XVIII. SPEECH COMMUNICATION*

Prof. M. Halle†	Dr. T. T. Sandel	H. Fujisaki
Prof. K. N. Stevens	Jane B. Arnold	M. H. L. Hecker
Prof. J. B. Dennis	P. T. Brady	J. M. Heinz
Dr. A. S. House	O. Fujimura‡	D. L. Hogan

RESEARCH OBJECTIVES

The objectives of our work are to further our understanding of (a) the process whereby human listeners decode an acoustic speech signal into a sequence of discrete linguistic symbols, such as phonemes; and (b) the process whereby human talkers encode a sequence of discrete linguistic symbols into an acoustic signal.

Current research activities related to these objectives include studies, aided by a digital computer, of methods of speech analysis and automatic speech recognition, accumulation of data on the acoustic characteristics of utterances corresponding to phonemes in various linguistic contexts, studies of the perception of speechlike sounds, measurements of the rapid changes in certain vocal-tract dimensions in syllabic utterances, and experiments on the generation of speech by electrical analog speech synthesizers.

K. N. Stevens, M. Halle

A. SPEECH ANALYSIS

1. ANALYSIS OF VOWEL SPECTRA

A method of speech analysis based on a principle that we call "analysis-by-synthesis" has been described previously (1). This general method has been implemented on the TX-0 computer (2) and in the past quarter has been used to measure vowel formants in diverse consonantal environments.

In brief, the computer programs are based on a conception of the speech-production mechanism between vocal folds and lips as a linear acoustic circuit excited by one or more sources. According to the theory of such linear circuits, a quasi-stationary speech spectrum can be expressed analytically as the product of a source function and a transfer function. The source spectra, or at least the envelopes of these spectra, are smooth and are relatively invariant from speaker to speaker and from speech sound to speech sound within a given phonetic class. The transfer function, on the other hand, depends on the articulatory configuration, and is determined completely if the locations of its poles and zeros in the complex frequency plane are known. According to the theory, therefore, it is possible to construct any speech spectrum from

*This research was supported in part by the U.S. Air Force (Air Force Cambridge Research Center, Air Research and Development Command) under Contract AF19(604)-6102; and in part by National Science Foundation.

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

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

information on the type of source and from knowledge of the poles and zeros of the articulatory configuration.

The speech materials for analysis consist of bisyllabic nonsense utterances of three adult males. Each nonsense word consists of an unstressed (carrier) syllable /hə/ followed by a stressed syllable having the form consonant$_1$-vowel-consonant$_2$ (where consonant$_1$ = consonant$_2$). Eight vowels, /iɪɛæɑʌʊu/, and 15 consonants, /pbtdkgfvθðszʃʧʤ/, were used in the materials.

The recorded speech materials were passed through a set of 36 simple-tuned filters covering the range 150-7000 cps. The filter outputs were rectified, smoothed, sampled every 8.3 msec, quantized by an analog-to-digital converter, expressed in 1-db steps, and stored in the computer. Subsequently the stored material was punched out in the form of perforated paper tape for future analysis. The TX-0 computer is used for these procedures, as well as the subsequent analysis.

The experimental procedures involved selecting three centrally located samples in each vowel by inspection of spectrograms, and comparing each of these spectra with synthetic spectra generated by the computer. The latter were specified by sets of poles whose locations were adjusted to yield the spectrum under analysis, or best fit with it. At the present stage of our research, formant frequencies for vowels can be determined within ±20 cps for the first formant and ±40 cps for the second formant. The derived data have proved to be systematic in a fashion hitherto unknown to the literature of experimental phonetics. The remainder of this report will illustrate some general trends shown by the measurements.

In Fig. XVIII-1, for example, values for eight vowels for each subject are shown

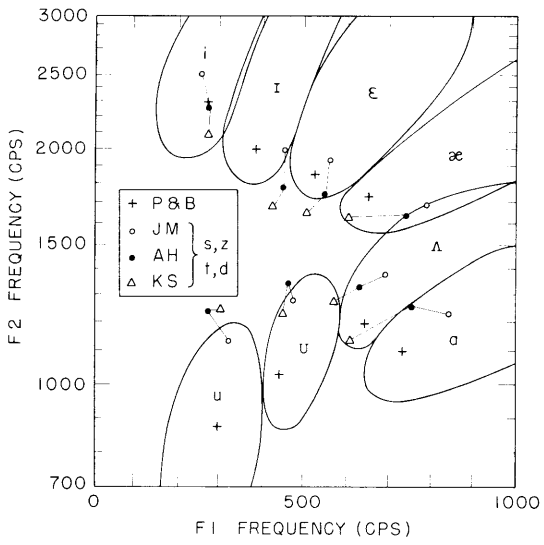

Fig. XVIII-1. Plots of average formant frequencies (F_1 versus F_2) for eight vowels.

separately, plotted in a typical manner. The vowel measurements were taken in the four phonetic environments /s-s/, /z-z/, /t-t/, /d-d/, and averaged. The solid lines represent the vowel areas reported by Peterson and Barney (3) and the crosses indicate those authors' average values for male subjects. This figure demonstrates the tendency of our three talkers to provide systematically differing data, as well as the further tendency for their averaged responses to move toward areas representing "neutral" articulations relative to the Peterson and Barney data.

The results of the complete study, of which the data in Fig. XVIII-1 represent a sample, will be described in detail in a paper that is being prepared for publication.

K. N. Stevens, H. Fujisaki, A. S. House

References

1. K. N. Stevens, J. Acoust. Soc. Am. 32, 47-55 (1960).

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

3. G. E. Peterson and H. L. Barney, J. Acoust. Soc. Am. 24, 175-184 (1952).

2. ANALYSIS OF VOWEL DURATION

The spectrograms of the words described in Section XVIII-A.1 include sampling pulses needed in the computer programs, and hence it is relatively easy to estimate the durational characteristics of speech sounds from these pulses. The durations of eight vowels produced by three talkers in a variety of symmetrical consonantal environments have accordingly been estimated. Measurements are made at the nearest sampling pulses, spaced 8.3 msec apart. Vowel onset and offset criteria included initiation and cessation of voicing and formant structure; aspiration was not included as part of the vowel. The measurements, therefore, represent information pertinent to changes in source excitation during syllable articulation, but do not necessarily reveal the point at

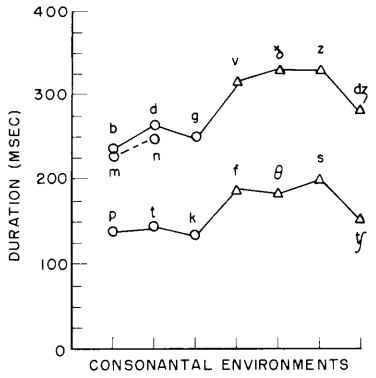

Fig. XVIII-2. Average durations of eight vowels in various symmetrical consonantal environments. Phonetic symbols indicate the consonant that preceded and followed each vowel. Data from 3 subjects; each point represents 24 vowels.

which a so-called vowel articulation per se began or ended.

When durations are averaged over vowels the influence of consonantal environment becomes apparent, as can be seen in Fig. XVIII-2. The vowels occurred in CVC syllables, such as /pip/, /pup/, etc. The measurements demonstrate that, when manner and place of articulation are held constant, vowels in voiced environments are longer than vowels in voiceless environments; this finding is true for all cognate pairs. In addition, Fig. XVIII-2 indicates that vowels in stop-consonant environments are, on the average, shorter than vowels in fricative environments, and that the influence of an affricative environment is more like that of a stop than a fricative environment. These general findings are in agreement with those reported earlier by House and Fairbanks (1).

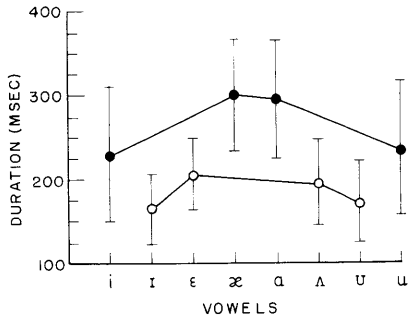

Fig. XVIII-3. Average durations of vowels in 14 consonantal environments. Open circles are "short" vowels; closed circles are "long" vowels. Vertical lines show variations in duration attributable to voiced-voiceless consonant cognate environments.

The measurements also demonstrate the characteristic durations of the various vowels in the vocabulary that is under study. Figure XVIII-3, for example, is a plot of the mean durations for eight vowels averaged over the consonant environments of Fig. XVIII-2 (excluding /m-m/ and /n-n/). The data are plotted in terms of the long and short vowels suggested by Peterson and Lehiste (2); the solid circles represent "long" vowels and the open circles represent "short" vowels. The vertical lines indicate the variation in duration associated with voicing; the highest point is the average value in a voiced consonant environment and the lowest point is the average value in a voiceless consonant environment.

These data constitute further demonstration of the importance of temporal information in the specification of linguistic units such as phonemes, and indicate, furthermore, that the influence on vowel duration is a function of phonetic environment.

A. S. House

References

1. A. S. House and G. Fairbanks, J. Acoust. Soc. Am. 22, 457-459 (1950).
2. G. E. Peterson and I. Lehiste, J. Acoust. Soc. Am. 32, 693-703 (1960).

3. ANALYSIS OF FRICATIVE CONSONANTS

Spectral curves that closely resemble the spectra of spoken fricative consonants were generated on the basis of an acoustical theory of speech production (1). A computer program with manual control described previously (2) was used to generate the spectral curves and to obtain comparisons with the speech data. The speech materials were taken from those described in Section XVIII-A. 1.

The approach used was based on the fact that the poles that characterize the spectrum of a fricative consonant are the same as those that characterize a vowel produced with the same articulation. Since the articulatory configuration changes continuously and relatively slowly with time during connected speech, it should be possible to locate the frequency positions of the poles in the spectrum of a fricative consonant adjacent to one side of a boundary between regions characterized by noiselike and periodic excitation by the relatively easier task of finding the poles of the vowel just to the other side of that boundary. Once the poles are known, the zeros may be added in a manner consistent with an acoustical model of production in order to obtain the desired spectral curve. After the pole-zero pattern for the fricative is found in the transition region, the changes in this pattern can be traced back into the "steady-state" region.

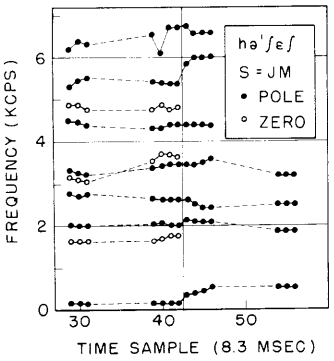

Fig. XVIII-4. Pole-zero patterns obtained from spectral matches over a portion of the initial /ʃ/ and /ɛ/ of the word /həˈʃɛʃ/. The vertical line in the middle of the graph marks the location of consonant-vowel boundary. The abscissa identifies sample numbers indexed from the beginning of the utterance.

Figure XVIII-4 illustrates pole-zero patterns obtained from spectral matches over a portion of the word /həˈʃɛʃ/. The vertical line in the middle of the figure indicates the location of the boundary between the initial /ʃ/ and the /ɛ/, the consonant being at the left of the boundary and the vowel at the right. The first seven formants were traced through a time interval of 33 msec on both sides of the boundary. Separate spectral matches were obtained for samples 8.3 msec apart. Three zeros were needed to obtain acceptable matches in the fricative portion of the word. Not shown in the figure are a factor needed to correct for the effect on the spectrum of poles and zeros located outside the frequency range considered here, and three real-axis zeros near the origin; it is

necessary to insert the real-axis zeros in order to simulate a flat source spectrum. Figure XVIII-4 also shows pole-zero patterns for three samples in the steady-state portion of the fricative and for the first four formants in the vowel steady-state region. Some of the poles and zeros in the fricative spectrum occur close together and it would be expected that such pairs would not have much effect on the spectrum. On the other hand, poles and zeros that are widely separated should exhibit a prominent effect on the corresponding spectrum.

Figure XVIII-5 shows two spectral matches taken 8.3 msec apart. The spectra of Fig. XVIII-5a are taken from the vowel side of the CV boundary, while the spectra of Fig. XVIII-5b are from the fricative side. The points lying near the 0-db line show the difference between the calculated curves and the speech spectra, which in this case is never more than 3 db. It should be noted that the speech spectra shown here have all been given a 6 db/octave pre-emphasis. The poles of each spectrum are approximately the same, but the spectral shape differs in the case of the fricative because it is modified by the presence of zeros. Upon close examination, it is possible to find peaks corresponding to each of the poles in the two spectra, but different peaks are prominent in each case. In the fricative spectrum the first formant has dropped somewhat in frequency so that its effect is partially canceled by the real-axis zeros near the origin. In addition, the first formant is highly damped and so has almost disappeared from the spectrum. The second formant is almost canceled by the first zero, while the third formant at 2600 cps stands out prominently, a characteristic of /ʃ/ spectra in general. In this example, the fourth and fifth formants are fairly well canceled by zeros and the higher formants do not contribute significantly. In many spectra, however, there is another prominent peak in addition to the third formant.

Shown in Fig. XVIII-6a is a sample spectrum taken in the "steady-state" portion of /ʃ/ in the word /hə'ʃɛʃ/ (sample 31 of Fig. XVIII-4), together with the matching spectrum. Again, the first prominent peak is due to the third formant and occurs at 2700 cps. In this case, there is another prominent peak at 5500 cps corresponding to the sixth formant. The fifth formant can also be seen at 4350 cps.

An example of a spectrum obtained for the initial /s/ in the word /hə'sɛs/ is shown in Fig. XVIII-6b. All of the lower formants are fairly well canceled by zeros and the first major peak is due to the sixth formant at 5800 cps. The seventh formant has moved just beyond the end of the graph and may also be significant in some cases.

Figure XVIII-7 shows the pole-zero patterns as a function of time; again the method of moving from the vowel back into the consonant has been employed. The figure points out the close association of each of the lower formants with a zero. This results in cancellation of all of the lower formants by zeros and is probably attributable to the very narrow constriction characteristic of /s/-articulation. Considered somewhat differently, since the articulatory constriction is quite narrow, there is very little coupling

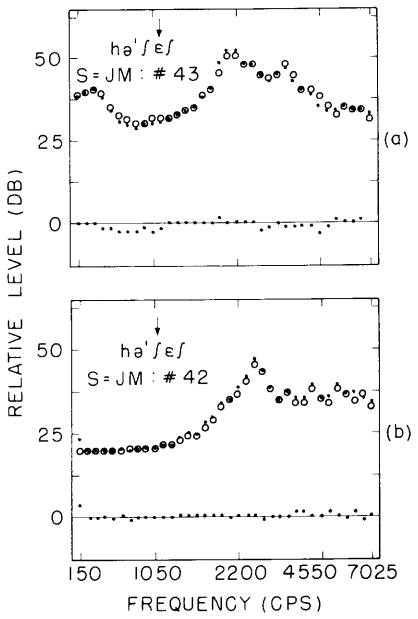

Fig. XVIII-5. Comparisons of speech spectral data for the word /hə'ʃɛʃ/ (small dots) with corresponding calculated points (open circles): (a) sample immediately on the vowel side of the ʃ-ɛ boundary (sample 43 in Fig. XVIII-4); and (b) sample taken 8.3 msec earlier, immediately on consonant side of the ʃ-ɛ boundary (sample 42 in Fig. XVIII-4). The points near the 0-db line give the difference between the speech data and the calculated points.

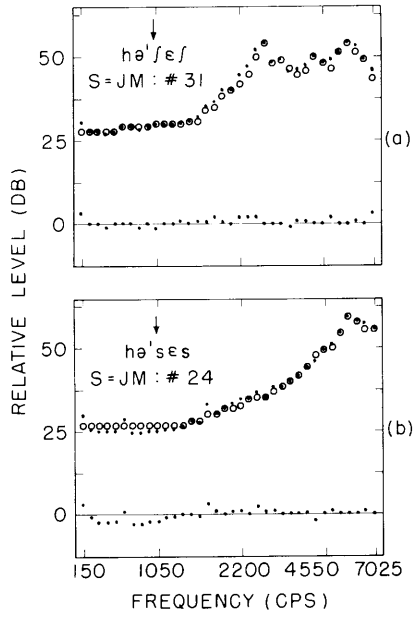

Fig. XVIII-6. Comparisons of speech spectral data (small dots) with corresponding calculated points (open circles): (a) sample in the middle of the initial /ʃ/ in the word /hə'ʃɛʃ/ (sample 31 of Fig. XVIII-4); and (b) sample in the middle of the initial /s/ in the word /hə'sɛs/. The points near the 0-db line give the difference between the speech data and the calculated points.

into the back cavity, and therefore one would not expect any formant for which the back cavity is important to be prominent. The lower formants are all dependent on the back cavity and so do not show up as peaks in the spectrum.

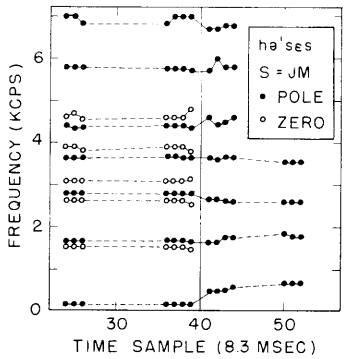

Fig. XVIII-7. Pole-zero patterns obtained from spectral matches over a portion of the initial /s/ and /ɛ/ of the word /hə'sɛs/. Sample 40 in the middle of the graph marks the location of the consonant-vowel boundary.

Below 5000 cps, all poles and zeros in Fig. XVIII-7 occur in pairs except for the "unpaired" zero around 3100 cps. Above 5000 cps, there are two poles that are not paired with zeros, one around 5800 cps and one around 7000 cps. A good approximation to the spectrum can be realized by using only the unpaired zero and the two poles above 5000 cps, since each pole-zero pair contributes little to the spectrum. The same result can be obtained from an approximation in the articulatory domain that assumes no coupling to the back cavity.

Similar results have been obtained with spectra of /f/. In the case of an /f/ spectrum, each pole is closely associated with a zero, and the resulting spectrum has no major peaks within the frequency range considered here.

J. M. Heinz

References

1. C. G. M. Fant, Acoustic Theory of Speech Production (Mouton and Company, 's-Gravenhage, in press).

2. C. G. Bell, H. Fujisaki, J. M. Heinz, A. S. House, and K. N. Stevens, Speech analysis, Quarterly Progress Report No. 57, Research Laboratory of Electronics, M.I.T., April 15, 1960, p. 121.

4. ANALYSIS OF NASAL CONSONANTS

The acoustic characteristics of nasal consonants have been investigated by use of the computer program which has been described previously (1, 2). The experimenter can

specify the frequencies and bandwidths of any number of poles and zeros by means of a flexowriter connected to the TX-0 computer. The computer calculates the output spectrum and the experimenter uses the oscilloscope display to determine the adequacy of the match. As speech samples, meaningless words of the form /hə'mVm/, /hə'nVn/, and /hə'rVŋ/ with various vowels V were drawn from our library of punched tapes. Spectral samples taken at 8.3-msec intervals throughout the intervocalic /m/ and /n/ were matched with calculated spectra, and the locations of poles and zeros that gave satisfactory matches were recorded. For each of these nasals, five or six words containing different vowels were matched for one speaker, as well as a few each for two other speakers. A similar method was used for /ŋ/ uttered in various vowel contexts by the three speakers and also for some samples of final /m/ and /n/.

Figure XVIII-8 illustrates some examples of typical matches that were obtained, as photographed from the oscilloscope display of the TX-0 computer. The pattern of poles and zeros required to achieve these matches is in good agreement with theoretical formulations discussed below. Figure XVIII-8a shows a spectral sample approximately 35 msec after the beginning of the first nasal /m/ of /hə'mim/, spoken by KS. Frequency from 150 cps to 3000 cps is plotted on the abscissa, and the ordinate is marked off in 5-db steps. The larger dots represent the calculated values of the outputs of the 24 filters in this frequency region, and the smaller dots give the measured values. The curve along the abscissa represents the difference of the two values. One way of evaluating the accuracy of matching is in terms of the sum of the squared difference values, which is 23 $(db)^2$ in this case, as can be seen at the left margin of the figure. (All other numbers in the column can be ignored for the present purpose.)

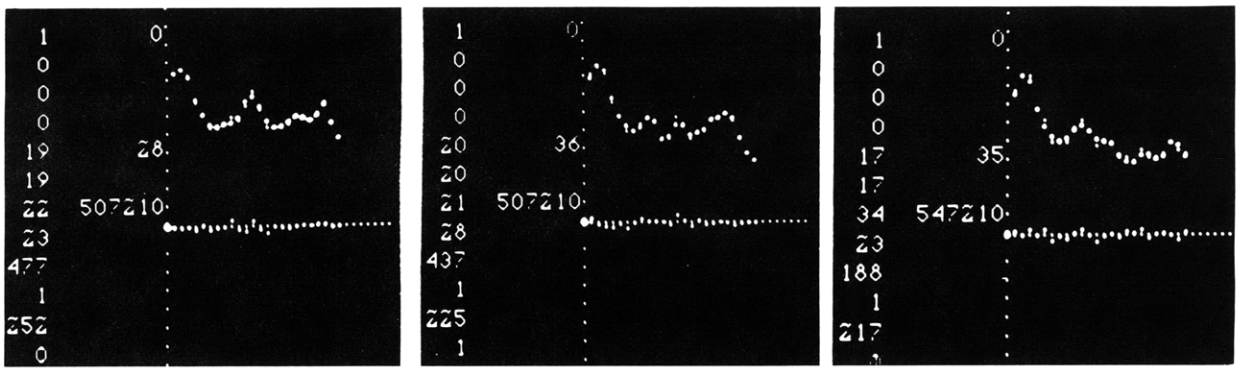

Fig. XVIII-8. Examples of the TX-0 computer oscilloscope display, by which the experimenter judges the accuracy of the match: (a) from the first /m/ of /hə'mim/, spoken by KS, approximately 35 msec after the beginning of the nasal; (b) from the nasal, approximately 65 msec after the previous sample; and (c) from the first /n/ of /hə'nɑn/, spoken by KS.

Figure XVIII-8b shows the result of matching for another sample within the same consonant. This sample was taken 65 msec after the previous one, and the /m/ ends about 40 msec after this sample. The sum of the squared differences is 28 $(db)^2$ in this case. Figure XVIII-8c shows an example of the spectrum of /n/. This sample was taken from the first /n/ of /hə'nɑn/, spoken by the same subject.

In order to perform this analysis-by-synthesis process effectively, it is important to have some knowledge about the distribution of the poles and zeros within a given frequency region. This prediction can be supplied by a simple evaluation of the dimensions of the acoustic tubes and through a graphical estimation of the transfer characteristics. The graphical estimation can be done in a manner similar to that reported previously in the discussion of nasalized vowels (3, 4). In the case of nasal consonants with mouth closure, the pharynx and the nasal passage can be regarded as the main tract, and the mouth cavity can be considered as a coupled side branch. Thus for nasals, the internal susceptance of the main tract, observed at the coupling point (velum), will be compared with the driving-point susceptance of the mouth cavity in order to determine the natural frequencies of the coupled system. Figure XVIII-9 shows an example of plots of these susceptances. The solid lines represent the sum of the susceptances looking into the pharynx and the nasal pharynx, respectively; the broken lines represent the negative of the susceptance looking into the mouth cavity. The poles of the combined system (closed circles) are given by the intersections of the two kinds of curves, whereas the poles for the uncoupled system (arrows) are given by the intersections of the solid curves with the abscissa (as in the case of /ŋ/, approximately).

In the example illustrated in Fig. XVIII-9 which is typical of the /n/ configuration, the antiresonance occurs at 1700 cps. It is seen that by introducing the coupling the two lowest formants are shifted slightly downwards. The distribution of poles is perturbed considerably in the vicinity of the antiresonance. In the case of /n/ it appears that the third pole of the uncoupled system is replaced by a group of one zero and two poles, while the other poles are displaced only slightly.

In the case of /m/, the first antiresonance occurs at a lower frequency, generally in the vicinity of the second pole of the uncoupled system. The lowest pole and the third pole of the uncoupled system remain relatively constant. The frequency region above 2300 cps may be appreciably influenced by the second antiresonance in this case.

Some of the data obtained in this experiment are illustrated in Fig. XVIII-10. For each of the words containing different vowels, three spectral samples were selected to represent the mouth-closure period of the intervocalic nasal consonant. The first of each three points represents the second spectral sample within this interval, the second point is the mid-point, and the third point is the next to the last sample in the interval. The closed circles represent the poles (formants), and the open circles represent the zero (antiformant). For each of the intervocalic nasals spoken by one subject (KS), four

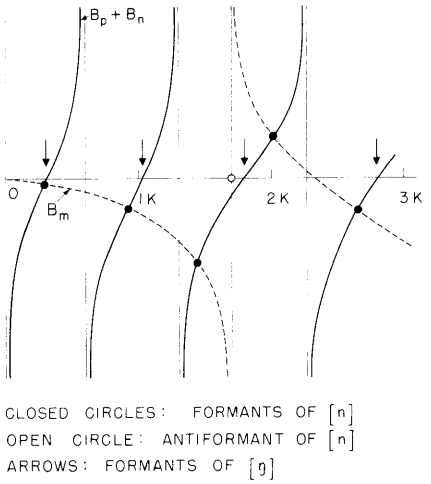

Fig. XVIII-9. Schematic susceptance curves for the evaluation of the pole-zero distribution for /n/. The curve labeled B_m is the negative of the driving-point susceptance of the mouth cavity measured at the coupling point; the curve labeled $B_p + B_n$ is the sum of the driving-point susceptances looking into the pharynx and the nasal passage, respectively. Closed circles give the formant frequencies of /n/, and the open circle gives the antiresonance frequency. Arrows indicate the formant frequencies for /ŋ/, for which B_m can almost be replaced by the abscissa in this frequency region.

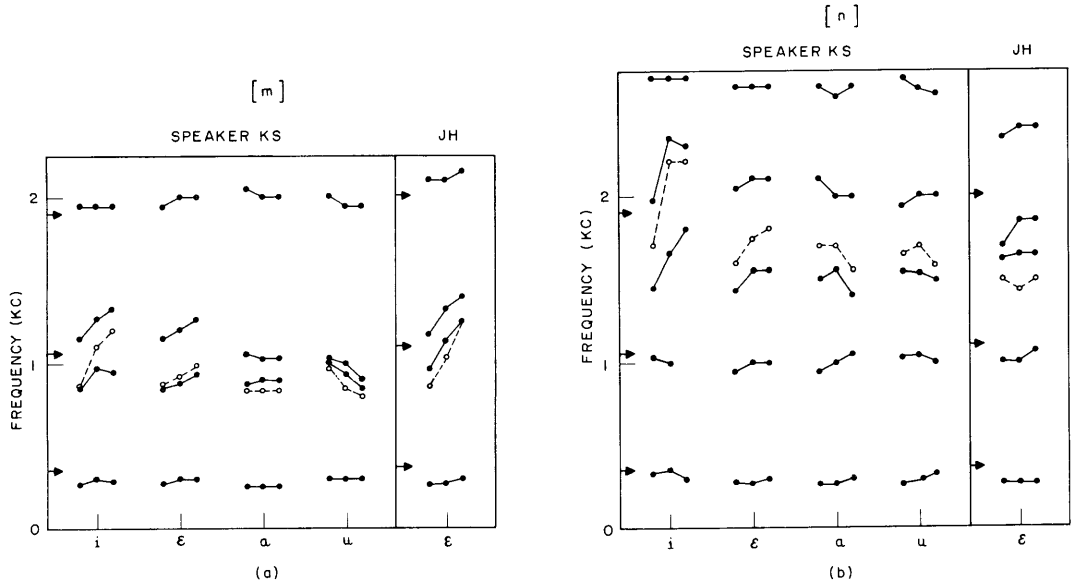

Fig. XVIII-10. Summary of the data of the pole-zero frequency distribution for intervocalic /m/ and /n/. The three points for each curve are spectral samples near the beginning, middle, and end of the consonant, respectively. The vowel symbols identify the vowels that were contained in the sample words of the form /hə'mVm/ or /hə'nVn/. The arrows show the typical locations of the formants of /ŋ/ spoken by the subjects.

words containing a variety of vowels are included in the illustration. One word spoken by another subject (JM) is shown for each of the nasals. A typical distribution of the poles of the final /ŋ/ for the pertinent speaker is indicated by arrows.

It is seen in these illustrations that the variation of the locations of the poles from one vowel environment to another is significant only within the limited frequency region in which the antiresonance occurs. In this region, the distribution of the two poles and one zero appears to be highly susceptible to the change of the articulatory configuration that occurs in the consonant as a result of the influence of the adjacent vowel. It is evident, when we compare the data for /m/ to those for /n/, that the location of this particular frequency region distinguishes /m/ from /n/ or /ŋ/. The location of the antiresonance, consequently, can be regarded as the most important factor that separates the nasals within the class, as far as the nasal murmur is concerned. On the other hand, the precise structure of the pole-zero distribution, including the order of the two poles and one zero along the frequency axis, depends on the particular articulatory anatomical conditions in detail. The data obtained for the three male speakers also indicate a considerable personal variation in the frequency region above 2000 cps.

It was observed that the formants of the nasals in general have wider bandwidths compared with those of vowels, but the bandwidth is not a simple function of frequency. For all of the three subjects, the fourth formant of /ŋ/ had a relatively narrow bandwidth (typically 100 cps to 150 cps), whereas the third formant showed a wider bandwidth (200 cps to 300 cps, or more). The corresponding formants in /m/ and /n/ showed the same tendency. This fact can be explained by assuming different amounts of dissipation for different parts of the acoustic system. The bandwidth of the antiresonance is much wider for /n/ (typically 500 cps) than for /m/ (mostly 100 cps, or less). The higher damping of the antiresonance of /n/ can be attributed to the wedgelike shape of the front end of the mouth cavity which gives rise to a better matched termination to the resonance tube.

O. Fujimura

References

1. C. G. Bell, J. M. Heinz, G. Rosen, and K. N. Stevens, Automatic resolution of speech spectra into elemental spectra, Quarterly Progress Report No. 54, Research Laboratory of Electronics, M.I.T., July 15, 1959, pp. 161-167.

2. C. G. Bell, H. Fujisaki, J. M. Heinz, A. S. House, and K. N. Stevens, Speech analysis, Quarterly Progress Report No. 57, Research Laboratory of Electronics, M.I.T., April 15, 1960, p. 121.

3. O. Fujimura, Spectra of nasalized vowels, Quarterly Progress Report No. 58, Research Laboratory of Electronics, M.I.T., July 15, 1960, pp. 214-218.

4. C. G. M. Fant, Acoustic Theory of Speech Production (Mouton and Company, 's-Gravenhage, in press).

B. DYNAMIC ANALOG OF THE NASAL CAVITIES

Acceptable nasality in synthetic speech generated by articulatory analog equipment may be achieved by appropriately coupling a circuit representing the nasal cavities to the circuit representing the vocal tract proper. From the point of view of articulation, movements of the velum must occur roughly in synchronism with changes in vocal tract configuration, i.e., the degree of velopharyngeal coupling must be a function of time. Through the use of a dynamic analog of the nasal cavities (DANA), in conjunction with the existing dynamic analog of the vocal tract (DAVO) and its associated programming device, many aspects of nasality can be studied in greater detail than was previously possible.

The geometrical dimensions and damping characteristics of DANA are based on the static nasal analog described by House (1). DANA consists of nine electrical π-sections. No section exceeds 1.5 cm in length, so that the analog is valid for frequencies up to 3 or 4 kcps. Four sections are variable; the first two sections, representing the nasopharynx, have an electronically variable cross-section area, while the areas of the third and eighth section can each be manually adjusted to one of 5 values. The remaining sections are of fixed dimensions. The approximate acoustical representation of DANA is shown in Fig. XVIII-11.

Electrically, the first two sections make up a single chassis which incorporates dynamically variable inductive and capacitive elements – two saturable reactor units and two Miller-effect amplifier-attenuator units (2). Sections 3-9 are housed in a second chassis and built with the use of specially prepared precision elements. The damping of the nasal cavities is distributed over the entire length of the analog model. The resistive losses inherent in the dynamically variable elements represent damping in the region of the velopharyngeal musculature, while damping in the remainder of the nasal tract is realized by means of appropriate plug-in resistors. Two rotary switches control the representative cross-section areas of sections 3 and 8.

A third chassis contains the control circuits for the bisectional unit. These include inverter stages for obtaining the control voltages for the Miller-effect amplifier-attenuators and the saturable reactors from a single-input voltage, with provisions for adjusting the tracking, and nonlinear current driver stages for the saturable reactors. A trapezoidal waveform, obtained directly from a function generator of the DAVO programming device, drives the DANA control circuit and thus specifies the degree of velopharyngeal coupling. With this arrangement, the envelope characteristics of velar activity can be chosen as desired and can be fully synchronized with other temporal functions controlling DAVO.

A program for the evaluation of DANA has been initiated, with preliminary testing procedures as well as formal listening tests used. The question of the compatibility of

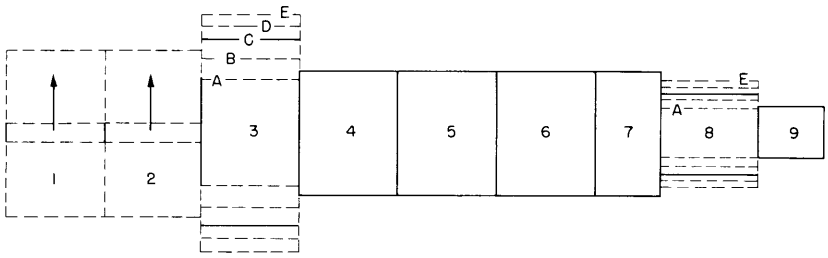

Fig. XVIII-11. Approximate acoustical representation of DANA. The cross-section area of sections 1 and 2 is dynamically variable from approximately 0.05 cm^2 to 5.0 cm^2. The areas of sections 3 and 8 are manually selected (2, 4, 6, 8 or 10 cm^2 for section 3; 0.4, 0.8, 1.2, 1.6 or 2.0 cm^2 for section 8). Sections 4, 5, 6, and 7 have a fixed area of 2.6 cm^2, while section 9 has an area of 0.42 cm^2.

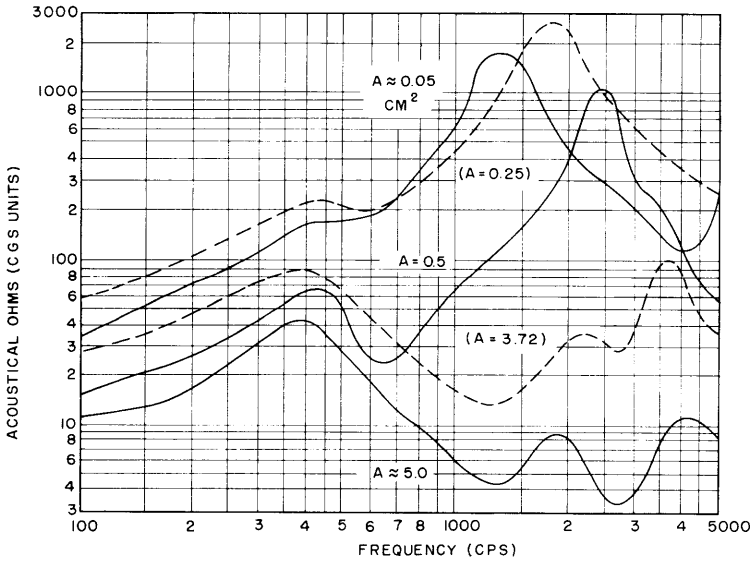

Fig. XVIII-12. Acoustical driving-point impedances of DANA (solid curves) and static analog (broken curves) for indicated degrees of velopharyngeal coupling.

this analog with the aforementioned static analog model is partially answered by comparing the acoustical driving-point impedances (at the velum) of the two models for various degrees of coupling. This comparison is shown in Fig. XVIII-12, in which the solid curves show the magnitude of the impedance for the extreme values of coupling of DANA, and the broken curves show the impedance for the extreme values of coupling of the static analog.

M. H. L. Hecker

References

1. A. S. House, Analog studies of nasal consonants, J. Speech and Hearing Disord. 22, 190-204 (1957).

2. G. Rosen, A prototype section for a dynamic speech synthesizer, Scientific Report No. 1, Acoustics Laboratory, M.I.T., 1955.

C. THE LEARNING OF ENSEMBLES OF SPEECHLIKE SOUNDS

A series of experiments was designed to investigate the information transmission in situations that require subjects to identify the members of ensembles of multidimensional acoustic displays. The experiments utilize stimuli that are speechlike in various degrees, and examine the performance of the subjects during the time in which they are

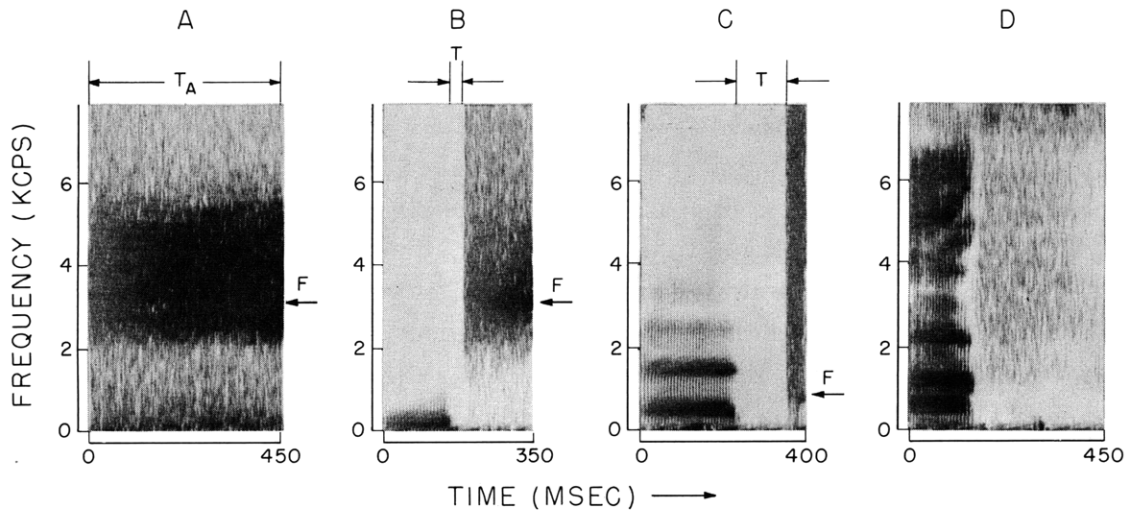

Fig. XVIII-13. Sound spectrograms of stimuli representative of the four classes of ensembles. The values for the temporal variable T and the frequency variable F are indicated. The variations of the intensity variable I, which is the intensity of the noise portion of the stimulus, are not adequately reproduced by the spectrograms because of limitations in dynamic range.

learning to put the stimuli into categories and to associate them with a set of buttons on a response box.

The present series represents an extension of the experiments already reported (1), in which the number of physical parameters was varied without changing the general form of the basic stimuli. In the new series of experiments the form of the stimuli used in various ensembles was modified in order to increase or decrease the resemblance of the stimuli to actual speech materials.

The four classes of stimuli used in the experiments are epitomized in Fig. XVIII-13. The stimuli in the second series (identified as B) are identical with the stimuli of the earlier experiment (2). The stimuli of series A, considered to be less speechlike than those of series B, were generated by exciting a bandpass filter with white noise. Two different stimulus ensembles of eight stimuli each were used in this series – a unidimensional case with center frequency of the passband F as the parameter, and a tridimensional case with F, total duration T, and over-all intensity I as the parameters.

The series C stimuli, considered to be more speechlike than those of series B, were generated in the same general manner as the B stimuli. Their initial portion, however, was characterized by three resonant frequencies, the lowest of which was modulated in time to resemble the first-formant transition that is appropriate for a vowel-stop consonant articulation (3). In addition, a rising inflection was imposed on the vowel-like portion, and the noise portion was generated with a circuit more appropriate for consonant production. Two ensembles were used: a unidimensional case with the center frequency of the noise portion F as the parameter; and a tridimensional case in which the parameters were the center frequency F, the interval between the two segments T, and the over-all intensity of the second (noisy) portion of the stimulus I.

The most speechlike series, D, consisted of actual vowel-consonant syllables produced by one male talker. Only a tridimensional ensemble was used in this case, and the stimuli were the various combinations of two vowels /ɪʌ/ and four consonants /fspt/.

Twelve subjects participated in the experiment. Each subject was required to identify the stimuli in each of the twelve ensembles and to repeat the first three ensembles that he attempted. Each ensemble was presented in one testing session, and the order of presentation of ensembles was varied after the fashion of a 12×12 latin square with all immediate and distant order effects counterbalanced. In each session the stimuli were presented in a quasi-random order adjusted so that each stimulus occurred twice in each block of 16 presentations. After each presentation the subject was required to "identify" the stimulus by pressing one of 8 buttons on a response box. After each response and before the next stimulus was presented, an indicator light on the box correctly associated the last stimulus with a button. The experiment continued until 128 (or 64 in the D series) responses were obtained.

The following discussion of the results of these experiments is based on data

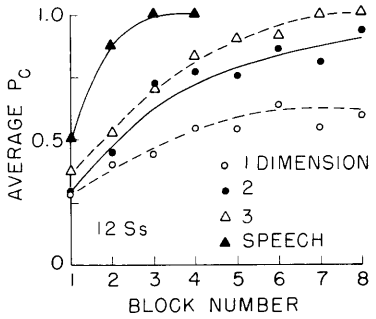

Fig. XVIII-14. Learning curves for uni-, bi-, and tridimensional stimulus ensembles (series B) and for spoken syllables. Responses for 12 subjects are averaged over successive blocks of 16 stimulus presentations. Curves are fitted by inspection.

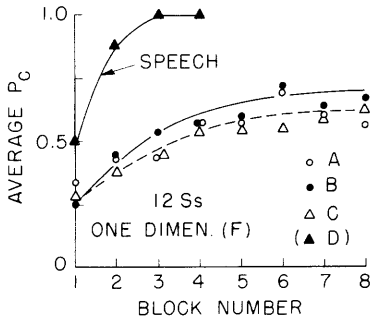

Fig. XVIII-15. Learning curves for unidimensional stimulus ensembles in which frequency is the parameter. A curve for the learning of spoken syllables is included for comparison.

collected in the last 12 sessions for each subject; a general learning or familiarization effect was eliminated. In general, the results demonstrate faster learning of stimulus ensembles as the number of parameters increases. Figure XVIII-14, for example, shows the average percentage of correct identifications for the uni-, bi-, and tridimensional ensembles, as indicated; the responses to the speech (D) ensemble is included for reference.

The other major result of the experiments is epitomized by the data of Fig. XVIII-15 which show the learning of unidimensional ensembles in which frequency is the parameter. It can be seen that no advantage in identification accrues from the speechlike nature of the various stimuli. On the contrary, the items of the C ensemble – the most speechlike materials – were learned at a slightly less rapid rate than those of the A and B ensembles. All materials, however, reached approximately the

same level of identification within the test period. Once again, the actual speech stimuli were learned at the fastest rate.

The experiment and its implications will be discussed more fully in a report that is being prepared for publication.

K. N. Stevens, T. T. Sandel, A. S. House, Jane B. Arnold

References

1. Jane B. Arnold, M. Halle, T. T. Sandel, and K. N. Stevens, Perception of speech-like sounds, Quarterly Progress Report No. 54, Research Laboratory of Electronics, M.I.T., July 15, 1959, pp. 168-169.

2. Ibid., Fig. XIV-6, p. 168.

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).