XVI. SPEECH COMMUNICATION*

Prof. M. Halle	Dr. T. T. Sandel	M. F. Dichter
Prof. K. N. Stevens	Dr. W. S-Y. Wang	O. Fujimura†
Prof. J. B. Dennis	Jane B. Arnold	H. Fujisaki
Dr. A. S. House	P. T. Brady	M. H. L. Hecker
Dr. G. Rosen	J. R. Carbonell	J. M. Heinz

A. VOWEL ANALYSIS

In previous reports we described a technique for utilizing the TX-0 computer in the analysis of speech spectra. Our basic approach is to store within the computer rules for constructing speech spectra. A series of such spectra is generated according to a predetermined strategy, and the spectrum yielding the best fit with each spectrum that is under analysis is identified. The rules for generating the spectra are based on the acoustic theory of speech production, which considers a speech signal as the result of excitation of a linear acoustic circuit by a source with a flat or smoothly sloping spectrum envelope.

Recently, we made an analysis of the principal sources of error inherent in this spectrum-analysis method as applied to vowel sounds. The largest source of error stems from the fact that the present approximate method of spectrum synthesis superposes elemental resonance curves that have each undergone modification by the filter bank used for the analysis of the signal, whereas for the speech spectra under analysis, the elemental curves are first superposed before being processed by the filters. Other inaccuracies are attributable to the discrete harmonic structure of the glottal spectrum, and to the fact that the speech data and the elemental curves are quantized in amplitude and frequency.

In order to minimize these errors, the original computer program for vowel analysis has been modified in several ways. The number of elemental curves available for constructing speech spectra was increased from 30 to 84, with the result that formant frequencies can now be determined within 25 cps. The strategy that determines the order in which synthesized spectra are to be compared with the data that are under analysis was modified to minimize the probability that false answers will be found. Various other features, such as automatic print-out and punch-out of the results of the analysis, were incorporated into the program.

The modified vowel-analysis program has been used to determine the formant

---

*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 from the Research Institute of Communication Science, University of Electro-Communications, Tokyo, Japan.

frequencies for a number of utterances by several speakers and seems to provide data that are more accurate than those obtainable by other methods used in the past. The program is now being used to gather systematic data of formant frequencies of vowels in various consonant contexts, both in the medial portions of the vowels and in the transitions.

H. Fujisaki

B. SPECTRA OF NASALIZED VOWELS

The behavior of poles and zeros of the transfer function of the vocal tract when the tract is coupled to the nasal tract has been studied by a graphical method. The present study is based on an idealized model of the acoustic system for the production of nasalized vowels, as illustrated in Fig. XVI-1. It is assumed that the three cavities, that is, the cavities of the nose, mouth, and pharynx, can be regarded as one-dimensional acoustic transmission lines, all of which join at a coupling point. This assumption is valid for the frequency region that is important for determining the vowel quality. The degree of nasalization is varied by changing the opening area at the innermost end of the nasal tract, as represented by the coupling gate in Fig. XVI-1.

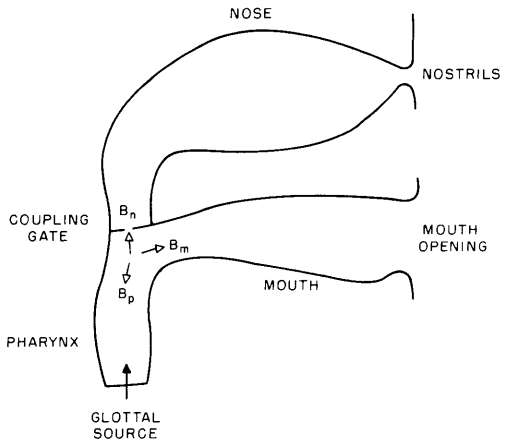

Fig. XVI-1. Simplified model of the acoustic system for production of nasalized vowels. The degree of nasalization is varied by changing the opening area of the coupling gate. The geometry of the nasal cavity is independent of the particular vowel articulation, whereas both pharynx and mouth cavity depend upon the vowel.

For purposes of qualitative discussion, we shall neglect the effect of the damping, and inspect the driving-point susceptance of the nasal tract B_n and the internal susceptance of the vocal tract as seen from the coupling point. Although it is one of the perceptual cues of nasalization (1,2,3), higher damping would not appreciably affect the locations of the formants that are under discussion. The internal susceptance of the vocal tract, which will be designated B_i, is equal to the sum of the susceptances of the mouth and the pharynx, each of which is measured at the coupling point. When we plot the two

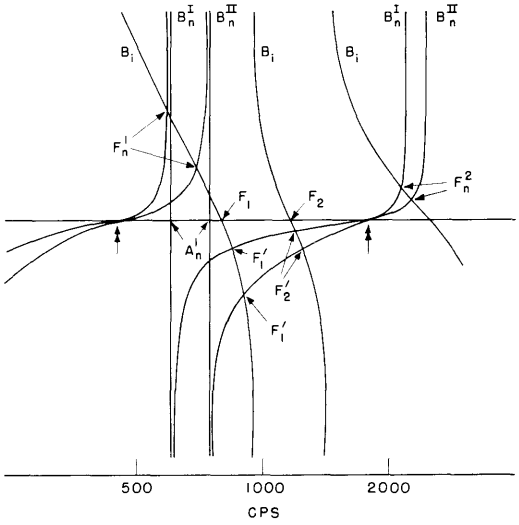

Fig. XVI-2. Sketch of the driving-point susceptance of the nasal tract B_n with two degrees of nasalization B_n^I and B_n^{II} and the internal susceptance of the vocal tract B_i (inverted). The vocal tract configuration is appropriate for the vowel [a]. The zeros and poles of these functions in this example were estimated from unpublished data of analog experiments by House and Stevens. F_1, F_2 represent first and second oral formants; F_1', F_2', shifted oral formants; F_n^1, F_n^2, nasal formants; and A_n^1, the antiformant to be observed in the output spectrum at the mouth. Double arrows indicate the characteristic frequencies of the nasal tract.

curves B_n and $-B_i$ against frequency as shown in Fig. XVI-2, the frequencies of the oral formants of non-nasalized vowels are given by the points where the $-B_i$ curve crosses the abscissa. The intersections of the two curves, B_n and $-B_i$, give the new poles of the combined system, that is, the formant frequencies of the nasalized vowels. In Fig. XVI-2, two curves for B_i represent two degrees of nasalization. (If we use the same B_i for the nasalized and the non-nasalized vowel, we neglect the possible effect of the slight change in the geometry of the vocal tract caused by lowering the velum. This approximation, which is adopted in the model of Fig. XVI-1, may not be appropriate for some heavily nasalized vowels.) The antiresonances that can be observed in the spectrum of the sound radiated at the mouth are given by the singularities of B_n. These are designated A_n^1 in Fig. XVI-2. If the degree of nasalization is slight, the locations of these antiresonances will be approximately the same as those in the sound spectrum observed at a distance from the speaker, that is, the combined output contributed by both the mouth opening and the nostrils. If, on the other hand, the sound radiated from the

nostrils is comparable in magnitude to that radiated from the mouth opening, the combined output (4) would have slightly higher locations for the antiresonances than these A_n.

When we decrease the degree of coupling continuously from a finite coupling area to zero coupling area, each antiresonance approaches one of the formants of the combined system and will finally be annihilated. Since we assume that the change in the nasal-tract geometry takes place only in its innermost portion, it can be shown that as the coupling is decreased, each pole of B_n, that is, the antiresonance A_n, moves downwards until it reaches a zero of B_n (indicated by a double arrow in Fig. XVI-2). This pole and zero of B_n can be considered to be associated. The zeros remain at the same positions as the degree of nasalization is varied. This fact and the knowledge that both B_n and B_i are monotonically increasing functions of frequency, lead to several general rules concerning the behavior of the poles and zeros of the transfer function.

Some of the principal conclusions of this study are:

a. Nasalized vowels, in general, have two kinds of formants which we shall call "nasal formant," and "shifted-oral formant." Each nasal formant is paired with an antiresonance observed in the mouth output. The antiresonance may be called an "antiformant." The nasal formant and the antiformant in each pair approach each other, and finally an annihilation results when the coupling of the nasal tract to the vocal tract reaches zero. When a vowel is heavily nasalized, however, the antiformant can be closer to one of the other formants than to its own mate. The shifted oral formant always joins a formant of the non-nasalized vowel without discontinuity.

b. The lowest formant of the coupled system can be either a nasal formant or a shifted oral formant; the choice depends on the location of the original first formant for the non-nasalized configuration. If the original first formant is of higher frequency than a certain critical frequency, the lowest formant is a nasal formant; and if the first formant is of lower frequency, the lowest formant is an oral formant. Physically, the critical frequency is the lowest resonant frequency of the nasal tract when it is closed at the coupling end, and it depends on the geometry of the nose, particularly on that of the anterior section. It can be seen, therefore, that the lowest formant is always a nasal formant, regardless of the vowel configuration, if the nostrils are closed.

c. The shifted oral formants are always higher than the formants of the non-nasalized vowel with the same vocal-tract configuration.

d. The frequency of a nasal formant is always higher than the frequency at which this formant would be annihilated when the vowel is denasalized (that is, higher than the characteristic frequency of the nasal tract).

e. For slightly nasalized vowels, the nasal formants always appear just above the characteristic frequencies of the nose, whereas the oral formants are always just above the non-nasalized formant frequencies. When the coupling is increased, a formant frequency can even exceed the original frequency of the next higher formant. The ordering

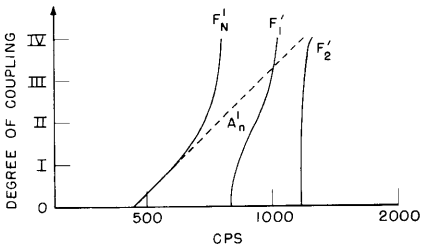

Fig. XVI-3. Shift of formant or antiformant frequency when the degree of nasalization is changed. These curves are derived from the sketches similar to Fig. XVI-2, based on the House-Stevens model. The spectra obtained in the analog experiments reported by House can be explained from this form of presentation. In terms of the coupling area, the four coupling conditions are: I, 0.25 cm^2; II, 0.71 cm^2; III, 1.68 cm^2; and IV, 3.72 cm^2.

of the formants, however, cannot be changed. Thus, for example, if for a slightly nasalized vowel, the first zero of B_n, the first zero of B_i (i.e., F_1), the second zero of B_i (i.e., F_2), the second zero of B_n, and so on, are arranged in this rising order in frequency, the positions of the corresponding formants will shift in various ways when the degree of nasalization is changed, but no formant can meet or overtake another.

f. This invariance of order does not hold for the antiresonances. There can therefore be an annihilation of the nasal formant and its conjugate antiformant for a certain finite degree of nasalization. An annihilation can also occur for a shifted oral formant and an antiformant. Generally, annihilation occurs when a pole of B_n coincides with a pole of B_i, regardless of whether the formant to be annihilated is nasal or oral. It is quite possible, for example, that the first formant of [ɑ̃] may disappear, and the lower nasal formant may appear conspicuously as if it were the first formant shifted downward. If the degree of nasalization is continuously decreased, however, the origin of the formant should be traced correctly. This phenomenon, not only for [ɑ̃], but also for other non-close vowels, can be clearly observed in the data of House and Stevens, who used an electrical analog system. Figure XVI-3 shows an example of the behavior of the formants and antiformants. An annihilation of the shifted first formant by the first antiformant occurs at approximately 1000 cps for a certain degree of coupling.

g. The location of the lowest formant of a nasalized vowel is always higher than the lowest characteristic frequency of the nasal tract and lower than the first formant frequency of the non-nasalized vowel. Consequently, the range of frequencies in which this lowest formant of nasalized vowels can be located is much more limited, compared with the range of the first formant, for different non-nasalized vowels (3). It is also worth mentioning that the sound transmission through the cavity wall may be appreciable in the form of spectrum emphasis at very low frequencies. This would result in an apparent

downward shift of the lowest formant frequency. The determination of the formant frequency of observed spectra in this region is ambiguous, in any case, because of the harmonic structure and the unknown characteristic of the source spectrum. (In the House-Stevens analog (2), the first zero of Y_n was slightly above 400 cps. In the data for Japanese subjects previously reported (3), it was probably around 300 cps, or lower.)

h. If the first characteristic frequency of the nasal tract is close to the first oral formant, these two formants, together with the first antiformant, will form a cluster and will have the appearance of one formant when the nasalization is weak. When the vowel is nasalized further, the components will separate, and their individual effects will appear in the spectrum. It would still be difficult, however, to identify each formant as nasal or oral.

i. For weakly nasalized vowels, in general, the nasal formant is close to its associated antiformant in the mouth's output spectrum, and the pair will hardly be detectable if the coupling is very small. In the output at the nostrils, however, the lowest antiresonance ascribed to the resonance of the mouth cavity occurs usually at well above this nasal formant frequency, and consequently this formant is more conspicuous in the spectrum, although the total output level is comparatively low (3).

j. From rules (d) and (g), it can be deduced that if we have a nasal formant as the lowest formant, then there should be no nasal formant between the (original) first formant and the second characteristic frequency of the nose, which is probably around 2000 cps for normal conditions. Thus, for non-close vowels, it is unlikely that we have a nasal formant between, say, 800 cps and 1500 cps. The minor components in this region of the output spectrum, which are often observed in this region (3), are most likely the result of the increased damping.

O. Fujimura

References

1. K. Nakata, J. Acoust. Soc. Am. 31, 661-666 (1959).

2. A. S. House and K. N. Stevens, J. Speech and Hearing Disord. 21, 218-232 (1956); A. S. House, op. cit. 22, 190-204 (1957).

3. S. Hattori, K. Yamamoto, and O. Fujimura, Nasalization of vowels in relation to nasals, J. Acoust. Soc. Am. 30, 267-274 (1958).

4. C. G. M. Fant, Acoustic Theory of Speech Production, Technical Report No. 10, Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden, 1958 (to be published by Mouton and Company, The Hague, 1960).

C. PERCEPTION OF SOUNDS GENERATED BY TIME-VARIANT RESONANT CIRCUITS

Experiments have been carried out to study the perception of brief sounds generated by periodic impulsive excitation of time-variant resonant circuits. Such stimuli resemble certain speech sounds that are characterized by rapidly moving

formants. A typical stimulus consisted of the waveform produced by driving a simple-tuned series resonator with a fixed number (n) of pulses 10 msec apart. Four values were used for n: 2, 3, 5, and 6. The bandwidth of the resonator was 100 cps, and thus the time constant of the decay after each pulse was approximately 3 msec.

While the resonator was being pulsed, its center frequency varied linearly with respect to time. Four different combinations of starting frequency (f_1) and terminating frequency (f_2) were used:

f_1 = 1000 cps, f_2 = 1500 cps f_1 = 1500 cps, f_2 = 2000 cps

f_1 = 1500 cps, f_2 = 1000 cps f_1 = 2000 cps, f_2 = 1500 cps

Figure XVI-4 shows a schematic representation of each of the 16 different stimuli used. The heavy lines indicate the path of the single resonator, and the short vertical lines at the base represent the positions of pulses. The numbers next to the heavy lines show the order in which the sounds were presented to each subject during the test.

The subject heard a single stimulus repeatedly, at a rate of approximately once per second. He could operate a two-position switch to select either the time-variant test stimulus or a comparison signal. The comparison signal was generated by excitation of a stationary tuned circuit with the same number of pulses as that used for generating the test stimulus. The subject could adjust the resonant frequency of the comparison

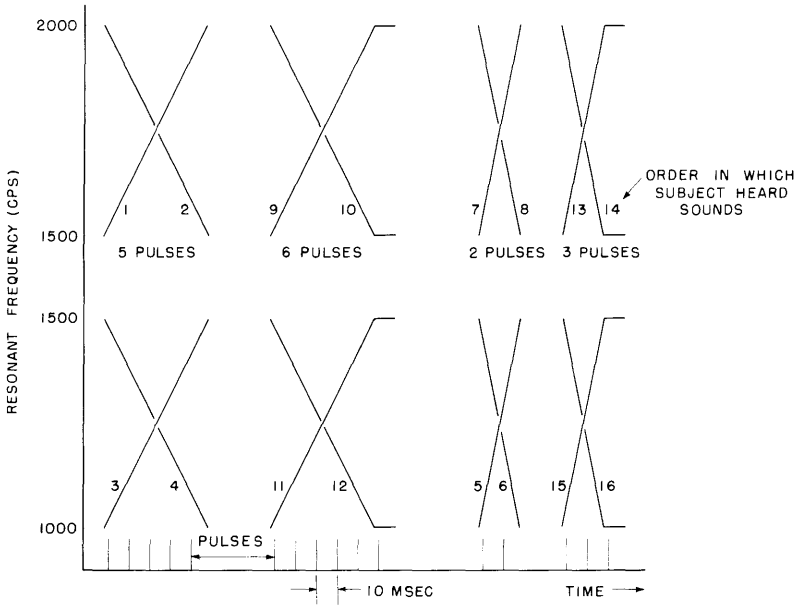

Fig. XVI-4. Stimuli with moving resonances. Paths of the center frequency of the resonator are shown as a function of time. The locations of pulses used to excite the resonator periodically are shown at the bottom.

signal by means of a potentiometer mounted adjacent to the switch. His task was to adjust the stationary resonance so that the comparison signal sounded as much like the test stimulus as possible. At any time, he was free to select either the test stimulus or the comparison sound.

Previous experiments, and also informal listening, have shown that such a match is indeed possible. In many cases, the time-variant character of the sound with the moving resonance was not detectable in a stimulus duration as brief as 50 msec, or especially 20 msec.

The sounds were generated by using a series RLC circuit; the "C" was formed with a Miller-effect electronic circuit. A more thorough description of the equipment for generating the stimulus can be found in Appendix A of the author's thesis (1). The subjects heard the sounds through Permoflux PDR-8 earphones in D-shaped cushions (6 cc) at a level of approximately 90 db re 0.0002 dyne/cm^2. This level may seem high, but the sounds were quite brief, well within the range in which subjective loudness depends on duration, as well as sound pressure level. Twelve subjects took part in the experiment.

The results of the tests indicate that:

a. When subjects match a sound of a stationary resonance with one whose resonance moves linearly from f_1 to f_2, the stationary resonance is always selected at a frequency that is very close to f_2.

b. As the duration of the stimulus is decreased, the subject usually sets the steady resonance closer to f_2. There is also greater agreement among subjects – that is, their deviation from a mean value is less in this case.

c. It is more difficult to make a judgment when the moving resonance is descending than when it is ascending. Subjects were sometimes quite dissatisfied with their settings when the resonance was descending, while they often thought that they had found an exact match for ascending resonances.

These findings suggest that there may be marked differences in the perception of the formant transitions at the beginning and end of a vowel. We plan to extend the present experiments to include sounds that are more speechlike.

P. T. Brady

References

1. P. T. Brady, Perception of Rapid Frequency Changes in Sine Waves and Speech Formants, S.M. Thesis, Department of Electrical Engineering, M.I.T., June 1960.

D. REACTION TIME TO CONSONANT-VOWEL SYLLABLES IN ENSEMBLES OF VARIOUS SIZES

In a previous report (1) an experiment was described in which consonant-vowel syllables were presented to subjects who were instructed to repeat each stimulus accurately,

but quickly. Reaction times were measured for sizes of stimulus ensembles from 2 to 12. The results of this experiment, which showed that reaction time was essentially independent of ensemble size, are in contrast with other data that show a systematic rise in latency as the ensemble size increases.

The earlier study has been extended to include ensemble sizes of 32, and it has been shown that the reaction time does not increase substantially with ensemble size, even for the larger ensemble sizes. For relatively untrained subjects, the average reaction time from the initiation of the stimulus to the beginning of the response is 280 msec. The present experiments also include a series of tests in which the ensemble size was held constant at 16, but the subjects were instructed to give verbal responses only when they heard stimuli that were members of a subset of the original group of 16. Subensemble sizes of 1, 2, 4, and 8 were used in these experiments. The reaction times showed a marked increase as the size of subensemble was increased, and in all cases were larger than the reaction times when responses were required for all stimuli. The greatest reaction time, an average of approximately 1100 msec for two subjects, was obtained for a subensemble of size 8. The data indicate that when a subject is required to respond only to members of a subensemble, a substantial amount of time is required for making a decision to respond, and this additional decision time increases monotonically with the information content of the signal. When the subject is instructed to respond to every stimulus, however, he is not required to make such a decision, and the time required to mimic the stimulus is apparently independent of the size of the ensemble from which the stimulus is selected.

M. Dichter

References

1. M. G. Saslow, Reaction time to consonant-vowel syllables in ensembles of various sizes, Quarterly Progress Report, Research Laboratory of Electronics, M.I.T., July 15, 1958, pp. 143-144.