Articulatory model for the study of speech production

P. Mermelstein

Bell Laboratories, Murray Hill, New Jersey 07974

(Received 8 April; revised 11 July 1972)

Study of midsagittal x-ray tracings reveals that the vocal-tract outline can be accurately represented by means of variables specifying the positions of the jaw, tongue body, tongue tip, lips, velum, and hyoid. As the articulators move, they modify the vocal-tract cross-sectional area and the vocal-tract transfer function computed thereform. The speech signal may be synthesized by concatenating the responses to repeated excitation of the quasistatic vocal tract. Vowels are specified in terms of variables denoting the positions of the jaw, tongue body, lips, and velum. Consonants are implemented as transformations on the underlying vowel-derived articulatory states that satisfy given constraints on the position of an articulator relative to the fixed structures. The set of states which satisfy the given constraint corresponds to the allowed productions of the consonant. Coarticulation effects control the selection of the underlying state and thus determine the particular consonant produced. Vowel-consonant-vowel sequences generated with the aid of rules for articulator movement and the articulator-position to vocal-tract cross-sectional-area transformation are intelligible and exhibit coarticulation in agreement with acoustic measurements.

Subject Classification: 9.2, 9.4, 9.7.

INTRODUCTION

Speech synthesis by rule, when implemented in articulatory terms, incorporates a dynamic model of the articulatory system. Articulatory representations of isolated phonemes are combined, modified and sequenced in time to generate the trajectory in articulatory space appropriate for the production of the specified speech signal. The dynamic model defines the rules governing this procedure.

The momentary state of the articulatory system may be represented in terms of the positions of the individual articulators—jaw, tongue body, tongue tip, lips, velum, hyoid, and the maxilla. The excitation conditions, glottal or fricative, are assumed independently specified. A static vocal-tract model uses the articulatory variables to determine the length and cross-sectional area function of the tract. These in turn suffice for generation of a synthetic version of the speech signal appropriate to the particular configuration of the stationary articulatory system. Time variation of the articulatory variables results in a quasistatic simulation of the speech event where the output signal is obtained as a sequence of responses to different stationary vocal-tract configurations.

This paper first describes the static model—the description of the vocal-tract shape in terms of articulatory variables. These variables, individually or pairwise, specify the position of the jaw, hyoid, tongue body, tongue blade, lips, and velum in the midsagittal plane. The temporal variations in variable values can be determined from measurements on x-ray tracings obtained at regular time intervals during a spoken utterance, and in turn serve for an accurate reconstruction of the time-varying midsagittal vocal-tract outline. Articulatory information so derived has been combined with fundamental frequency and amplitude information determined from the natural speech signal to yield a synthesized speech signal and thereby indicate the adequacy of the method for speech generation purposes.

The second part of the paper postulates rules for temporal variation of the articulatory variables in particular phonetic contexts. As yet, the rules have been systematically formulated only for vowel–voiced stop/nasal–vowel sequences. Similar rules are considered applicable to wider contexts. The rules allow temporal overlap in the movement of the individual articulators in contexts where they are free to move independently, and they impose delays where constraints between articulators or with respect to the fixed structures must be satisfied for specified intervals. The vowel–consonant–vowel (VCV) sequences synthesized with the aid of the rules yield high consonant identification scores.

The fundamental point of the postulated rules is that the invariant property of a consonant is best expressed as some constraint on the position of an articulator relative to the fixed structures. Examples of such constraints are tongue-tip closure, lip closure, and velar closure. There exists no one-to-one mapping between phonemes and articulatory states. Rather the collection of such states satisfying the given constraints corresponds to the allowed productions of the phoneme. When consonants are coded in such terms coarticulation across phonemes, across syllables, and even across words (Daniloff and Moll, 1968; Amerman, Daniloff, and Moll, 1970) can be readily implemented.

I. PREVIOUS MODELS

The picture of the state description of the articulatory system undergoing modification under the influence of a set of goals or targets was succinctly presented by Henke (1967). His solution in terms of separately controlled flesh-points on the surface of the tongue and lip is unsatisfactory due to the complexity of the constraints between such points. Coker and Fujimura (1966) introduced a model with parameters assigned to the tongue body, tongue tip, lips, and velum. This model, when appropriately controlled, suffices for the generation of all English phonemes and has been used to generate contextually complex senstences. Further developments of the model by Coker make use of time constants with which the individual articulators respond to particular commands. He further introduces the notion of "priority" of certain consonantal characteristics to ensure that the most pertinent articulatory features of individual consonants survive drastic smoothing effects.

Concurrently with the present work, Lindblom and Sundberg (1971) focused attention on the role of the jaw in an articulatory model. The jaw position influences the vocal-tract cross-sectional area in both the tongue and lip regions. It allows the isolation of parameters pertaining to the tongue and lips alone, contributing to a simpler picture of the articulatory mechanism for vowel generation.

Our work considers the lips, jaw, tongue, velum, and hyoid as movable structures. The tongue body and tongue blade are considered as separate but interdependent articulators. The position variables for all articulators except the tongue body correspond to measurable points on midsagittal x-ray tracings; the tongue-body coordinates are those of the apparent center point of the approximately circular tongue-body outline.

The parameter set adopted is not minimal—a smaller number of parameters may be sufficient to generate an adequate static representation of the vocal-tract outline. Tongue movement in a fixed coordinate system may take place as a result of jaw movement only—in which case the tongue remains fixed relative to the jaw. Alternatively, the jaw may remain fixed and the tongue move relative to the maxilla and the mandible. These two modes are differentiated to allow independent expression of contextual constraints. Articulations equivalent with respect to tongue height (tongue-body position relative to the maxilla) can then be generated with different tongue-body and jaw positions. The choice allows a reduction in the required travel of any one articulator in the given context, and thus faster attainment of the articulatory goal is made possible.

II. REPRESENTATION OF THE MOVABLE STRUCTURES

Parameters assigned to the movable structures are position variables which indicate the position of the structure in fixed space or relative to some more massive structure to which it is attached. The position of the jaw and hyoid are expressed directly in the fixed coordinate system. Lip and tongue-body positions are specified with respect to the moving jaw. Tongue-tip position is specified relative to the tongue body. This representation provides for an active mode of movement for an articulator by changes in the corresponding variable. Alternatively, a passive movement may be executed in the fixed coordinate system as a result of movement of the structure to which the articulator is attached, but relative to which its position remains unchanged.

Limits on the articulatory variables may be specified in absolute terms, such as maximal jaw or velar opening. More frequently, the limits are imposed by the (nominally) fixed structures of the maxilla and rear pharyngeal wall. These limits are attained only during specified time intervals, for example, that of tongue-tip closure. To attain these limits the articulatory state is modified by means of a consonantal gesture. The modification takes place through changes in the variable values reflecting movement of the participating structures required in order to achieve the requisite articulatory result.

The model variables are determined by reference to x-ray tracings introduced into an interactive computer system. J. Perkell was most generous in making available to us the data used in his study (Perkell, 1969). The experimenter, by adjusting the variables, generates a vocal-tract outline that is brought into registration with the x-ray tracing. Significant deviations noted by the experimenter are used to modify the model. Automatic extraction of the articulatory variables is feasible; however, manual matching with careful inspection allowed a better focus on the defects of a model under development.

In the following vocal-tract representation a rather abundant set of variables is used in order to allow detailed independent consideration of the individual articulatory structures. Some of the variables may be treated as constants in most cases. In other places relationships between the variables noted in the analysis

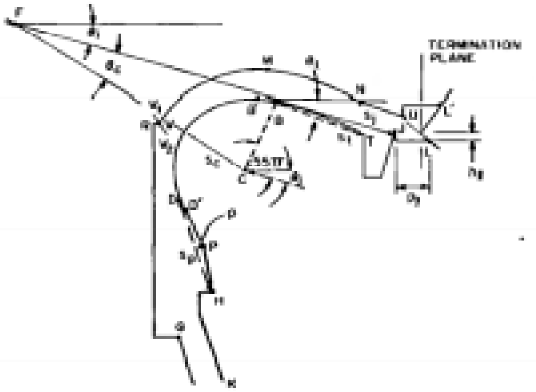


FIG. 1. Model-generated vocal-tract outline.

are used to reduce the number of free variables of the model.

A. The Jaw

In general, the jaw may be assumed to execute rigid rotation about a point fixed with respect to the stationary structures. Let the jaw position be indicated by the point J at the top of the lower incisors in the sagittal plane (see Fig. 1). Its location relative to the fixed point F is given by polar coordinates (s_j,θ_j). The distance s_j remains constant at 11.3 cm for our subject in most cases. In a few cases, such as the fricatives /s/ and /v/ it is observed to decrease by less than 1 cm indicating a slight retraction of the mandible relative to the upper incisors.

B. The Hyoid

The hyoid is allowed to move in the midsagittal plane, and its position is expressed by means of vertical and horizontal position coordinates representing the integrated action of the muscles connecting the hyoid to the jaw, tongue, styloid process, and infrahyoid structures. The hyoid position is denoted by the point H, located on the sagittal x-ray tracing at the apparent intersection of the anterior edge of the epiglottis with the top edge of the hyoid bone. The point H represents the junction of curve sections such that below H the curve is a function of H only and above it a function of H and the tongue body. The point K, an estimate of the anterior extremity of the larynx, is found to move up and down roughly synchronously with H, and is located 2.7 cm below it. As a rough estimate, anterior–posterior movement of K is found to be one-half that of H. Thus the curve KH changes orientation slightly with anterior–posterior movement of the hyoid.

C. The Tongue Body

Following Coker and Fujimura (1966), we represent the tongue-body outline as a circle with a moving center and fixed radius. At first the radius was allowed to vary, but good matches to most x-ray tracings were found with a constant radius of 2 cm. In a few articulatory situations such as the elevated tongue body just prior to achieving palatal closure, the superior outline of the tongue body was observed to deviate from a circular arc, but the extent of deviation was deemed small enough to be generally negligible.

The position of the tongue-body center C is given with reference to the jaw-based line FJ by polar coordinates (s_c,θ_c). Since the jaw rotates about point F, the angular coordinate with respect to the fixed system θ_{jc} is given by the sum $\theta_j+\theta_c$.

The coordinates specify the position of the tongue with respect to the fixed reference system and the jaw. Identification of variables with unique muscles is not intended. The coordinate assignment is purely functional in that it permits simple descriptions of tongue movement in two distinct modes: (a) as fixed relative to the moving jaw—primarily executed by the muscles of the jaw, and (b) as moving relative to a fixed jaw controlled by the intrinsic and extrinsic muscles of the tongue.

Analysis of our data indicates that the anterior outline of the pharynx is controlled by the tongue body and hyoid bone positions. Let s_p be the distance from H to a point D on the tongue-body circle such that HD is tangent to it on the posterior side. The actual outline deviates from this line to the anterior or posterior side depending on the height of the tongue body. Let P be a point on the normal bisector of HD such that the straight lines HP, PD′ best represent the observed outline. PD′ is the tangent to the tongue-body circle through the point P. The offset p from HD has been determined for various vowel and consonant articulations in our x-ray data, and the observed values are plotted against s_p in Fig. 2. The data indicate that a straight line adequately approximates the variations in the offset as a function of tongue-body–hyoid distance s_p. This relation is used in the model to determine the pharynx outline from the tongue-body and hyoid positions. It was further observed that the rear pharyngeal wall moves backwards as s_p increases—it is maximally anterior for /a/ and maximally posterior for /i/ and /u/. This might be considered an indication of the contraction of the sphincter muscle in the lower pharynx for the low back vowel.

D. The Tongue Blade

The tongue tip and tongue blade are considered attached to the tongue body and move with respect to a point B nominally on the tongue-body surface. The line CB is oriented at an angle $\theta_j+0.55\pi$ with respect to the horizontal. When executing up–down movements, the tongue tip appears to rotate about B; therefore its position is expressed by polar coordinates (s_t,θ_t) with respect to B. Active extension or retraction of the tongue tip relative to the tongue body is expressed by variations in s_t. The angle of elevation of the tongue tip θ_t has components $\theta_j+\theta_{tc}$, reflecting the position of the jaw and the tongue body. A further component θ_{te} reflects active elevation of the tongue tip and is zero for static vowels. Again, a functional separation is made to allow tongue-tip elevation with fixed tongue body or involuntary tongue-tip movement due to tongue-body movement.

The actual tongue-blade outline is approximated as a smooth curve tangent to the tongue-body circle at B′ and passing through T, the point on the tongue blade farthest from the tongue-body center C. B′ is located on the tongue body circle such that the angular difference between CB and CB′ is proportional to the intrinsic tongue blade elevation θ_{te} defined below. The tongue blade decreases in thickness and increases in flexibility as one moves along its surface from the tongue body to the tongue tip. We can represent the curve modelling the tongue-blade surface by a radial coordinate about the tongue-body center C which varies as the square of the angular difference with respect to the starting point B′ on the tongue body.

The angular coordinate of the tongue tip θ_t is observed to vary with tongue fronting. In particular, for the /i/ vs /u/ distinction a tongue-body contribution to tongue-blade orientation θ_{tc} can be defined such that

$$\theta_t=\theta_j+\theta_{tc}+\theta_{ti}.$$

θ_{ti}, an intrinsic tongue-blade elevation variable, is assigned a zero value for stationary vowels—defined as the approximately stationary vowel targets in the [hə′CV] articulations. Our analysis reveals that for these cases

$$\theta_{tc}\simeq 0.004(s_c-8.6),$$

where θ_{tc} is in radians and s_c in centimeters. This allows a specification of tongue-blade orientation for vowels that is a function of the jaw and tongue-body coordinates.

The tongue-body–tongue-tip distance s_t is roughly constant at 3.4 cm for the observed vowels. Articulations with a retroflexed tongue tip where s_t can be expected to be reduced have not been studied.

E. The Lips

The lips are allowed to open (close) or protrude (retract) relative to the jaw and maxilla. The position of the lower lips is denoted by the point L having coordinates (p_l,h_l) with respect to the point J which is fixed with respect to the jaw. Although for vowels in English the jaw opening and either h_l or p_l permits a fair approximation of the other one (Mermelstein, Maeda, and Fujimura, 1971), for continuous articulatory variations containing labial consonants both coordinates need to be independently specifiable. Use of h_l and p_l as separate variables allows lip closure with spread or rounded lips. Furthermore lip opening and rounding have different characteristic rates of change, and this difference would be masked if only one coordinate were used. The protrusion and elevation of the upper lip relative to the upper incisors are assumed equal to those of the lower lip relative to the lower incisors. Total lip opening is again given by two components—one a function of the jaw opening θ_j, and the other lip opening h_l relative to the jaw. Again, we provide for differentiation between lip closure by jaw movement with lip muscles inactive or with the aid of the lip muscles alone and the jaw stationary.

F. The Velum

The state of the velum is represented by the position V of the apparent tip of the uvula moving along a straight line V_1V_2. V_1 is empirically defined as the position noted for the most elevated point of the velum and V_2 as the lowest observed position. The velar opening area is assumed proportional to the squared distance $(V-V_1)^2$. Where a nonzero velar area is noted, an appropriate nasal sidebranch is assumed coupled to the pharyngeal–oral tract. The glottis-lips transfer function is replaced by one which determines the total pressure due to radiation from nostrils and lips for a unit glottal velocity pulse. The velar position modifies the vocal-tract cross-sectional area as it controls the position and length of segment RV, a normal from V to the real pharyngeal wall, and curve VM approximating the outline of the soft palate.

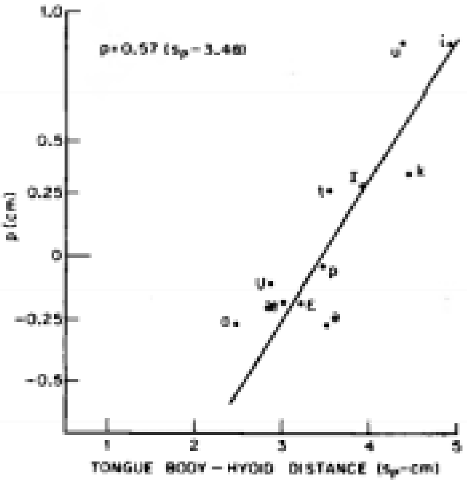


FIG. 2. Dependence of offset from tongue-body–hyoid line on tongue-body–hyoid distance.

G. Maxilla and Rear Pharyngeal Wall

The posterior–superior vocal-tract outline was matched with the aid of variables controlling the highest point of the periarytenoid G, the horizontal coordinate of the rear-pharyngeal wall position R, the velum position V, the highest point on the maxilla M, the point U on the upper incisors and the point L′ on the upper lip. Point N is located on the straight line MU such that distance MN is twice the distance NU. Circular arcs VM and MN are drawn with centers on a vertical line through M. G, the center of the grid system to be discussed later, is also located on this line 3.85 cm below M.

The variations in the rear pharyngeal wall and maxilla positions were considered significant for analysis purposes. However, as noted further below, for synthesis purposes the posterior–superior outline (except for the upper lip) may be considered fixed if slight adjustments are made in the jaw and tongue variables of the vowels to maintain the relative separation between the two sagittal outlines.

III. ANALYSIS OF X-RAY DATA

Mid-saggittal x-ray tracings of eight bisyllabic utterances [ha'CV] consisting of roughly 20 x-ray frames each, and one sentence of 120 frames (2.5 sec) have been analyzed with the aid of the model. The articulatory variables were adjusted with the aid of programmed knobs to achieve best registration on the computer display between the stored x-ray tracing and the repeatedly regenerated model-derived outlines. The adjustments were carried out in the order: rear-pharyngeal wall, maxilla, velum, hyoid, jaw, tongue, body, tongue tip, lips. Results of this matching procedure are illustrated in Fig. 3. Frames 1340 and 1343 show examples of a back and front vowel from the sentence. Frames 1344 and 1353 are examples of alveolar and palato-velar stops.

The experimenter tried to attain particularly small deviations in the acoustically important regions—for example, the point of maximum constriction on the tongue body. Deviations between the solid (x-ray derived) and dotted (model) curves noted in the pharyngeal and tongue blade regions are due to the approximations used in deriving the articulatory shapes. The deviations are not considered acoustically significant—the errors they introduce are exceeded by the errors involved in estimating the areas in the sections through tract normal to the center line.

IV. DETERMINATION OF THE VOCAL-TRACT LENGTH AND CROSS-SECTIONAL AREA

The vocal-tract cross-sectional area is determined in terms of the areas of coronal sections whose projections on the midsagittal plane form a grid system as shown in Fig. 4. Grid lines are 0.5 cm apart where parallel and 10° apart where radial. The anterior–inferior outline of the vocal tract in the midsagittal plane is determined from the articulatory parameters as discussed above. The posterior–superior outline shown in Fig. 4 is a sequence of straight line and arc segments approximating the rear pharyngeal wall, the soft and hard palates, the upper incisors and upper lips. The two outlines intersect the grid system as shown. Let the jth grid segment delimited by these two outlines have length g_j, and midpoint C_j. The vocal-tract center line is assumed to be in the midsagittal plane and is approximated as the sequence of straight-line segments joining the C_j. Assume the direction of wave propagation at C_j is given by $\frac{1}{2}[\mathrm{ang}(C_{j+1}-C_j)+\mathrm{ang}(C_j-C_{j-1})]$, which deviates from the normal to the jth grid line by angle α_j. The estimated area at grid-line j is then given by

$$\hat{a}_j=t(j,g_j)\cos\alpha_j,$$

where $t(j,g_j)$ is a transformation that maps the jth midsagittal projection g_j into the cross-sectional area in the plane of the grid line. The grid system is fixed with respect to the maxilla to limit the variations in the transformation with changes in the articulation. Recent data of Maeda (1972) show that improved accuracy may be obtained by including variations in the transformation as a function of the articulation, i.e., allowing $\hat{a}_j$ to depend not only on j and g_j but also on g_i, $i\neq j$, but these effects were considered minor and were not implemented here.

The transformation $t(j,g_j)$ is based on previously published data. Following Heinz and Stevens (1964) the cross-sectional area in the pharyngeal region is approximated as an ellipse with g_j as one axis and the other increasing from 1.5 to 3 cm as one moves upward from the larynx tube to the velopharynx. In the oral region we approximate the data of Ladefoged *et al.* (1971). In the soft-palate region the area (cm²) is taken as $2g_j^{1.5}$, in the hard-palate region as $1.6g_j^{1.5}$, and between the alveolar ridge and incisors as $1.5g_j$ for $g_j<0.5$, $0.75+3(g_j-0.5)$ for $0.5<g_j<2$, and $5.25+5(g_j-2)$ for $g_j>2$. In the labial region the area is assumed elliptical with width in centimeters given by $2+1.5(s_l-p_l)$ where p_l is the lip protrusion and s_l the vertical lip separation (Mermelstein *et al.*, 1971).

In order to include articulations such as /l/, the transformation to cross-sectional area must be modified in a few special cases. To allow for /d/ vs /l/ distinction, which is presumably not possible on the basis of midsagittal data alone, the area values in the tongue-tip region are modified on the basis of additional information regarding the shape of the tongue in the frontal plane near the point of constriction.

The front termination plane of the vocal tract is determined from the upper and lower lip positions as follows. Tangents are drawn to the lip surfaces horizontally, and inclined to the horizontal at +45° and −45°. Points L and L′ are located on the lower and upper lips at the intersections of the respective tangent lines (see Fig. 1). The point of intersection of the two tangents drawn inclined to the horizontal is the apex of a 90° sector formed by the lips in the midsagittal plane and approximates the location of the frontal plane where the source for radiation from the tract is located. It can be shown that this simple definition of the vocal-tract termination point results in a dependence on jaw opening closely approximating that given by Lindblom and Sundberg (1971).

The area values $\hat{a}_j$ represent nonuniform samples of the vocal-tract cross-sectional area separated by the straight-line distance between center points C_j. The area function is converted to a sequence of sections of uniform area $\hat{\hat{a}}_k$ by allowing the section impedance to approximate the varying impedance integrated over the extent of the section. The tract is continued to the nearest integral multiple of the section length, 0.875 cm, in a parabolic horn of area $\hat{a}_l[1+(x-x_l)/(\hat{a}_l/\pi)^{\frac{1}{2}}]^2$, where $\hat{a}_l$ is the estimated lip area at distance from the glottis $x=x_l$. Under these conditions, the radiation impedance looking from the tract at x_l is replaced by a

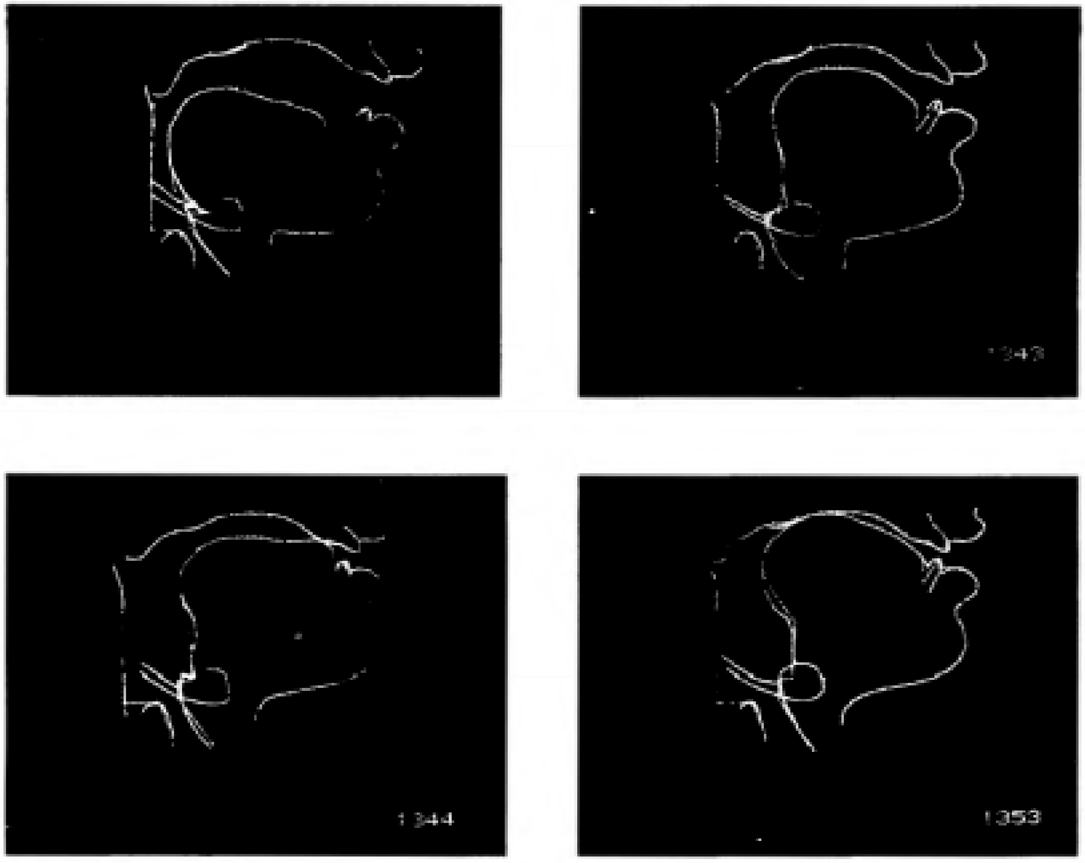


FIG. 3. Comparison of midsagittal x-ray tracing with model-generated match. Solid lines—x-ray tracing; dotted lines—model.

cascade of horn section and radiation impedance of equivalent total input impedance (Sondhi and Atal, 1970).

For voiced excitation, the set of values $\hat{a}_k$ contains the requisite articulatory information for signal synthesis or format computation. For fricative excitation the source is assumed located just anterior to the point of cross-sectional area minimum. Methods for computing the vocal-tract transfer function are given in Mermelstein (1971, 1972).

V. REPRESENTATION OF THE SPEECH EVENT IN TERMS OF TIME-VARYING ARTICULATORY VARIABLES

The mathematical framework described above allows one to represent the momentary articulatory state in terms of the respective variables. The extent to which such a static representation is adequate for the representation of the dynamic sequence of events may be explored by synthesis of the speech signal from variable values determined at closely spaced points in time through the utterance. Articulatory variables were determined for the sequence of x-ray tracings corresponding to the utterance "Why did Ken set the soggy net on top of his deck?" Cross-sectional area values were computed for each frame using the model-derived vocal-tract outlines as previously discussed. Pitch and pitch-period energy data were derived from the naturally spoken version of the sentence using linear predictor

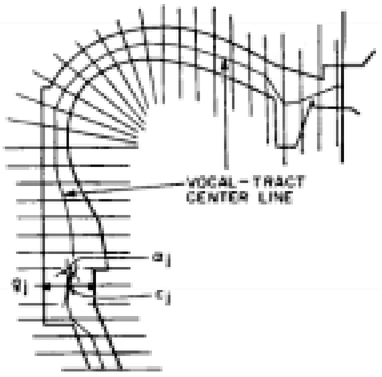


FIG. 4. Grid system for conversion of midsagittal dimension to vocal-tract cross-sectional area.

TABLE I. Comparison of model-derived formant frequencies with those determined from the natural speech signal.

Frame No.	Model-derived F1	Model-derived F2	Model-derived F3	Speech-derived F1	Speech-derived F2	Speech-derived F3
1340	666	1291	2537	717	1261	2373
1343	436	1683	2373	480	1846	2533
1348	382	1841	2650	354	1813	2544
1356	452	1833	2369	472	1839	2592
1358*	476	1686	2503	514	1835	2485
1372	492	1611	2308	509	1649	2581
1381	314	1385	2971	395	1367	2705
1390	602	1367	2491	632	1260	2312
1398*	334	2177	2929	305	2215	2832
1403*	391	1737	2575	336	1819	2574
1405	569	1498	2577	594	1698	2537
1417	480	1259	2681	536	1233	2848
1426	648	1290	2671	619	1197	2352
1433	306	1211	2571	527	1225	2421
1438	369	1741	2486	392	1754	2460
1446	546	1618	2551	472	1710	2553
Average absolute error (%)	10.3	4.9	5.5			

analysis (Atal and Hanauer, 1971). The vocal-tract area at the onset of each pitch period was estimated by interpolation of the area values corresponding to the neighboring frames.

A linear predictor analysis of the spoken sentence (Atal and Hanauer, 1971) yielded formant values for comparison purposes. The model-derived formant values are defined as the poles of the transfer function expressed as a ratio of polynomials in the delay parameter z^{-1}. The formant comparisons are given in Table I. Frames marked with an asterisk indicate a lowered velum, and the corresponding nasalized formant values are computed using a nasal tract with a fixed prototype area based on published data (Fant, 1960).

Some of the formant differences, for example, the large difference in the first formant for frame 1433, are considered to be due to tracing errors. The mean absolute error of about 5% in the second and third formants illustrates the extent of validity of the assumptions regarding the geometry of the vocal tract and the acoustics of wave propagation there, for example, that the tract is hard-walled within its interior. The differences are considered to arise primarily from a lack of detail regarding the shape of the tract in cross-sectional sections normal to the center line of the tract. Informal perceptual observations comparing the natural and reconstructed speech signals reveal no phonemic confusions, though improvements in naturalness can be expected.

The formant differences are plotted in F1–F2 space in Fig. 5. The model-synthesized signal formants appear generally more centralized than the natural-signal formants. The low F2 for front vowels and high F2 for back vowels indicates that for this subject the rate of change of cross-sectional area with sagittal segment is possibly too small in the hard-palate region.

Some of the articulatory parameter trajectories used in the sentence reconstruction are illustrated in Fig. 6. The data appear to have quite complex structure. It is noteworthy to remark that the consonants are normally associated with extremes in some parameter. The values associated with vowels are intermediate and rarely stationary for any length of time. To obtain a better understanding of the data we must consider more simply structured speech events, such as the nonsense syllables discussed below.

The advantages of separate jaw and tongue representations compared to one monitoring only the tongue in the fixed reference system can be observed by reference to the angular measures of jaw and tongue position plotted in Fig. 6. The vowel /a/ in both its productions manifests a marked jaw lowering. In the first case, marked (1), the tongue simultaneously anticipates the stop /g/ and moves to a high position. In the second case, marked (2), anticipating the labial stop, it remains stationary. Separation of the variables simplifies the dy-

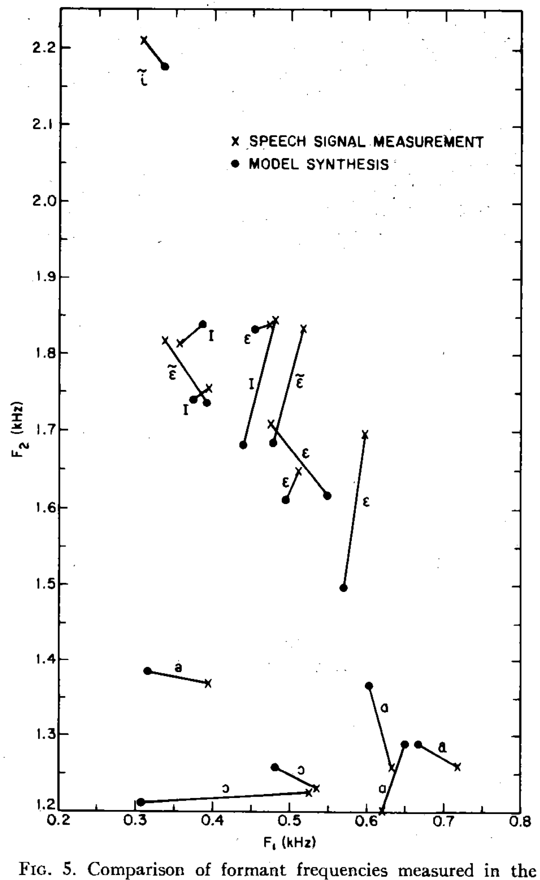


FIG. 5. Comparison of formant frequencies measured in the natural speech signal with the corresponding model-synthesized values.

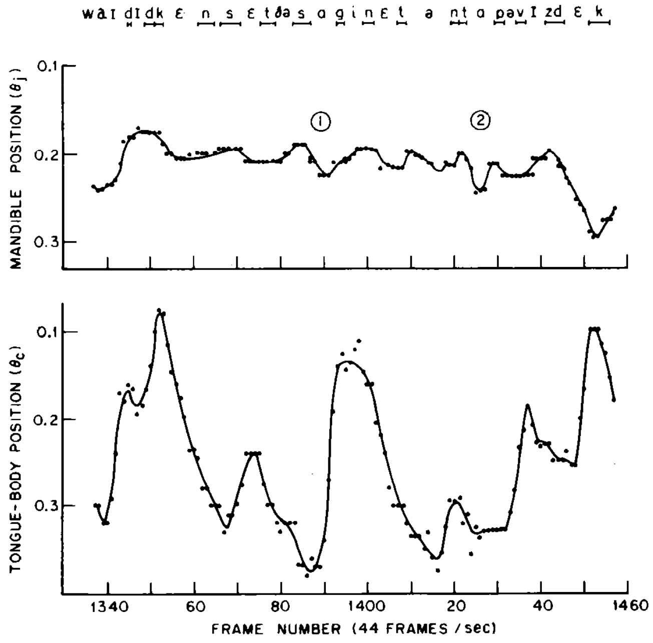


FIG. 6. Articulatory parameter trajectories for sentence "Why did Ken set the soggy net on top of his deck?"

namic picture. The range of jaw positions and jaw velocity are much smaller than those of the tongue even when both angular measures are converted to displacement values. The corresponding parameter values appear, in general, independently controlled yet carefully coordinated in time to yield a smoothly flowing speech output.

Eight [hə′CV] utterances have been analyzed in terms of the articulatory model variables. Tongue-body position is always determined relative to the jaw position, and this has enabled us to separate tongue-body movement from labial movement for the stops studied. The labial stop is executed through the action of jaw and lips, the alveolar and palatal stops with the aid of the jaw and tongue. Figure 7 illustrates tongue-body-center coordinate s_c variation in different environments. For [hə′pɛ] the tongue body moves smoothly between the vowel targets. For [hə′kɛ] it moves first to a relatively posterior target, advances anteriorly maintaining closure, and then moves rapidly to an appropriate stationary value. For the alveolar stop the tongue body appears to have a target that shows considerable variation with the succeeding vowel. Following the release, it undergoes an exponential-like transition to the final vowel with a characteristic time constant of roughly 75 msec. Study of the other variables shows that their release gestures can also be approximated with the aid of exponentials with time constant appropriate to the articulator; lips—30 msec, tongue tip—50 msec, and jaw—75 msec.

The vowel-transition gestures, i.e., those not involved in the stop production are smooth and possess a characteristic transition function (Houde, 1967, and our observations). The same transition function with a somewhat reduced peak velocity is appropriate for transitions during dipthongs as well (Kent, 1970). We approximate the vocalic transition function as

$$p(t)=\tfrac{1}{2}\{p(t_a)+p(t_b)+[p(t_a)-p(t_b)]\cos f(t)\}, \qquad (1)$$

where $f(t)$ increases linearly with time from 0 to π between the times t_a and t_b appropriate for vowel targets $p(t_a)$ and $p(t_b)$.

Figure 8 illustrates smoothed tongue-body center trajectories for the utterances [hə′tV] and the points

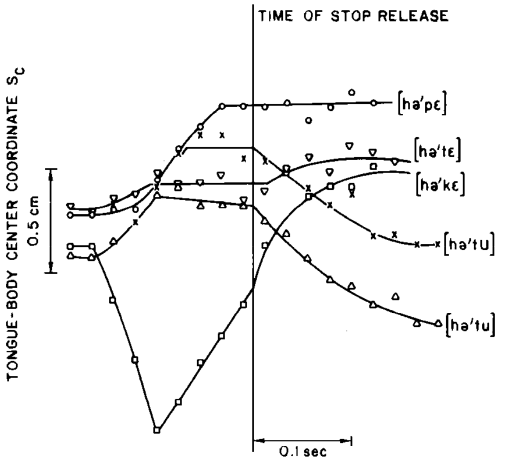


FIG. 7. Tongue-body center coordinate (s_c) trajectories for utterances [hə′CV].

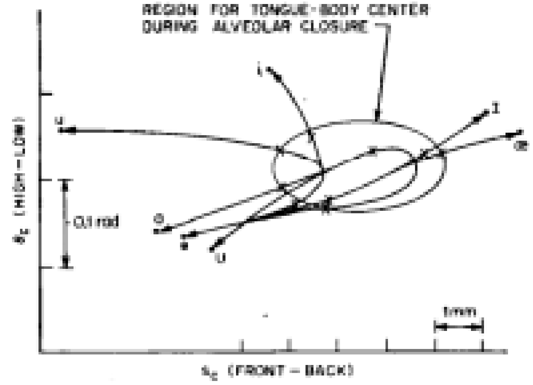


FIG. 8. Tongue-body center trajectories for utterances [hə'tV] in (s_c,θ_c) space.

on the trajectories delimiting the interval of tongue-tip closure. One observes that while tongue-tip closure is maintained the tongue-body position can take on values within a region of the (s_c,θ_c) space. Once the tip is released, the tongue body is apparently free to move to the position appropriate for the following vowel. During closure, the tongue body moves in the general direction of its ultimate target. The data appear to support a model that specifies that the vowel–vowel tongue-body trajectory is sufficiently deflected in the direction of a tongue-body target for an alveolar stop, so that tongue body is contained within a region consistent with tongue-tip closure.

The concept that an articulatory parameter is restricted to a range of values under certain conditions may be further generalized. The palatal stop /g/ may be articulated with a range of jaw opening values. Ohman (1966) gives examples for the Swedish utterances [ygy], [aga], and [ugu] where the jaw opening for the stop varies with that of the vowel context. We implement such variations with the aid of the following rule. Let $\theta_J^c \pm \Delta\theta_J$ be the range of jaw opening values permitted during closure. If $\theta_J^{V_1}$ and $\theta_J^{V_2}$ are the jaw opening values for the preceding and succeeding vowels, and $\alpha(t)$ controls the rate of articulatory change with time, the vowel–vowel transition is given by

$$\theta_J^1(t)=\theta_J^{V_1}+\alpha(t)(\theta_J^{V_2}-\theta_J^{V_1}).$$

Define the normalized distance from the mean consonantal value as

$$d=|\theta_J^1(t)-\theta_J^c|/\Delta\theta_J.$$

During closure, if $d>1$, set

$$\theta_J(t)=\theta_J^1(t)+[\theta_J^c-\theta_J^1(t)]\frac{d-1}{d},$$

otherwise

$$\theta_J(t)=\theta_J^1(t). \tag{2}$$

One may observe that explicit control of the jaw in this manner manifests itself in a particularly fluent type of coarticulation appropriate for equally stressed vowels. In the Perkell (1969) data, where the stops are in prestressed position, the jaw is seen to be stationary until the constriction is released.

VI. DYNAMIC ARTICULATORY MODEL FOR VCV PRODUCTION

As an initial step in exploring synthesis by rule with the aid of the above vocal-tract model, the results of our analysis were incorporated into control rules for articulator movement, and the resulting speech materials were evaluated. Observations based on the limited [hə'CV] context were generalized to hold for all VCV's.

For the production of every consonant, particular articulators are of primary importance. In general these are the articulators effecting closure. In Table II the corresponding variables are marked 1—primary pertinence. Other articulators contribute to a lesser extent. These variables are constrained to take on positions within a given range and marked 2, as being of secondary importance. Articulators not involved in the consonant production at all have their variables marked 3. In order to generate a vowel–consonant–vowel sequence, a planning program examines the articulatory trajectory between the preceding and succeeding vowel and determines points on the trajectory using Eq. 1 for the time values delimiting consonantal closure. These articulatory states represent the underlying vowel articulations at the given times and must be altered to satisfy the specified level 2 and level 1 constraints. If variables marked 2 are outside the permissible range, they are modified using Eqs. 2. Finally, variables marked 1 are moved in specified directions in variable space until closure relative to the fixed structures is attained.

For labial closure the lower lip height is set to one-half the maxilla–jaw separation previously determined—the upper and lower lips are assumed to move symmetrically. Since no explicit jaw opening target has been specified, and in fact the actual opening value is context dependent, the lips are made to execute whatever adjustment is necessary for closure—i.e., the variable is made to take on a value *a priori* not known. For the alveolar stop the tongue-body and jaw positions are adjusted first. With these variables set to their acceptable ranges no undue extension or contraction of the tongue blade

TABLE II. Pertinence of articulatory variables to the production of stops and nasals.

	Labial	Alveolar	Velar
Lip height h_l	1	3	3
Lip protrusion p_l	3	3	3
Jaw opening θ_J	2	2	2
Tongue tip s_t, θ_{t}	3	1	2
Tongue body s_c, θ_c	3	2	1

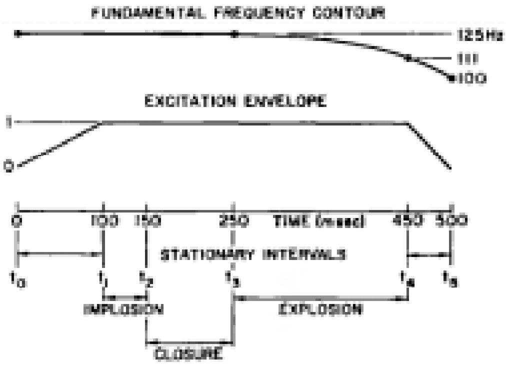


FIG. 9. Timing, excitation envelope, and fundamental frequency contour for synthesized VCV utterances.

is required to achieve closure. The tongue tip is moved in the direction of a target defined to lie above the alveolar ridge. Since closure takes place before the target is reached, the point of closure varies with the direction of approach. Following Henke (1967), a similar virtual tongue-body target is specified for the velar stop. The tongue body moves in the direction of the target—in this case in the fixed reference frame—until closure is reached. In the absence of further data, the required change in θ_{jc} is arbitrarily subdivided into equal changes in θ_j and θ_c. If necessary, the tongue blade is lowered to insure the absence of contact in the alveolar region.

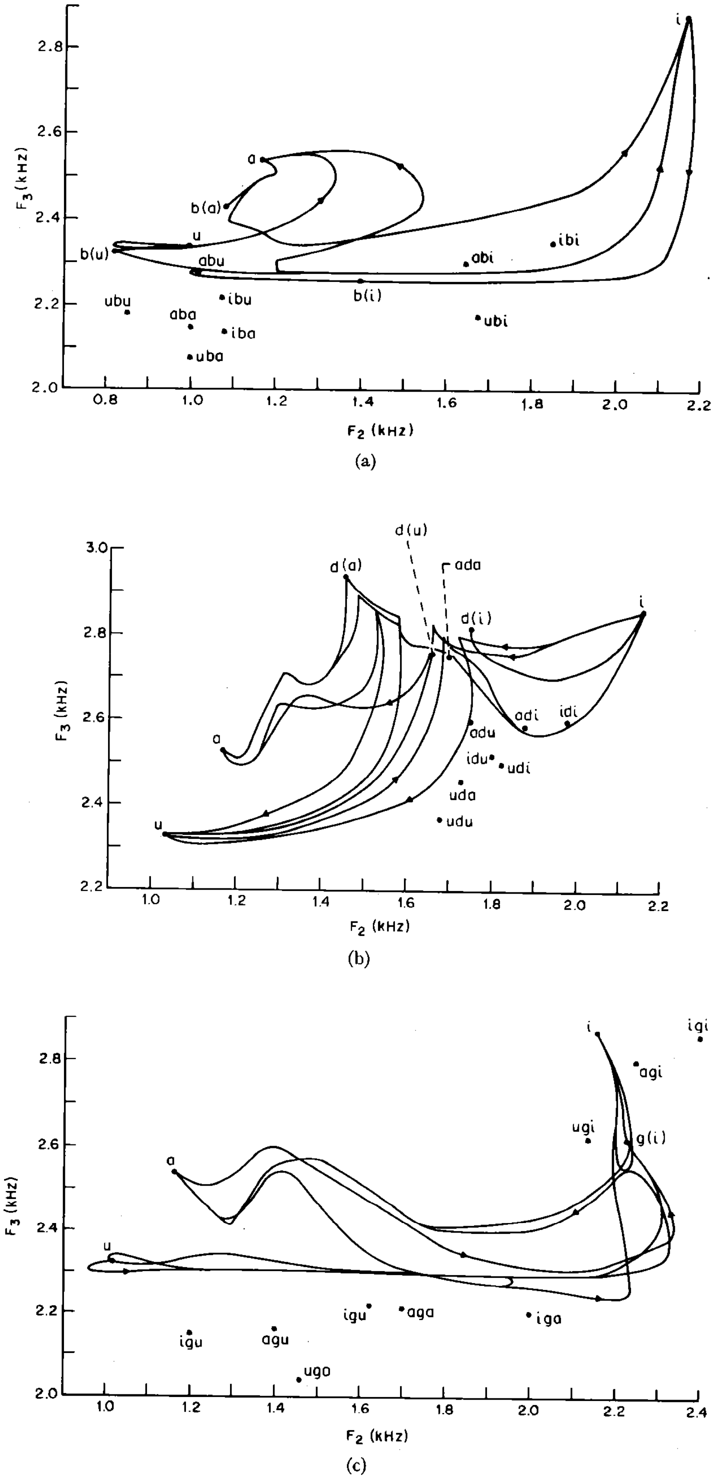
The time-course of events for the VCV stimuli synthesized was as shown in Fig. 9. Excitation envelope and fundamental frequency contour were set arbitrarily to provide smooth amplitude onset and decay, and acceptable intonation. All consonants were assigned an implosion interval of 50 msec, a closure interval of 100 msec, and an explosion interval of 200 msec, in rough agreement with the observed [hə'CV] productions. A set of reference vowels was defined in terms of the variables $\{s_c, \theta_c, \theta_j, \rho_t\}$, measured with respect to a stationary hyoid, rearpharyngeal wall and maxilla. As our rules assume those structures fixed, the variables measured for the stationary target vowels of the [hə'tV] utterances were corrected to preserve the relative spacing with respect to the fixed structures.

Execution of the planning program results in a sequence of partially specified articulatory states at the marked interval boundaries. Articulators of pertinence 3 are not explicitly specified by the planning program and are interpolated using Eq. 1 and $f(t)=\pi(t-t_1)/(t_4-t_1)$, $t_1<t<t_4$ (see Fig. 9—for the time definitions). The other variables are independently, and linearly interpolated between the determined values, except for the explosion interval where exponential transitions with time constants appropriate to the articulators are used. Thus the actual transition may in fact be practically completed long before the end of the given explosion interval is reached.

The structure allows for consonant cluster production as well, although this has not been implemented as yet. The timing of the consonants and therefore the selection of the underlying vowel articulation would be different. Articulators not pertinent to the production of any member of the cluster are free to move as determined by vowel context. Articulators not pertinent to the production of the initial member of the cluster are free to anticipate their future positions as long as they do not close the tract. Articulators not pertinent to the production of the remaining members of the cluster are free to move towards their vocalic values.

The articulatory data of Houde (1967) as well as the acoustic data of Öhman (1966) and Menon *et al.* (1971) indicate a directional symmetry between the [Vg] and the [gV] transition for the same V. The initial consonantal target, the articulatory state when closure is first attained, has a tongue-body position that is posterior to that of the final consonantal target where closure terminants. In acoustic terms, for symmetric vowel contexts the second formant frequency at the time of release exceeds that at the time of closure. Coker (1969) models this phenomenon by assigning a larger time constant and therefore slower movements to tongue-body movement in the horizontal direction than in the vertical. Houde's data show no such asymmetry for tongue-body movement in [VbV] context—transitions between vowel targets follow the same transition function irrespective of the orientation of the target difference. Apparently timing differentiation with direction of movement has to be suppressed when tongue-body closure is not involved. We follow Houde's formulation by applying a horizontal perturbation to a virtual target position for the tongue body and separate the consonant initial and consonant final targets by 7 mm. The effect of the separation is to differentiate between the closure directed and release directed transition directions in articulatory space—a differentiation consistent with, but not necessarily caused by the lack of precise antagonism between the participating muscles.

Figures 10a, 10b, and 10c illustrate the formant trajectories in the F2–F3 plane predicted by the VCV model for the voiced stops /b/, /d/, and /g/, respectively. For comparison, the formant frequencies for consonant targets preceding transition to the final vowels measured by Menon *et al.* (1971) are also given. Their formant frequency measurements are the averages from five English speakers uttering the same VCV sequence. Qualitative agreement is observed between the transition directions, although the actual target frequencies do differ considerably. Absolute differences may be due to vocal-tract size factors. The articulatory model allows tracking the formants through closure with the aid of the assumption that a very small cross-sectional area (say 0.01 cm^2) can replace the completely closed tract-section and thereby simulate the coupling between the separated parts. Ideally the measured formant



(a)

(b)

(c)

FIG. 10. Formant trajectories in F2-F3 space for model-generated VCV's. Marked points are corresponding consonant targets preceding the same final vowels measured in natural speech by Menon *et al.* (a) Labial stop /b/; (b) alveolar stop /d/; (c) palatal stop /g/.

targets would be located on these trajectories at times corresponding to the release of closure.

The vowel–consonant–vowel sequences were evaluated for consonant identification by generating the consonants /b/, /d/, /g/, /m/, and /n/ in all possible vowel contexts chosen from the set /i/, /a/, and /u/. Nasals were generated identically to stops, except for a lowered velum. Four subjects listened to 45 stimuli and reported a total of five errors in consonant identification, an error rate of 2.8%. None of the stimuli was missed more than once. The precise numerical results for such a limited variety of stimuli are of no great significance. They do demonstrate, however, the utility of simple articulatory rules in place of the complex acoustic rules that would be required to achieve similar results.

VII. CONCLUSIONS

In constructing a model that accurately reflects the behavior of the articulatory system and produces intelligible synthetic speech, one strives for a compromise between including the myriad of complexities observed in human behavior and the hopefully fewer sufficient details that contribute to the naturalness and intelligibility of the synthesized result. The model described here is sufficiently general to produce all English phonemes and the rules for VCV production illustrate its use in a dynamic environment. Although the data for the model are derived from only one speaker, the principles of movement differentiation according to articulator function on which the model is based are considered to be generally applicable.

The variables, when determined from an x-ray frame sequence, carry the requisite articulatory information for synthesis of the speech signal. In conjunction, excitation information indicating the presence or absence of glottal excitation of specified excitation frequency, plosion, or frication must also be supplied. These excitation control signals operate synchronously with the articulatory control mechanisms, indeed in many cases directly follow as a result of the articulatory trajectory.

The organization of independent vowel parameter values into continuous control signals reflects the contextual modification of stationary articulations when embedded in a dynamic environment. The generation VCV utterances illustrates a simple first step along this road. The parametric vocal-tract representation, since it characterizes the positions of physical articulators rather than, say, individual points along the vocal tract, possesses advantages of simplicity and compactness. It can be viewed as a language for description of articulatory phenomena.

Inspection of the midsagittal x-ray data in terms of the model reveals that otherwise independent articulators become constrained with respect to each other under certain conditions. For the final stationary vowels, jaw opening angle θ_j and tongue-body angle θ_c appear to be independent variables and must be separately specified. Some vowels are differentiated primarily by θ_j—/ɔ/ vs /a/—others by the pair (x_c,θ_c)—/a/ vs /U/. When executing the palatal (velar) stop /k/ the tongue body and jaw become constrained with respect to each other by the closure condition. Similar relationships hold for the other stops. Consonants are not assigned unique values for the articulatory variables—the actual values are determined by context.

The range of values introduced as appropriate for a variable in a particular context is a somewhat different point of view than that of targets used by earlier workers. Ohman (1967) computes an "ideal consonantal target" shape and considers deviation from it as due to coarticulation. For an arbitrary production, he requires exact agreement with the target over some parts of the tract, and over others he allows undershooting of the target to a variable extent. Implicit in the notion of target is that under certain conditions it will be attained. We broaden the concept to a target region that must be attained for the production. Through separate consideration of the individual articulator we may differentiate among those with exact targets (pertinence 1 in our discussion)—those with target regions (pertinence 2), and those with no target regions at all (pertinence 3).

Our work represents a reanalysis of Perkell's data using an articulatory model. One may ask whether the model incorporates Perkell's conclusions regarding an "extrinsic muscle system" producing the vowels and an "intrinsic muscle system" that assists the first in the production of consonants. We differentiate among the roles that the articulator may play in different contexts by assignment of different control rules for its variables. Our differentiation generally follows the same lines as Perkell. However, since the same articulator may be effective in producing either a consonant or a vowel, such as the tongue body, we see no logical necessity for assigning separate muscle systems for the articulator's control. The tongue body can move fast—65 msec for the implosion in [hə'kɛ]. This figure is about the same as for the tongue blade in [hə'dɛ]. However, the tongue body is normally slower than the tongue blade when executing a stop release. The difference for the same articulator may be due to the participation of different muscles, or due to different control over the same muscle. As an example of the latter, closure may be signaled by tactile feedback rather than proprioceptive sensory processes.

The dynamic model for VCV production is based on the following principles:

(1) The midsagittal vocal-tract outline is modeled in terms of nine selected variables describing the position of the participating articulators.

(2) Stationary vowels are represented in terms of four variables, two describing tongue-body position, one the jaw position, and one the lip position. Movement from vowel to vowel, expressed as changes in the variable values, is slow and precisely controlled.

(3) Representation of consonants requires additional control of tongue-tip elevation, lip height, and velar opening. Tongue body or jaw closure is specified by the variables pertinent to vowels.

(4) Consonants are not defined directly in terms of variable values but by constraints on articulator position relative to the fixed structures. Articulators independent of the specific constraints are free to take on positions independent of the consonant under production subject to the requirement that they do not otherwise constrict the vocal tract.

(5) Stop consonants are released by rapid movement of the constricting articulator.

ACKNOWLEDGMENTS

O. Fujimura supplied numerous valuable suggestions for the construction of the jaw-based tongue model and assisted with the preparation of the manuscript. The representation of the lips with reference to the jaw is due to S. Maeda. B. S. Atal contributed to the transfer-function representation of the vocal-tract used to generate a synthetic speech signal for the evaluation of the model.

J. D. Amerman, R. Daniloff, and K. L. Moll, "Lip and Jaw Coarticulation for the Phoneme /ae/," J. Speech Hear. Res. **13**, 147–161 (1970).

B. S. Atal and S. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," J. Acoust. Soc. Am. **50**, 637–655 (1971).

C. Coker and O. Fujimura, "Model for Specification of the Vocal-Tract Area Function," J. Acoust. Soc. Am. **40**, 1271 (1966).

C. Coker, unpublished memorandum (1969).

R. Daniloff and K. Moll, "Coarticulation of Lip Rounding," J. Speech Hear. Res. **11**, 707–721 (1968).

G. Fant, *Acoustic Theory of Speech Production*, (Mouton, s-Gravenhage, Netherlands, 1960).

J. M. Heinz and K. N. Stevens, "On the Derivation of Area Functions and Acoustic Spectra from Cineradiographic Films of Speech," J. Acoust. Soc. Am. **36**, 1037 (1964).

W. Henke, *Preliminaries to Speech Synthesis Based Upon an Articulatory Model*, preprints of 1967 Conference on Speech Communication and Processing, Office of Aerospace Research (United States Air Force, Cambridge, Mass., 1967).

R. A. Houde, "A Study of Tongue Body Motion During Selected Speech Sounds," Ph.D. thesis, University of Michigan, (1967).

R. D. Kent, *A Cinefluorographic-Spectrographic Investigation of the Component Gestures in Lingual Articulation* (University Microfilms, University of Iowa, Ames, Ia., 1970).

P. Ladefoged, J. Anthony, and D. Riley, "Direct Measurements of the Vocal Tract," J. Acoust. Soc. Am. **49**, 104 (1971).

B. E. F. Lindblom and J. E. F. Sundberg, "Acoustical Consequences of Lip, Tongue, Jaw, and Larynx Movement," J. Acoust. Soc. Am. **50**, 1166–1179 (1971).

S. Maeda, "Conversion of Midsagittal Dimensions to Vocal Tract Area Function," J. Acoust. Soc. Am. **51**, 88 (1972).

P. Mermelstein, S. Maeda, and O. Fujimura, "Description of Tongue Lip Movement in a Jaw-Based Coordinate System," J. Acoust. Soc. Am. **49**, 104 (1971).

P. Mermelstein, "Calculation of the Vocal-Tract Transfer Function for Speech Synthesis Applications," in Proc. Seventh International Congress on Acoustics (Akadémiai Kiadó, Budapest, 1971), Vol. 3, pp. 173–176.

P. Mermelstein, "Speech Synthesis with the Aid of a Recursive Filter Approximating the Transfer Function of the Nasalized Vocal Tract," in Proc. 1972 International Conference on Speech Communication and Processing, Boston, Mass. (1972).

K. M. N. Menon, P. V. S. Rao, and R. B. T. Thosar, "Perception of Stop Consonants," Proc. 7th Int. Congress on Acoustics (Akadëiai Kiadó, Budapest, 1971), Vol. 3, pp. 13–16.

S. E. G. Ohman, "Coarticulation in VCV Utterances; Spectrographic Measurements," J. Acoust. Soc. Am. **39**, 151–168 (1966).

S. E. G. Ohman, "Numerical Model of Coarticulation," J. Acoust. Soc. Am. **41**, 310–320 (1967).

J. S. Perkell, *Physiology of Speech Production: Results and Implications of a Quantitative Cineradiographic Study* (MIT Press, Cambridge, Mass., 1969).

M. M. Sondhi and B. S. Atal (personal communication) (1970).