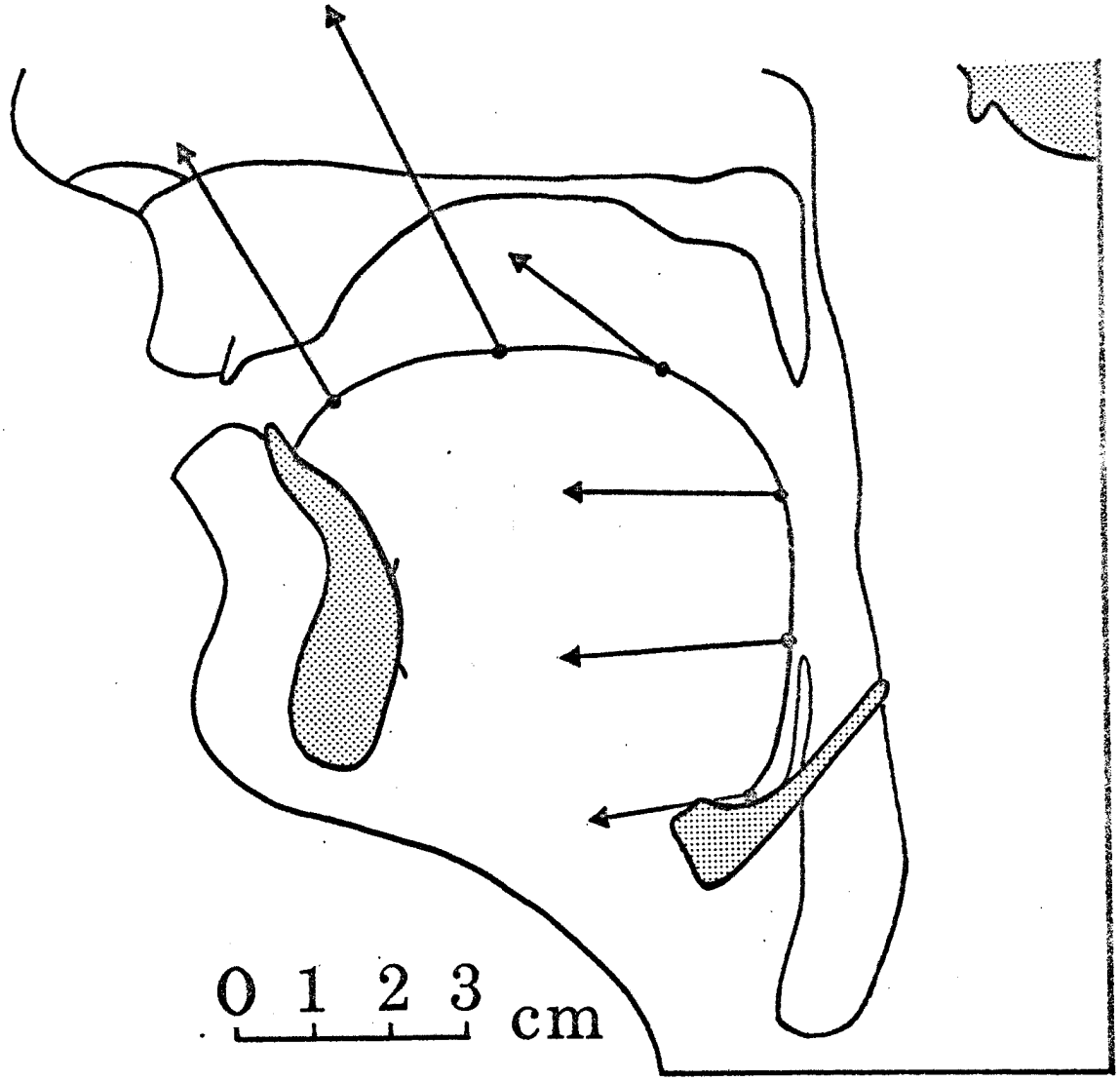



Figure 4. Vectors indicating the degree and direction of influence of one factor on each point.

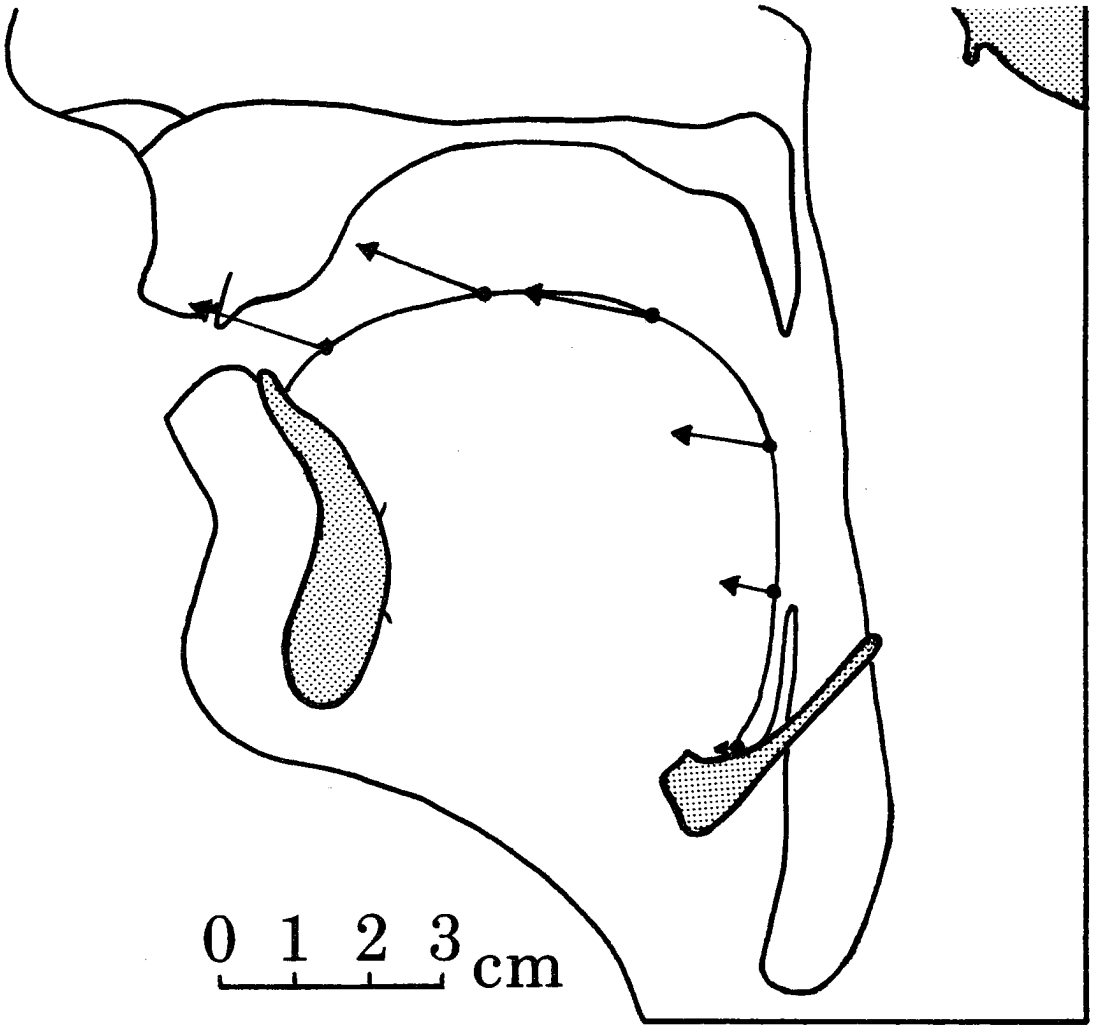



Figure 5. Vectors indicating the degree and direction of influence of a second factor on each point.

The results of all our analyses so far seem to suggest that the shape of the tongue cannot be readily described in terms of a small number of factors which correspond to the actions of individual muscles. If the description of the shape of the tongue is to be made in terms of a reasonably small number of parameters, then these parameters will reflect complex interacting forces, rather than individual muscle actions.

References

Harshman, R. (1970), "Foundations of the PARAFAC procedure: Models and conditions for an 'explanatory' multi-modal factor analysis," *Working Papers in Phonetics No. 16*, 84 pp.

Lindblom, B. and Sundberg, J. (1971), "Acoustical consequences of lip, tongue, jaw and larynx movement," *Papers from the Institute of Linguistics University of Stockholm* 2, 1-41.

*Parameters of Tongue Shape*

Peter Ladefoged, Joseph L. DeClerk*, and Richard Harshman

[Abstract of a paper to be presented at the
Speech Symposium, Budapest, August 1971]

This paper will discuss the shape of the tongue during the pronunciation of ten vowels of American English by six speakers. The primary data consists of a cineradiology film from which the midline of the tongue was traced in appropriate frames. Analyses have been made of the relative positions of the tongue with reference to coordinate systems fixed with regard to (1) the individual's skull, (2) the mandible, and (3) one of the individual's tongue shapes. There is a great deal of variety in the way in which different individuals produce the tongue shapes required for these ten vowels. Most of the subjects make a clear distinction between tense vowels in which the tongue is bunched up longitudinally, and lax vowels with a flatter tongue. The differences within the group of lax front vowels are made largely by lowering the jaw. For the complete set of vowels it appears to be difficult to characterize the tongue shapes in terms of the actions of separate muscles.

---

* Joseph L. DeClerk is at the U.S. Army Electronics Laboratory, Fort Monmouth, New Jersey.

# *Glottal Pulse Waveform Effects in Line Analog Speech Synthesis*

Lloyd Rice

[Paper presented at the 81st meeting of the Acoustical Society of America]

At the Phonetics Laboratory at UCLA we recently undertook the challenge of constructing a line analog vocal tract model which would operate in a small computer environment. This initial project has been successfully completed in the form of a program for the Digital Equipment Corporations PDP-12 computer (Rice 1971). The vocal tract model is operated in an interactive mode where the operator may determine both the glottal source waveform and the tract area function. The area function is dynamically controlled by a sequence of motion-picture like frames which are individually set up using a mid-saggital x-ray view of the tract displayed on the computer screen. Before an utterance is computed, a glottal waveform may be specified as a hand-drawn curve using a waveform editor, or a mathematically generated curve using DEC's interactive language, FOCAL, or by a combination of the two techniques.

To demonstrate the capabilities of this system, I have prepared a set of utterances using a variety of glottal source waveforms. The mathematical functions used as a basis for these curves were taken from the article by Rosenberg (1971). I have identified these curves by reference to Rosenberg's lettering scheme.

In selecting an utterance for the demonstration, we were constrained by the lack of a friction source generator and a nasal tract in the present model. Accordingly, we selected the phrase, "Where were you a year ago?", to ask the question which Bell Telephone Laboratories has been answering for several years now. The Digital Equipment Corporation has provided us with a demonstration machine so we will now be able to hear the computer produce this phrase.

In the upper part of Figure 1 the volume velocity function is shown. This is a polynomial function using the parameters of Rosenberg's curve B. In the lower part of this figure you see a wide-band spectrographic section of the pressure function generated by this volume velocity pulse.

In the next figure (Figure 2) is shown the same polynomial curve except that the critical sharp corner has been rounded somewhat. Note the faster decay of energy in the frequency section. In this sample you may hear a slight softening of the quality.

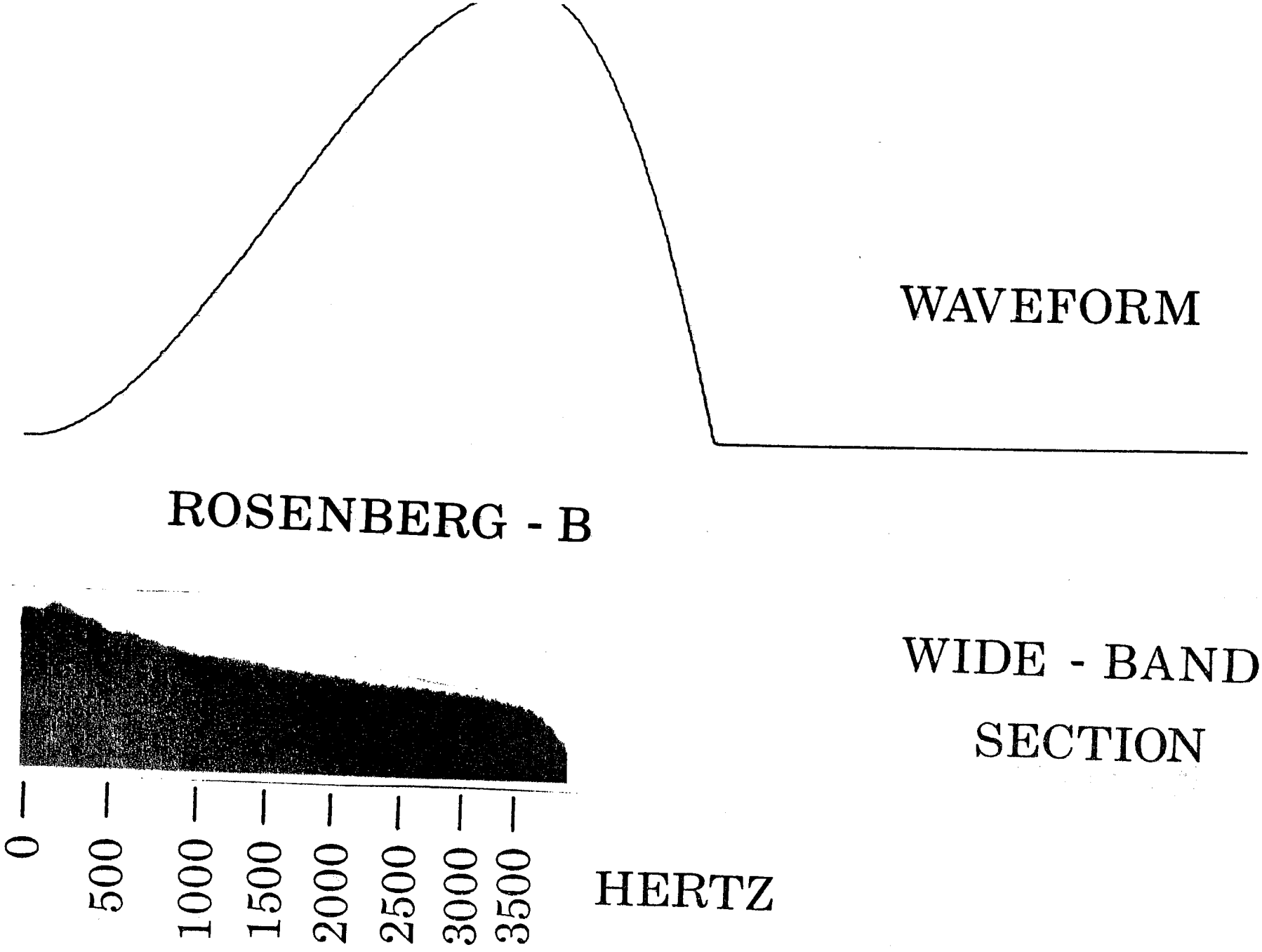



Figure 1. Rosenberg B polynomial function.

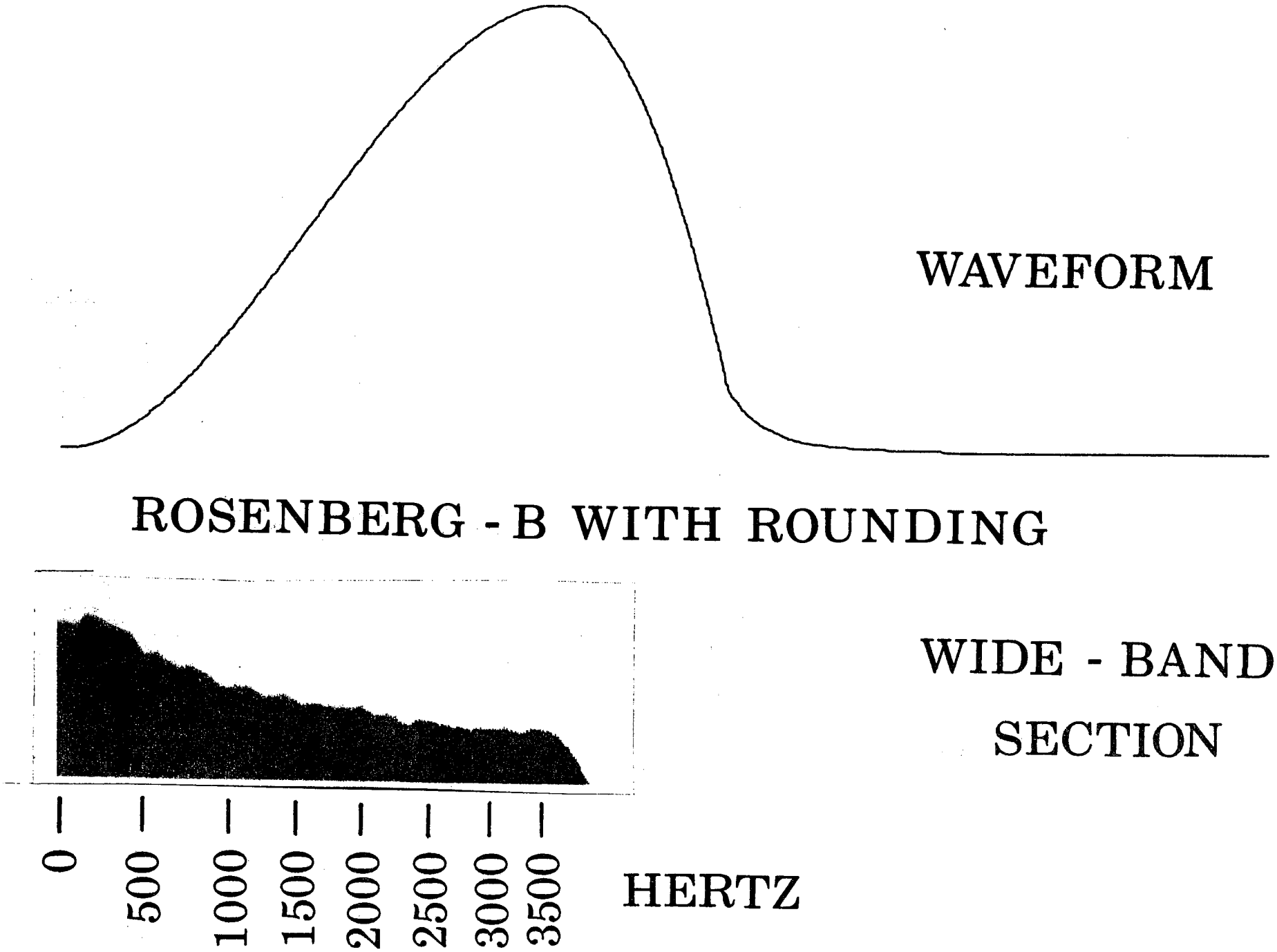



Figure 2. Rosenberg B polynomial function with rounded corner.

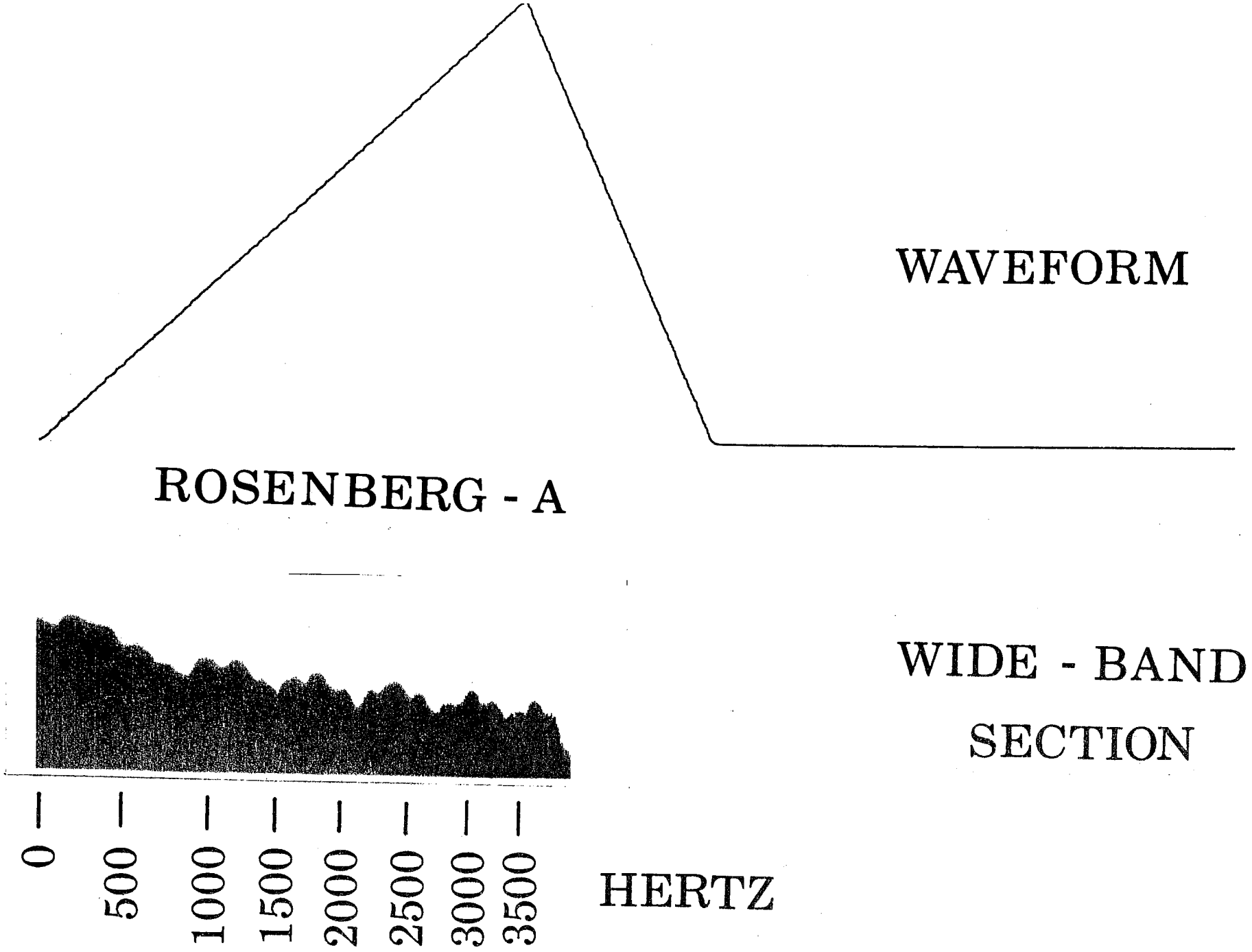



Figure 3. Rosenberg A triangular function.

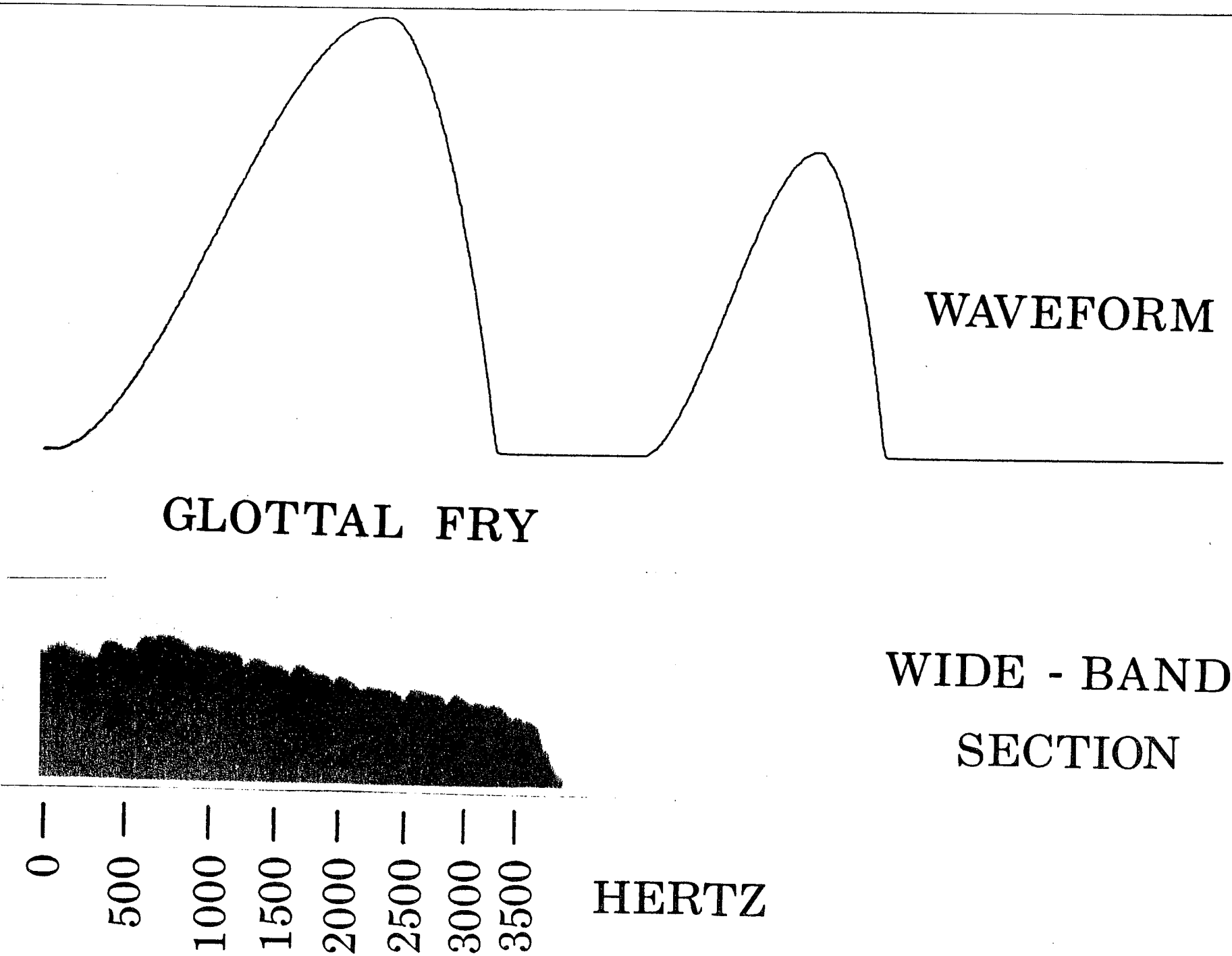



Figure 4. Double phonation with Rosenberg B polynomial functions.

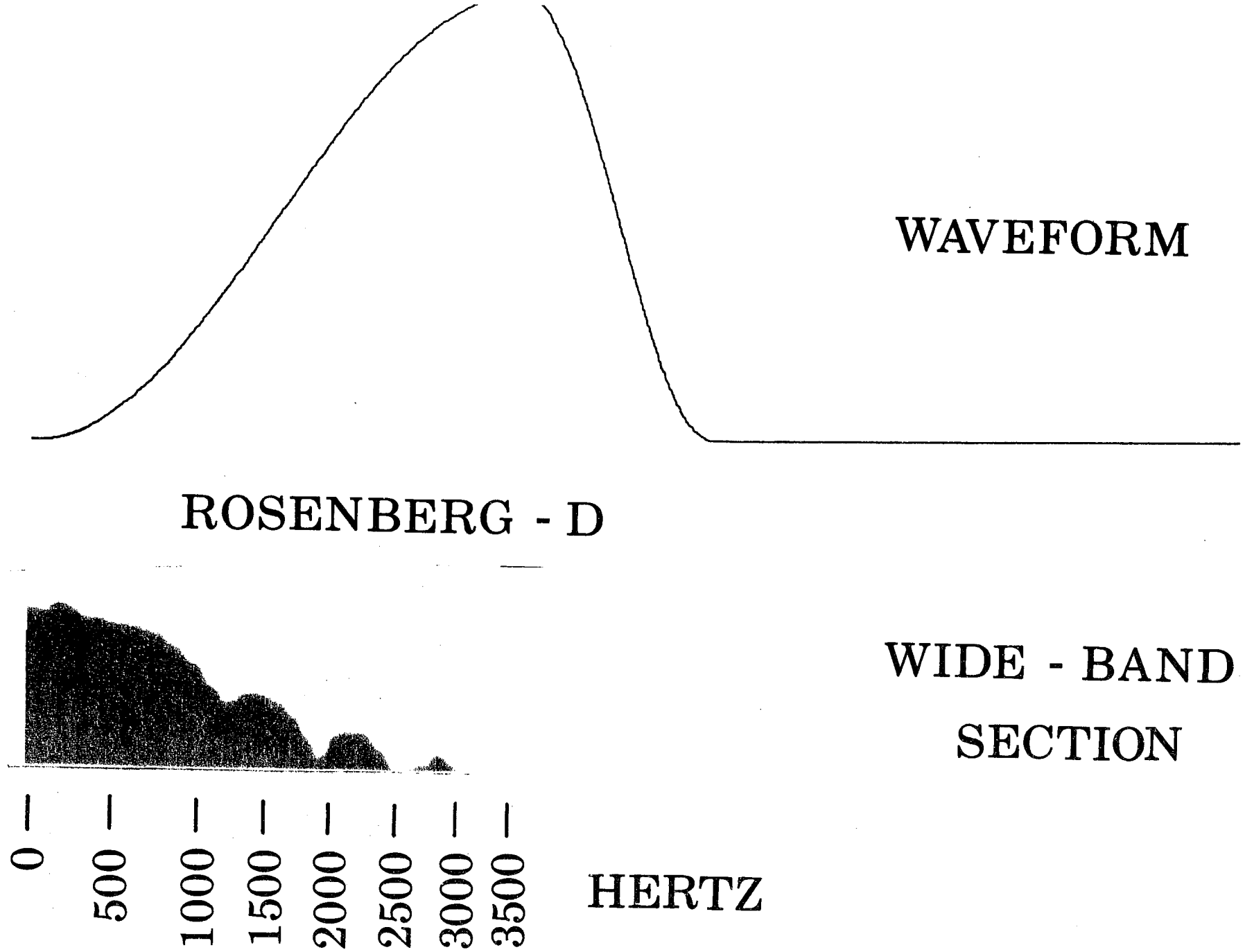



Figure 5. Rosenberg D trigonometric function.

In Figure 3 is presented a triangular pulse. Notice the periodic structure in the frequency section. Speech generated using this waveform seems to have a hollow quality.

With the waveform shown in Figure 4 I have attempted to imitate the characteristics of double phonation thought to be associated with creaky voice or glottal fry. Each part of this waveform is a Rosenberg-B polynomial function. The sound, however, does not seem to have the characteristic snap of glottal fry and sounds to me somewhat like that of the triangular pulse.

In the next figure (Figure 5) we see the trigonometric function, Rosenberg's curve D. It has a very rapidly falling frequency section curve and also has a smoother derivative than the other curves presented here. It has a quality similar to that of breathy voice or murmur.

I have attempted to demonstrate the feasibility of reasonably good quality synthesis using a line analog model within the budget of a university phonetics laboratory. The program will generate approximately 1 1/2 seconds of speech, requiring about 4 minutes to compute a full length utterance.

The system has been very useful for teaching the relationship between articulatory gestures and vocal tract resonances and appears to be a useful tool for investigation of glottal source characteristics and voice quality.

References

Rice, L. (1971), "A new line analog speech synthesizer for the PDP-12," *Working Papers in Phonetics No. 17*, 58-75.

Rosenberg, A. (1971), "Effect of glottal pulse shape on the quality of natural vowels," *Journal of the Acoustical Society of America, Vol. 49, No. 2*, 583-590.

# *Syllable Structure and Sentence Rhythm*

George Allen*

[Paper presented at the 81st meeting of the Acoustical Society of America]

## Introduction

The rhythm of language is usually taken to mean the organization of stresses and syllables into phrase groupings, in analogy to the metric feet one encounters in poetry. The terms "syllable timed" and "stress timed" are thus applied to languages whose phrase rhythms seem to be organized either around all of the syllables of the phrase or around just the stressed syllables, respectively. However, if one listens to two languages of the same type, for example English and German, both of which are stress timed, one is struck by the rhythmic *differences* that remain. These differences force us to consider other possible sources for the rhythmic character of languages. Instead of looking just at how syllables are organized into phrases, we must consider as well how segments are organized into syllables, phrases into breath groups, breath groups into paragraphs, and paragraphs into discourse. The first of these, that is, the rhythmic effects of the way segments are organized into syllables, is the focus of the research reported here.

## Structure of the study

We chose to study the interrelationship between the initial consonant-vowel sequence of a syllable and the final segments of the preceding syllable, since it seemed to us that the articulatory organization of this intersyllabic region might contain some of the information

---

* George Allen is now at the University of North Carolina, Chapel Hill, North Carolina, 27514. The paper at the Acoustical Society had Peter Ladefoged listed as co-author; but the editor of *Working Papers in Phonetics* has decided that he had not done enough work on this paper to justify calling him a co-author. Some of the work on the PDP-12 and the original data processing at UCLA was done by George Papçun; and we would also like to acknowledge the generous donation of time on the PDP-12 computer of the Information Science Division of the Department of Community Health Science, Duke University.

that gives the syllables of a language their particular rhythmic flavor. Earlier work has shown, for example, that the rhythmic beat associated with a stressed syllable in English precedes the onset of the nuclear vowel of the syllable by an amount that is positively correlated with the length of the initial consonant cluster of the syllable; that is, the longer the duration of the initial cluster, the earlier in the syllable will the perceived rhythmic beat appear to fall. It is perhaps intuitively reasonable to hypothesize, on the other hand, that segments in *other* syllables should have little effect, if any, on perceived beat location; that is, the beat should fall in the same place in a syllable regardless of the nature of other syllables nearby.

In French, however, the situation is apparently different: the final consonants of a syllable are said to group with the initial consonants of the next syllable to form a grand cluster. That is, where in English we would have "my stick" and "mice tick," French would have only "my stick." Finally, Polish should be more like English than like French with respect to this consonant grouping phenomenon.

We therefore composed a set of three nonsense utterances that would exploit this presumed difference between English, Polish, and French, as shown in Figure 1. At two different stressed syllable locations in the sentences, three consonant sequences were contrasted around the juncture boundary between syllables. In utterance one, for example, we have the syllables /sap fáʃ/, containing the sequence vowel-consonant-juncture-consonant-vowel, as compared to the other two utterances, which contain the sequences vowel-juncture-consonant-vowel, in utterance number two, and vowel-juncture-consonant-consonant-vowel, in utterance three. The structures associated with the last stressed syllables of the three utterances are analogous, but more complex, involving an additional /l/ segment. According to the description of sequential syllable structure given earlier, English and Polish should have the rhythmic beat for the first stressed syllable fall in similar locations in utterances 1 and 2, ignoring the presence or absence of the final /p/ in the preceding syllable; these two beat locations should contrast with that for utterance three, where the presence of the /p/ in the initial cluster should displace the rhythmic beat forward in the syllable. In French, however, the beats should be similar in utterances 2 and 3, indifferent to the location of the juncture, and utterance 1 should be the different one, the beat being relatively later in the syllable.

Two native speakers of each language repeated these three sentences a number of times, treating each sentence as if it were a nonsense utterance in his own language, and from these repetitions one example was chosen that contained no hesitations or awkward pronunciations. Each of these eighteen utterances, that is, three languages times two speakers of each language times three sentences by each speaker, was copied onto a four second tape loop for presentation to listeners.

1. a la kárt, sap fáʃ, də lɛ pláʃ.

vc#cv́

2. a la kárt, sa fáʃ, də lɛ spláʃ.

v#cv́

3. a la kárt, sa pfáʃ, də lɛs pláʃ.

v#ccv́

Figure 1 Broad transcription of experimental utterances.

## Experimental apparatus

The experimental apparatus is shown schematically in Figure 2. Listeners sat in a sound treated room and heard the utterances binaurally through headphones. They were directed to tap their finger on a copper plate in time to the rhythm of an utterance, and no attempt was made to constrain their manner of tapping. During each revolution of a tape loop, a timing pulse on track one shortly preceding the onset of the utterance on track two reset the computer's millisecond clock to zero. Each time the listener tapped his finger, the pulse generator signaled the computer to store the time of occurrence of that tap. At the end of a certain number of revolutions the sound was turned off, the data for that utterance were stored on computer tape, and a new utterance was begun. The resulting data were tap locations, in milliseconds, relative to the timing pulse on track one of each loop.

Listeners for this study were native English, French, and Polish speakers recruited either through the classroom or from a list in the foreign-student center at UCLA. They were paid for their services.

## Results

The data generated by these subjects are not very pretty. Listeners differed greatly in their ability to perform the task, some giving well defined clusters of taps easily interpretable as defining a beat, others tapping so variably that no smoothing and averaging could be performed. As a result, only six subjects' data were analyzed, two native speakers of each language. Nevertheless, there are some results of interest.

The first result concerns the variability with which the listeners tapped. As we just noted, there were great differences in the abilities of the different subjects, but the subjects showed a *pattern* to their variability. Using as our measure of variability the statistical variance of the time locations of the several taps associated with a given syllable, we find that a listener tapped with *less variability*, that is, smaller variance, on the utterances that had been spoken by native speakers of his own language than on the utterances that had been spoken by foreigners. It was as though he had an immediate intuitive "understanding" for the rhythm of his own language that did not extend to other languages, even though he might in fact speak these other languages. This result is highly reminiscent of the data reported by Wish, wherein foreign students appeared not to have a very good idea of the stress or rhythm patterns of English words and phrases.

The second result of importance concerns the locations of the subjects' taps. On the basis of the hypotheses suggested earlier,

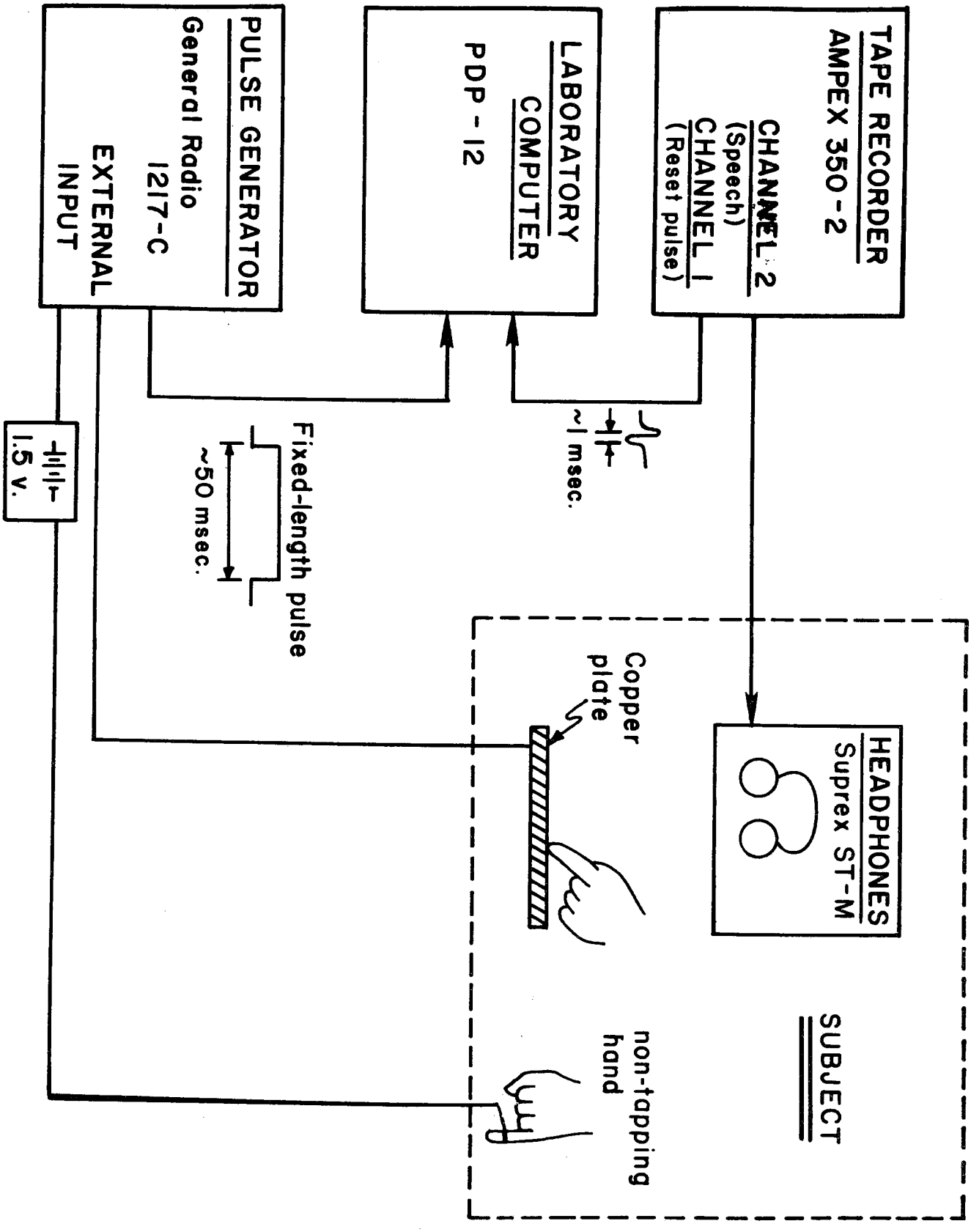



Figure 2 Experimental apparatus

we had hoped to observe some differences between the locations of the various listeners' taps on the different stimulus utterances. There were in fact no obvious language related differences, and so the second preliminary conclusion stated in the abstract for this paper (*J. Acoust. Soc. Amer.* forthcoming) is not supported by these six subjects' data.

There were, however, consistent differences in average tap location for the different phonetic types of syllable structure represented in the stimulus utterances. The average tap location for the syllable /fa/ in utterance 2 was about 40 or 50 milliseconds later than the average tap location for the comparable syllables of the other two utterances. That is, the presence of the extra /p/, either before or after the juncture, served to displace the felt beat location forward in time by 40 or 50 msec. The situation was similar for the presence vs. absence of the /s/ in the last beat location of the utterances, except that the range was a 30 to 60 msec. difference instead of 40 to 50 msec. As noted the location of the juncture appeared to have no effect on beat location, in any of the languages, a counter-intuitive result.

## Summary

The results of this study are suggestive in two ways. First, the differences within subjects in variability of tapping to utterances in their own vs. other languages indicates that there are indeed rhythmic differences between languages that can be perceived and that are difficult to overcome. Second, the differences in average tap location for the different utterances verify that segmental sequences do influence the rhythmic character of a syllable, though not in ways we might expect. The absence of language-related effects was disappointing, but might mean only that our French speakers spoke too carefully or that their language had been contaminated by their English speaking environment. Or, the supposed consonant shift might not actually exist in French. Certainly the variability within and between listeners indicates that ours is not the best possible methodology, and we would be most appreciative of suggestions as to alternative experimental schemes.

# *Simplicity is a Complicated Question*

Victoria A. Fromkin

[Abstract of a paper read at the First Annual California Linguistics Conference, Berkeley, May 1-2, 1971]

Despite all the problems which have been raised regarding the proposed Evaluation Metric (hereafter, EM) it continues to be a central notion in linguistic theory. It is interesting to note the changing character of this concept over the years. In *Syntactic Structures* (1957), Chomsky suggested that the only reasonable goal for our theory was that it provide an *evaluation procedure* for the selection of the *better* of two grammars; both a *discovery procedure* and a *decision procedure* (i.e. means of selecting the *best* grammar of a language) were considered beyond our means. Yet with the spelling out of the competence/performance distinction, and the insistence on the 'psychological reality' of the linguists' grammars, it is clear that the EM must serve as a decision procedure to select *the* descriptively adequate, psychologically real, innate grammar which models the set of rules in the ideal speaker-hearer's mind. With the extension of the 'classical' theory of generative phonology to include a theory of Markedness, the EM again seems to have taken on a new role, i.e. to explicate the notion of 'naturalness'. While this can be seen as an outgrowth of the second phase, it is clear that the EM no longer has as its main task the selection of one particular grammar of a language.

This paper will therefore discuss the concept of the EM as related to its various functions. It will attempt to show that serious problems arise when trying to apply and develop the metric and the theory of Markedness. Specific reference will be made to a number of languages, in evaluating the evaluation metric and some of the proposed markedness conventions.

*The Interpretation of Phonological Features in Assimilation Rules*

Theo Vennemann

[Abstract of a paper read at the First Annual California Linguistics Conference, Berkeley, May 1-2, 1971]

Germanic phonology offers a number of problems, involving consonant-vowel interactions, which have never found a systematic explanation. Central among these is the Old High German monophthongization, traditionally given as follows:

$$\left\{\begin{array}{ll} ai \rightarrow \bar{e} \;/\; \_\_ & \left\{\begin{array}{l} h \\ r \\ w \\ \# \end{array}\right\} \\ au \rightarrow \bar{o} \;/\; \_\_ & \left\{\begin{array}{l} h \\ \text{all dentals} \\ \# \end{array}\right\} \end{array}\right\}$$

The above is, of course, nothing more than a list. However, all attempts to convert this list into a uniform formulation of immediate explanatory value have failed, the last one in a specialized study by George Williams of M.I.T. in 1970.

I offer the following explanation. First, I note that the change occurred in the following phonetic long vowel and diphthong system (umlaut variants omitted):

| | | |
|---|---|---|
| $\bar{i}$ | iu,io | $\bar{u}$ |
| $\bar{e}$ | | $\bar{o}$ |
| ai | $\bar{a}$ | au |

Next, I take into consideration that ai and au, where they did not monophthongize, were spelled *ei*, *ou* at an early date. Therefore, monophthongization was preceded by an assimilatory change ai → ɛi, au → ɔu. Monophthongization consists, therefore, of the following changes:

$$\left\{\begin{array}{lll} \varepsilon i \rightarrow \overline{\varepsilon} \;/\; \_\_\_ & \left\{\begin{array}{l} h \\ r \\ w \\ \# \end{array}\right\} \\ \\ ɔu \rightarrow \overline{ɔ} \;/\; \_\_\_ & \left\{\begin{array}{l} h \\ \text{all dentals} \\ \# \end{array}\right\} \end{array}\right\}$$

It was, in essence, a lowering of the glidal vowels: $\varepsilon i \rightarrow \varepsilon\varepsilon$/ ..., $ɔu \rightarrow ɔɔ$/ ... (where VV and $\overline{V}$ are just notational variants, in the case of tautosyllabic long vowels):

$$\left\{\begin{array}{lll} i \rightarrow \varepsilon \;/\; \varepsilon \_\_\_ & \left\{\begin{array}{l} h \\ r \\ w \\ \# \end{array}\right\} \\ \\ u \rightarrow ɔ \;/\; ɔ \_\_\_ & \left\{\begin{array}{l} h \\ \text{all dentals} \\ \# \end{array}\right\} \end{array}\right\}$$

The common portion of the environment (disregarding # from here onward) is h , r . It is well-known that Gothic has the following rule:

$$\left\{\begin{array}{l} i \rightarrow \varepsilon \\ \\ u \rightarrow ɔ \end{array}\right\} / \_\_\_ \quad \left\{\begin{array}{l} h \text{ (and } h^w) \\ \\ r \end{array}\right\}$$

i.e. $$\left[\begin{array}{l} V \\ -\text{long} \\ +\text{stress} \end{array}\right] \rightarrow [+\text{low}] \;/\; \_\_\_ \quad \left\{\begin{array}{l} h \ (h^w) \\ r \end{array}\right\}$$

The frequency of this change in other languages indicates that it is assimilatory. The environment in the Gothic rule must therefore be low. Indeed, both h (a post-velar spirant) and r (an alveolar trill) are low in the region where vowels are formed:

$$\left[\begin{array}{l} V \\ -\text{long} \\ +\text{stress} \end{array}\right] \rightarrow [+\text{low}] \;/\; \_\_\_ \quad \left[\begin{array}{l} C \\ +\text{low} \end{array}\right]$$

This explanation is then extended to the OHG case:

$$\begin{bmatrix} V \\ -stress \end{bmatrix} \rightarrow [+low] \; / \begin{bmatrix} V \\ +low \\ +stress \end{bmatrix} \_\_ \begin{bmatrix} C \\ +low \end{bmatrix}$$

The interpretation of this rule requires, however, the following convention:

> Assimilatory features in a rule environment must be interpreted as relative to the corresponding features in the assimilable (i.e. the assimilatorily affected) segment.

The velar glide w , usually characterized as high, is high only in the back vowel region but is low relative to front vowels: ɛi → ɛ̄ / ___ w, but not *ɔu → ɔ̄ / ___ w. Conversely, dentals are formed with the blade of the tongue raised but with the tongue-body sloping down toward the back region of the oral cavity. They are, therefore, relatively lower for back vowels than for front vowels: ɔu → ɔ̄ / ___ dentals, but not *ɛi → ɛ̄ / ___ dentals.

It is a consequence of this convention (which is part of the theory of grammar rather than of the grammar of OHG) that the natural classes in the environment of an assimilation rule are defined relative to each particular assimilable segment and need not, therefore, be identical.

This new convention (and, by implication, my explanation of the OHG rule) is confirmed by a large number of similar changes both inside and outside Germanic. Conversely, all these changes are recognized for the first time as a class of rather simple assimilation rules. I list a few of these.

1. Old Icelandic: au → ó, ai → á / ___ $\begin{Bmatrix} h \\ r \end{Bmatrix}$

   Later: ei → é / ___ h

2. Old Icelandic: ai → æ (i.e. ɛ̄) / ___ w

   Actually: ɛi → ɛ̄ / ___ w

3. Old Icelandic: $\begin{Bmatrix} i \rightarrow e \\ u \rightarrow o \end{Bmatrix}$ / ___ h

with e further lowered if preceded by w :

$$ehC \rightsquigarrow \begin{Bmatrix} æCC & / & w\ \_\_\_ \\ éCC & \text{elsewhere} \end{Bmatrix}$$

4. Old Icelandic: eu → $\begin{Bmatrix} \text{jó} & \text{before dentals} \\ \text{jú} & \text{elsewhere} \end{Bmatrix}$

5. OHG: eu → eo / $\_\_\_\ C_1 \begin{bmatrix} V \\ -\text{high} \end{bmatrix}$

But: In Upper German only if $C_1$ is either dental or h (which is exactly the same natural class as in OHG ou → ō).

6. Swiss German (Kiparsky 1968):

$$O \rightarrow o \ / \_\_\_ \begin{Bmatrix} r \\ \text{dentals} \\ \text{palatals} \end{Bmatrix}$$

Kiparsky writes:

$$\begin{bmatrix} V \\ -\text{high} \\ +\text{back} \end{bmatrix} \rightarrow [+\text{low}] \ / \_\_\_ \begin{bmatrix} +\text{consonantal} \\ -\text{grave} \\ -\text{lateral} \end{bmatrix}$$

This formulation misses the assimilatory nature of the change. Notice that the same natural class occurs here as in OHG ou → ō: Swiss German has no postvocalic h , and OHG had no palatals after back vowels. The rule is:

$$\begin{bmatrix} V \\ -\text{high} \\ +\text{back} \end{bmatrix} \rightarrow [+\text{low}] \ / \_\_\_ \begin{bmatrix} C \\ +\text{low} \end{bmatrix}$$

7. OHG "secondary umlaut":

Umlaut: $V \rightarrow [-\text{back}] \ / \_\_\_ \ X \begin{bmatrix} -\text{consonantal} \\ +\text{high} \\ -\text{back} \end{bmatrix}$

condition: no # in X.

Umlauted a is spelled *e* except before certain consonants: *ht*, *hs*, *rC*, *lC*, *h*. Here, *a* is written; in MHG, ä .
Explanation:

i ü ← u where ü, ö, æ only before i, r̄, j.

e ö ← o

æ ← a

Raising: $\left\{\begin{matrix} e \\ æ \end{matrix}\right\} \rightarrow \underset{\cdot}{e} \; / \; \_\_ \; C_1 \; \left[\begin{matrix} -\text{consonantal} \\ -\text{back} \\ +\text{high} \end{matrix}\right]$

except where consonants intervene which we have recognized as low on independent grounds.

8. Modern Spanish, as described by Tomás Navarro (1966):

i u

close and open allophones

ị ụ

e o

ẹ ọ

(a) ị ụ ẹ ọ / ___ $\left\{\begin{matrix} x \\ \bar{r} \end{matrix}\right\}$

(b) (α) ị ụ ọ / r̄ ___

(β) ẹ / r̄ ___ unless followed by a tautosyllabic dental obstruent (or nasal)

(c) (α) ị ụ ọ in closed syllables

(β) ẹ in closed syllables unless the closing consonant is a dental obstruent (or nasal)

[(d) ẹ ọ in ẹi̲ , ọi̲ ]

(e) ọ / a ___ $\left\{\begin{matrix} r \\ l \end{matrix}\right\}$

Notes: I. r̄ is an alveolar trill.

II. x is a postvelar spirant.

III. Dental obstruents are higher in Spanish than in Germanic because of their forward articulation. We observe the complement of the OHG and Old Icelandic changes: Instead of lowering back vowels, Spanish dental obstruents raise the front vowel e .

9. In Quechua, i and u are lowered to e and o by uvulars but not by velars (William Bright and Lyle Campbell, personal communication).

10. In Diegeño, which has a phonemic contrast between dentals and alveolars, only the alveolars have a lowering influence on i and u (Langdon 1970).

Finally, I draw some theoretical conclusions from my study. Contrary to wide-spread belief among transformational-generative phonologists, detail information, such as universal height differences between dentals and alveolars, velars and uvulars, and also language-specific height differences such as between l and r , must be specified in the grammar before certain phonological rules (which are not themselves detail rules) apply. Furthermore, at least two detailed height values must be specified for consonants involving tongue movement. In a theory fully accounting for the changes outlined here, rules may take forms such as the following:

$$\begin{bmatrix} V \\ . \\ . \\ . \\ \alpha\ \text{back} \\ m\ \text{height} \end{bmatrix} \rightarrow [(m\text{-}n)\ \text{height}] \ /\ \ldots \begin{bmatrix} C \\ p\ \alpha\text{-height} \end{bmatrix} \ \ldots$$

where "α-height" ( = "height" for vowels) is the height scale in the region of [α back] vowels, and $p < m$ (m, n, p may themselves turn out to be variable). In any event, my study suggests that considerably more phonetic information has to be made available for the formal representation of phonological processes than is assumed in current models of phonology.

References

Kiparsky, Paul (1968), "Linguistic universals and linguistic change," in *Universals in Linguistic Theory* (E. Bach and R. Harms, eds.), New York: Holt, Rinehart and Winston, 170-202.

Langdon, Margaret (1970), *A Grammar of Diegueño: The Mesa Grande Dialect*, Berkeley: University of California Press.

Navarro, Tómas (1966), *Manual de Pronunciación Española* (6th ed.), New York: Hafner.

Williams, George (1970), "Germanisches ai und au im Altsächsischen und Althochdeutschen," *Zeitschrift fur Dialektologie und Linguistik* 1, 44-57.

# *Is the Left Hemisphere Specialized for Speech, Language, or Something Else?*

George Papçun, Stephen Krashen, and Dale Terbeek

Abstract

Experienced Morse code operators showed significant left hemisphere lateralization for the perception of dichotically presented Morse code letters. Rapid monotically presented words were not significantly lateralized. Subjects who did not know Morse code tended to show left lateralization when the stimuli were restricted in duration and presented with a relatively low intensity. The naive subjects did not show significant lateralization when a list including longer stimuli with greater intensity was presented. The results support the hypothesis that articulability is not a necessary property of stimuli lateralized to the left in dichotic listening.

## Introduction

In experiments over the last ten years normal right handed subjects have consistently shown a right ear superiority for the perception of competing dichotically presented verbal material. Kimura (1961) first showed ear asymmetry in normal subjects by demonstrating a right ear superiority for the perception of dichotically presented digits. This right ear superiority for verbal material is related to the left hemisphere of the brain by two further lines of research -- aphasiological research (as summarized in Kimura 1967) and neuroanatomical research (e.g., Rosensweig (1957) demonstrated that the contralateral evoked response in the cat is stronger than the ipsilateral response).

Since Kimura's work there has been a series of efforts to determine the precise nature of hemispheric lateralization. Kimura (1964) demonstrated that it is not the case that all auditory stimuli are perceived lateralized to the left hemisphere as she showed a left ear advantage for the perception of dichotically presented melodies. These findings are supported by Milner (1961) with experiments on brain damaged subjects. Also, Curry (1967) showed a left ear advantage for the perception of dichotically presented environmental sounds, (e.g., car starting, toilet flushing, etc.). Although it is generally supposed that dichotic presentation is essential to reveal ear dominance, Bakker (1967) found left ear dominance when he presented Morse code like signals to 6 to 12 year old children "to each ear separately."

Not all sounds have been found to be lateralized one way or the other. Schulhoff and Goodglass (1969) were unable to demonstrate any ear advantage either way for dichotically presented click sounds. Knox and Kimura (1969) obtained similar results with animal sounds presented to children. Zurif and Sait (1969) found that strings of nonsense words read off with list intonation did not produce significant laterality effects, while the same words with normal English intonation and morphology imposed on them produced a significant right ear superiority.

Curry (1967) and Kimura (1967) found left hemisphere lateralization for nonsense words. This, together with Zurif and Sait's results (op. cit.), indicates that meaningfulness is not a necessary condition for lateralization.

In two experiments, Shankweiler and Studdert-Kennedy (1967a and 1967b) showed a right ear superiority for the perception of CV syllables as a whole, but not for synthetic steady state vowels or for real speech vowels in C- context, the latter two showing no significant lateralization to either side. In a later paper, Shankweiler (1968) hypothesized that the degree of encoding is the crucial property that influences lateralization, that is, in the formant transitions in the CV sequence both the consonant information and the vowel information are encoded in parallel. Only the left hemisphere, he theorizes, can decode this parallel information.

In an effort to discover whether articulability is necessary for left hemisphere lateralization, Kimura and Folb (1968) experimented with dichotic backward speech. They found left hemisphere lateralization, but in view of the fact that the sounds, though "quite unusual and unfamiliar," seemed to resemble a Slavic language, Kimura and Folb remained unclear on whether they had showed anything about lateralization and articulation.

Using an experimental design which involved determining from which side a signal would interfere with the perception of other signals, Tsunoda (1969) found that in two cases out of three runs for Morse code operators, Morse code was dominant in the left hemisphere.

This search for the mechanism of lateralization, then, is the background for our experiments on the perception of dichotically presented Morse code signals. We hypothesize that they will be lateralized to the left hemisphere for Morse code operators, showing that articulability is not a crucial factor in lateralization but that language rather than speech is what is lateralized. We hypothesize that Morse code signals will not be lateralized to the left hemisphere for Morse code naive subjects. We want to show then that the two groups will treat the same acoustic stimuli differently, depending on whether or not the stimuli were considered to be language.

Procedure

A PDP-12 computer was used to control two electronic oscillators to produce a two-channel tape of synchronous Morse code signals. The oscillators were set at 707 Hz. and 1000 Hz. so that the pitch interval between the signals would be a diminished fifth, a dissonant musical interval. This interval minimizes any masking of one stimulus by the other due to shared overtones. Low print through tape was used to avoid echo effects.

The pitch of the signals was randomized with respect to the channels, i.e., signals of the higher pitch might appear on Channel One a few times in succession, and then signals of the lower pitch might appear on Channel One, the signals of the higher pitch then appearing on Channel Two.

All pairs of stimuli were matched in length; for example, if on one channel the signal were ·-, on the other channel it might be -· . A dot was counted as one unit long, a space as one unit, and a dash as three units. These are the ratios used in Morse code practice. However, the actual lengths of the dots and dashes as measured directly from the tape differ somewhat from the intended ratios. To measure the lengths of the dots, dashes, and spaces, the tape was dipped in "Magna See," a liquid consisting of a fine magnetic powder suspended in carbon tetrachloride. When magnetic tape is dipped in this liquid, the powder adheres to magnetized portions of the tape, thus allowing one to "see" the magnetism. We hypothesize that the discrepancies between the intended ratios and those actually produced may be accounted for by the inability of the relays to drop out as quickly as they could be pulled in. In all cases, however, stimuli began and ended simultaneously.

Three lists of stimuli were prepared.

1. All the English alphabet International Morse letters (Hereinafter ALLETTER)

2. A restricted list in which the longest stimulus was seven units long (Hereinafter RLIST)

3. A list of words (Hereinafter WORD)

Each list was prepared in a range of speeds as indicated in Figure 1, A being the slowest and F the fastest. Speed A is considerably slower than the normal Morse code sending and receiving rate; speed F is near the upper limit of most Morse code operator's capacity. There were forty test items in each list with additional items at the beginning for pretraining.

The tapes were played at 7 1/2 i.p.s. into stereo earphones from

a two channel tape recorder. The amplitudes of the channels were balanced with a Brüel and Kjœr model 2409 voltmeter reading a full track 1 KHz. tone.

| Speed number | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| Measured length on tape of dot | .80 | .58 | .48 | .43 | .38 | .27 |
| " " " dash | 2.28 | 1.67 | 1.36 | 1.18 | 1.01 | .69 |
| " " " space | .69 | .47 | .37 | .32 | .26 | .19 |
| Calculated time in seconds at tape speed of 7 1/2 i.p.s. of dot | .107 | .077 | .064 | .057 | .051 | .036 |
| " " " dash | .304 | .223 | .181 | .157 | .135 | .092 |
| " " " space | .092 | .063 | .049 | .042 | .035 | .025 |
| Calculated time of longest stimulus on ALLETTER (---) | 1.30 | .935 | .754 | .654 | .561 | .387 |
| " " RLIST (--) | .7 | .509 | .411 | .356 | .305 | .209 |

Figure 1. Measured Stimulus Length

Two groups of subjects were used. Six experienced Morse code operators were tested at speeds B, C, D, E, and five Morse code naive subjects. The Morse code operators were Morse code instructors at the U.S. Naval Training Center in San Diego.* The Morse code naive subjects were students in introductory linguistics classes at UCLA.**

---

* We thank Lt. Donald F. Hagey and Chief Bezotte, who accomodated us very helpfully, and the Morse code instructors, who were extremely diligent and attentive subjects.

** We also thank the students who voluntarily participated in this experiment.

One of the Morse code operators was left-handed and one was ambidextrous. In neither case did their scores differ in direction or significantly in degree from the scores of the other Morse code operators. All the Morse code naive subjects were right-handed.

Subjects pressed a remote start switch for the tape recorder to hear a pair of stimuli. The experimenter, who was monitoring on a pair of headphones parallel to the subject's earphones, turned off the recorder after each pair, thus subjects proceeded at their own pace, calling up each pair of stimuli when they were ready for it.

For the Morse code operators, the task was to type the letters they heard. For the Morse code naive subjects, the task was to write what they heard, with dots and dashes corresponding to short and long tones.

Stimuli were presented in quarters of ten each. The order of presentation of the quarters was randomized. Writing their responses on a form numbered from one through forty in blocks of ten, subjects were instructed to try to write (or in the case of Morse code operators, type) what they heard in both ears. Within each block of ten, they were instructed as to which ear they were to write down first. The ear of first response was alternated from block to block; earphones were switched after two blocks of ten; first ear of response and initial headphone position were randomized from subject to subject. When more than one list of stimuli was presented to subjects, the order of presentation of the lists was balanced across subjects.

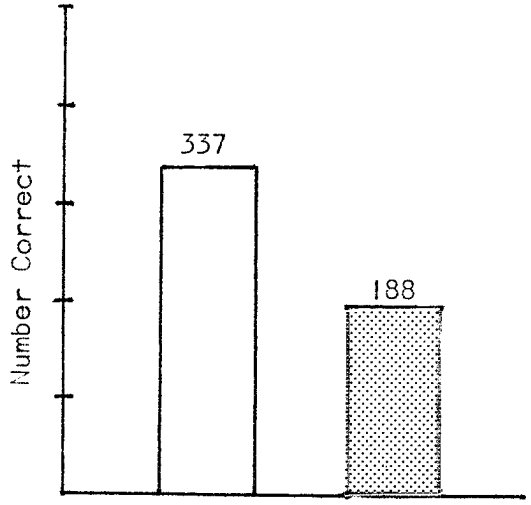
## Results

Morse code words presented monotically at a very rapid speed to Morse code operators are not significantly lateralized. But Morse code signals presented dichotically to the same group are significantly lateralized to the right-ear, left-hemisphere ($p < 0.005$) as shown in Figure 2. At speeds B, C, and E all six subjects showed a right ear superiority; at speed D four of the six subjects showed a right ear superiority. This group was not tested at speed A, because this rate was reported to be uncomfortably slow.

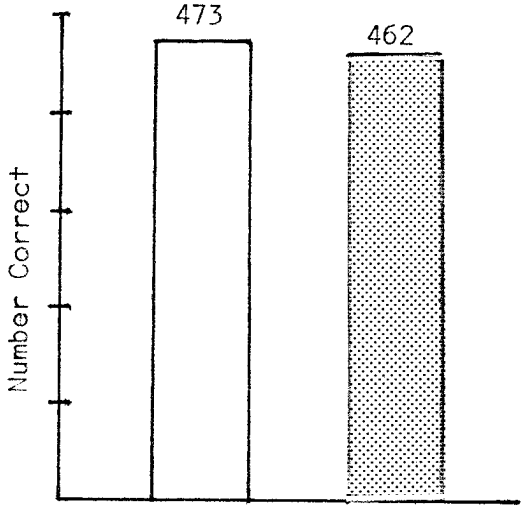
There are more letters on the left hand side of the typewriter keyboard than on the right side. Therefore, it is possible that an apparent right ear dominance could be caused by a contralateral hand-ear interaction.* However, in only 6 of 17 cases did the left-ear, right-hand score exceed the left-ear, left-hand score. In this tendency we see the

---

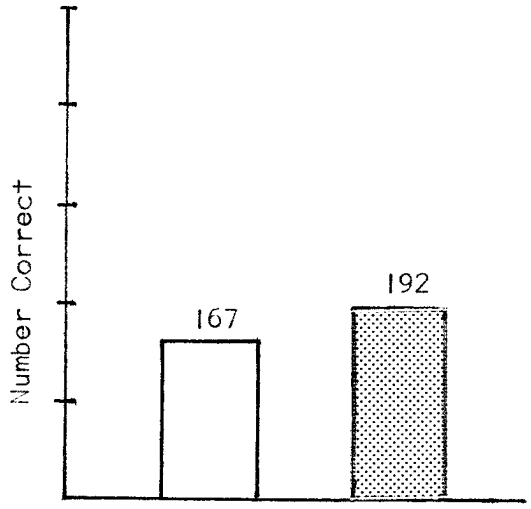
* Dr. Charles I. Berlin suggested to us that we check this possibility.



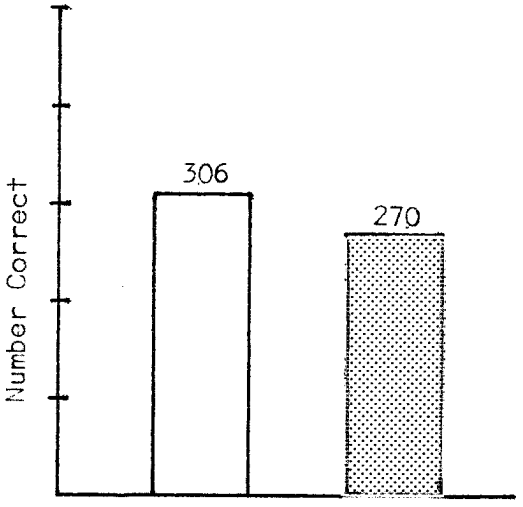

a. Morse code operators ($p < .005$)



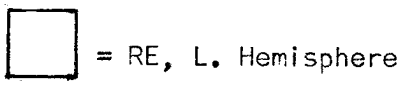

b. Naive subjects (partial scoring method, NS)




c. Naive subjects (ALLETTER, partial scoring method, NS)




d. Naive subjects (RLIST, partial scoring method, $p < .05$)



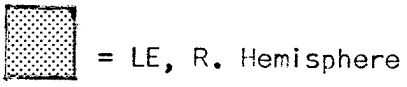




Figure 2

influence of the greater number of letters typed by the left hand, but no contralateral hand-ear interaction. What might appear to be a contralateral hand-ear interaction in the right ear can be accounted for by a combination of right ear dominance and the greater number of letters typed by the left hand. In short, there does not appear to be a contralateral hand-ear interaction.

Responses of Morse code naive subjects showed (not surprisingly) a far larger number of errors. Accordingly they were scored two ways: first by counting only completely correct responses; and then by giving partial credit. In the partial credit scoring scheme, one point each was accorded to correct first and final elements, and three points were given for a completely correct response. Thus a response could receive one, two, or three points. For all groups of subjects the partial scores were positively correlated with the right-wrong scores, but gave higher significance ratings. Therefore, we accept partial credit scores as measuring the same thing as the total right-wrong scores, but more sensitively.

Viewing the Morse code naive subjects as a whole, we do not find any significant lateralization (Figure 2b). Note, however, that the Morse code naive subjects were presented with two separate lists of stimuli -- ALLETTER and RLIST. Also the lists were presented under two different conditions: ALLETTER was played at a relatively high intensity and RLIST was played at a relatively low intensity, but both were at a comfortable loudness level. We have no reason to suppose that amplitude has any effect on lateralization (provided both channels are balanced) but we cannot yet be certain.

Morse code naive subjects could not be shown to lateralize ALLETTER to either hemisphere (Figure 2c). They tended to lateralize RLIST stimuli to the left hemisphere (scoring on a right-wrong basis, $p < 0.10$; using the partial credit scoring method, $p < 0.05$; see Figure 2d).

## Conclusions

The results from the Morse code operators support our hypothesis that articulability is not a factor in lateralization; that is, it is not the perception of *speech* which is the cause of lateralization in the left hemisphere. Were it not for the possible left hemisphere lateralization of the RLIST by the naive subjects, we might suppose that it is *language* that is lateralized in the left hemisphere -- a conclusion supported by the results of the naive group as a whole.

Our tentative results with Morse code naive subjects on RLIST raise the question of whether certain types of pattern-recognition may be the essence of language perception, and, if so, what those types may be.

Because of the surprising nature of some of our results and the small number of subjects yet involved, we feel it is necessary to continue our experimentation to see if these patterns of observations are sustained, and to test the following hypotheses:

Left hemisphere lateralization is caused by:

a. Patterns of a particular maximal length

b. Patterns of a particular maximal complexity

c. Patterns which are recognized as wholes

d. Patterns presented at relatively low volume.

References

Bakker, D.J. (1967), "Left-right differences in auditory perception of verbal and non-verbal material by children," *Quart. J. Exp. Psychol.* 19, 334-336.

Curry, F.K.W. (1967), "A comparison of left-handed and right-handed subjects on verbal and non-verbal dichotic listening tasks," *Cortex* 3, 343-352.

Kimura, D. (1961), "Cerebral dominance and the perception of verbal stimuli," *Canad. J. Psychol.* 15, 166-171.

----- (1964), "Left-right differences in the perception of melodies," *Quart. J. Exp. Psychol.* 16, 355-358.

----- (1967), "Functional asymmetry of the brain in dichotic listening," *Cortex* 3, 163-178.

Kimura, D. and Folb, S. (1968), "Neural processing of backwards-speech sounds," *Science*, 395-396.

Knox, C. and Kimura, D. (1969), "Cerebral processing of nonverbal sounds in boys and girls," *Neuropsychologia* 8, 245-250.

Milner, B. (1961), "Effects of temporal lobectomy on auditory discrimination in man." Paper read at Eastern Psychological Association, Philadelphia, April, 1961.

Rosensweig, M.R. (1951), "Representations of the two ears at the auditory cortex," *Amer. J. Physiol.* 167, 147-159.

Schulhoff, C. and Goodglass, H. (1969), "Dichotic listening, side of brain injury, and cerebral dominance," *Neuropsychologia* 7, 149-169.

Shankweiler, D. (1968), "An analysis of laterality effects in speech perception," Haskins Laboratory Status Report on Speech Research, January-June 1968, 9-26.

Shankweiler, D. and Studdert-Kennedy, M. (1967a), "Identification of consonants and vowels presented to left and right ears," *Quart. J. Exp. Psychol.* 19, 59-63.

----- (1967b), "An analysis of perceptual confusions in identification of dichotically presented CVC syllables." Paper presented at 73rd Meeting, Acoustical Society of America.

Tsunoda, T. (1969), "Contralateral shift of cerebral dominance for non-verbal sounds during speech perception," *J. Aud. Research* 3, 221-229.

Zurif, E.B. and Sait, P.E. (1969), "The role of syntax in dichotic listening," *Neuropsychologia* 8, 245-250.

*Notes on Some Recent Computer Programs*

Lloyd Rice

## I. Waveform data software

### SAMPLE

This program reads an analog input signal from one of the PDP-12 input channels, converts that signal to a digital pulse stream and stores the pulses on a PDP-12 LINCtape. The program differs from the usual data sampling routine in that a LINCtape write operation is begun as soon as 256 samples have been stored. Sampling and tape operations are then overlapped using all available memory as a circular buffer until a sample store is attempted into a location which has not yet been transferred to the LINCtape. This mode of operation allows significantly longer segments of speech to be sampled and recorded. The sample pulse rate may be specified up to approximately 12000 or 13000 pulses per second. At a pulse rate of 12000 PPS, approximately 2 seconds of data may be sampled as a unit; at a pulse rate of 8000 PPS this is increased to about 6 seconds. Sample pulses are stored in 9-bit format, one sample per 12-bit word. As presently written the program requires 12K of PDP-12 memory, however it may be easily modified for more or less memory.

language: Linc code

requires: 12K memory
1 LINCtape
A-D converter

### PLABAK

This program is the complement to the program SAMPLE, allowing reconstruction of an analog signal from LINCtape stored 9-bit digital samples. It is virtually identical to SAMPLE in organization and operation. The analog voltage is available at the vertical scope deflection output on the standard front panel connector.

language: Linc code

requires: 12K memory
1 LINCtape
Y axis output on scope

FORMANTS

This program was set up primarily for manual formant tracking on LINCtape stored digital speech data. It includes a display of the data waveform similar to that in the DEC program MAGSPY, a 512 point Fast Fourier Transform with log scale conversion, and a data output routine to write on LINCtape the spectral log amplitude of selected frequency points on the spectral display. The program requests that the user type in the data pulse rate and the frequency scale is then computed accordingly.

language: Linc code

requires: 8K memory
2 LINCtapes

MERGE

This program is used in conjunction with the following one, DUBLPLAY, to produce stimulus tapes for dichotic listening experiments. MERGE allows starting pointers to be positioned at two points on the waveforms of normal single channel sampled data. These pointers represent the starting points of the Channel 1 and Channel 2 data, respectively. The program then merges the two single channel files into one dichotic file consisting of alternating Channel 1 and Channel 2 pulses.

language: Linc code

requires: 8K memory
1 or more LINCtapes

DUBLPLAY

This program is used to reproduce dichotic analog signal pairs from a dichotic digital data file as described above. The Channel 1 signal is available at the horizontal scope deflection output and the Channel 2 signal at the vertical output. The program is structured very much like PLABAK with the difference that the effective pulse input rate from tape is twice as high for a given output pulse rate, severely limiting the time of operation. At an output rate of 12000 PPS one may play back simultaneously two signals, each of approximately 600 msec, so as to form a dichotic pair. The program is currently configured for 12K of memory.

language: Linc code

requires: 12K memory
1 LINCtape
X and Y axis outputs on scope

VARDFB

This is a short program which uses the input A-D and scope output D-A converters in real-time to produce a delayed auditory output of the analog input signal. The delay time is adjustable via a panel knob as is the degree of reverberation or percentage of the output signal added into the input. The delay time range is 0 to roughly 1/2 second.

language: Linc and PDP code

requires: 8K memory
A-D converter
Y axis output on scope

## II. Text data software

FRAGMENT

This program accepts from the user up to 8 lists of character strings and produces output consisting of items made up by concatenating one string from each list. The lists are numbered 1 through 8. The program outputs all possible combinations by taking each string from list 1 followed by each string from list 2 followed by each string from list 3, etc.

language: Linc code

requires: 4K memory
teletype

TAPRECON

This is a binary to symbolic code reconstructor which reads input from LINCtape blocks of any size up to 256 words and produces both a source text file in the DIAL working area and a listing on the teletype or line printer. Intermixed

Linc and PDP code may be reconstructed with automatic and/or user controlled mode changes.

language: Linc code

requires: 4K memory
2 LINCtapes
teletype or LP08 line printer

## III. System I/O software

DUMP

This is a Fortran callable subroutine which saves a restartable core dump of the currently running Fortran program on LINCtape. A calling argument specifies a sense switch number and the dump process occurs only if that sense switch is set. Versions of the routine are available for both the DECUS CPS and the DEC PS-8 Fortran systems.

language: Fortran and SABR code

requires: 1 page of memory in either CPS or PS-8 Fortran system
1 LINCtape

CONSIO

This is a collection of Fortran callable subroutines for performing a number of console I/O functions within the PS-8 Fortran system. The primary function of this package is a routine to access the Grafpen XY reader and return the coordinates as Fortran integer values. In addition CONSIO does the following:

a. Load a 12 bit number into the MQ register. This register is not disturbed by most operations of the Fortran system and thus serves as a console numerical readout for indices, iteration counts, etc.

b. Read a given sense switch and return with integer 0 or 1.

language: SABR code

requires: 1 page in PS-8 Fortran system
Grafpen XY reader

PLOTSUBS

This is a set of subroutines for operating the Hewlett-Packard 7035 analog XY plotter. It exists currently in Linc code callable form and future plans include a PS-8 Fortran callable version. It includes routines for pen control, plotting single points and plotting straight lines as well as calibrate mode which moves the pen to any corner of the plot surface.

language: Linc code

requires: 350 (octal) words of memory
Hewlett-Packard XY plotter

QACALL

This is a set of interface routines to allow more convenient use of the DEC question and answer subroutine QANDA. Either octal or decimal answers are packed into standard binary word form and the routine automatically performs the advance to the next answer field. Single characters may be read utilizing the same index register.

language: Linc code

requires: 117 (octal) words of memory

## IV. Mathematical subroutines

LOG1Ø

This is a fast single precision Linc code logarithm routine. Input is a 2 word (23 bit positive) fraction and output is the base 10 logarithm in 1 word with the binary point to the right of bit 3. Thus the output may vary over the range -7.776 to -0.002 (octal). A 4-term log series approximation is used so that the output is accurate to $\pm$ 0.002 octal. Execution time is approximately 300 microseconds.

language: Linc code

requires: 134 (octal) words of memory

DPDIV2, DPDIV3, DPDIV4

These are extensions to the basic LINC-8 division subroutine

DIVSUB with multiple precision arguments. DPDIV2 is essentially the same as the older LINC-8 DPDIV reassembled in PDP-12 code. These routines then provide the following combinations:

| | num | denom | quotient | remainder |
|---|---|---|---|---|
| DIVSUB | 24 bit | 12 bit | 12 bit | 24 bit |
| DPDIV2 | 24 bit | 12 bit | 24 bit | 24 bit |
| DPDIV3 | 24 bit | 24 bit | 12 bit | 24 bit |
| DPDIV4 | 24 bit | 24 bit | 24 bit | 24 bit |

language: Linc code

requires: Approximately 130 (octal) words of memory each

## V. Miscellaneous

OCTALMOD

This is a very useful program for making octal modifications on LINCtape data. The contents of a tape block are displayed in octal format (1/2 block at a time) and modifications may be typed in octal format. The modified data may be rewritten onto the same tape block.

language: Linc code

requires: 8K of memory
1 LINCtape

# *An Opinion on "Voiceprints"*

Peter Ladefoged

[The following is the text of a letter in response to a request for my opinion]

You are correct in thinking that I have been much concerned over the use of spectrograms in legal proceedings. I used to be very much opposed to their use; but in the past year various events have made me cautiously reconsider the possibility.

The first event was the publication of Oscar Tosi's report.* I have read this very carefully, and consider it to be an excellent piece of work, well designed and carried out with true scientific objectivity. It successfully removes some of my objections to the method; previously there had been no large scale experiments on the validity of "voiceprint" identification in circumstances in any way similar to those found in the majority of court cases. Tosi showed that in his experiments, when his judges had to match spectrograms made from utterances recorded a month apart, and where the words to be matched were spoken in random contexts, there were approximately 18% errors, of which 12% were failures to make a match when there was one, and 6% were misidentifications. As Tosi points out, his judges were not allowed to say that they did not know; if they had been allowed this possibility, the number of false identifications would have been less. It is worth noting that an experienced responsible investigator, Sgt. Ernest Nash of the Voice Identification Section of the Michigan State Police, stated in court that in over half the cases presented to him he decided that he was unable to give an opinion. In addition it is probably true that the error rate would fall if the investigator was given as much time as he required, and was able to make his own spectrograms of as many parts of the recordings as he wished; and he might also be helped by being able to listen to the two recordings. All these things normally happen in legal cases, and would lessen the chance of misidentification.

Tosi's experiments did not involve female speakers; and he did not investigate the success with which one person can mimic another person's voice, nor the possibility of successfully disguising the voice. All these points clearly need investigating in the future. It seems extremely probable that women's voices will be harder to identify because a

---

* Tosi, Oscar, Oyer, H., Lashbrook, W. Pedrey, C., and Nicol, Julie, *Voice Identification through Acoustic Spectrography*, Michigan State University.

higher pitched voice results in the formant pattern being less well defined. But the majority of legal cases involve men, not women; in only one case that I know of has the defense claimed that somebody was mimicking the defendant's voice; and I have never heard of the prosecution claiming that they were identifying a disguised voice. So the crucial question is: can we conclude that in a court case not involving women, mimics, or disguised voices, there is at most a 6% chance of a misidentification, and possibly, considering the additional factors normally applicable in legal cases such as the increased time available and the fact that responsible investigators will not give an opinion when they are uncertain, there is perhaps only a 3% chance of a misidentification?

Before answering this question we must note that the errors in the responses in the experiments were often (according to Tosi's statement in court) due to certain pairs of voices being highly confusable. It was not the case that each voice was equally likely to be confused with each of the other voices. The crucial question therefore becomes: what are the odds against finding a pair of confusable voices in a random sample of suspects of a crime? If Tosi's 250 male students represent a valid sample of possible suspects, then his figure of 6% (or my reduction of it to 3%) is correct. But there seems to me to be no way of knowing in advance the likelihood of coming across two confusable voices. Judging from the report in the *New York Times** supplemented by comments by Lou Gerstman, it seems fairly certain that there already has been a case of a wrong "voiceprint" identification involving two people who were both policemen, both having a similar socio-economic background, and both having a similar physique. There is clearly always a risk of this possibility occurring.

We will be able to make a better estimate of the likelihood of coming across two similar voices when someone has conducted a slightly different kind of experiment in which everybody in a given community is recorded, and the number of confusable voices found. At the moment we do not know if different communities are equally likely to contain a similar number of confusable voices. There may be a larger number of similar voices in the group of young middle-class suburbanites living in a Chicago housing tract, or in the group of militant black nationalists living in Watts, than in Tosi's 250 students, who almost certainly contained people from different backgrounds. This possibility might more than counterbalance the reduction in the error rate discussed above, so that Tosi's 6% figure should be taken as a minimum rather than a probable maximum. When we consider other possible groups of suspects, such as a small neighborhood gang, or a group of dropouts from the same high school, the degree of similarity among the voices may become even greater.

---

* *New York Times*, 27 March 1971, page 57 column 2.

If I were asked to testify on the validity of the system, I would have to emphasize that we do not at the moment know the probable error rate. But I would accept a minimum of 6% as a rough estimate of the possibility of making a misidentification (assuming, of course, that there was no question of women, mimics, or disguised voices being involved, and that the identification had been made by an experienced, responsible, investigator). Then it becomes a legal, not a scientific, matter as to whether this is "beyond a reasonable doubt" or not.

Various other events in addition to the Tosi study have been conducive to my change of opinion. My own informal experiments have shown me that there are seemingly definable characteristics of voices which are difficult to hear, but which are easily seen on spectrograms. Typical of these are the rate of transition from a voiced stop consonant into a vowel, and the duration and quality of the affrication/aspiration after a voiceless stop. These aspects of sounds seem fairly difficult to mimic.

I have also been involved in other cases, and now have much more respect for the methods used by proponents of the technique such as Sgt. Nash. I think Larry Kersta did himself a great disservice by some of his extreme statements, and his continual references to "voiceprints" being like fingerprints. Clearly "voiceprint" identification is nothing like fingerprint identification; it is much more like handwriting analysis. It might even be advisable to stop misleading judges and juries by dropping the term "voiceprints", and substituting a more neutral term such as "spectrographic analysis".

I still remain worried about some points. It seems possible that if the use of "voiceprints" becomes more common, we will have more cases in which recordings are used to incriminate people; and we all know how easy it is to edit a recording so as to completely change the meaning of what was said. I am sure I could take recordings of President Nixon's speeches and edit them into a new speech so that even he could not tell that he had not declared war on China. In addition I am alarmed over the prospects of spectrographic analysis being used by a law enforcement agency which is also investigating a suspect in other ways. An investigator is liable to be biased even by the knowledge that a pair of voices represent an unknown criminal and a suspect; he cannot help thinking that there must be some reason for considering the person to be a suspect. If the investigator is aware of other incriminating evidence it will be extremely hard for him to make a judgment in an unbiased way. I am impressed by the procedures used by the Voice Identification Section of the Michigan State Police, where every care is taken to avoid having any knowledge of a case beyond the fact that two tapes, identified only by code numbers, may or may not be recordings of the same person. I can only hope that this standard of integrity will prevail in all legal cases.

I hope that spectrographic analysis will never be used by unscrupulous, self-proclaimed, experts. Unless we get some standards of expertise established, it will be only too easy for a glib pseudo-scientist to gull a jury. But of course the fact that something could be misused is no reason to disparage its legitimate use.

Thank you for giving me this opportunity to state my opinion.