

X. SPEECH COMMUNICATION*

Prof. M. Halle†
Prof. K. N. Stevens
Prof. J. B. Dennis
Dr. A. S. House

Dr. T. T. Sandel
Jane B. Arnold
P. T. Brady
O. Fujimura‡

M. H. L. Hecker
J. M. Heinz
D. L. Hogan
A. P. Paul

A. ANALYSIS OF NASALIZED VOWELS

In a previous report (1), a theoretical explanation of the spectral structure of nasalized vowels was given, and rules that govern the locations of formants and antiformants were described. An experimental study has been performed on vowels of this type, with the use of a computer program and speech materials reported on previously (2). The speech samples in this study consisted of the vowel portions of nonmeaningful words of the forms $[h\theta'mVm]$ and $[h\theta'nVn]$, in which the vowel V was one of $[i]$, $[ɪ]$, $[ε]$, $[æ]$, $[a]$, $[ʌ]$ and $[u]$. For each combination of the two consonants and seven vowels, uttered by a phonetically trained talker of American English (JM), the central portions of the vowels have been matched sample by sample (8.3 msec apart), in a manner similar to that described in a previous report (2) on nasal murmurs. Also, several analogous utterances by two other talkers have been studied by the same methods. The complete results of the study are being prepared for publication and will not be presented here. In general, the results compare very well with the theory previously advanced. In some cases, the vowel in the central portion of the utterance does not seem to be nasalized at all, and in such cases synthesized spectra specified by conjugate-pole pairs only give satisfactory matches with the vowel spectra. In many cases, on the other hand, it is necessary to introduce an additional pole-zero pair in order to obtain reasonable matches. (Later portions of the vowels are always nasalized; this indicates a strong anticipation of the final consonant.) The locations of the poles are, in general, well explained by the theory. The theoretical predictions are based on the observed location of the antiformant and the formant locations of the non-nasalized vowels. The latter were estimated from previous measurements of the pertinent vowel uttered by the same speaker in a similar word containing a voiced-stop consonant $[b]$ or $[d]$ in place of the nasal $[m]$ or $[n]$, respectively. The degree of nasalization, which is indicated by the location of the antiformant, seems to vary from utterance to utterance.

Some typical examples taken from the central portions of vowels are illustrated in

*This work was supported in part by the U. S. Air Force Command and Control Development Division under Contract AF19(604)-6102; and in part by National Science Foundation.

†On leave, 1960-61, as Guggenheim Fellow at the Center for Study of the Behavioral Sciences, Stanford University.

‡On leave from the Research Institute of Communication Science, University of Electro-Communications, Tokyo, Japan.

(X. SPEECH COMMUNICATION)

Fig. X-1. The consonant and the vowel of each word are identified by the phonetic symbols at the bottom of the figure. No effect of nasalization was found in the sample from the word [hə'nɪn]. As predicted by the theory, the shifted oral formants are generally

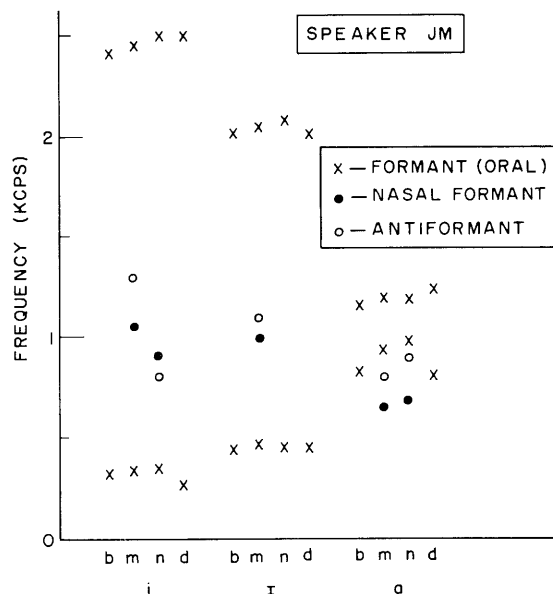


Fig. X-1. Frequencies of the formants and antiformants of three vowels in nasal and non-nasal contexts. A typical sample in the central portion of the vowel was taken from words of the form [hə'CVC]. The consonant C is given in the upper row under the graph, and the vowel V is given in the lower row.

slightly higher in frequency in nasal contexts than in the corresponding non-nasal contexts. When the first formant is low, as in [i] and [ɪ], the nasal formant often appears below the antiformant. This happens when the antiformant occurs at a frequency higher than a certain frequency (the singularity of the internal admittance, see Fujimura (1)) that is characteristic of the particular vocal-tract configuration. When the first formant is high, as in [a], a lower formant appears below the non-nasal first formant, and the first formant is shifted upward.

O. Fujimura

References

1. O. Fujimura, Spectra of nasalized vowels, Quarterly Progress Report No. 58, Research Laboratory of Electronics, M. I. T., July 15, 1960, pp. 214-218.
2. O. Fujimura, Analysis of nasal consonants, Quarterly Progress Report No. 60, Research Laboratory of Electronics, M. I. T., Jan. 15, 1961, pp. 184-188.

B. ANALYSIS OF THE GLIDES

The glides [w] and [j] are generated by glottal excitation of a time-variant non-nasalized vocal-tract configuration. The acoustic output for these consonants can be characterized, therefore, by a set of changing poles or resonances representing the transfer function of the vocal tract. The frequencies and bandwidths of the lowest three resonances have been traced through utterances in which the glides [w] and [j] occur, by using a computer analysis procedure that has been described previously (1). It has been noted elsewhere that this analysis procedure permits a more precise measurement of the resonant frequencies than that achievable by conventional spectrographic techniques.

The speech materials were drawn from an inventory of utterances prepared as part of a more general analysis program. The utterances containing examples of [j] consist of an unstressed syllable followed by a stressed syllable in which the initial consonant is [j]; an example is [hə'jæ]. The utterances containing examples of [w] are of the same form, except that the stressed syllable terminated in the consonant [ʒ]. Utterances with five different stressed vowels were studied: [iɪæʊu]. These utterances were generated by three talkers, although not all vowel environments were analyzed for each talker; detailed measurements were made for 18 of the words.

The computer technique was used to obtain the frequencies and bandwidths of the lowest three resonances as a function of time from the onset of voicing in the initial unstressed syllable to the end of the stressed vowel. On the average, measurements were made every 30 msec throughout these portions of the utterances.

An example of the data obtained for one utterance is displayed in graphical form in Fig. X-2. The formant frequencies for the word [hə'wæʒ] are plotted as a function of time, and the approximate bandwidths are indicated by the two lines drawn symmetrically about the curves representing the paths of the formants. This example shows a substantial broadening of the bandwidth of the second formant during the portion of the utterance in which the vocal tract is relatively constricted for the production of [w].

The data for the first two formant frequencies, as well as for [hə'jæ] by the same speaker, are displayed in a different form in Fig. X-3. The frequency of the second formant is plotted as a function of the frequency of the first formant, equal intervals of time being indicated by the points on the curves. The characteristic frequency locations for the "target" positions [w] and [j] are clearly evident from these curves, as are the frequency locations appropriate for the vowel [æ] and the motion of the formants toward a characteristic [ʒ] position in the word [hə'wæʒ].

Also shown in Fig. X-3 are three points, labeled [i], [u] and [æ], which represent values of the first- and second-formant frequencies averaged over a large number of utterances of these vowels by the same speaker in a variety of consonant contexts. The fact that the [w] and [j] target positions are displaced from the [u] and [i] positions

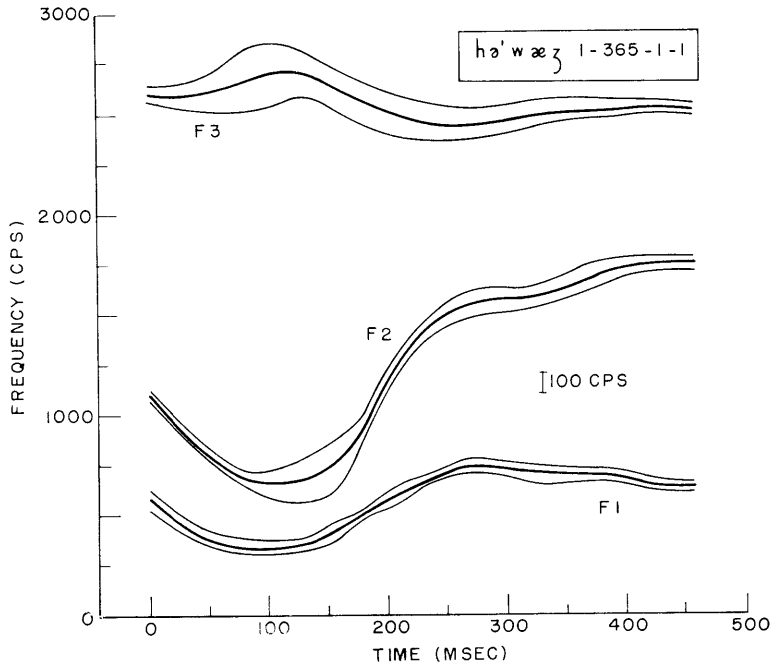


Fig. X-2. Formant frequencies and bandwidths as a function of time for a portion of the nonsense word as indicated; the talker is Subject JM. Zero time represents the beginning of voicing in the unstressed vowel; the curves terminate at the end of the stressed vowel.

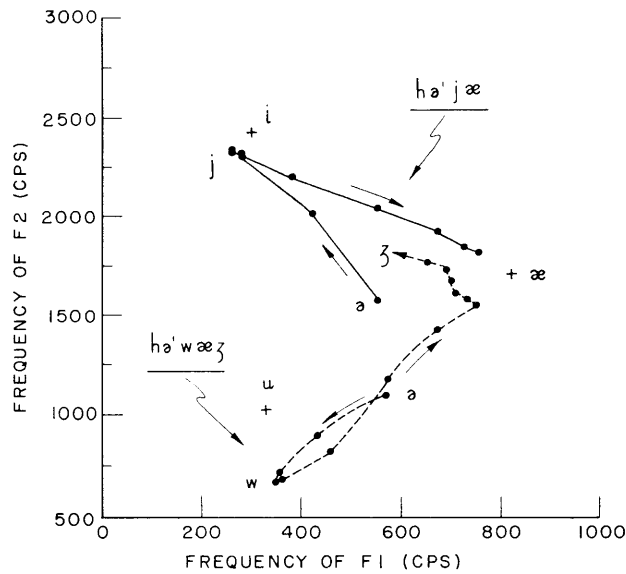


Fig. X-3. Plots showing variations in the frequencies of the first two formants during the course of two nonsense utterances by Subject JM. The filled circles represent measured samples 33 msec apart. The crosses represent values of the first- and second-formant frequencies of the vowels indicated, averaged over a number of utterances by the same talker in a variety of consonantal contexts.

Table X-1. Values of the first three formants of [j] and [w] followed by [æ] for three subjects. Measurements made in appropriate 'target' portions of the syllable. (All values in cps.)

	<u>JM</u>	<u>AH</u>	<u>KS</u>
j	250	290	300
	2340	2170	2150
	3200	3150	3200
w	350	380	310
	660	760	650
	2680	1950	1950

provides evidence that [w] is produced with greater lip rounding than [u], and that the vocal-tract configuration for [j] is more palatalized than that for [i].

In general, the formant frequencies at the target positions for [j] and [w] did not show a great deal of variation from one stressed vowel context to another. The variation was less than that observable in the formant frequencies for a vowel in different consonant contexts. Values of the frequencies of the first three formants for [j] and [u] followed by [æ] are given in Table X-1. These data fail to show a consistent rank-ordering of the formant frequencies from one talker to another. For these talkers, previous data on vowel formant frequencies have indicated a rank ordering from JM with the highest formant frequencies (and, presumably, the shortest vocal tract) to KS with the lowest formant frequencies (and, presumably, the longest vocal tract). Apparently, for the highly constricted or rounded configurations such as those for [j] and [w], the usual simple rules for scaling cannot be applied.

The measurements of formant bandwidths consistently show a broadening of F2 for [w], similar to that seen in Fig. X-2, and a broadening of either F2 or F3 (depending on the talker) for [j]. This increased damping is due, apparently, to the losses that occur in the highly constricted portions of the vocal tract.

One of the features that characterizes the glides is the timing of the vocal-tract movements from the target position for [w] or [j] to the approximate target position for the following vowel. Measurements of this duration can be made from plots of the type shown in Fig. X-3, in which the locations of the appropriate articulatory targets are reasonably evident. This duration was in the range 130-170 msec for all of the utterances studied.

K. N. Stevens, A. S. House

References

1. C. G. Bell and others, Quarterly Progress Report No. 57, Research Laboratory of Electronics, M. I. T., April 15, 1960, p. 121.

(X. SPEECH COMMUNICATION)

C. DYNAMIC ANALOG OF THE NASAL CAVITIES

Some design considerations and over-all characteristics of a dynamic analog of the nasal cavities (DANA) have been described previously (1). This analog, when operated in conjunction with the dynamic analog of the vocal tract (DAVO) and a common programming device (2), permits the synthesis of utterances that have nasalized portions. The performance of the composite electrical analog synthesizer has been evaluated in a series of formal and informal listening tests. For each of the nasal consonants [m] and [n], five consonant-vowel syllables were generated, using vocal-tract configurations appropriate for the vowels [iɪæʊ] and [u] during the vowel portion of the utterances. For the nasal consonant [ŋ], five vowel-consonant syllables were similarly generated with the same set of vowels. These syllables were produced with a variety of vocal-tract configurations having bilabial, postdental, and velar closures for the consonant portions of the utterances, and with appropriate timing patterns specifying the temporal course of such parameters as vocal-tract configuration, glottal excitation, and velopharyngeal coupling.

The first set of formal listening tests contained as stimuli the aforementioned syllables, generated with only two timing patterns corresponding to consonant-vowel and vowel-consonant articulations. The results indicate that for the consonant [m] those bilabial consonant configurations similar to the configuration for the following vowel were generally preferred to the other choices; for the consonant [n] the postdental consonant similar to the configuration for the vowel [i] was generally preferred; for the consonant [ŋ] those velar consonant configurations similar to the configuration for the preceding vowel were preferred. These findings are in general agreement with predictions based on phonetic theory.

In the second set of formal listening tests, the most suitable vocal-tract configurations for the nasal consonants, as determined by the previous tests, were used to generate the syllables [mæ], [næ] and [æŋ] with a number of different timing patterns representing reasonable deviations from the timing patterns previously employed. The results indicate that the most suitable timing patterns for the syllables [mæ] and [næ] differ only with respect to the rate of reduction in velopharyngeal coupling; a 150-msec transition in coupling is preferred for [mæ], and a 50-msec transition is preferred for [næ]. The most suitable timing patterns for the opening of the velum in the syllable [æŋ] require a transition in the coupling of approximately 150 msec.

Details on the construction of these tests and the interpretation of the results are recorded in the author's thesis (3) and are being prepared for publication.

M. H. L. Hecker

(References on following page)

References

1. M. H. L. Hecker, Dynamic analog of the nasal cavities, Quarterly Progress Report No. 60, Research Laboratory of Electronics, M. I. T., Jan. 15, 1961, pp. 189-191.
2. G. Rosen, Dynamic Analog Speech Synthesizer, Technical Report 353, Research Laboratory of Electronics, M. I. T., Feb. 10, 1960.
3. M. H. L. Hecker, Construction and Evaluation of a Dynamic Analog of the Nasal Cavities, S. M. thesis, Department of Electrical Engineering, M. I. T., May 20, 1961.

D. MOTION-PICTURE STUDIES OF ARTICULATORY MOVEMENTS

The application of the stroboscopic technique to a motion-picture study of the movements of the lips during speech production has been described in another report (1). In the earlier experiment, a frame rate of 240 per second was employed for the examination of labial articulations associated with bilabial stop and nasal consonants. A second series of experiments has been performed by using a frame rate of 60 per second to study the movements of the lips and the mandible during articulation of various nonlabial consonants. The speech material consisted of both meaningful and nonmeaningful words. In one part of the study, nonmeaningful words of the form [hə'CVC] were made by using each of the consonants [k], [t], [tʃ], [f], [θ], [s], [ʃ], [l] as C in combination with each of the vowels [i], [e], [æ], [a], [o], [u], [ʌ] as V. These words were uttered by two phonetically trained speakers of American English. In addition, one of the talkers uttered words of similar structure containing [g] and [d] as the consonants, and the other talker uttered the same stressed syllables without the initial unstressed syllable [hə], e. g., [gig], [dod]. In the second part of the present experiment, English monosyllabic words containing stop consonants in the initial and final positions were pronounced by the two talkers, as well as by a third speaker of American English who had no phonetic training. Words like "kick," "taught," etc., with symmetric structure, were used for comparison with the nonmeaningful words. Word pairs like "big"- "pick" were also spoken to see the effect of the tense-lax opposition of the consonants on the articulation of the stressed vowels. A third part of the inventory consisted of dissyllabic English words like "echo," "veto," "cookie," "okay," etc., in which significantly different labial articulations are required for the vowels preceding and following the medial consonant. Part of the data has been processed, and some of the results obtained concerning labial articulations of nonlabial stop consonants are described briefly.

Usually, a stop consonant is characterized by a feature that is primarily relevant to a part of the articulatory system. In the case of bilabial stops, for example, the lip closure is the primary feature, and other parts of the articulatory system seem to have some freedom of variation. This variation, which is due to coarticulation, sometimes results in substantial changes in the characteristic formant frequencies associated with

(X. SPEECH COMMUNICATION)

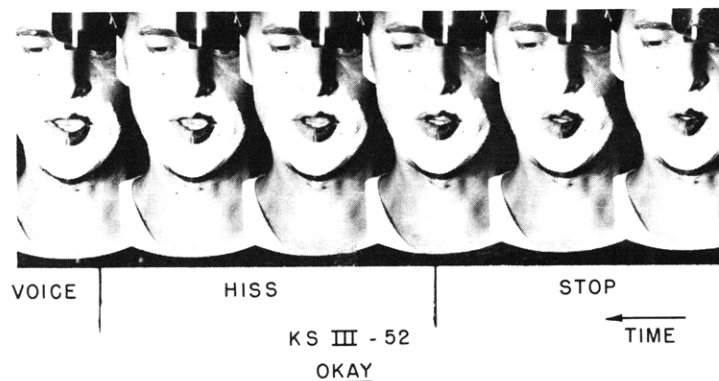


Fig. X-4. Full-face pictures during the production of a portion of the word "okay." Time runs from right to left, and the frame interval is 17 msec. The small dots at the left bottom are time-marking pulses which were simultaneously recorded on the film and the magnetic sound tape. Sound spectrograms were used to identify the changes in source characteristics for the [k].

the articulatory configuration. The primary purpose of this part of the experiment is to specify the nature and degree of the variation that can be deduced from the concomitant labial articulation during the production of nonlabial stop consonants.

In Fig. X-4, full-face frames obtained for an utterance of the word "okay" are shown. Time runs from right to left; the stop period of [k], the succeeding hiss portion, and the beginning of the voiced period of the second vowel [e] are indicated in the figure. The implosion of the stop from the preceding vowel [o] is not included in this illustration; it occurs approximately 2 frames (i. e., approximately 30 msec) before the frame appearing at the extreme right. It is seen that by the end of the stop period, the condition of the lips is considerably changed, and at the end of the hiss portion the lip condition is almost ready for the following spread vowel [e].

For the stop consonants [k] and [t] in a similar phonetic environment, a considerable and relatively rapid change in the labial articulation during closed intervals is always observed. The anticipation of the following vowel is generally strong, but the effect of the preceding articulation of a rounded vowel tends to persist for a short time after the implosion into the stop. Some utterances of the words "okay" and "cookie" showed minimum mouth-openings within the stop periods rather than in the vowel portions. In general, for a vowel with lip rounding, it seems that the narrowest labial constriction is articulated toward the end rather than in the central portion of the vocalic period. (This is true for nondiphthongized vowels.) Further experimental evidence pertinent to this point has been found in the data of monosyllabic words with symmetric phonemic structures ([gug], etc.).

Figure X-5 shows the horizontal and vertical dimensions of the mouth-opening for

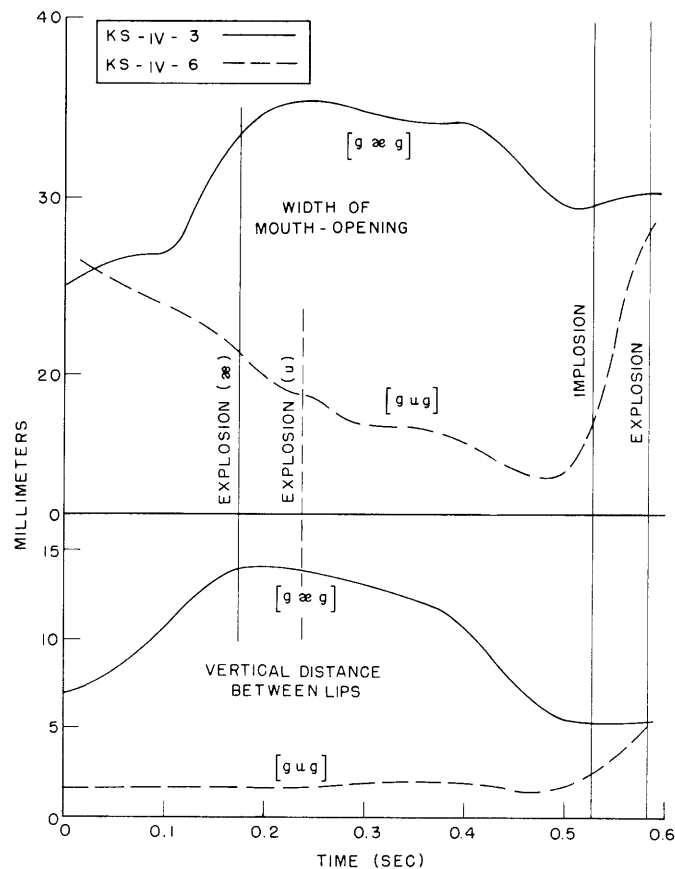


Fig. X-5. Horizontal and vertical dimensions of the mouth-opening for utterances of [gug] and [gæg]. (The closure period of the final stop, i. e., the interval between the implosion and the explosion, happened to be exactly the same for the two utterances.)

utterances of the words [gug] and [gæg]. The minimum mouth-opening for [u] occurs near the end of the vowel portion, whereas the maximum mouth-opening of [æ] takes place shortly after the explosion of the initial stop. In both words, during the explosion of the initial [g], the labial articulation is already highly characterized by the articulatory conditions for the following vowel. The labial condition at the explosion of the final stop, on the other hand, seems to be almost free from the influence of the preceding vowel articulation. At the end of the vocalic portion, that is, at the implosion into the final stop, the labial condition is still very much influenced by the vowel, and therefore a marked change is observed, in particular, in the width of the mouth-opening, during the closure period of the final [g].

In connection with the notion of the locus proposed by Delattre, Liberman, and Cooper (2) it is of particular interest to know the acoustic consequences of these various

(X. SPEECH COMMUNICATION)

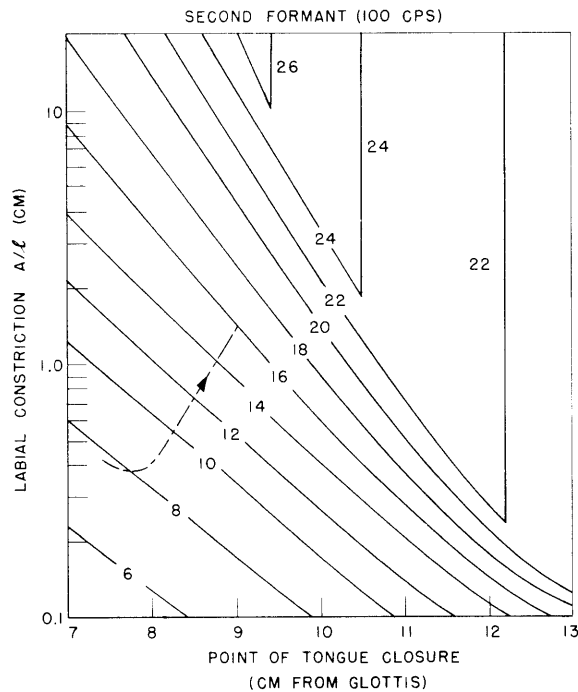


Fig. X-6. Contours of constant frequency of the second formant for various idealized articulatory conditions with a complete closure (replotted from Stevens and House (3)). The broken arrow indicates the estimated trajectory of the articulatory change during the closure period of the utterance of "okay" (see Figs. 1 and 4).

articulatory conditions during the stop periods. Estimations of formant frequencies can be made from data published of Stevens and House (3). In Fig. X-6 their data are replotted for the condition of complete closure at various points along the vocal tract. The location of the point of articulation is given in terms of the distance from the glottis (abscissa); the labial condition is given in terms of the ratio of the mouth-opening area A to the effective length l of the orifice (ordinate). Each curve represents a contour on which the second-formant frequency has the value indicated. If, for example, the mass of the tongue shifts from a retracted position toward a front position while keeping the closure characteristic of [k], and the lips, which were rounded for a back vowel, spread in anticipation of a front vowel, the formant frequency may rise rapidly. The path of the broken arrow in the figure is based on the observed mouth-opening condition for the stop period of the word "okay."

In Fig. X-7, which gives some data for this utterance, the mouth-opening area is represented by the upper solid curve. The area was estimated from the measured height and width of the mouth-opening. A simple assumption was made for the shift of the point of articulation during the stop period, and this assumption is shown in the middle

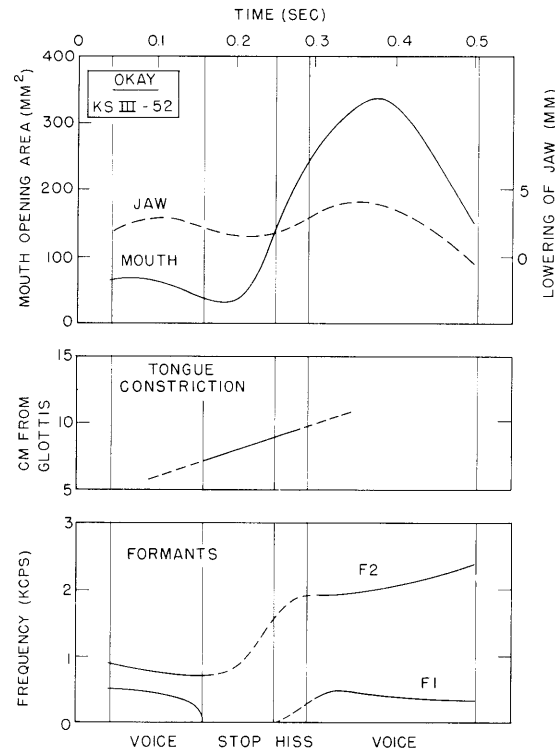


Fig. X-7. Measured data for the utterance "okay" (the upper section); a shift of the point of articulation assumed for the estimation of the formant frequencies (middle); and the estimated frequencies of the first and second formants F1 and F2 (bottom). The second formant for the stop period of [k] (broken line) and the first and second formants for the hiss period (dotted lines) were obtained by use of Fig. 3 and the original Stevens-House data, respectively. The curves were extended into the voiced part of the utterances by inspection of a spectrogram.

portion of the figure by a straight line connecting two points that represent the central parts of the two vowels. These central parts were located, in this case, by observing the maxima of the jaw angle (upper broken curve). Based on this crude model and also assuming a fixed length of the mouth constriction, the graph of Fig. X-6. was consulted to obtain an estimation of the second-formant frequency. The broken line in the F2 curve in the lower part of Fig. X-7 represents the estimated shift of the formant during the closure period. Similar estimations of the first and second formants were made for the hiss period by consulting the Stevens and House data, and these are represented by the dotted lines. The solid lines were extended for the voiced portions by inspecting the formant bars of the spectrogram of the utterance. According to this estimation, a shift of the second formant from approximately 800 cps to approximately 1600 cps takes place during the stop period. A substantial part of this shift is attributable to the change in the labial condition, as can be verified by inspecting Fig. X-6.

(X. SPEECH COMMUNICATION)

Under normal circumstances, the second formant is not affected by the labial condition during the closure period of [t] or [d]. In the region in which straight vertical contours appear in Fig. X-6, the formant represents a resonance of the cavity behind the closure, and theoretically there is no effect of the anterior constriction. The labial articulation changes considerably during this phonation, nonetheless, and this has an effect on the starting points of higher formant frequencies and on the second-formant transition as the area at the tongue constriction becomes larger.

In general, the data concerning stop consonants in various phonetic environments show that the changes in various physical dimensions of the articulators, such as the height and the width of the mouth-opening, the lowering of the mandible, etc., are not necessarily in synchronism. Nor, apparently, are these movements always related to the discontinuities in the excitation, such as the onset and cessation of voicing.

O. Fujimura

References

1. O. Fujimura, High-speed motion-picture study of the movement of the lips, Quarterly Progress Report No. 56, Research Laboratory of Electronics, M. I. T., Jan. 15, 1960, pp. 174-180; O. Fujimura, Bilabial stop and nasal consonants: A motion picture study and its acoustical implications, J. Speech Hearing Res. (in press).
2. P. C. Delattre, A. M. Liberman, and F. S. Cooper, Acoustic loci and transitional cues for consonants, J. Acoust. Soc. Am. 27, 769-773 (1955).
3. K. N. Stevens and A. S. House, Studies of formant transitions using a vocal tract analog, J. Acoust. Soc. Am. 28, 578-585 (1956).

E. AUTOMATIC RECOGNITION OF A LIMITED SET OF SYLLABLES

A method for the automatic recognition of utterances containing a limited number of phonemes is the subject of the author's thesis submitted to the Department of Electrical Engineering, M. I. T., June 1961, in partial fulfillment of the requirements for the degree of Electrical Engineer. The recognition procedure involves processing of the speech signal in two stages. The speech wave is first reduced to a representation in terms of signals with low information rate, and this simplified representation is then used as a basis for identifying a sequence of discrete symbols.

The low-information-rate signals at the output of the first stage describe the behavior of the vocal-tract resonances as a function of time and indicate the positions of vowel-consonant boundaries. The extraction of the resonances for vowels is accomplished by a spectrum-matching technique developed previously (1). This technique has been extended to provide approximate data on vocal-tract resonances during nonvowel portions of the utterances, and to provide a means for indicating boundaries between vowel

(X. SPEECH COMMUNICATION)

and consonantal regions. The vowel-consonant boundaries were established on the basis of both the over-all energy in the spectrum and the success with which a fit was obtained by using the spectrum-matching technique for vowels.

Fifty-six utterances by one speaker were used to test the adequacy of the spectrum-matching techniques and to serve as material from which to evolve rules for identification of phonemes from the resonances and segmentation data. These utterances included the eight vowels [i εæ ^] in six symmetrical consonant environments, using [bdgv z] and in the context [w-]. An iterative recognition scheme made tentative identification of the individual phonemes on the basis of spectrum-matching data in regions influenced primarily by those phonemes. Ambiguities in the preliminary identification were resolved on the basis of the identity of the adjacent phonemes and of a more detailed examination of resonance behavior at the vowel-consonant boundaries. The final step of the recognition scheme involved limited application of contextual constraints.

In order to test the adequacy of the recognition scheme, an additional group of 14 speech samples was obtained from the same speaker. Perfect identification was achieved on these samples. The analysis and testing were performed with the aid of the TX-O computer. Methods for refining and extending the recognition procedure are suggested in the thesis.

D. L. Hogan

References

1. H. Fujisaki, Vowel analysis, Quarterly Progress Report No. 58, Research Laboratory of Electronics, M. I. T., July 15, 1960, pp. 213-214.

F. ERRATA

In a report entitled "Some Synthesis Experiments on Stop Consonants in the Initial Position" by O. Fujimura, published in Quarterly Progress Report No. 61 (pages 158-162), the following errors and omissions should be corrected.

Page 161. In Figs. XI-2 and XI-3, onset of buzz is expressed in milleseconds.

Page 161. In Fig. XI-3, the curve labels should be reversed.

A. S. House

