A STATIC MODEL OF THE TONGUE  
DURING VOWEL ARTICULATION

by

JAMES TALBOT WILLIAMS

B.S., Stanford University  
(1964)

SUBMITTED IN PARTIAL FULFILLMENT OF THE 
REQUIREMENTS FOR THE DEGREE OF  
MASTER OF SCIENCE  
at the  
MASSACHUSETTS INSTITUTE OF TECHNOLOGY  
August, 1965

Signature of Author .................................................................
Department of Electrical Engineering, August 23, 1965

Certified by ................................................................. 
Thesis Supervisor

Accepted by ................................................................. 
Chairman, Departmental Committee of Graduate Students
A STATIC MODEL OF THE TONGUE
DURING VOWEL ARTICULATION

by

JAMES TALBOT WILLIAMS

Submitted to the Department of Electrical Engineering on August 23, 1965 in partial fulfillment of the requirements for the degree of Master of Science.

ABSTRACT

The purpose of this thesis is to present a simplified model of the tongue during vowel articulation. It is a static model in that it considers the tongue only at the midpoints of vowels. The features of interest are assumed to be its shape and position relative to the other articulatory organs. The form of the data used is cineradiographic films of two different speakers. For the purposes of this model, therefore, the tongue is assumed to be specified by its shape in the midsaggital plane.

The model arrived at defines, for a single speaker, a constant shape which his tongue maintains during the production of all vowels. This shape is allowed to translate and rotate to different positions in the vocal tract to produce the various vowels. Certain deviations from the vowel shape exist. The position of the tip is dependent upon the location of the mandible. However, the tip is found to have little effect on the acoustic form of the vowel and great latitude is permitted in its specification. The shape of the tongue in the vicinity of the lower pharynx is seen to be quite variable and these deviations significantly affect the acoustic output. Its variations are, however, systematic and may be predicted at least qualitatively from the location of the tongue body. We thus arrive at a static model of the tongue during the production of vowels for one speaker. This model is applied in a limited way to another speaker in another language and is found to be valid as far as it is tested. A short discussion of the implications of this static model to the dynamic workings of the vocal mechanism is presented. In conclusion, the work done here is related to the models some others have proposed.

Thesis Supervisor: Kenneth N. Stevens
Title: Professor of Electrical Engineering
Acknowledgement

The author wishes to thank the Speech Communication Group of the Research Laboratory of Electronics at the Massachusetts Institute of Technology. Their aid and equipment made this work possible. In particular, he wishes to express his gratitude to Professor Kenneth N. Stevens for his supervision, suggestions, and ideas. Professor John M. Heinz and William L. Henke were of great help in the study reported in chapter one. The latter also provided encouragement throughout the preparation of this report.

The work in chapter one was supported in part by Project MAC, an M. I. T. research program sponsored by the Advanced Research Projects Agency, Department of Defense, under the Office of Naval Research, Contract Number Nonr-4102(01). The entire study was supported in part by the National Institutes of Health (Grant NB-04332-03) and by the U. S. Air Force Cambridge Research Laboratories, Office of Aerospace Research, Contract AF19(628)-3325.
# Table of Contents

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>6</td>
</tr>
<tr>
<td>Chapter One - Acoustic Analysis</td>
<td>11</td>
</tr>
<tr>
<td>Chapter Two - Formulation of the Model</td>
<td></td>
</tr>
<tr>
<td>Section A - Discussion of the Data</td>
<td>24</td>
</tr>
<tr>
<td>Section B - Presentation of the Model</td>
<td>28</td>
</tr>
<tr>
<td>Section C - Comparison of Vowel and Consonant Shapes</td>
<td>38</td>
</tr>
<tr>
<td>Chapter Three - A Test of the Model for a Second Speaker</td>
<td>41</td>
</tr>
<tr>
<td>Chapter Four - A Study of the Validity of the Model During a Diphthong</td>
<td>48</td>
</tr>
<tr>
<td>Conclusion</td>
<td>53</td>
</tr>
</tbody>
</table>
List of Figures and Tables

Fig. 1. - Formant Frequency Changes From Vocal Tract Perturbations. 14
Fig. 2. - Tongue Perturbations. 16
Fig. 3. - Formant Frequency Changes From Tongue Perturbations. 17
Fig. 4. - Tip Perturbations. 19
Fig. 5. - Base Perturbations. 23
Fig. 6. - Relative Tongue Locations and Shapes for Seven Vowels. 26
Fig. 7. - Overlayed Tongue Shapes for Seven Vowels. 30
Fig. 8. - Vowel Tongue Body Shape and Base Central Position. 32
Fig. 9. - Relative Tongue Translation and Rotation Data. 33
Fig. 10. - Mandible Location Data. 35
Fig. 11. - Relative Tongue Locations and Shapes for Seven Consonants. 39
Fig. 12. - Relative Tongue Locations and Shapes for Eight Vowel Samples from a Second Speaker. 42
Fig. 13. - Overlayed Tongue Shapes for Eight Vowel Samples from a Second Speaker. 43
Fig. 14. - Vowel Tongue Body Shape and Base Central Position for a Second Speaker. 44
Fig. 15. - Relative Tongue Translation and Rotation Data for a Second Speaker. 45
Fig. 16. - Relative Tongue Locations and Shapes During a Diphthong. 49
Fig. 17. - Relative Tongue Translation and Rotation Data and Mandible Rotation Data During a Diphthong. 50
Table 1 - Formant Frequency Data for Tongue Tip and Base Perturbations. 20
Introduction

When an engineer attempts to analyze a complex system, one of his first jobs is to simplify the problem into one with which he is able to deal. This often requires that some dubious assumptions be made and many details neglected, but it gives a point of departure—a theory into which more of the system complexities may be incorporated as the work progresses. Just such a procedure is helpful in developing an understanding of the human vocal mechanism.

Man's speech production apparatus consists of a number of moving parts including the larynx, epiglottis, tongue, jaw, velum, and lips. The movement of any one part is not necessarily independent of the others and, to further confuse our understanding of this system, deficiencies in one part can be compensated for by actions of another. Furthermore, there are many variations in the articulatory structures of different people. The result is a system of great complexity and one about which we have relatively little knowledge.

Perhaps our lack of understanding would alone be sufficient to justify an interest in a simple model of the tongue during vowel production. It may be of use in ordering some of our knowledge about the articulatory system and provide something to which further work might be related. However, such a model could also have some value in itself. Two important examples are the design and construction of speech synthesizers and recognizers that make use of
articulatory parameters. For such purposes the tongue is particularly important. As Ladefoged states, "... the tongue has a different position for at least sixteen sounds. From the point of view of signalling power, it is two or three times more important than the other parts of the speech apparatus. Yet we have very little real knowledge about the methods of producing the different gestures of the tongue."

There are many indications that a model may be constructed making use of only a few parameters to specify the position and shape of the tongue. Stevens and House show that over the length of the tongue an approximation to the vocal tract area function can be defined through knowledge of two parameters, the location of the point of maximum constriction in the tract and its degree. They then proceed to develop an idealized vocal tract configuration in terms of these two measures and one other, an approximation to the acoustic impedance of the portion of the vocal tract between the tip of the tongue and the lips. This is given by the ratio of its average cross sectional area to its length. By using this model on a vocal tract analog, they have shown that it produces acceptable sounding vowels. Further, Heinz and Stevens have found that, at least for the single case they studied, a twenty-two percent variation in successive two cm lengths of the vocal tract area functions they have calculated from articulatory data causes at most an error of four percent
in the frequencies of the first, second, and third formants. While this is far from conclusive, it suggests that small errors generated in our approximation may not have an overly large effect on what would be the acoustic output of the model. This will be discussed in more detail in the following chapter.

The greatest encouragement for the development of such a model, however, comes from an observation made by Mr. Joseph Perkell. In working with what was to become the primary data source for the study, he observed that for the one speaker being considered, the posterior two-thirds of the tongue had a somewhat constant shape during the temporal midpoint of all the vowels observed. This led him to hypothesize that for vowels the tongue might be completely characterized by simply specifying the position of this constant shape and another parameter which would locate the tip. It was this discovery that formed the basis for the model.

In this and the following discussions it must be remembered that the system being dealt with is subject to many variations. The way in which a speaker articulates a phoneme is influenced by many things. The most obvious of these are the effects of context or coarticulation. If for no other reason than that the articulatory system is a physical one composed of finite and nonzero masses, forces, and time intervals, two phonemes spoken in rapid succession are bound to influence each other. Other factors, some
perhaps more psychological than physical, will undoubtedly affect the way in which a particular sound is made at different times.

Some idea of the variations to be expected within this system may be had by looking at studies of vowel formant frequencies. Stevens and House,8 Peterson and Barney,9 Potter and Steinberg,10 and Lindblom11 all have relevant information. Potter and Steinberg is probably the most interesting for this purpose. They present some information concerning the changes in fundamental and formant frequencies between different utterances of the same vowel by a trained phonetician. Some of their results will be given in the next chapter.

The model to be discussed here has a number of weaknesses. It was developed primarily for only one speaker since he is the only one for whom sufficient data was readily available. It is a static model and only a few conjectures can be made concerning its significance on the dynamic process of actually forming the articulatory configurations. Because of a lack of sufficient data and time, only a limited number of vowels were actually treated. To get the statistical data one would desire, the model should be tested for many vowels for the same speaker and for many more speakers.

The data used was in the form of cineradiographic films showing the vocal tract of a speaker while he was articulating certain utterances. The methods for making
these films and bringing the information they contain into useful form are reported in detail elsewhere.\textsuperscript{12,13} Suffice it to say that the films are made and the utterances recorded on tape with provisions for frame by frame audio-visual synchronization. Spectrograms of what was said may then be made and a correspondance between the articulatory configurations shown on the x-rays and the acoustic output may then be generated.

To make quantitative measures on the appropriate frames it is then usually best to make line tracings of the pictures and make the measurements from these. This is required because of the lack of contrast in the x-rays themselves. Studies using this method have been carried out and are reported elsewhere.\textsuperscript{7,14}

With these procedures and some of their limitations in mind, let us proceed with the development of the model itself. This will be done in essentially two steps. First, an acoustic study of some of the approximations to be made will be presented. Second, the mechanics of the model itself will be given. In conclusion, some of its implications will be discussed.
Chapter One

Acoustic Analysis

The purpose of this acoustic analysis is to discover which parts of the tongue are important in vowel production. Its relevance is predicated on the belief that the factor of interest in any model of the vocal system is its acoustic output. It is this output which actually contains the information that is processed by any receiving device. This study should then make it possible to construct the model in a more intelligent way.

For vowels, the acoustic features of primary concern are the center frequencies of the first three formants. Here we shall be interested mainly in the second or $F_2$. There are a number of reasons for this. (1) Intelligibility experiments in which speech signals were passed through filters of constant bandwidth and variable center frequency have shown that the midrange frequencies carry the most information. (2) The human ear is apparently most sensitive to frequencies in the approximate area of $F_2$. (3) $F_3$ characteristically has a much higher frequency, and therefore a much shorter wavelength, than $F_2$. Its center frequency is then much more sensitive to minor variations in the vocal tract than is that of $F_2$. $F_3$ is, therefore, more a characteristic of individuals than the particular phone they are articulating. (4) $F_1$ has a much smaller range than $F_2$, 200-1200 cps as compared to 500-3500 cps. (5) Finally, $F_1$ transitions into consonants are always upward in frequency.
while those of $F_2$ may be either up or down. The last
two reasons mean that $F_2$ is able to carry a greater informa-
tion load than $F_1$.\textsuperscript{15}

While none of these arguments is in itself conclusive, taken
together they give a strong indication that $F_2$ is indeed the more important of the three parameters.

The basis for this acoustic study is the work done by
Heinz and Stevens on the conversion of x-ray data in the
midsaggital plane to vocal tract area functions. This is
described elsewhere and the details will not be presented
here.\textsuperscript{6} The speaker and data that Heinz and Stevens used
are the same as those used for this study so their work is
directly applicable.

For reasons of convenience, the computer program used
to calculate formant frequencies from the vocal tract area
function was not that of Heinz and Stevens but rather one
written by Henke.\textsuperscript{16} The results it yields are essentially
the same as those of the former program. It approximates
the vocal tract by a series of cylinders, in this case each
one-half cm long, and calculates the complex transfer
function of the tube using the acoustic impedance of each
section and various other loss and load factors. From the
transfer function it then locates the transmission poles.
Thus to see the acoustic change brought about by a change
in the tract configuration, one must merely vary the cross-
sectional area of the appropriate cylinders.

To determine which areas of the vocal tract were
particularly critical in the determination of formant frequencies, a particular vowel, /I/, was chosen and the vocal tract approximation systematically perturbed in successive two centimeter lengths starting at the tongue tip and working back toward the glottis. (This is essentially a repeat of some of the work reported by Heinz and Stevens.) The results are shown in Fig. 1. The abscissa represents the position along the vocal tract of the center of the perturbed section and the ordinate the change in the formant center frequency away from its normal or unperturbed value. The curves labeled "2" indicate that here the change consisted of doubling the cross-sectional area. Similarly, in those marked "1/2," the cross-sectional area was halved in the indicated cylinders.

Figure 1 contains some very revealing information. For each of the formants there are portions of the tract in which substantial changes in the area function have only minor effects on the center frequency. (This effect is predictable from the acoustic theory of resonating tubes. Velocity node points, or pressure anti-node points, give the locations at which the formant frequencies are least sensitive to changes in the cross-sectional area of the tube.) The existence of such areas means, perhaps, that certain parts of the tongue may be very loosely specified in the model. Caution must be exercised in drawing this conclusion, however. The functional relationships between cross dimension and vocal tract area are quite different in
Figure 1. - Formant Frequency Changes From Vocal Tract Perturbations. Horizontal scale represents distance along vocal tract from glottis in cm. Vertical scale represents change in the indicated formant frequency from unperturbed value in cps.
the various parts of the tract. Therefore, similar displacements of the tongue over different segments of its length will have different effects on the cross-sectional area.

To obtain a better idea of the effect of variations in the position of different parts of the tongue, another study was made. This time four sections of the tongue were successively displaced by the same distance. See Fig 2. (In this and all succeeding tracings, the scale is 1.4 times life-size. All lengths stated in this paper are life-size unless otherwise stated.) For each trial, one part of the tongue was displaced either up or down to the dotted line—a distance of 0.36 cm. All four segments are of approximately equal length, 3 cm, on the dorsal surface of the tongue. The results are shown in Fig. 3. The vertical coordinates represent change in formant frequency while the horizontal scales indicate which section was displaced.

Taking particular note of $F_2$ for the reasons discussed above, these results seem to indicate that the tip of the tongue has a negligible influence of the acoustic characterization of the vowels. Its opposite member, the base, seems to play an important role. Since it was already known that these two areas, the tip and the base, would be the hardest to specify in a systematic way, more work was done with them.

For vowels, it is known that the tongue tip is always low and is often in contact with the maxilla. For all the vowels available for the speaker considered, the tracings
Figure 2.-Tongue Perturbations. Heavy lines indicate maxilla and mandible overlays.
Figure 3.—Formant Frequency Changes for Tongue Perturbations. Horizontal scale indicates perturbed section. Vertical scale represents change in indicated formant frequency from unperturbed value in cps.
were overlayed so that the maxillae were aligned. This process defined a range of allowable positions for the tip with respect to the maxilla.

Three vowels were then selected for study, /i/, /I/, and /a/. The first two are high front and the last is a low back vowel. /I/ is lax while /i/ and /a/ are tense. For each the tip of the tongue was allowed to take on four equally spaced positions over the range of those allowable. See, for example, Fig. 4. The length of the displaced section was in each case approximately 2 cm. The formant frequencies for each variation were then found. The results are given in Table 1 along with \( \Delta F_t \), the individual formant frequency range due to changes of the tip. The underlined letters indicate the unperturbed position for the tongue tip.

To see what these figures mean, let us take a brief look at two studies from the literature. Potter and Steinberg\(^{10}\) report a range of formant frequencies for the same speaker, a phonetician, for a number of successive repetitions of the same vowel. These are given in Table 1 in the rows marked P. and S. r. Stevens and House\(^{8}\) give some figures of interest. These are the standard deviation of the first two formant frequencies for a particular vowel. Their sampling, however, is taken across a range of different contextual environments. They are relevant here because the values shown happen to be for the same speaker as was used in this work. The values are given in Table 1 as S. and H.\(\sigma\)\(^{5}\).
Figure 4: Tip Perturbations. Example shown is for /I/.
<table>
<thead>
<tr>
<th>/i/</th>
<th>F₁</th>
<th>F₂</th>
<th>F₃</th>
<th>/i/</th>
<th>F₁</th>
<th>F₂</th>
<th>F₃</th>
<th>/a/</th>
<th>F₁</th>
<th>F₂</th>
<th>F₃</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>473</td>
<td>1807</td>
<td>2590</td>
<td>b</td>
<td>297</td>
<td>2039</td>
<td>3049</td>
<td>c</td>
<td>639</td>
<td>1299</td>
<td>2818</td>
</tr>
<tr>
<td>b</td>
<td>485</td>
<td>1820</td>
<td>2593</td>
<td>d</td>
<td>324</td>
<td>2069</td>
<td>3140</td>
<td>d</td>
<td>647</td>
<td>1301</td>
<td>2805</td>
</tr>
<tr>
<td>c</td>
<td>496</td>
<td>1821</td>
<td>2556</td>
<td></td>
<td>335</td>
<td>2081</td>
<td>3175</td>
<td></td>
<td>657</td>
<td>1303</td>
<td>2773</td>
</tr>
<tr>
<td>d</td>
<td>507</td>
<td>1814</td>
<td>2487</td>
<td></td>
<td>344</td>
<td>2090</td>
<td>3197</td>
<td></td>
<td>667</td>
<td>1301</td>
<td>2728</td>
</tr>
<tr>
<td>ΔF₁</td>
<td>34</td>
<td>14</td>
<td>106</td>
<td>ΔF₁</td>
<td>47</td>
<td>51</td>
<td>148</td>
<td>ΔF₁</td>
<td>28</td>
<td>4</td>
<td>90</td>
</tr>
<tr>
<td>P+Sr</td>
<td>100</td>
<td>250</td>
<td>250</td>
<td>P+Sr</td>
<td>50</td>
<td>150</td>
<td>450</td>
<td>P+Sr</td>
<td>100</td>
<td>100</td>
<td>250</td>
</tr>
<tr>
<td>S+Ho</td>
<td>24</td>
<td>105</td>
<td>---</td>
<td>S+Ho</td>
<td>25</td>
<td>40</td>
<td>---</td>
<td>S+Ho</td>
<td>34</td>
<td>54</td>
<td>---</td>
</tr>
<tr>
<td>ΔF₁</td>
<td>43</td>
<td>109</td>
<td>61</td>
<td>ΔF₁</td>
<td>32</td>
<td>158</td>
<td>24</td>
<td>ΔF₁</td>
<td>38</td>
<td>131</td>
<td>184</td>
</tr>
<tr>
<td>A</td>
<td>489</td>
<td>1810</td>
<td>2562</td>
<td>A</td>
<td>305</td>
<td>2008</td>
<td>3066</td>
<td>A</td>
<td>687</td>
<td>1198</td>
<td>2574</td>
</tr>
<tr>
<td>B</td>
<td>470</td>
<td>1868</td>
<td>2595</td>
<td>B</td>
<td>293</td>
<td>2071</td>
<td>3079</td>
<td>B</td>
<td>669</td>
<td>1295</td>
<td>2693</td>
</tr>
<tr>
<td>C</td>
<td>446</td>
<td>1919</td>
<td>2623</td>
<td>C</td>
<td>273</td>
<td>2166</td>
<td>3090</td>
<td>C</td>
<td>649</td>
<td>1329</td>
<td>2758</td>
</tr>
</tbody>
</table>

Table 1.-Formant Frequency Data for Tongue Tip and Base Perturbations. Lower case letters indicate tip variations. Upper case letters indicate base variations. Underlined letters indicate the unperturbed or the approximation to the unperturbed case. ΔF₁ and ΔF₂ are the ranges of the individual formant frequencies due to tip and base variations respectively. P+Sr and S+Ho are explained in the text.
It is seen, then, that for each vowel all the formant frequency ranges are less than those given by Potter and Steinberg. All the ranges for $F_2$ are approximately equal to or less than the standard deviations of Stevens and House. It can be concluded from this that the tongue tip, as defined in Fig. 4, has very little role in the characterization of vowels. Rather large variations in its position are of only minor significance. Except for the requirement that it be low and within the specified position range, it may be neglected within the model.

A similar study was conducted for the base of the tongue in order to determine whether essentially the same conclusion could be drawn about it. This was not to be the case.

The same three vowels were chosen for this work. In order to simplify the data reduction process, it was assumed that the tongue midline could be represented as a straight line in the area of the lower pharynx, i.e. for the lower 3.5 cm. This was a rather gross approximation as can be seen by comparing the formant frequencies for "B" in Table 1, in each case the approximation to the unperturbed tongue shape, with those for the unperturbed tip. However, since the variation in formant frequencies and not their absolute values was of primary interest, this approximation was assumed to be of value.

To determine the range of base shapes, a procedure different than the one used for the tip was followed. The vowel tracings were overlayed so that the tongue shapes
coincided as well as possible and straight lines were used to represent the two shapes most extreme from the center. The resulting approximations were then transferred to the vocal tract tracings of the vowels to be analyzed so that they maintained a constant relationship to the tongue body. An example is shown if Fig. 5. The formant frequencies were then calculated for the resulting shapes. Their values and $\Delta F_p$, the frequency range caused by the pharynx variation, are also given in Table 1.

It is observed that many of the ranges are greater than the standard deviations given and in some cases are larger than Potter and Steinberg's observed ranges. This is especially true for $F_2$. This means that for the purposes of the model, the tongue base must be specified much more accurately than the tip.

Through the use of this acoustic analysis we have, then, seen that there are some areas in the vocal tract which appear to be non-critical in the specification of vowel formant frequencies. Further, we have been able to isolate the effects of the various changes in the articulatory configuration. And, finally, two particular parts of the vocal mechanism, the tip and base, have been investigated to see what importance they might have in vowel production.
Figure 5: Base Perturbations. Example shown is for /i/.
Chapter Two

Formulation of the Model

Section A-Discussion of the Data

The data used for the construction of the model is in the form of x-ray motion pictures of the vocal tract of a single, American English speaker. The films were discussed in the introduction. The tracings that resulted show the shape of the tongue's dorsal surface in the midsagittal plane. For the purposes of this model, it was assumed that this was sufficient to characterize the entire tongue. This may only be considered valid as a first approximation. The lateral surface shape is certainly not flat. It varies with both distance along the length of the tongue and the absolute position of the tongue. It may well be that these deviations are systematic and, therefore, could easily be introduced into the model with the addition of only a few parameters. No work was done to assess the validity of this statement.

The use of the x-ray data itself introduced other approximations. The pictures lack contrast and the position of a certain organ was often difficult to determine. This is especially true when that organ is hidden behind a hard body such as a jaw bone or the teeth. In such cases one can only estimate the hidden portions by judging from what he is able to see. The inaccuracies thus introduced into the tracings are on the order of 1 to 2 mm. (This is about 0.7 to 1.4 mm in the life-size scale.) At least
a small amount of practice is required before one is able to pick out the important features and trace them in a repeatable way.

In making this model it was often necessary to compare the tongue shapes for different frames of the film. This was done by simply overlaying the tracings involved so that the tongues were aligned with each other. In so doing, it would seem desirable to match the shapes so that the same physical points of the tongue are overlaying each other. This requires a series of reference points to be placed on the tongue when the films are being made. For the film used here, only one such point was available—a small lead pellet located approximately one-third the length of the tongue back from the tip. See Fig. 6 for an example. Therefore, when tongue shapes were compared, the only physical point that could be aligned was at the location of the pellet. It had to be assumed that the rest of the points would then coincide with one another. This procedure effectively ignores the possibility of variable amounts of stretching in the tongue's dorsal surface. If such deformations do occur, their effect is difficult to appraise.

The method of selecting the frames to be studied is straightforward. All of the vowels to be considered in this chapter occupied the final position in bisyllabic utterances of the form /hətV/. Sonograms were available for each of the utterances. These also included information
Figure 6. Relative Tongue Locations and Shapes for Seven Vowels. Coordinate system and vowel diagram are explained in text.
about the location of the relevant frames on the film. The formant structure of the vowels was observed. Frames were chosen which were as long after the final consonant as possible yet before the point where the speaker started to relax and the vowel degenerated. At this point the articulatory organs should have been closest to their target positions. 17 This is the position of interest here.

In order to determine the location of the tongue with respect to the other organs, some reference system had to be defined. Since the Speech Communications Group at MIT is currently using a coordinate system in which the position of the maxilla is considered to be fixed, this is what was done here. 18 For this speaker, a set of overlays had previously been developed on which were drawn tracings of either the mandible or the lower parts of the maxilla. Incorporated in each were two reference points fixed with respect to the body which the overlay represented. Since the positions of the tongue and mandible were to become of primary concern, all that it was necessary to trace was the tongue and four points giving the reference locations.

Figure 2 demonstrates this method for the frame representing the vowel /I/. Tracings of the overlays are also included.

With these procedures defined and some of their limitations recognized, it is now possible to demonstrate the model.
Section B-Presentation of the Model

For the purposes of this model, the tongue will be divided into three parts, the tip, body, and base. These terms are used in the same way they were in the acoustic study, i.e. the tip is the anterior 2 cm, the base the posterior 3.5 cm, and the body the remaining length. These terms are merely for ease in reference; no anatomical considerations are implied.

The basic idea of the model is that the tongue body maintains an essentially constant shape, the vowel shape, for all vowels. This shape is allowed to translate and rotate with respect to the maxilla in order to produce the various vowels. The tip is relatively unconstrained. It must merely be low and come in contact with or approach a range of positions just behind the lower incisors on the mandible. This is the same range that was defined in the previous chapter. The base is not so well behaved and its shape may vary. However, its perturbations away from a central position, defined along with and with respect to the body, are systematic and qualitatively predictable from the location of the body.

Let us now see how the model works for one speaker. In doing so, it will probably be helpful to the reader if he would make a tracing of Fig. 8 so that he may overlay it and compare its shape with the tongue shapes in other figures.

Figure 6 shows the relative locations and shapes of the tongue for the seven vowels studied, /i/ (it), /i/ (eye),
/ɛ/ (met), /ɜ/ (foot), /ə/ (father), /æ/ (at), and /u/ (boot). The maxilla is considered as having a fixed location for each. The circles represent the positions of the reference pellets.

Jones has developed a diagram which shows the approximate relationship of the tongue positions for various English vowels to those for a series of cardinal vowels. This was done using a series of still x-ray photographs of the vocal tract during the articulation of steady state vowels. A portion of this diagram is also shown in Fig. 1. If the tongue positions for the tracings are taken as being specified by the locations of the reference pellet, it would be expected that the relative positions of the pellet would correspond closely to the locations given by Jones. This is easily seen to be the case.

The results of aligning all of the tongue shapes for the best possible fit is shown if Fig. 7. The areas considered as the tip, body, and base are indicated. The alignment is done by eye.

An average error correcting method was used to generate the vowel body shape and base central position. From a drawing such as Fig. 7 a guess at the average shape was made. This shape was then overlayed on the tongues for each of the individual vowels and the best fit obtained. The difference between the two was then measured and recorded at 1 cm (to the scale of the tracings or 0.7 cm life-size) intervals on each side of the pellet. The
Figure 7: Overlaid Tongue Shapes for Seven Vowels
average error over all the vowels was computed and this was
added to the vowel shape to form the corrected version.
Since this method still involves the matching of curves
by eye, it is quite subjective. In order to see if the
method was at all successful, the corrected version was cor-
rected a second time and the difference between the two
corrected versions was found to be minor. The twice
revised version is shown in Fig. 8.

Because of the extremely good match between the different
shapes here, this amount of detail in generating the vowel
body shape and the base central position does not seem
necessary. However, it may prove useful in future studies
where the fits may not be as good. Presumably the procedure
of aligning the various curves for the best fit could
easily be implemented on a computer with a form of graphical
input. It would merely require the calculation and mini-
mization of a mean-square-error matrix.

Figure 9 shows the translations and rotations required
to move the shape from a reference position to the location
it assumes for the individual vowels. The reference here
is the coordinate system shown if Fig. 6 with the pellet
location for /a/ taken as the origin. It was desired
to keep the x axis approximately horizontal with respect
to the maxilla. For this reason, this axis is rotated 0.9°
clockwise from the direction defined by the two
maxilla reference points shown in Fig. 2.

With the body located, it is then necessary to define
Figure 8. - Vowel Tongue Body Shape and Base Central Position
Figure 9.—Relative Tongue Translation and Rotation Data. Rotation angles are counterclockwise.
the tip position. As was discussed in the previous chapter, the tip is merely required to be within a certain range of positions rather than at a particular one. It must be low and either pointing toward or in contact with an area, the contact area, on the posterior side of the lower incisors and the floor of the mouth. As before, this region is defined to be the smallest area into which the tip fell for all vowels studied. The contact area is the curvilinear rectangle depicted in Fig. 10.

In order to determine the position range for the tip, it is now necessary to locate the maxilla with respect to the mandible. Figure 10 shows the location of the lower medial incisor with respect to the upper jaw for the seven vowels considered. It was found that the location of the tip of the medial incisor formed, to a very good approximation, a circle. Its center is also shown and is used as a reference point in locating the lower jaw. To specify the position of the mandible, then, it is required only that an angle with respect to a predetermined line through the center be known. This line was taken, for consistency, to be the one connecting the center and the position of the tip of the medial incisor for /a/. These angles were measured and the results are also shown in Fig. 10.

For comparison purposes, the angle of tongue body rotation is shown along with that of mandible rotation in Fig. 10. It might be expected that as the lower jaw drops, the tongue would tend to rotate in a clockwise direction.
Figure 10.—Mandible Location Data. Dots show position of tip of lower medial incisor for indicated vowel.
This would appear as an inverse correlation between the two columns. Such a relation does not seem to exist. One might then draw the conclusion that there must be at least one other means for rotating the tongue body besides relocation of the mandible.

To conclude this formulation of the model, it must now be shown how the variation of the base away from its central position can be related to the translation and rotation of the body. The deformation seems to be the result of two effects, one arising from each type of motion. In order to follow this explanation, it is almost mandatory that the suggestion previously made of tracing Fig. 8 and overlaying it on the various tongue shapes be followed.

The base of the tongue should be viewed as being anchored to its surrounding structures in the vicinity of the letter "A," i.e. just above the hyoid bone, in Fig. 8. In this area there is a certain amount of tongue mass. When the body is low and attempts to rotate in the counterclockwise direction, this mass cannot be compressed and a bulge develops on the posterior side of the central shape. This is particularly illustrated for the vowels /ɛ/ and /æ/. They have approximately the same location but differing amounts of rotation. As the height of the tongue increases, a different effect becomes noticeable. Certain muscles are contracting to force the position of the hump upward and an S shaped deviation from the central
shape arises. As the body moves up, the variation becomes more pronounced. A comparison of /I/ and /i/ will illustrate this. They will also show that at these heights the previously noted rotation effect is still present. This description is only qualitative. Presumably, with enough data so that these effects could be studied on a statistical basis, one would be able to quantify the trends that have been noted here.

We have, then, arrived at a model for the tongue during the midpoints of vowels. It requires four parameters, the x and y coordinates of body translation, the angle of body rotation, and the angle of mandible rotation. These and a predetermined shape specify the relevant tongue shape in the midsaggital plane. Before seeing how well it might work for other subjects, let us see how the vowel tongue shapes we have observed are applicable to those for some related consonants.
Section C-Comparison of Vowel and Consonant Shapes

The purpose of this section is to see if consonants may also be characterized by a constant body shape. If this were to be so, the model might be generalized to include them.

To compare the tongue shapes for consonants, a different procedure had to be used to select the frames to be studied. Seven consonants were used, /d/, /p/, /t/, /k/, /h/, /s/, and /z/. For the first four, the frame immediately preceding release was chosen; the frame showing maximum constriction of the vocal tract was used for the last three. It was felt that these would give the shapes most characteristic of their respective consonants. The tongue shapes are shown in Fig. 11.

The consonant with the greatest deviation from the vowel body shape is obviously that for /k/. Part of this is probably due to the extreme position to which the tongue must be moved in order to form complete closure. For the rest of the consonants the body fit is rather good in the region posterior to the pellet. However, as one moves anteriorly from the pellet, the correspondence becomes worse. For /p/, /h/, and /s/ the agreements with the vowel shape are poor from the pellet forward. For /z/, /t/, and /d/ the fits become poor as the tip is approached.

These observations suggest that what we should consider as being characteristic of vowels is the length of the portion of the body which maintains a constant shape.
Section C-Comparison of Vowel and Consonant Shapes

The purpose of this section is to see if consonants may also be characterized by a constant body shape. If this were to be so, the model might be generalized to include them.

To compare the tongue shapes for consonants, a different procedure had to be used to select the frames to be studied. Seven consonants were used, /d/, /p/, /t/, /k/, /h/, /s/, and /z/. For the first four, the frame immediately preceding release was chosen; the frame showing maximum constriction of the vocal tract was used for the last three. It was felt that these would give the shapes most characteristic of their respective consonants. The tongue shapes are shown in Fig. 11.

The consonant with the greatest deviation from the vowel body shape is obviously that for /k/. Part of this is probably due to the extreme position to which the tongue must be moved in order to form complete closure. For the rest of the consonants the body fit is rather good in the region posterior to the pellet. However, as one moves anteriorly from the pellet, the correspondence becomes worse. For /p/, /h/, and /s/ the agreements with the vowel shape are poor from the pellet forward. For /z/, /t/, and /d/ the fits become poor as the tip is approached.

These observations suggest that what we should consider as being characteristic of vowels is the length of the portion of the body which maintains a constant shape.
Figure 11.-Relative Tongue Locations and Shapes for Seven Consonants
It appears that for some consonants also the tongue may assume the vowel shape over a part of its length. If a way for describing the tip and the front part of the body could be formulated, a model for certain consonants might be developed similar to the one shown here for vowels.
Chapter Three

A Test of the Model for a Second Speaker

In an attempt to test the generality of the model, it was decided to try it on another speaker in another language. The model was developed using English vowels. Here the utterances are in Swedish.

The data for this case were in a slightly different form than for the former one. The film was of significantly poorer quality than the first and accurate tracings were, therefore, more difficult to make. The reference pellet was in a position similar to that of the previous case but was often difficult to see. Its location then had to be approximated. The context in which the different vowels were located was not constant. The position of the frame to be considered as characteristic of each vowel was taken as being midway between the surrounding consonants. This was determined from the acoustic recordings of the utterances.

The results are shown in Fig's 12, 13, 14, and 15. The illustrations are similar to those presented in the previous chapter. The vowels used are: Swedish orthography, /ɛ/ /ɔ/ /o/, /i/, and /u/. Two examples of the first and three of the last are given. Fig. 12 illustrates the position and shape of the tongue for the various vowels. The maxilla is considered as the fixed point of reference. The small x's on the right hand side of the drawing indicate the position of the tip of the lower medial incisor. Figure 13 shows the tongue shapes superimposed and fitted to each other. The resulting shape for the body and base central position
Figure 12.-Relative Tongue Locations and Shapes for Eight Vowel Samples from a Second Speaker. Coordinate system is explained in text. x's show position of tip of lower medial incisor for indicated vowel.
Figure 13.- Overlaid Tongue Shapes for Eight Vowel Samples from a Second Speaker.
Figure 14.-Vowel Tongue Body Shape and Base Central Position for a Second Speaker.
Figure 15.-Relative Tongue Translation and Rotation Data for a Second Speaker
is shown in Fig. 14. Figure 15 gives the translations and rotations required for the body to assume the various tongue positions. These are relative to the coordinate system shown in Fig. 12. As before, the origin is taken as the location of the pellet for one of the phones, /o/. In order to keep the x axis approximately horizontal, it was rotated counterclockwise so that it formed an angle of 2.9° with the line determined by the tips of the subject’s upper right medial incisor and third molar. These were the two reference points used to locate the position of the maxilla.

The systematic base variations discussed before are even more marked here. Keeping in mind the previous description, compare the body and central base shapes for the group of three vowels most extreme from the origin. They all have approximately the same translation coordinates but different angles of rotation. /ɛ/₂ has the greatest angle of rotation among the three. While the base has somewhat of an S shape, its deviation from the central base shape is mostly a bump on the posterior side of the central shape over the entire length of the base. /i/ has a smaller angle of rotation than /ɛ/₂; its deformation from the normal shape is of the S form with a good sized posterior hump at the lower end. The vowel with the least rotation of the three, /ɛ/₁, has only a small posterior hump in the base region. Their bases then vary away from the central position in the above mentioned manner.
Figure 15 illustrates the phenomenon of undershoot in the articulatory domain. There are three examples of the vowel /ä'/; /ä'/₂ and /ä'/₃ are of approximately the same length, 60 msec, while /ä'/₁ is more than twice as long, 159 msec. The locations and rotations of the tongue body are similar for the former two but very different from those for the latter example. Presumably the configuration for the first and longest sample, /ä'/₁, is the closest one to the target position. /ä'/₂ and /ä'/₃ then represent time limited approaches to the target position.

It is seen, then, that the body does indeed maintain a somewhat constant shape across all vowels. The base does behave as described previously. And the position of the mandible can be at least roughly approximated by its rotation on a circular arc. While success in applying this model to two speakers is certainly no guarantee of general validity, it does indicate that further studies of it may be worthwhile.
Chapter Four

A Study of the Validity of the Model During a Diphthong

Although this model is intended to deal only with the static target positions of vowels, it is of interest to discuss what implications it might have when the vocal tract is in motion. In order to do this, let us look at an approximation to a diphthong, the vowel pair /iə/ as uttered by the first speaker.

In Fig. 16 tracings of the tongue during the four frames representing the transition from /i/ to /ə/ are shown. Again, the maxilla is considered as being fixed. The frames are 23 msec apart. Observe that the tongue body maintains an approximately constant shape as it changes from the configuration for /i/ to that for /ə/. One possible interpretation for this is that the intrinsic tongue muscles are holding the tongue in a constant shape while the extrinsic musculature is responsible for the translation and rotation of this shape. The tongue translation data and the rotation data for both the tongue and mandible as functions of time are shown in Fig. 17. It is of interest to note that $x(t)$ and $y(t)$ are approximately linear with time while both $\theta(t)$ and $\phi(t)$ show sharp accelerations toward the end of the transition. The relation between the last two indicates that the tongue body does rotate as the mandible is lowered. However, as was previously stated, their are probably other ways to rotate the body besides dropping the lower jaw.
Figure 16.-Relative Tongue Locations and Shapes During a Diphthong
Figure 17.-Relative Tongue Translation and Rotation Data and Mandible Rotation Data During a Diphthong. Horizontal scale represents time in msec. \( \phi \) = angle of mandible rotation. \( \theta \) = angle of tongue rotation.
In relation to the above mentioned hypothesis that the tongue's intrinsic muscles may be primarily responsible for maintaining its shape and the extrinsic musculature for its position, Mac Neilage and Sholes present some relevant data.\textsuperscript{20} They have placed surface electrodes on the dorsal surface of the tongue and made electromyograms for various tongue muscles during vowel production. Different trends were reported for the different phones studied. One might be tempted to think that if the model presented here were accepted, the electrodes on the surface of the body should show similar patterns for all vowels and that this data thus contradicts the model. However, such is not necessarily the case. It is unclear whether the electrodes are measuring a potential which is related to the amount of stretch in a muscle or the amount of force exerted by that muscle. If the first alternative is true, then the previous conclusion could be correct. If, however, the second alternative is the right one, the validity of this model is still conceivable. As the tongue is moved to different positions, it is to be expected that the extrinsic muscles will tend to change its shape. This will mean that the forces required of the intrinsic muscles to retain the body shape will vary with tongue position although their deformation from their relaxed position will remain constant.

If the tongue body does maintain a constant shape over all vowels, this may imply that the speaker has some way of knowing what the shape of his tongue is. The existence
of appropriate proprioceptors in the tongue would then be suggested. These would give him the necessary information so that he might bring his tongue to the vowel shape.

While the discussion presented above is very limited and tentative, it is indicative of the types of information that might be gained from a further study of this model.
Conclusion

To conclude, let us look at how the study presented here might relate to the work done by others, some further implications of the model, and a few suggestions for additional investigation.

The model of Stevens and House which was cited in the introduction might appear similar to the one presented here. They use a tube of varying radius to approximate the vocal tract area function. Over the length of the tongue, the relation between the radius of the tube to the distance along it may be specified by two parameters, the position of the maximum constriction in the tract, $d_0$, and the radius at this point, $r_0$. The former is used to define a parabolic function between radius and distance and the latter is used to locate the vertex of this curve. Thus, for a constant $r_0$, a graph of this relation could be viewed as a parabola, a constant shape, which moves to different positions along the tract with variations in $d_0$. This model is somewhat like the one presented here, but the differences are significant. The shape of the curve does change as $r_0$ varies. This model deals with the approximating vocal tract radius while here we are concerned with the vocal tract crossdimension. Since the cross dimension to area transformations are not simple square functions, an acoustic correspondence between the constant shapes in the different domains is not guaranteed.

Ohman has presented a model which relates the time
varying tongue shapes for intervocalic stop consonants to the shapes for the surrounding vowels. Consonantal motion is viewed as being superimposed on diphthongal transitions between the two vowels. The work done here may make it easier to specify the preceding and succeeding vowel shapes. If the vowel tongue shape description can be extended to include consonants as previously indicated, his entire model might be considerably simplified.

Henke is developing a dynamic articulatory model based on an operator-state dichotomy, i.e. a distinction between the state of the vocal tract at any time and the operators which work on that state to transform it to its future configuration. In the specification of the operators he makes use of the target positions for the various phonemes. The constant shape concept presented here might simplify the description of these target positions for vowels. It might also be useful for defining some properties of the various operators which would be found to be constant for all vowels.

Finally, this work is very directly related to some suffestions made by Ladefoged. He states:

If we neglect, for the moment, the tip and blade of the tongue (perhaps regarding it as down behind the lower front teeth, as it often is in vowel sounds) we can describe the remainder of the vocal tract by stating the position of the center of the body of the tongue in an arbitrary coordinate system, providing we also state the length of the arc or pseudo-diameter at the center.... It [this description]...reflects the anatomical constraints on the shape of the vocal tract arising from the fact that the general position of the body of the tongue (the
locus of the center point) can be moved in the mid-
sagittal plane by the action of three or four ex-
trinsic tongue muscles, which operate independently
of the intrinsic muscles controlling the shape of
the body of the tongue (roughly specified by the
length of the arc).

Obviously significant differences exist, but if he were
to evaluate his model, it might be expected that the
pseudo-diameter would be related to the deformation of the
base away from the central position. The arc length would
then be related to the translation and rotation coordinates
established here. His position data should be approximately
the same as those given in Chapter Two. Thus a very
definite correlation might be found between the two
models.

A number of articulatory correlates of vowels exist.
Two are that the velum must be up and the vocal tract must
be unobstructed so that there is a free, i.e. nonturbulent,
flow of air throughout its length. This study suggests
another correlate—that the tongue body must assume a specified
shape. To a certain extent it seems that this same result
may hold for consonants. If this could be generalized, it
might be shown that the same correlate holds for nonvowels
as well, that is for all speech.

Much needs to be done on this model of vowel produc-
tion before it is in a really useable form. As discussed
earlier, certain ranges in acoustic cues are allowed in
the specification of most phonemes. A corresponding set
of ranges must also exist in the characterization of
articulatory parameters. Investigation of these could
be done in two ways.

Further acoustic studies of the type carried out here might be conducted. This would require a suitable definition of allowable acoustic tolerances, generation of the appropriate area functions found using the model itself under a set of systematic perturbations, and finally the solution of the area functions for the formant frequencies. The ranges of the parameters could also be established by further study of cineradiographic films showing different examples of the same vowels.

Even though the model is still rather incomplete, the goals given in the introduction have been at least partially achieved. A representation of the tongue during vowel production has been devised that is easy to use. Developing it has involved a number of approximations and numerous questions have arisen in the process. Work at solving these problems will undoubtedly generate new ones. But we now have something on which further discussion may be based.
References


