Generating vocal tract shapes from formant frequencies

Peter Ladefoged, Richard Harshman,[a] Louis Goldstein, and Lloyd Rice

Phonetics Lab, Linguistics Department, University of California, Los Angeles, California 90024

(Received 26 January 1978; revised 27 June 1978)

An algorithm that uses only the first three formant frequencies has been devised for generating vocal tract shapes as seen on midsagittal x-ray diagrams of most English vowels. The shape of the tongue is characterized in terms of the sum of two factors derived from PARAFAC analysis: a front raising component and a back raising component. Stepwise multiple regression techniques were used to show that the proportions of these two components, and of a third parameter corresponding to the distance between the lips, are highly correlated with the formant frequencies in 50 vowels. The recovery algorithm developed from these correlations was tested on a number of published sets of tracings from x-ray diagrams, and appears to be generalizable to other speakers.

PACS numbers: 43.70.Gr, 43.70.Ve, 43.70.Qa, 43.70.Bk

INTRODUCTION

Whenever a speaker produces the vowel /i/ as in "heed" the body of the tongue is always raised up towards the hard palate. Whenever anyone produces the vowel /ɑ/ as in "father" the tongue is always low and somewhat retracted. In fact, with a few exceptions, the articulatory positions used by different speakers producing a given speech sound are always very similar to one another. Because there is such a constant relationship between articulations and acoustic properties, it should be possible to generate characteristic shapes of the vocal tract from the acoustic properties of the corresponding speech sounds.

In trying to recover articulations from acoustic data, certain limits must be set on the problem. We decided to try to specify the vocal tract in terms of the positions of 18 points, equally spaced along the lower surface of the vocal tract. The position of each point would be specified in terms of its distance along a reference line originating at a fixed point on the upper surface of the vocal tract. Thus the data we are trying to recover correspond to the sagittal dimensions as seen on an x ray rather than the area functions of the vocal tract that are often used in discussions of articulatory-acoustic relations. We decided not to try to recover area functions because there would have been no way of accurately assessing our degree of success in such an endeavour.

We also decided to use no more than three formant frequencies as input data. Accordingly, we could not use techniques such as those suggested by Wakita (1974). These techniques are useful for giving a rough approximation of the shape of the vocal tract in terms of about six or eight cross-sectional areas. But they cannot be used to specify 18 points on the tongue without knowing the frequencies and bandwidths of at least nine formants. With our present techniques it seems very unlikely that anybody will be able to get reliable measures of the bandwidths of up to nine formants. Obviously, if one could use more than three formant frequencies it should be possible to get more accurate articulatory predictions. But because the acoustic structure of a vowel is fairly completely determined by the first three formant frequencies, it should also be possible to recover a plausible vocal-tract shape for a vowel from a knowledge of just these frequencies.

A major problem in the determination of vocal-tract shapes from formant frequencies is that there is an infinite set of vocal-tract-like tubes that could produce a given set of three formant frequencies. Thus a tube with the shape shown on the left in Fig. 1 will produce the same formant frequencies (at least in the range below 3 500 Hz) as the vocal-tract shape shown on the right. Of course, the tube shape shown on the left could never be replicated by a human being. For anatomical reasons the vocal tract is constrained so that it can assume only a limited range of shapes. Before attempting to recover vocal-tract shapes from formant frequencies, we must first determine the parameters that will appropriately constrain the set of shapes to those that could be produced by a normal speaker of the language being investigated.

I. PARAMETERS OF TONGUE SHAPE

We have described elsewhere (Harshman, Ladefoged, and Goldstein, 1977) a procedure for analyzing a set of x rays of tongue shapes so as to develop a simple description of all the tongue shapes in the data set in terms of a minimum number of underlying components. The procedure uses a three-way factor-analysis (or, in this

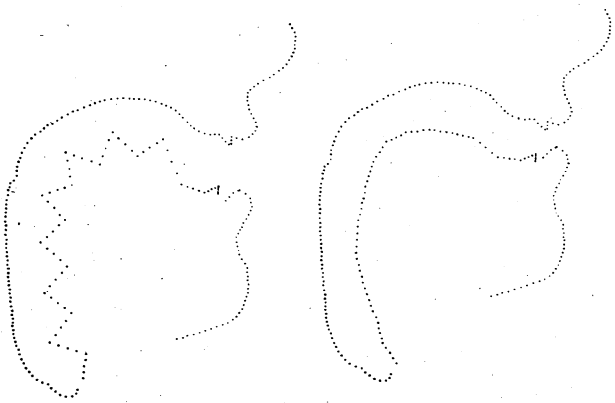

FIG. 1. Two tubes that will produce sounds with the same first three formant frequencies.

[a] Present address: Department of Psychology, University of Western Ontario, London, Canada N6A 3K7.

TABLE I. The values of the sets of constants t_1 and t_2 that can be used in determining displacements of points on the tongue. The values of n are used for determining a reference position.

Section	t_1	t_2	n (cm)
1	0.0000	0.0000	1.23
2	0.1300	0.6000	1.67
3	0.3500	1.0500	1.92
4	0.5949	1.1900	1.72
5	0.8817	1.0940	1.46
6	1.0500	0.9461	1.31
7	1.1640	0.5893	1.36
8	1.1370	0.0056	1.46
9	0.9540	−0.4594	1.50
10	0.5536	−0.9281	1.45
11	−0.1349	−1.2440	1.35
12	−0.8719	−1.4180	1.43
13	−1.2590	−1.3670	1.69
14	−1.4060	−1.2420	1.89
15	−1.3190	−0.9281	1.83
16	−0.9063	−0.4905	1.75

case, component-analysis) technique, known as PARAFAC (Harshman, 1970; compare also Carroll and Chang, 1970). The number of components underlying the data, and the characteristics of these components, as determined by PARAFAC, can be interpreted as providing an empirical estimate of the degrees of freedom of the motion of the tongue. By applying this procedure we can determine appropriate constraints that allow description of observed shapes without allowing too much freedom of shape (i.e., more dimensions than required).

The procedure used for determining this system of constraints or parameters can be briefly recapitulated as follows. High-quality audio recordings and cinefluorograms were made of five subjects saying sentences of the form "Say h(vowel)d again." The vowels in the frame included /i, ɩ, eɩ, ɛ, æ, a, ɔ, oɷ, ɷ, u/ as in "heed, hid, hayed, head, had, hod, hawed, hoed, hood, who'd." Wide-band spectrograms were made of all 50 utterances (five subjects, each saying ten test words), and an appropriate point in the vowel in the test word was marked. In the case of the vowels in "hid, head, had, hod," and "hawed," a steady-state part of the second formant was selected. For the more diphthongal vowels in "heed, hayed, hoed, hood," and "who'd," a point shortly after the first consonant was selected. The corresponding frame in the film was then located and traced.

The vocal tract was divided by a series of grid lines into 18 approximately equal sections, with section 1 being just above the glottis and section 18 being between the lips. A line was drawn through the center of each section approximately perpendicular to the surface of the tongue. A given position or shape of the tongue was then characterized in terms of the amounts that each tongue point (intersection point of the tongue surface and a grid line) had been displaced with reference to an arbitrary mean position for that point. Thus, tongue shapes were characterized as displacements of the surface of the tongue from an arbitrary reference line.

The set of 50 tongue shapes was then submitted to PARAFAC analysis. The results indicated that for sections 4–16, which constituted the main body of the tongue, two underlying components were adequate to describe the entire set of data. As shown in Harshman, Ladefoged, and Goldstein (1977), additional components probably account only for random effects in the data. The two components represented two underlying patterns of displacement which could be combined in differing amounts to produce any particular displacement corresponding to a given person's tongue shape during the utterance of a particular vowel. The measurements for each vowel could be estimated by taking an appropriate weighted sum of the two components t_1 and t_2 which can be specified in terms of the two sets of constants shown in Table I. The values of the constants for points 1–3 were extrapolated from those found by the PARAFAC analysis for points 4–16 on the assumption that point 1, which is just above the glottis, remains fixed. The set of 16 constants can be used to determine the positions of points on the tongue, by combining a weighted amount of each value from set 1 with its corresponding (and usually differently weighted) value from set 2.

For a particular person k, producing a particular vowel j, a point i has an estimated displacement $\hat{d}_{ijk}$ from its reference position as determined by the relation

$$\hat{d}_{ijk} = (t_{1i} v_{1j} s_{1k}) + (t_{2i} v_{2j} s_{2k}) \ , \qquad (1)$$

where v_{1j} and v_{2j} are the weights for the vowel j on the components t_1 and t_2, respectively, and s_{1k} and s_{2k} are scaling constants for the particular kth vocal tract with respect to components t_1 and t_2, respectively. So that actual shapes of the vocal tract may be determined, Table I also specifies the mean position of the tongue used as a reference position. The fourth column, n_i, shows the distances measured in centimeters along the grid lines between each point on the tongue in this reference position and an arbitrary, standardized upper surface of the vocal tract. The estimated distance $\hat{x}_{ijk}$ of each point on the tongue from the upper surface of the tract is then determined by the equation

$$\hat{x}_{ijk} = n_i + \hat{d}_{ijk} \ . \qquad (2)$$

Examples of some of the tongue shapes that can be produced using these equations are given in Figs. 2 and 3.

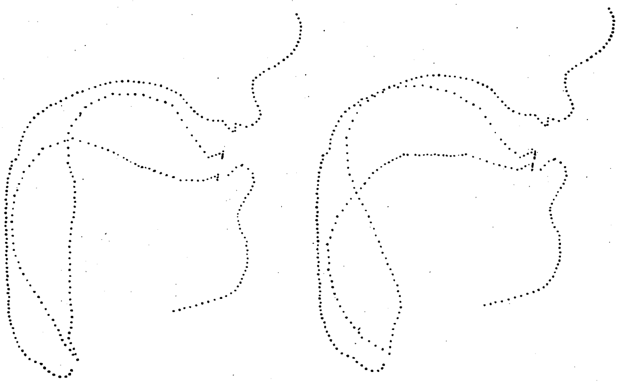

FIG. 2. The effect of varying the weighting of t_1 (left) and t_2 (right).

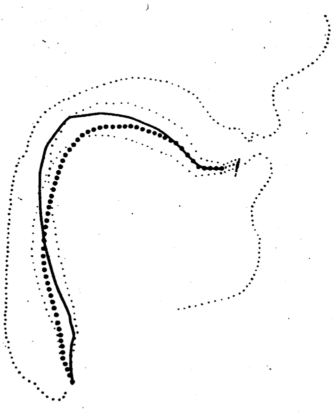

FIG. 3. Reconstruction of the vowel /u/ as in "who" (solid line). The heavy dotted line indicates the reference position for the tongue, and the two light dotted lines indicate the deviations from the reference line of the two components that sum to give the deviation for /u/. The two dotted lines cross near the uvula.

The scaling constants s_1 and s_2 are assumed to have a value of 1.0. First consider what happens when v_2 is zero, so that t_2 contributes nothing to the tongue shape. If the scaled value of v_1 is 1.0, the tongue points will all be displaced from the reference position by the amounts (in centimeters) shown in Table I for t_1. Accordingly, the vocal tract will have the raised tongue shape shown in the left-hand diagram in Fig. 2. If the scaled value of v_1 had been − 1.0, the displacement would have been in the other direction as shown by the lower tongue position in this diagram. Other vowels in which v_1 is zero (so that only t_2 is involved) are shown in the right-hand diagram in Fig. 2. The raised tongue position shows the vocal-tract shape when v_2 has a scaled value of 1.0, and the lowered tongue shows the shape when it has a scaled value of − 1.0. Comparison of the two diagrams in Fig. 2 indicates that t_1 may be thought of as a front-raising component, whereas t_2 is more of a back-raising component.

An example of the reconstruction of the tongue shape of a particular English vowel is given in Fig. 3. In this case the heavy dots represent the reference tongue position. The solid line is the best approximation for the vowel /u/ as in "who'd." The contributions of the two components are the light dotted lines deviating from the heavy dotted reference line. The solid line representing the vowel is always the arithmetic sum of the two deviations. It may be seen that in the front of the mouth the two components deviate from the reference line in opposite directions, so that the solid line representing the vowel is very close to the reference line. Near the uvula both components deviate from the reference line in the same direction, so that the solid line is considerably raised towards the roof of the mouth. In the pharynx both deviations are again in opposite directions, so that the solid line is between them. The position of the lips is estimated from other data (Fromkin, 1964), and the position of the lower teeth is the mean of the position of the lips and the point on the tongue nearest to the lips.

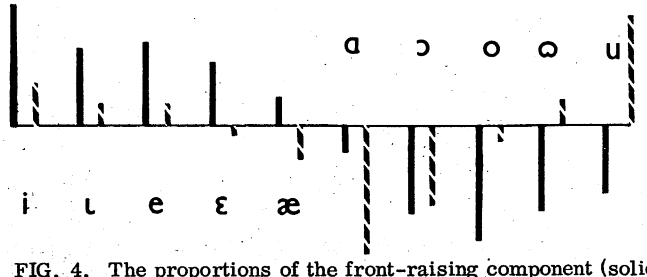

FIG. 4. The proportions of the front-raising component (solid bar) and the back-raising component (striped bar) required for ten American English vowels.

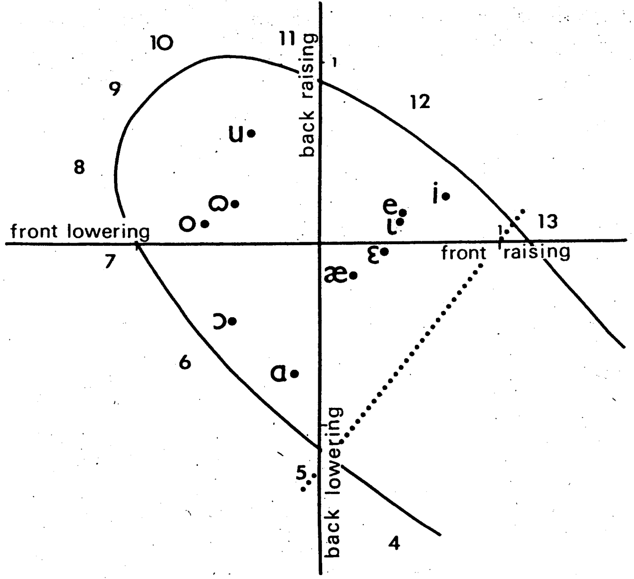

FIG. 5. The range of possible values for the two components (see text).

The two components—front raising and back raising—can be combined together to form the tongue positions that occur in all nonrhotacized English vowels. Front raising has its maximum value in the vowel /i/ and back raising in the vowel /u/. Figure 4 shows the different proportions of these components that occur in some American English vowels. The solid bar represents the proportion of the front-raising component, and the striped bar that of the back-raising component. It may be seen that all the front vowels /i, ɪ, e, ɛ, æ/ involve positive amounts of the front-raising component. The low back vowel /ɑ/ is distinguished by having a very large negative value of the back-raising component. The other back vowels /ɔ, o, ʊ, u/ have increasing amounts of this component, and fairly large negative values of the front-raising component.

The range of possible values of v_1 and v_2 for the two components is shown in Fig. 5 [assuming that the scaling constants s_1 and s_2 in Eq. (1) have a value of 1]. When the values fall outside the area bounded by the solid line, the tongue will be displaced so that there is a closure in the vocal tract at the distance from the glottis (in centimeters) indicated by the number around the periphery. When the values fall beyond the area bounded by the dotted line, the two components would combine to produce an impossible tongue shape with an inflection in the curvature of the middle of the tongue. The arrangement of the data is obviously very reminiscent of a (reversed) classical vowel diagram. The major difference is that the front vowels /i, ɪ, e, ɛ, æ/ are closer together than the back vowels /ɑ, ɔ, o, ɷ, u/. The much larger differences in tongue positions among back vowels have less auditory effect because of the counteracting effect of the differences in lip rounding among these vowels. It is also noteworthy that the vowel /e/ has a slightly higher mean tongue position for our five subjects than the vowel/ι/. The mean first-formant frequencies (500 Hz for /e/ and 480 Hz for /ι/) might lead one to expect the reverse of these tongue positions.

The similarity between the data in Fig. 5 and the classical vowel diagram might be taken to suggest that the component of tongue movement that we have called front raising might more properly be called front–back movement, since the so-called front vowels are distinguished from the so-called back vowels in terms of their values on this axis. To do this, however, would be to continue to perpetuate the traditional errors in the description of vowels. The classical vowel diagram reflects an auditory arrangement of vowels, despite the traditional labeling of the axes in terms of tongue positions. It is not really true that, for example, the highest point of the tongue is at the back of the oral cavity for [u]. If we really want to describe vowels in articulatory terms, we must learn to do so in terms of the amount of front raising and the amount of back raising involved.

II. CORRELATION BETWEEN ARTICULATORY AND ACOUSTIC PARAMETERS

The research reported above shows that the tongue shapes of the ten vowels of the five subjects could be described fairly precisely in terms of the sums of certain proportions w_1 of the front-raising component t_1, and w_2 of the back-raising component t_2. The proportions w_{1jk} and w_{2jk} for vowel j as produced by speaker k are the products $v_{1j}s_{1k}$ and $v_{2j}s_{2k}$. We may now consider the relation between these proportions and the formant frequencies of the corresponding vowels.

The formant frequencies were measured both from spectrograms and from the outputs of two different LPC spectral analyses programs, one operating on the PDP-12 in the UCLA Phonetics Lab, and the other operating in the Speech Understanding Research Lab at Systems Development Corporation, Santa Monica. The different measures of the three formant frequencies of the 50 vowels were within 50 Hz of one another on 140 (out of the total of 150) occasions. On these occasions the median frequency was considered to be correct. On the remaining occasions, either the spectrograms were impossible to interpret reliably because one of the formants (usually $F3$) was not clearly displayed, or one or both of the computerized systems produced an obvious error, such as regarding a spurious peak in the vicinity of $F0$ as a formant, failing to separate $F1$ and $F2$, or failing to find $F3$ because it had too low an amplitude. On these occasions experimenter judgment was involved in deciding which of the three measurements should be considered to be correct.

In order to find the correlation between the proportions of each of the components and the formant frequencies we conducted a series of stepwise multiple regression analyses. The dependent variable in the first series was w_1, the proportion of the front-raising component for each of the 50 vowels in the data set, and for the second series was w_2, the proportion of the back-raising component. In each series there were the following 25 independent variables: $F1$, $F2$, $F3$, their paired cross products, the triple cross product, each of $F1$, $F2$, $F3$ divided by the paired cross products of the other terms, the reciprocals of all these items, and each of $F1$, $F2$, $F3$ divided by each of the other terms.

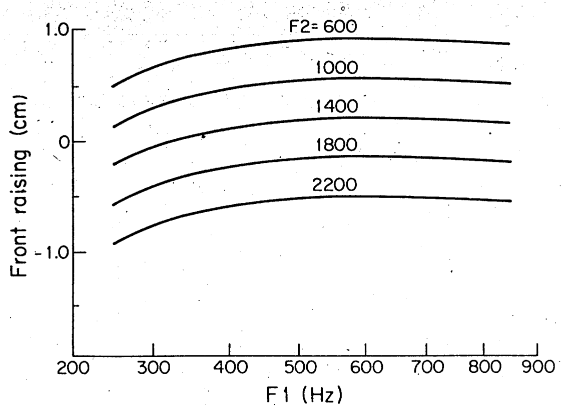

FIG. 6. The relation in Eq. (3) between front raising of the tongue and the first- and second-formant frequencies. The third formant is taken to be constant at 2600 Hz.

The best prediction of the proportion of the front-raising component using only three terms is shown in (3):

$$w_1 = c_1(F2/F3) + c_2(F1/F3) + c_3(F3/F1) + c_4 \ , \qquad (3)$$

where

$$c_1 = 2.309, \quad c_2 = 2.105, \quad c_3 = 0.117, \quad c_4 = -2.446 \ .$$

The correlation between the values of w_1 predicted from the formant frequencies and those previously determined by the PARAFAC analysis for the 50 vowels in the data set is 0.935. As we have shown above, the front-raising component corresponds to tongue movements from something like [o] to something like [i]. A graphical representation of the information in Eq. (3) is given in Fig. 6, which shows the amount of front raising associated with different values of $F1$ and $F2$, with $F3$ constant at 2600 Hz. It may be seen that the front-raising component is associated mainly with variations in the frequency of the second formant. The first formant has an effect on the determination of the front-raising component only when the frequency of this formant is comparatively low. Other graphs indicate that variations in the frequency of the third formant also have very little effect. The role of $F3$ is largely one of scaling the values of $F1$ and $F2$.

The best prediction of the proportion of the back-raising component using only three terms is

$$w_2 = c_5(F1/F2) + c_6(F2/F1) + c_7(F3/F1) + c_8 \ , \qquad (4)$$

where

$$c_5 = -1.913, \quad c_6 = -0.245, \quad c_7 = 0.188, \quad c_8 = 0.584 \ .$$

In this case the correlation is 0.902. A graphical interpretation of Eq. (4) is given in Fig. 7. It shows that increases in the proportion of the back-raising component, which corresponds to movements from something like [ɑ] to something like [u], are associated with variations in both the first- and the second-formant frequencies. When the frequency of the first formant is comparatively low, variations in the frequency of the second formant have very little effect, except when this formant has a comparatively high frequency.

In order to complete the specification of the midsagittal section of the vocal tract, the correlation between the distance between the lips and the formant frequencies was investigated in a similar way. In a third series of stepwise multiple regression analyses the distance between the lips x_{18} was used as the dependent variable with the 25 acoustic variables as the independent variables. The constants and the equations for the best prediction are

$$x_{18}=c_1F2+c_2F2F3+c_3(F1/F2)+c_4 \tag{5}$$

where

$$c_1=0.300\times10^{-3},\quad c_2=-0.343\times10^{-6},\quad c_3=4.143,$$

$$c_4=-2.865\ .$$

The correlation between the predicted distance between the lips and the observed distance for the 50 vowels in the data set was 0.78. Again, as may be seen from Fig. 8, both the first- and the second-formant frequencies are involved in determining this aspect of the position of the lips.

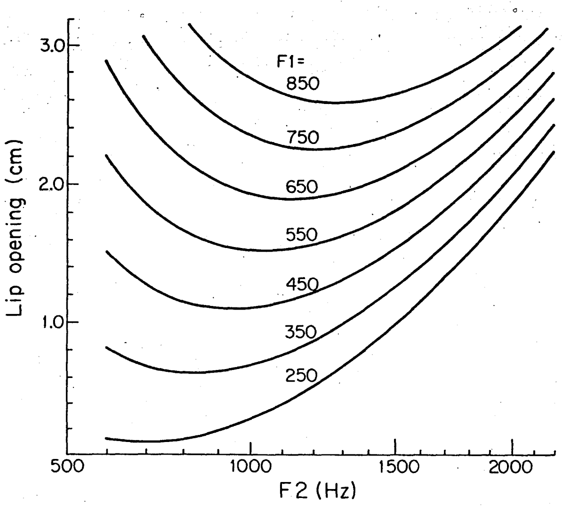

FIG. 8. The relation in Eq. (5) between lip rounding and the first- and second-formant frequencies. The third formant is taken to be constant at 2600 Hz.

The distance x_{17}, corresponding roughly to the distance between the upper and lower teeth, can be estimated from the previously determined distance as shown in (6):

$$x_{17}=\tfrac{1}{2}(x_{16}+x_{18})\ . \tag{6}$$

III. APPROPRIATENESS OF GENERATED ARTICULATIONS

The general aim of our research is to generate appropriate articulations simply from the formant frequencies. The six equations and sets of constants given above provide one possible way of doing this for a limited set of speech sounds. Assuming that a sound is an American English vowel without *r* coloring, then a sagittal diagram of a plausible vocal tract shape can be generated from three formant frequencies. No claim is made for the application of this algorithm to other kinds of sounds. It is obviously not general enough to cope with, e.g., front-rounded vowels.

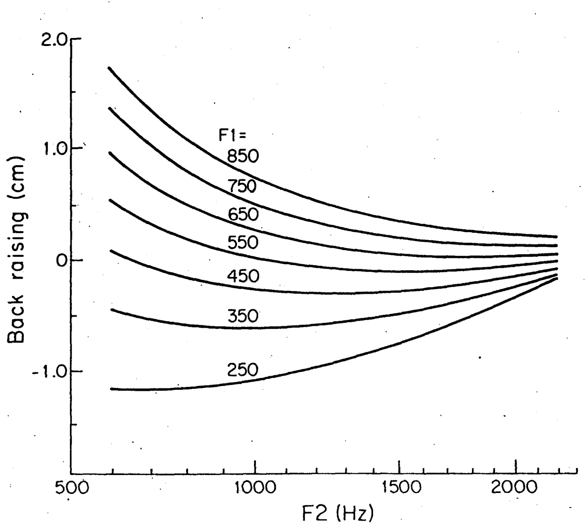

FIG. 7. The relation in Eq. (4) between back raising of the tongue and the first- and second-formant frequencies. The third formant is taken to be constant at 2600 Hz.

The appropriateness of the vocal-tract shapes can be assessed in various ways. In the first place we can consider a graphical representation of the relation between the original data and the corresponding shapes generated by the algorithm. Figure 9 shows the data for one speaker in terms of a standardized vocal tract, with a reference position as specified by the values in Table I. For this speaker many of the generated lip positions are slightly higher than in the original data. There are also a number of discrepancies in the positions of the lower part of the tongue. But, with the possible exception of the vowel [ɑ] as in "hod," the general shape of the vocal tract is similar to the original for each of these vowels.

One way to assess the appropriateness of the generated vocal-tract shapes is to quantify the accuracy of the predictions. There are two ways in which this can be done. In the first place we can consider the correlation (Pearson's *r*) between the original 18 measurements of the vocal-tract shapes of the ten vowels of the five subjects and the measurements that can be predicted from the formant frequencies. Alternatively, we can consider the rms error between the predicted and the reconstructed shapes. Table II shows both the correlation and the rms error for each of the five speakers for the body of the tongue (sections 4–16), for the whole tongue (sections 1–16), for the lips (section 18), and for the entire vocal tract (sections 1–18).

It may be seen that, as far as the tongue is concerned, the algorithm applies almost equally well to each of the five subjects. Speaker 3, who is the speaker represented

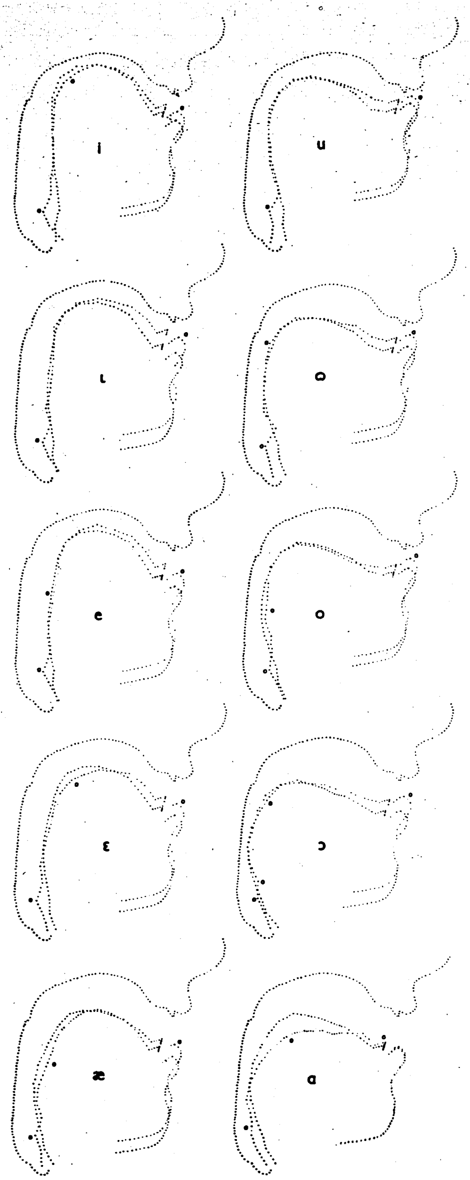

FIG. 9. The original measurements of the displacement of the lower surface of the vocal tract and the reconstructed vocal-tract shapes for speaker 3. Wherever there is a noticeable difference between the two shapes the original tract shape is indicated by a small o.

in Fig. 9, is the median speaker on two of the four measures, and is below the median on the other two.

The considerable decrease in the correlations that occurs when sections 1–3 are included may not be entirely due to a weakness in the algorithm. As noted in a previous paper (Harshman, Ladefoged, and Goldstein, 1977) measurements of the position of the tongue in the area just above the glottis are not very reliable, and a considerable part of the decrease in correlation may be due to measurement error.

The correlations and rms error for each vowel are shown in Table III. This table shows the limitations of the correlation statistic as an indicator of the degree of success in recovering articulations. With the exception of [ɑ], all the vowels have similar rms errors. But the correlations for [æ] and [ɛ] are much lower. This is because these vowels do not deviate very much from the mean vowel used as a reference, and most of the points on the tongue in each of them will be represented by comparatively small numbers. The generated vowel and the original vowel may be very similar, but there may be a poor numerical correlation between them.

The somewhat larger error for the vowel [ɑ] may be due to some inadequacy in the algorithm. It is possible, for example, that these vowels cannot be adequately described in terms of only two components of tongue shape. It is also possible that the discrepancy between the generated shapes and the originally observed shapes may be due to the fact that a given vowel may be produced with different vocal-tract shapes by different speakers. The algorithm that has been proposed may produce a plausible vocal-tract shape for the vowel [ɑ]. But since there may be many possible shapes for this vowel, the shape produced by the algorithm may not be highly correlated with the actual shape used by a particular subject.

The inability to generate correct lip positions for some vowels is probably due to the fact that the distance between the lips as seen on a sagittal x ray is not a good measure of lip opening. In considering articulatory–acoustic relations, the degree of approximation of the corners of the lips must also be taken into account. The current project is somewhat hampered by our having un-

TABLE II. The correlation (r) and the rms error (in centimeters) between the predicted and the observed vocal-tract shapes for the five speakers.

	Tongue body ($n = 130$)		Whole tongue ($n = 160$)		Lips ($n = 10$)		Whole tract ($n = 180$)	
	r	error	r	error	r	error	r	error
Speaker 1	0.8938	0.2809	0.7506	0.4254	0.9198	0.7842	0.6958	0.4563
Speaker 2	0.9452	0.3292	0.8879	0.4071	0.8002	0.4927	0.8837	0.4185
Speaker 3	0.8613	0.3081	0.8007	0.3553	0.8871	0.6394	0.7650	0.3898
Speaker 4	0.9138	0.3554	0.8350	0.3857	0.8910	0.4741	0.8276	0.3831
Speaker 5	0.9163	0.2671	0.8844	0.2919	0.9174	0.4640	0.8770	0.3053
All	($n = 650$)		($n = 800$)		($n = 50$)		($n = 900$)	
speakers	0.9020	0.3098	0.8332	0.3760	0.7775	0.5842	0.8132	0.3938

TABLE III. The correlations (r) and the rms error (in centimeters) between the predicted and the observed vocal-tract shapes for each vowel.

	Tongue Body ($n=130$)		Whole Tongue ($n=160$)		Lips ($n=10$)		Whole tract ($n=180$)	
	r	error	r	error	r	error	r	error
i	0.9418	0.3086	0.9018	0.3459	0.6898	0.6238	0.8773	0.3660
ι	0.8765	0.2850	0.7497	0.3495	0.7722	0.8140	0.6393	0.4044
e	0.8949	0.2845	0.7735	0.3533	0.6384	0.5016	0.7558	0.3669
ε	0.6721	0.2944	0.4790	0.3775	0.5756	0.4985	0.4887	0.3858
æ	0.5421	0.3321	0.5051	0.3892	0.2509	0.6878	0.6103	0.4213
ɑ	0.9013	0.4387	0.8627	0.4273	−0.3042	0.7164	0.8397	0.4869
ɔ	0.9516	0.3160	0.9189	0.3834	0.5883	0.4701	0.9078	0.3901
o	0.9365	0.2744	0.8748	0.3517	0.5645	0.5979	0.8538	0.3773
ω	0.9333	0.2113	0.8219	0.3426	0.6844	0.4267	0.8131	0.3469
u	0.8675	0.3048	0.7782	0.3767	−0.4052	0.3386	0.7912	0.3723

dertaken the task of trying to generate the equivalent of an x-ray diagram in which the vocal-tract shape is characterized in terms of 18 midline sagittal dimensions. The midline sagittal dimensions at the lips are very similar for [i] and [u], although the lip positions of these two vowels are very different.

The high correlations between the observed and the generated vocal-tract shapes for these five subjects may be inflated somewhat by the fact that the algorithm was devised by analyzing data from just these subjects. We must also consider the accuracy of the predictions when the algorithms are applied to other speakers. It is most appropriate to do this by reference to data that has already been published. The diagrams on the left of each pair in Fig. 10 show the shapes of the vocal tract in the four vowels illustrated in a paper by Ladefoged (1975). Here, as in the original article, the center of the tongue is shown for each of the vowels, except for [ɑ] for which the outlines of both sides of the tongue are shown. The diagrams on the right show the corresponding generated vocal-tract shapes represented in terms of the same standarized vocal tract that was used for the data in Fig. 9. The overall resemblance between the members of each pair is good, though in each case the region in the front of the mouth is not generated very precisely. But the position of the body of the tongue is fairly accurately represented, and there is an appropriate location of the major constriction within the vocal tract. The similarity of the two sets of vocal-tract shapes is a good indication that the algorithm has some validity when applied to at least one speaker not included in the data base used in determining the procedure.

Further validation of the algorithm is provided by the results shown in Fig. 11, in which a set of generated vocal-tract shapes can be compared with those published by Perkell (1969). As in Fig. 10, the diagrams on the left of each pair are based on x-ray tracings, and those on the right are the vocal-tract shapes calculated from the formant frequencies. Perkell does not give the formant frequencies corresponding to his x-ray tracings, but he was kind enough to provide us with copies of the original spectrograms so that we could measure them for ourselves. As is often the case, it was difficult to determine the centers of all three formants from the spectrograms of some of the back vowels. The formant frequencies that we could determine were very similar to those previously published for this subject (Stevens and House, 1963). Accordingly we decided to use these more reliable

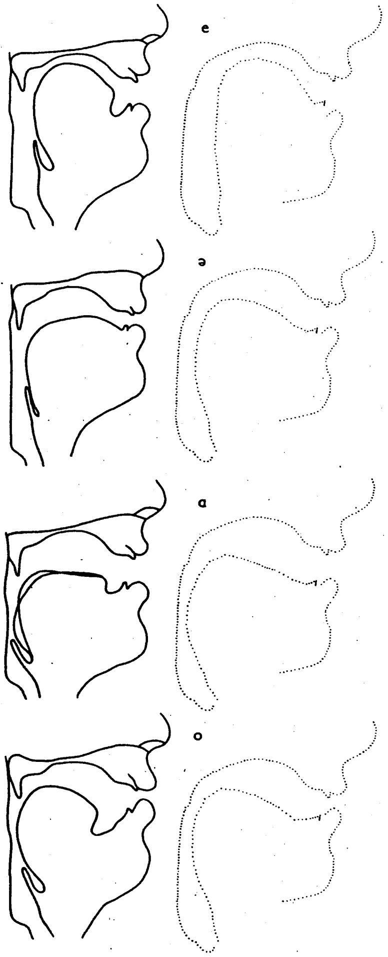

FIG. 10. Vocal-tract shapes during the pronunciation of four vowels as spoken by a speaker of British English on the left and the shapes recovered from the corresponding formant frequencies on the right.

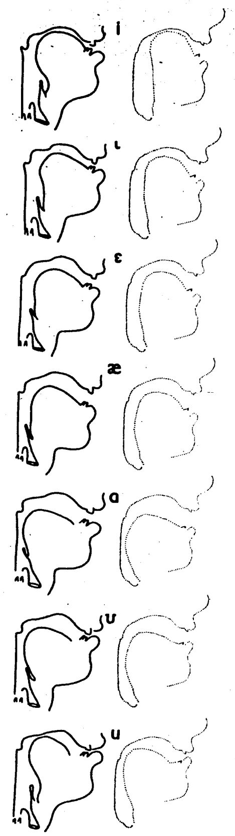

FIG. 11. Vocal-tract shapes during the pronunciation of seven vowels as spoken by another speaker of American English on the left and the shapes recovered from the corresponding formant frequencies on the right.

measurements for generating all the vocal-tract shapes of the subject shown in Fig. 11.

The same general comments apply to the data in Fig. 11 as were made about the data in Fig. 10. The modeling of the front cavity needs improvement, but for most of the vowels the overall vocal-tract shape is substantially correct. The main differences are in the lip positions. Some of the differences may be due to the fact that the reconstructed shapes are specified within a standardized head shape that may not have the same proportions as the original speaker. But it is also possible that the differences occur because different vocal tract shapes can produce the same formant frequencies. All these shapes are plausible shapes for these American English vowels. They were not exactly the shapes used by that speaker, but they might well have been used by some other speaker.

As a final test of the algorithm, we tried to generate the vocal tract shapes of a speaker of Russian. Using the well-known sets of x-ray tracings and formant frequencies published by Fant (1960) we produced the data shown in Fig. 12. Here again there is a close resemblance between the original and the generated data. There is a discrepancy in the case of the vowel [i], where the algorithm determines that the tongue must be pushed through the roof of the mouth in order to produce the given formant frequencies. Fant's value for the frequency of the third formant of this vowel is 3200 Hz. This is 600 Hz more than its value for any other vowel produced by that subject, and well beyond the range of our subjects (2600–2900) for the corresponding English vowel. The worst match is for the vowel [ɨ] where the front cavity is not generated.

The body of the tongue is in an appropriate position for each of the remaining vowels, and, with the possible exception of the non-English vowel [ɨ], the place of the maximum constriction is also correct. Once more we should note that the reference vocal tract used in the computer display is rather different from that of the original speaker. The application of our algorithm is something

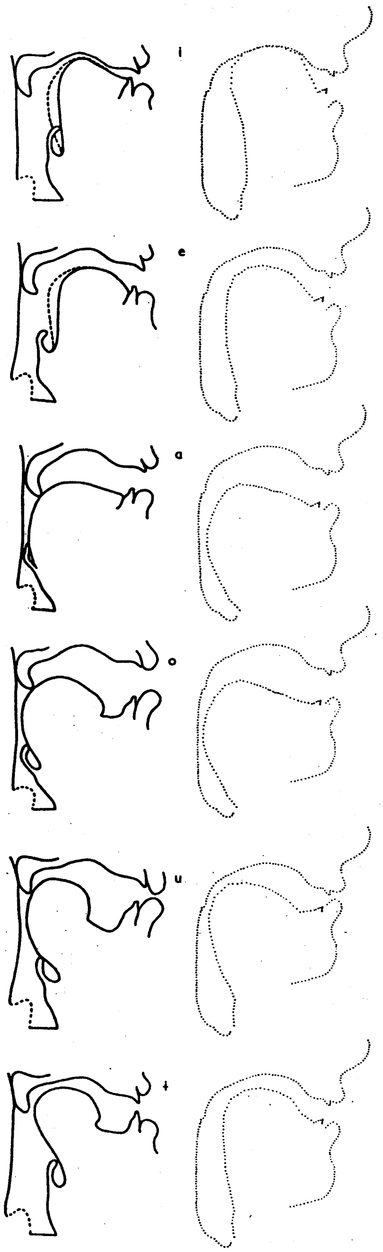

FIG. 12. Vocal tract during the pronunciation of six vowels as spoken by a speaker of Russian on the left and the shapes recovered from the corresponding formant frequencies on the right.

like putting one person's tongue in another person's mouth. It is not surprising that this may occasionally produce some deviant results.

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

Fant, C. G. M. (1960). *Acoustic Theory of Speech Production* (Mouton, The Hague).

Fromkin, V. (1964). "Lip positions in American English vowels," Lang. Speech **7**, 215–225.

Harshman, R. (1970). "Foundations of the PARAFAC procedure: Models and procedures for an 'explanatory' multi-modal factor analysis," UCLA Working Papers in Phonetics 16, University Microfilms No. 10085.

Harshman, R., Ladefoged, P., and Goldstein, L. (1977). "Factor analysis of tongue shapes," J. Acoust. Soc. Am. **62**, 693–707.

Ladefoged, P. (1975). "The specification of the languages of the world," Paper presented at the 8th Int. Congr. Phon. Sci. Appears in UCLA Working Papers in Phonetics 31, 3–21 (1976).

Perkell, J. (1969). *Physiology of Speech Production* (MIT Press, Cambridge, MA).

Stevens, K. N., and House, A. S. (1963). "Perturbation of vowel articulations by consonantal context: an acoustical study," J. Speech Hear. Res. **6**, 111–128.

Wakita, H. (1974). "A theory of linear prediction–acoustic tube method for estimating vocal tract area functions," Speech Commun. Res. Lab., Santa Barbara, CA.

Erratum: "Generating vocal tract shapes from formant frequencies"

[J. Acoust. Soc. Am. 64, 1027-1035 (1978)]

Peter Ladefoged, Richard Harshman, Louis Goldstein and Lloyd Rice

Phonetics Lab, Linguistics Department, University of California, Los Angeles, California 90024.

PACS numbers: 43.70.Gr, 43.70.Ve, 43.70.Qa, 43.70.BX

Correction: The scales on the ordinates in Fig. 6 on page 1030, and Fig. 7 page 1031 are inverted. In both cases the negative values should be at the top of the graph. In addition, the constant C_1 in (5) on page 1031 should be 0.300 X 10^{-2}, not 0.300 X 10^{-3}.