Analysis of Digital and Analog Formant Synthesizers

BERNARD GOLD, MEMBER, IEEE
LAWRENCE R. RABINER, MEMBER, IEEE

Abstract—A digital formant is a resonant network based on the dynamics of a second-order linear difference equation. A serial chain of digital formants can approximate the vocal tract during vowel production. In this paper, the digital formant is defined and its properties discussed, using *z*-transform notation. The results of detailed frequency response computations of both digital and conventional analog formant synthesizers are then presented. These results indicate that the digital system without higher pole correction is a closer approximation than the analog system with higher pole correction. Finally, a set of measurements on the signal and noise properties of the digital system is described. Synthetic vowels generated for different signal-to-noise ratios help specify the required register lengths for the digital realization. A comparison between theory and experiment is presented.

I. Introduction

The development, in recent years, of the theory of digital filters,[1]–[3] has made it feasible to simulate a wide variety of speech communication devices on a general purpose computer. The formant-type speech synthesizer is one of the devices that has been profitably simulated.[4]–[6] In this paper, digital filter theory is used to study the behavior of a serial formant synthesizer for generating vowel-like sounds; this type of synthesizer, using analog components, has been used in the OVE[7] series and in SPASS.[8] In the digital simulation of such devices, two new problems arise, namely, the sampling and quantizing problems. As is well known, a sampled-data filter is periodic in the frequency domain. Thus, a digital formant network obtained via simulation has a different frequency response than does an analog formant network. As we shall see, the periodic frequency response of a digital formant network is actually a desirable feature, since it eliminates the need for the higher pole correction used with analog synthesizers. The quantization present in the finite-register length computer creates two disturbances: inaccuracies in the formant positions,[9] and a wide-band "noise" caused by roundoff errors during the execution of the linear recursion.[10],[11] These effects place a lower limit on the length of the registers used and, therefore, must be seriously considered in simulating digital filters on computers with small register lengths. Also, the component advances in digital hardware raise the possibility that a special purpose all-digital speech synthesizer or formant vocoder could become a feasible device; clearly, the knowledge of register length constraints becomes major design information.

A widely held misconception is that difficulties arising in computer simulation of speech systems can be avoided by increasing the sampling rate. However, quantization problems will generally increase in severity as the sampling rate is raised. Thus, a sound theoretical understanding of the effects of both sampling and quantizing are necessary for the design of digital speech synthesis programs or special purpose digital hardware synthesizers.

In Section II of this paper, the digital formant network will be defined and discussed, and it will be shown that although linear analysis, using *z*-transform techniques, is applicable, in practice it is necessary to consider carefully the lengths of registers used in the computation. In Section III, the frequency response characteristics of digital formant synthesizers will be studied theoretically and experimentally, utilizing only the linear model. The primary purpose is to find the extent to which a digital synthesizer can approximate the vocal tract transfer function. In Section IV, we will derive the characteristics of the higher pole correction

Manuscript received August 31, 1967. This paper was presented at the 1967 Conference on Speech Communication and Processing, Cambridge, Mass.

B. Gold is with Lincoln Laboratory, Massachusetts Institute of Technology, Cambridge, Mass. (Operated with support from the U. S. Air Force.)

L. R. Rabiner was at the Massachusetts Institute of Technology, Cambridge, Mass. He is now with Bell Telephone Laboratories, Inc., Murray Hill, N. J.

network used in analog synthesizers. In Section V, the quantization problem will be reintroduced, and theoretical and experimental methods will be applied to study the register-length problem.

II. Digital Formants

The transfer function $H(z)$ of a digital formant can be defined, using z-transform terminology, as

$$H(z) = \frac{(1 - 2r\cos bT + r^2)z^2}{z^2 - (2r\cos bT)z + r^2} \tag{1}$$

where T is the sampling interval, and r and b are defined by reference to the z-plane pole-zero diagram shown in Fig. 1. The frequency response of the digital formant is obtained by setting $z=e^{j\omega T}$ in (1). Except for the frequency dependent scale factor in the numerator, this frequency response can be obtained geometrically from Fig. 1 by measuring the distance from any point on the unit circle (at an angle ωT) to the poles, the magnitude of $H(e^{j\omega T})$ being inversely proportional to the product of the distances from that point to the poles (and directly proportional to the product of the distances to the zeros which, in our case, are unity). The significance of r is illuminated by letting $r=e^{-aT}$, so that the parameter a may be interpreted as a half-bandwidth radian frequency. It can be seen from (1) that $H(1)=1$, which shows that the digital formant has the correct dc gain independent of the resonant frequency. This is accomplished by making the numerator dependent on the pole positions so as always to satisfy this condition on the dc gain.

The transfer function $H(z)$ can be approximately realized in a variety of ways; "approximately" because no indication of the quantization problem appears in (1). Thus, the recursive relation

$$\begin{aligned} y(nT) &= 2r\cos(bT)y(nT - T) - r^2 y(nT - 2T) \\ &\quad + (1 - 2r\cos bT + r^2)\, x(nT) \end{aligned} \tag{2}$$

permits the variables $x(nT)$ and $y(nT)$ to take on any real values, whereas in the computer these variables are always contained in finite-length registers. A convenient way of representing the computation of (2) is via the "network" of Fig. 2. The triangular boxes represent unit delays of time T, the rectangular boxes are the fixed multipliers, i.e., the coefficients of the recursive equation (2), and the sum is represented by the circle with the plus sign. These elements are the basic ones for any general system of linear recursions. Computationally, Fig. 2 [and (1)] is interpreted as follows. A new sample $x(nT)$ appears at the input. This signal is multiplied by the fixed number $(1+r^2-2r\cos bT)$; the multiplications indicated by the other two rectangular boxes are carried out, all the indicated products summed, and the appropriate register transfers performed, to fulfill (2). The system is now ready for a new input sample.

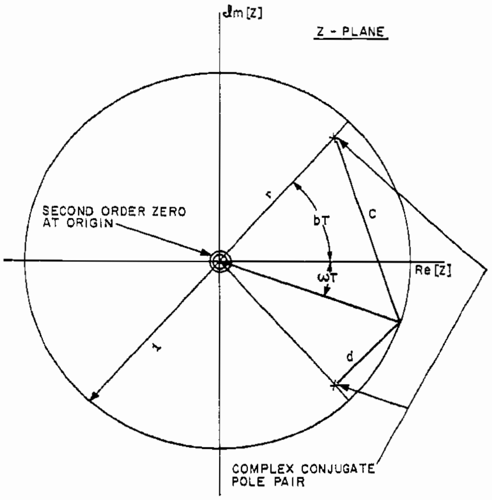


Fig. 1. Z-plane pole-zero diagram for digital formant.

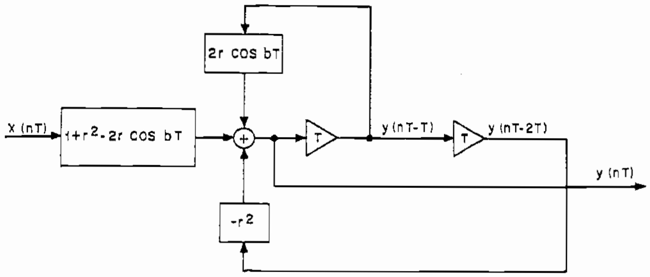


Fig. 2. First digital network representation of a single formant.

Because of the linearity of the network of (1), it is possible to exchange the sequence of operations. For example, Fig. 3 represents a different sequence of computations leading to the same transfer function $H(z)$ of (1). Although the difference between the networks of Figs. 2 and 3 may seem trivial, if one remembers that the actual computations involve finite register lengths, these differences may be significant. To illustrate, assume that $1+r^2-2r\cos bT=0.01$ for a given system. If an input sample $x(nT)$ of magnitude 20 appeared, the product is less than unity and would be truncated to zero. Thus, the system of Fig. 2 exhibits a noticeable nonlinear effect if the input signal level is too small. However, the same signal applied through the network of Fig. 3 might not exhibit such an effect, because the first portion of the network (up to the final multiplier) could have boosted the signal level to well above 100. Thus, although the "linear" behavior of the networks of Figs. 2 and 3 is identical, the actual behavior of the two could be markedly different.

In the remainder of this section and until Section V, the finite register length problem will be ignored, and the frequency response characteristic of the digital

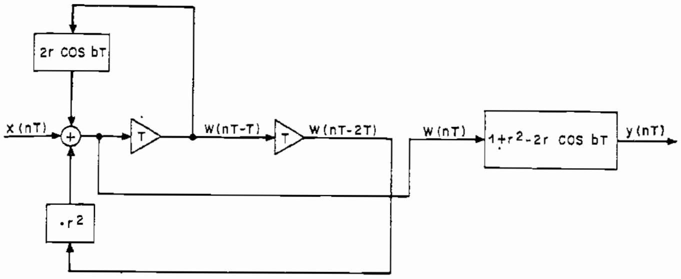


Fig. 3. Second digital network representation of a single formant.

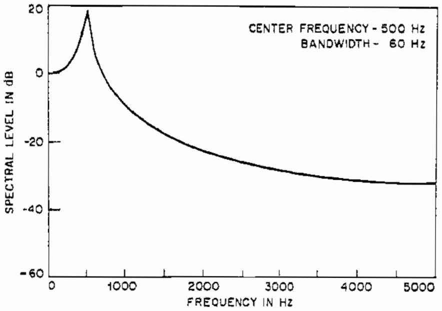


Fig. 4. Frequency response of a digital formant.

formant network will be studied, using (1) and Fig. 1 as the starting point. $H(z)$ actually has an infinity of poles, occurring at the frequencies $(\pm b/2\pi \pm nf_r)$ Hz with $n=0, 1, 2, \cdots$ and $f_r = 1/T$. Thus, the frequency response of the digital formant is periodic, with a period equal to the sampling rate f_r. This well-known property of sampled system is made explicit for the digital formant by writing $|H(e^{j\omega T})|$, that is, the magnitude of $H(z)$ at any angle ωT on the unit circle

$$|H(e^{j\omega T})| = \frac{1-2r\cos bT + r^2}{[1+r^2-2r\cos(\omega-b)T]^{\frac{1}{2}}[1+r^2-2r\cos(\omega+b)T]^{\frac{1}{2}}}. \quad (3)$$

$|H(e^{j\omega T})|$ is clearly periodic in the angle ωT with period 2π, and this is equivalent to periodicity in frequency with period f_r. Also, the resonant effect is clearly seen via the left side of the denominator, which becomes small when $(\omega-b)T=n\pi$, $n=0, \pm 1, \pm 2, \cdots$ yielding the type of result illustrated in Fig. 4.

III. Digital Formant Synthesizer

It is, of course, the repetitive nature of the frequency response of the digital formant network which suggests that it resembles more closely (than does the analog formant network) the repetitive frequency response of the vocal tract. The upper sketch of Fig. 5 indicates the frequency response of an acoustic tube excited at one end and open at the other end. (We have assumed equal bandwidths for all resonances.) This simple model is a representation of an ideal neutral vowel. If the sampling time T is chosen to be 0.5 millisecond, then a digital formant at 500 Hz has repetitive modes at the same frequencies as the tube, while a single analog formant at 500 Hz does not at all resemble the tube. The remaining sketches of Fig. 5 show the comparison between five formant analog and digital ($T=10^{-4}$ seconds) approximations to the tube. It is clear that, for this case, the digital system is a good approximation to the tube, whereas the analog system needs a correction network to compensate for the high-frequency falloff characteristics of cascaded analog formants.

A mathematical representation of the distributed parameter vocal tract system is quite difficult, and we are not able (nor have we really tried) to create a purely theoretical argument for choosing either the digital or analog formant as the better approximation to the actual vocal tract. However, it can be argued that an analog formant synthesizer, consisting of a large number of resonators and higher pole correction, can serve as a criterion for the correct frequency response characteristic of the vocal tract. The standard we have adopted uses 10 cascade resonators and an improved higher pole correction.[1] If we denote this standard configuration as system 1, then the remainder of this section presents and discusses experimental comparisons between system 1 and the following three systems:

System 2: 10-pole digital formant synthesizer using 20-kHz sampling;

System 3: 5-pole digital formant synthesizer using 10-kHz sampling;

System 4: 5-pole analog formant synthesizer with improved higher pole correction.

As indicated previously, we have guessed that a digital synthesizer does not need any higher pole correction, and no such network is used in systems 2 and 3.

Fig. 6 represents system 3. The resonant frequencies of F_1, F_2, and F_3 are variable and correspond to the three lowest resonances in the voiced speech spectrum, thus determining, for example, the particular vowel sound generated. The fixed resonators F_5 and F_4, with resonances at 4500 and 3500 Hz, help provide the correct overall spectrum shape. $S(z)$ represents a formant-like digital network which has been recommended as a suitable source filter, and the transfer function $1-z^{-1}$ approximates the mouth-to-transducer radiation. Each of the digital formant networks is of the form given in Fig. 2 or 3, and has a transfer function of the form of (1). Thus, the transfer function of the entire synthesizer is given by

$$F(z) = S(z)(1-z^{-1})\prod_{i=1}^{5} F_i(z)$$

[1] The nature of this improvement is examined in Section IV.

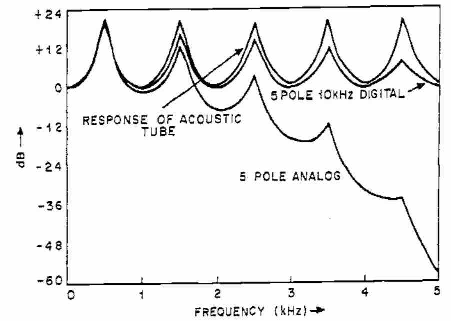


Fig. 5. Digital and analog approximations to transfer function of an acoustic tube; open at one end and closed at the other end.

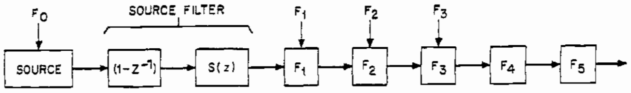


Fig. 6. 5-pole, 10-kHz digital formant synthesizer.

TABLE I

Formant Frequencies for the Vowels

Typewritten Symbol for Vowel	IPA Symbol	Typical Word	F_1	F_2	F_3
IY	i	(beet)	270	2290	3010
I	I	(bit)	390	1990	2550
E	ε	(bet)	530	1840	2480
AE	æ	(bat)	660	1720	2410
UH	ʌ	(but)	520	1190	2390
A	a	(hot)	730	1090	2440
OW	ɔ	(bought)	570	840	2410
U	ʊ	(foot)	440	1020	2240
OO	μ	(boot)	300	870	2240
ER	ɝ	(bird)	490	1350	1690

TABLE II

Analog and Digital Resonator Bandwidths and Center Frequencies

Resonator	Center Frequency (Hz)	Bandwidth (Hz)	Q
F_1	Variable	60	Variable
F_2	Variable	100	Variable
F_3	Variable	120	Variable
F_4	3500	175	20
F_5	4500	281	16
F_6	5500	458	12
F_7	6500	722	9
F_8	7500	1250	6
F_9	8500	2125	4
F_{10}	9500	4750	2

with

$$F_i(z) = \frac{(1 + r_i^2 - 2r_i \cos b_iT)z^2}{z^2 - (2r_i \cos b_iT)z + r_i^2}. \qquad (4)$$

For the 10-pole digital 20-kHz system 2, five additional digital formants at 5500, 6500, 7500, 8500, and 9500 Hz have been inserted into the chain of Fig. 6.

Each digital formant is specified by values of the parameters r_i and b_i. To change these parameters into frequencies, we use the relations $r_i = e^{-2\pi g_i T}$ and $b_i = 2\pi f_i$, so that f_i is the resonant frequency and g_i is the half-bandwidth expressed as a Herzian frequency. Table I shows the values of f_1, f_2, and f_3 chosen[12] for each of the 10 vowel sounds analyzed by us. Table II shows the bandwidths of all the formants; the same fixed values were used throughout for both digital and analog cases. The values and extrapolations for higher formants are based on data by Dunn.[13]

The analog formant synthesizers are the classical vowel synthesizer treated by Fant.[14] They consist of 5 (for Case 4) or 10 (for Case 1) analog resonators of the form

$$H(s) = \frac{s_1 s_1^*}{(s - s_1)(s - s_1^*)}, \qquad (5)$$

an additional analog resonator of center frequency 200 Hz and bandwidth 250 Hz for the source filter, a differentiator, and a higher pole correction (to be described in greater detail in Section IV).

Given the 10 vowels listed in Table I, a total of 40 frequency response curves had to be experimentally determined in order to compare systems 1, 2, 3, and 4. The measurement for systems 2 and 3 was made by passing a unit sine wave through a simulation of the system, and determining the peak output amplitude after the transient response of the system had subsided. The frequency of the input was varied from 50 Hz to 5000 Hz in 50-Hz steps. The data for systems 1 and 4 were theoretically calculated from the synthesizer system functions. Figs. 7 through 10 show results for the four systems for each of three vowels.[2] In these figures, the logarithmic magnitude (in dB) is plotted on a linear frequency scale. The contribution of the source filters is omitted from these curves and will be treated separately. No generality is lost thereby, since, as we shall see, it is a simple matter to combine the effects of the source and resonators.

Figs. 11, 12, and 13 show plots of the differences between spectral magnitudes of systems 2, 3, and 4 relative to the reference system 1 for each of the vowels IY, A, and OO. (Table I shows the IPA symbols and our typewritten equivalents for the vowels.) We see that the 10-pole 20-kHz digital system 2 is extremely close to the reference system. This strongly indicates that higher poles of the vocal tract transfer function are automatically and more or less correctly taken into account by the repetitive nature of the digital formant frequency

[2] All 40 curves will be made available in a forthcoming M.I.T. Research Lab. of Electronics report.

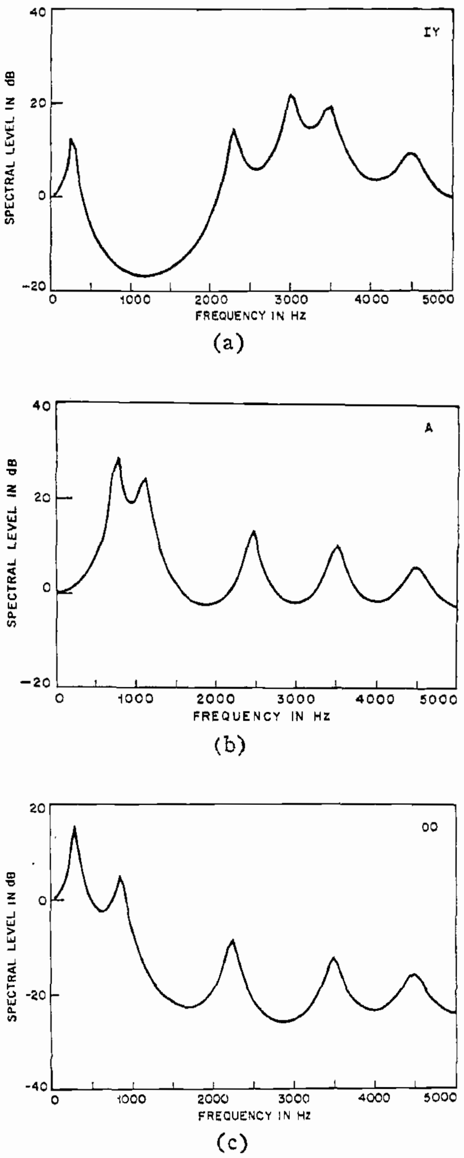


Fig. 7. System 1: 10-pole analog. (a) IY; (b) A; (c) OO.

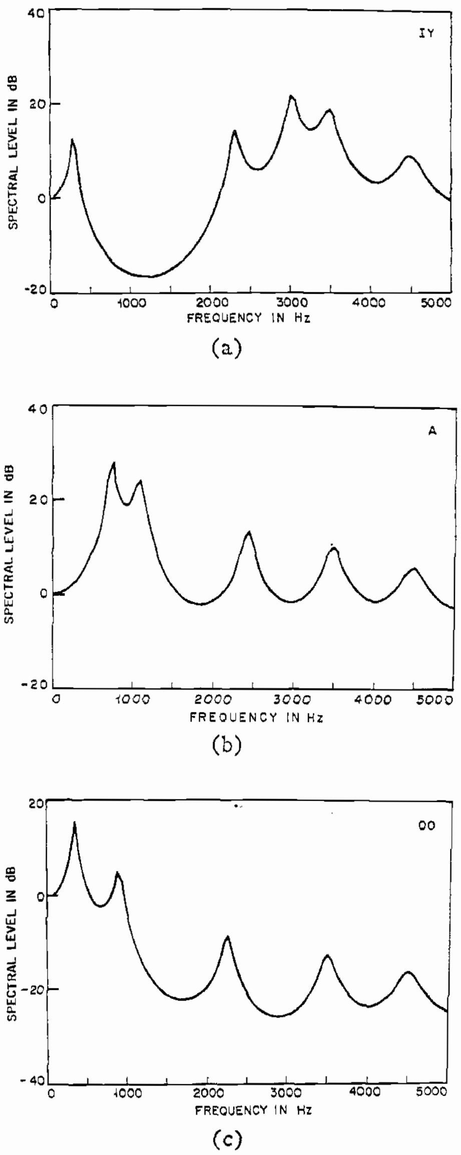
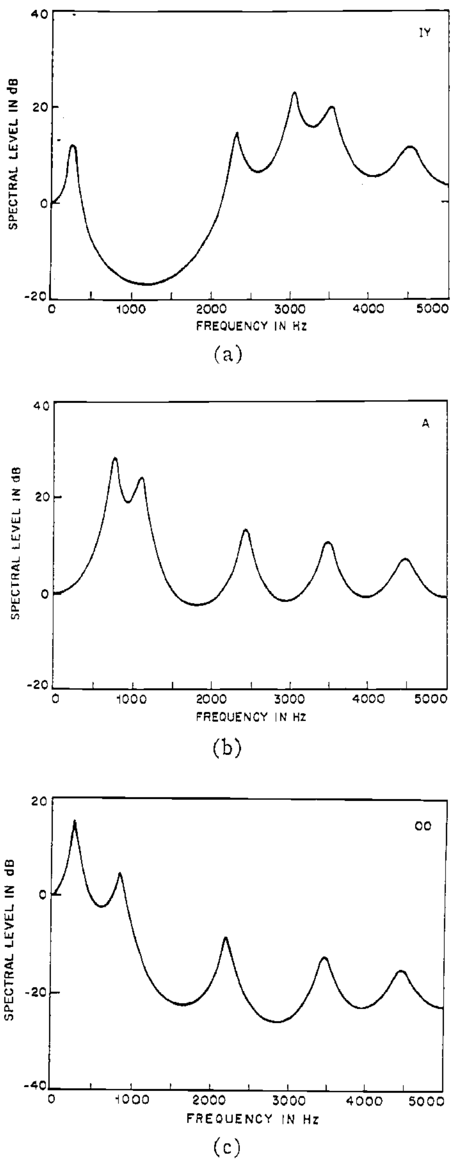


Fig. 8. System 2: 10-pole digital, 20-kHz sampling frequency. (a) IY; (b) A; (c) OO.

response. We also note that this intrinsic correction is actually more accurate than the quite good analog higher pole correction used in our computations. These results are generally valid for all the vowels.

Comparison of system 3 with the standard is of particular interest, since a 5-pole 10-kHz system appears to be a good compromise design for a possible hardware version of a digital formant synthesizer. The peak difference between the magnitude curves for systems 1 and 3 is listed in Table III for each vowel. On the basis of this result, it seems reasonable to expect that a 5-pole 10-kHz digital vowel synthesizer should produce synthetic vowels of quality comparable to a well-designed 5-pole analog vowel synthesizer which includes a higher pole correction. Informal listening reinforces this expectation.

Inclusion of the source filters for both analog and digital cases slightly increases the deviations of systems 2, 3, and 4 from the reference. Fig. 14 shows the frequency responses of the two digital and one analog source filters. (We have included the differentiator as part of the source filter.) The plots are normalized so the peaks are set to 0 dB for all three cases. With the inclusion of source filters, the frequency response of system 2 is within 1 dB of the reference for all vowels and all frequencies. The peak difference, in the worst case (for IY), between system 3 and the reference is 7.48 dB at 5 kHz. For all vowels except IY and for all frequencies below 4 kHz, the difference never exceeds 3.5 dB. It is possible that a digital source filter with slightly decreased bandwidth could bring the two results closer together.

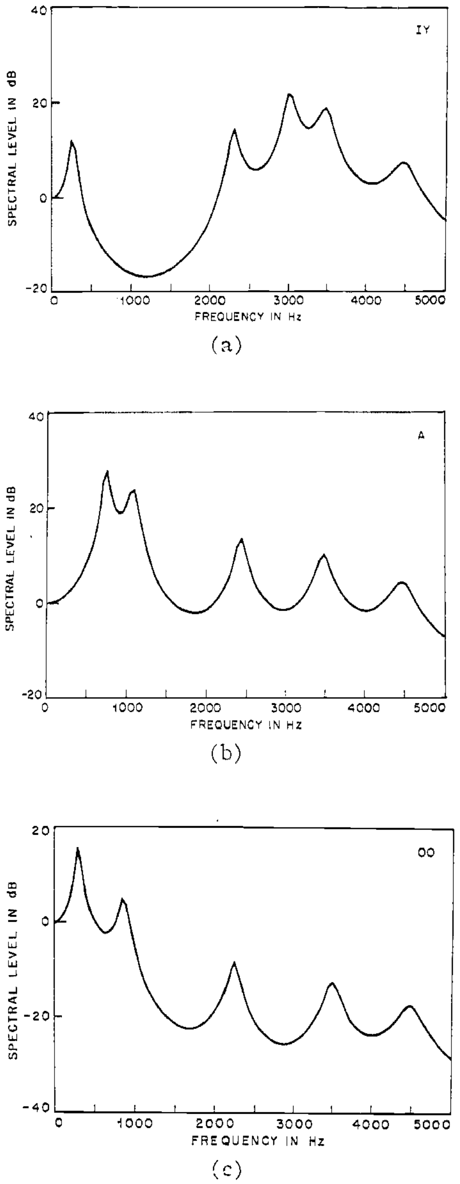


(a)

(b)

(c)

Fig. 9. System 3: 5-pole digital, 10-kHz sampling frequency. (a) IY; (b) A; (c) OO.



(a)

(b)

(c)

Fig. 10. System: 4-pole analog. (a) IY; (b) A; (c) OO.

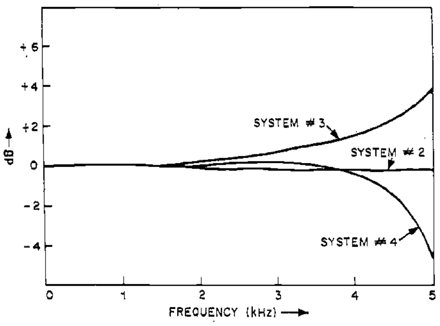


Fig. 11. Spectrum magnitude differences for IY.

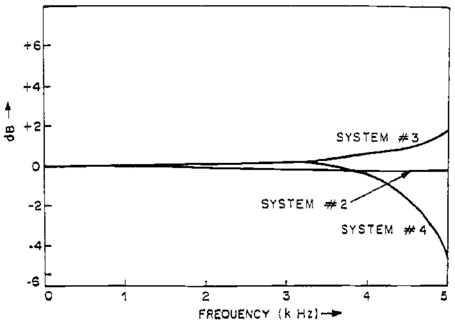


Fig. 12. Spectrum magnitude differences for A.

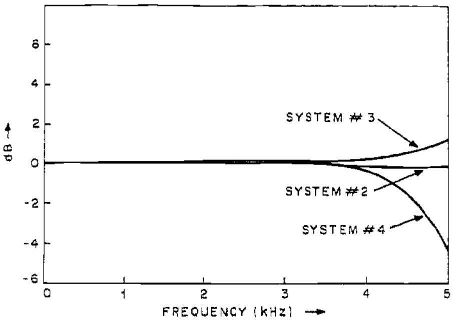


Fig. 13. Spectrum magnitude differences for OO.

TABLE III

Vowel	Peak Difference Between Systems 3 and 1
IY	3.69 dB
I	2.42 dB
E	2.18 dB
AE	2.00 dB
UH	1.56 dB
A	1.62 dB
OW	1.44 dB
U	1.25 dB
OO	1.16 dB
ER	0.65 dB
Average	1.80 dB

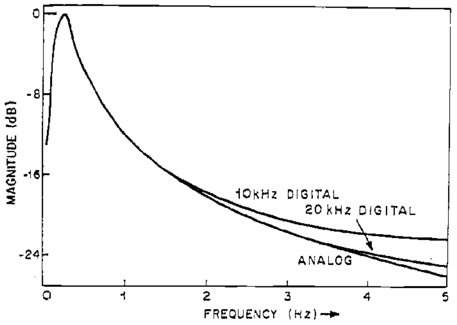


Fig. 14. Source filter characteristics.

IV. Higher Pole Correction for Analog Systems

The material to be presented in this section is incidental to the main line of development of this paper, and deals only with the question of the higher pole correction for analog formant synthesizers. The higher pole correction is used to approximate the higher modes of the vocal tract which are not explicitly present in the synthesizer. The frequency response magnitude of this network [to be referred to as $Q_k(\omega)$] was derived by Fant[14] to be

$$|Q_k(\omega)| \cong e^{(\omega^2/\omega_1^2)R_k}$$

with

$$R_k = \frac{\pi^2}{8} - \sum_{n=1}^{k} \frac{1}{(2n-1)^2} \tag{6}$$

where it has been assumed that k analog formant networks are used to approximate the vocal tract, and ω_1 is the radian frequency of the first formant. In order to make $Q_k(\omega)$ into a network with fixed rather than variable parameters, ω_1 is usually chosen to be an average, say $2\pi\times 500$ rad/s.

Our observations were that the 5- and 10-pole analog synthesizers, both using the $Q_k(\omega)$ specified by (6), nevertheless resulted in substantially differing frequency response curves. In fact, results which appeared qualitatively wrong were obtained. These results were that the 5-pole system was attenuated more with increasing frequency than was the 10-pole system. Given that the 10-pole system utilized rather wide bandwidths for formants 6, 7, 8, 9, and 10, and that the higher pole correction presumably corrects for higher modes having narrower bandwidths, we would presume that the reverse result should have been observed. We conjectured that the approximations leading to (6) were too gross, and herewith present a somewhat more refined formula for approximating the higher modes of the vocal tract for an analog formant synthesizer.

We begin with the same assumptions used by Fant in his original derivation: that the vocal tract filter during vowels can be represented in the frequency domain by the infinite product

$$\begin{aligned} P(j\omega) &= \prod_{n=1}^{\infty} \frac{s_n s_n^*}{(s-s_n)(s-s_n^*)} \\ &= \prod_{n=1}^{k} \frac{\omega_n^2}{[(\omega_n^2-\omega^2)^2+(2\sigma_n\omega)^2]^{1/2}} \\ &\quad \cdot \prod_{n=k+1}^{\infty} \frac{\omega_n^2}{[(\omega_n^2-\omega^2)^2+(2\sigma_n\omega)^2]^{1/2}} \\ &= P_k(j\omega)Q_k(j\omega). \end{aligned} \tag{7}$$

σ_n and ω_n are the damping term and resonant frequency expressed in radians per second, and $P_k(j\omega)$ represents those k formants which are explicitly constructed in the synthesizer. Thus, $Q_k(j\omega)$ appears as the product from

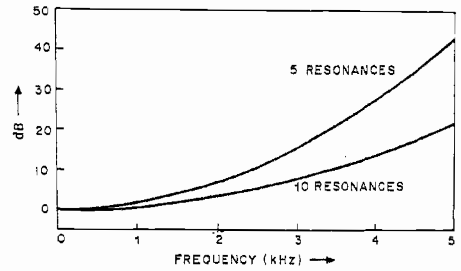


Fig. 15. First-order higher pole correction.

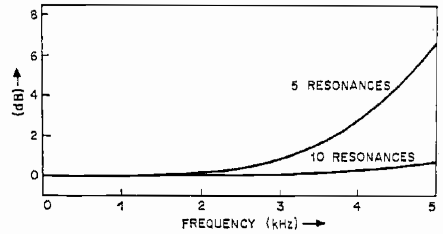


Fig. 16. Second-order improvement in higher pole correction.

$k+1$ to infinity of those formants which are not built into the synthesizer. To approximate $|Q_k(j\omega)|$, Fant first assumes that σ_n is small enough to be set to zero for all n. This yields

$$|Q_k(j\omega)| = \sum_{n=k+1}^{\infty} \frac{1}{\left|1 - \frac{\omega^2}{\omega_n{}^2}\right|} \tag{8}$$

and, taking the logarithm of both sides,

$$\ln |Q_k(j)| = - \sum_{n=k+1}^{\infty} \ln \left|1 - \frac{\omega^2}{\omega_n{}^2}\right|. \tag{9}$$

Fant then expands the logarithm as a power in $(1/\omega_n)^2$ series, and uses only the first two terms, which leads to (6). Our extension includes an extra term in this series, so that

$$\ln |Q_k(j\omega)| \approx \omega^2 \sum_{n=k+1}^{\infty} \frac{1}{\omega_n{}^2} + \frac{\omega^4}{2} \sum_{n=k+1}^{\infty} \frac{1}{\omega_n{}^4}. \tag{10}$$

If we now take the modes to be that of a straight pipe of length l, the values of ω_n are periodic and are $\omega_n = (2n-1)\omega_1 = (2n-1)(\pi c/2l)$, where c is the velocity of sound. Making use of the identities

$$\frac{\pi^2}{8} = \sum_{n=1}^{\infty} \frac{1}{(2n-1)^2} \quad \text{and} \quad \frac{\pi^4}{96} = \sum_{n=1}^{\infty} \frac{1}{(2n-1)^4},$$

we arrive at the result,

$$|Q_k(j\omega)| \approx e^{(\omega/\omega_1)^2 R_k + \frac{1}{2}(\omega/\omega_1)^4 L_k}$$

with

$$L_k = \frac{\pi^4}{96} - \sum_{n=1}^{k} \frac{1}{(2n-1)^4}. \tag{11}$$

The first term in (11) is the usual higher pole correction. Fig. 15 shows plots of the first term of (11) [or (6)] for the two cases $k=5$ and $k=10$. It is evident that both 5- and 10-pole systems need this standard higher pole correction. Fig. 16 shows plots of the second term in (11), namely, the expression $\exp[\frac{1}{2}(\omega/\omega_1)^4 L_k]$. We see that if a 10-pole synthesizer is used, this extra refinement is insignificant; but if 5 poles are used, a reasonably significant correction is added. It should be noted that, at frequencies above about 4 kHz, the cross modes of the vocal tract are of significance, thus diminishing the significance of this additional correction factor.

V. Quantization Effects in Digital Formant Synthesizers

The finite length of the registers containing the signals flowing through the networks of Figs. 2 and 3 influences the results in several ways. First, the coefficients of the difference (2) cannot, in general, be specified exactly, so that the true pole positions may be in error. This is a fixed error and easily computed by comparing the quantized and nonquantized coefficient values. Second, the signals are perturbed by quantization during each iteration of the computation. If signal level changes from one iteration to the next are large relative to an individual quantum step, then it seems reasonable to hypothesize[2],[10],[11],[15] that signal quantization behaves like additive noise, that all such sources of noise are uncorrelated, and that each sample of this noise is uncorrelated with past and future samples. Such an hypothesis greatly simplifies the formulation of the digital network quantization problem and makes it easier to interpret experimental results, but clearly some indication must first be had that valid predictions can be made on the basis of such a simple hypothesis. The first portion of this section is, therefore, devoted to a study of the validity of the simple additive noise model; the second portion discusses some of the results obtained, these results being illuminated by reference to the model.

Fig. 17 is a modified version of Fig. 3, wherein three noises e_1, e_2, and e_3 are added, corresponding to the roundoff or truncation errors implicit in each of the three multiplications. We assume that each noise sample produced at every recursion is uncorrelated with all other noise samples produced by the same noise generator during other recursions, and that $e_1(nT)$, $e_2(nT)$, and $e_3(nT)$ are mutually uncorrelated even for the same iteration.[3] Thus, all that need be known statistically are the one-dimensional probability distributions asso-

[3] Such an assumption is surely wrong, if, for example, any two coefficients in the recursive equation were exactly equal, so that our hypothesis will not include such special cases.

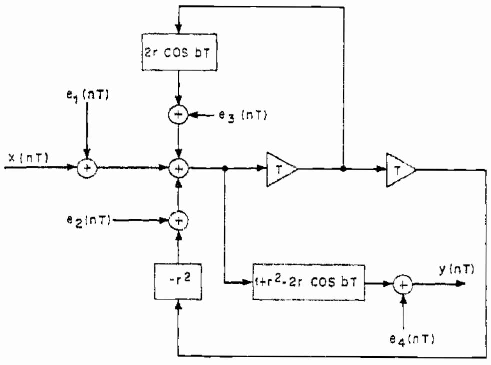


Fig. 17. Noise model formant network.

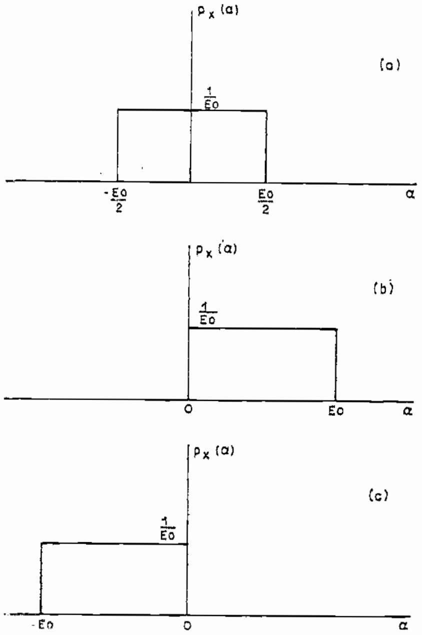


Fig. 18. Probability density functions of noise.

ciated with each of the three random variables. Again, a reasonable assumption is that e_1, e_2, and e_3 are uniformly distributed over a quantization interval and for fixed point arithmetic, independent of signal level. We also specify that quantization levels are uniformly spaced (linear quantization of the signals). Whether or not the probability distributions depend on the sign of the signal is determined by the precise manner in which quantization is effected. Let us examine this point more closely.

In a digital computation, the product of two numbers can occupy a register of twice the length of each of the numbers. For example, the product of the two 5-bit positive binary numbers 0.1011 and 0.1110 yields the 10-bit product 00.100|11010. To store the result in a 5-bit register requires that the five lower bits be removed, and this may be accomplished via truncation, wherein the low-level bits (after a 1-bit left shift to restore the original decimal point placement) are simply removed, yielding 0.1001. Alternatively, the result may be rounded off to the nearest quantization level, yielding, in this example, the product 0.1010. Now, this latter operation results in the uniform probability density shown in Fig. 18(a), while (b) holds for truncation of a positive signal, and (c) holds for truncation of a negative signal. Thus, truncation introduces a quasiperiodic component of the resultant noise. If a sign dependent truncation were performed which could lead to the result of either Fig. 18(b) or (c) regardless of signal sign, then only a dc component would be induced in the noise spectrum. The importance of raising these seemingly trivial points lies in the fact that different hardware configurations on different computer programs would be required, depending on how the extra bits were chopped off, and the programmer or designer ought to be cognizant of the effects on the resultant noise of these different realizations.

Returning now to the noise model of Fig. 17, let us consider the noise generated at the output of the digital filter caused by, say, $e_1(nT)$. The variance of this noise at any time nT created by a noise sample at $m=0$ is given by $\sigma^2h^2(nT)$, where σ^2 is the variance of $e_1(nT)$ and $h(nT)$ is the network unit pulse response. Similarly, the variance created by a noise sample at $m=1$ is $\sigma^2h^2(nT-T)$. Proceeding in this way, one can construct the formula for the output variance due to $e_1(nT)$ to be[4]

$$\sigma_{d1}^2(nT) = \sigma^2 \sum_{m=0}^{n} h^2(mT). \qquad (12)$$

The variance σ^2 of $e_1(nT)$ can be obtained by inspection of Fig. 18 and is $E_0^2/12$, where E_0 is the magnitude of a single quantization step. To obtain the total variance due to all the noise sources in Fig. 17, we need only add the contributions due to each noise source; this yields

$$\sigma_d^2 = \sigma_{d1}^2 + \sigma_{d2}^2 + \sigma^2 = E_0^2/6\left[\sum_{m=0}^{n} h^2(mT) + \frac{1}{2}\right]. \qquad (13)$$

To obtain the total variance due to all the noise sources in a more complex network, such as the three cascaded digital formant networks shown in Fig. 19, we again add the variances due to each source, using that unit pulse response which describes the passage of that particular source through the system. For example, $e_1(nT)$ in Fig. 19 passes through all three digital networks, whereas $e_7(nT)$ passes through only the final one; thus, the $h(nT)$ used to compute σ_{d1}^2 is different than the $h(nT)$ used to compute σ_{d7}^2.

In a digital system where all poles are within the unit circle, the summation (12) converges to a finite value, so that, if we let the upper limit n of (12) become infi-

[4] This formulation is explained in greater detail by Gold and Rader.[2],[11]

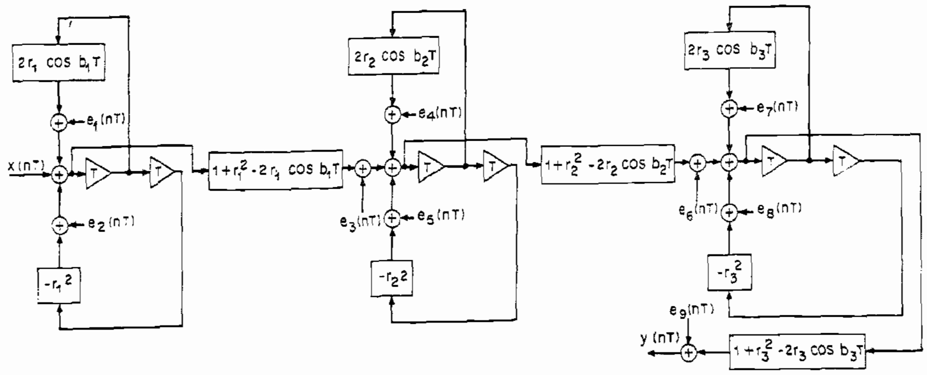


Fig. 19. Noise sources in a cascade of three formants.

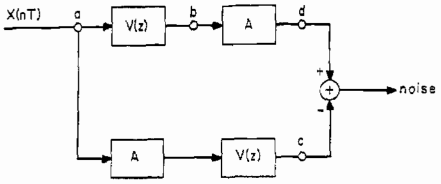


Fig. 20. Noise measurement on digital formant synthesizer.

nite, we have an expression for the "steady-state" variance of the system. Physically, one would expect this "steady-state" to be reached in a time which is about the same as the transient response time of the system. For this case, evaluation of (6) for specific networks is algebraically less cumbersome and, indeed, crude approximations can be made, which increase physical insight into the noise effects and may perhaps help suggest improvements in configurations. Before further elaboration of these statements, let us first describe an experimental method of measuring the noise in an arbitrary system, and then show some results comparing theory with experiment which tend to verify our noise model.

The digital transfer function $V(z)$ in Fig. 20 represents the complete 5-pole 10-kHz formant synthesizer described previously, including source and radiation transfer functions. A is an attenuator such that the output is a small fraction of the input, and $x(nT)$ is a periodic train of pulses with a duration of one sampling interval. Since the amount of quantization noise is not a function of the input signal level, points b and c of Fig. 20 contain approximately equal noise levels. Attenuating the signal from b to d should not change the signal-to-noise ratio at these points; therefore the noise at point c is appreciably larger than the noise at point d, although the signal levels are equal. Thus, subtracting the two signals should give a reasonable measure of the noise, especially if there is significant noise present.

In order to compare theory and experiment, the noise variance from $V(z)$ should be measured using the setup of Fig. 20, and this result should be compared with that obtained by application of (12) to the same system. This was done for the 10 vowels listed in Table I;[5] the precise cascading of the components of $V(z)$ is shown in Fig. 6;[6] the comparisons of the variances expressed as octal numbers are shown in Table IV. Although the agreement is not perfect, it is clearly close enough to encourage use of our simple noise model.

We now can return to the problem of crudely approximating the noise generated by a single digital formant. Using the result[10]

$$\sum_{n=0}^{\infty} h^2(nT) = \frac{1}{2\pi j} \oint H(z)H(1/z)z^{-1}dz \qquad (14)$$

where $H(z)$ and $h(nT)$ are a transform pair and the integral is around the unit circle, computation of (12) is easily performed using the calculus of residues if $H(z)$ is a digital formant network. The approximate result obtained when the poles are close to the unit circle is $E_0^2/12\epsilon$, where $\epsilon = 1 - r$. Since the gain at resonance of a digital formant network is also inversely proportional to ϵ, it follows that a network will amplify the noise proportionally to its resonant gain. From this, it follows that the noise generated by the digital formant network can be altered by rearrangement of the order of the chain. For example, since F_5 has a higher resonance gain than F_1, it should appear earlier in the chain, since, thereby, all the noise generated by the system following F_5 does not pass through F_5 and is not amplified as much.

[5] The variance was measured by averaging the sum of the squares of 3500 samples of the noise. Measurements showed the noise had zero mean.

[6] The wrong value of the damping term for the source filter was inadvertantly used in this experiment (60 Hz instead of 250 Hz), but this should have no effect on the general validity of this comparison between theory and fact.

TABLE IV

Comparison Between Theory and Experiment for Predicting Noise

Vowel	Measured Noise	Theoretically Determined Value
IY	702 664	735 547
I	125 574	114 717
E	101 110	104 036
AE	57 674	52 241
UH	51 414	52 241
A	51 050	55 346
OW	50 700	52 763
U	53 460	52 242
OO	41 044	42 123
ER	27 574	27 465

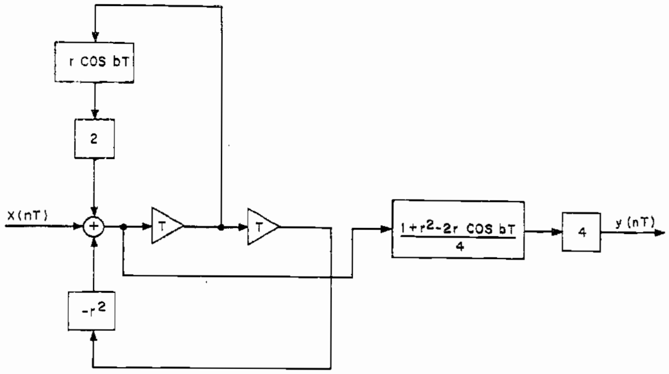


Fig. 21. One way of realizing a digital formant on TX-2.

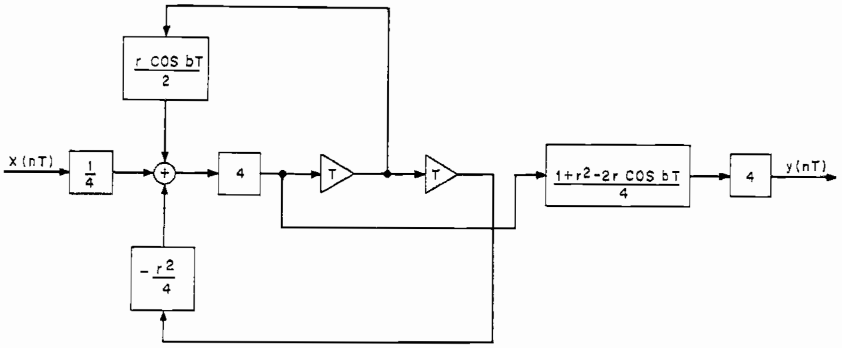


Fig. 22. Alternate method of realizing a digital formant on TX-2.

We see that quantization considerations using a simple noise model help us decide how the synthesizer is to be arranged, and in what order the formant networks should be arranged to keep the noise low. Other considerations also enter into such decisions. For example, it has been conjectured that the system is less sensitive to transient disturbances following formant frequency changes if the higher formant networks precede the lower ones. Intuitively, this argument resembles the noise argument and leads to the same or similar arrangement. Another consideration is dynamic range; the problems arising here are equivalent to those arising in analog systems wherein the formants are arranged so that the signal becomes neither too large nor too small. The comparisons of Figs. 2 and 3 allude to this problem.

A further benefit may be derived by closer examination of the precise way that the computation for a single digital formant is carried out. Often, the way the computation is performed depends on the computer; in what follows, we will illustrate by an example using a TX-2 computer program. TX-2 is a fixed-point computer with an automatic left shift after multiplication, so that if the decimal points directly follow the high level bit (as in the example earlier in this section), then the product will automatically have the same decimal point position. This makes it convenient to treat all numbers as decimal fractions. However, the coefficient $2r\cos(bT)$ in (2) is usually greater than unity, and the program must take this into account. Two ways of doing this are illustrated in Figs. 21 and 22. Multiplications by powers of two are, of course, only shifts, so that the above restriction of treating numbers as decimal fractions does not apply. We intuitively feel that the configuration of Fig. 21 leads to better signal-to-noise ratio, since the roundoff or truncation caused by the multiplications in either case are the same, but the signal levels in Fig. 21 are maintained higher. Experimental results indicate that the noise variance of the formant network using Fig. 21 is approximately double that obtained using Fig. 22.

Finally, we present experimental results which make it possible to specify the required register lengths needed for each of the data carrying registers in each of the networks. This is accomplished as follows. A given vowel is generated by setting formants 1, 2, and 3 to one of the rows of values in Table I; the digital synthesizer is excited by a periodic pulse train corresponding to the pitch (for most experiments the pitch was set to 125 Hz) and the magnitude of this excitation is systematically reduced until the effects of quantization are audible. Also, the signal-to-noise ratio (defined as the ratio of the rms of the output signal to the rms value of the noise) is measured. During the execution of the program, peak magnitudes are recorded for each register in the system. From this information, it is possible to construct a table

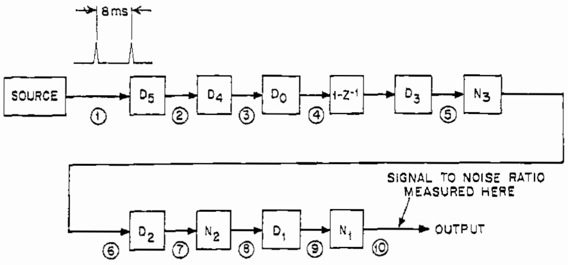


Fig. 23. 540321 sequence of digital formants.

for any given configuration, listing the number of bits needed for each register. Referring to Figs. 2 and 3, we see that only two registers per digital formant need be listed; for example, in Fig. 2, the input and output of the numerator multiplier.[7]

For convenience, we express each digital formant $H(z)$ as the ratio $N(z)/D(z)$. The chain drawn in Fig. 23 shows the sequence of operations in one particular run. Note that we have omitted the numerator factors N_5, N_4, and N_0. These are fixed multipliers, and should not be included since they introduce extraneous and unnecessary noise.

Table V shows the required register length associated with each member of the chain. The particular ordering of the chain was chosen to try to pass as little noise as possible through the high-gain formants; hence F_5 and F_4 were put at the beginning. The signal-to-noise ratio, defined as the rms signal divided by the rms noise, is listed in the last column of Table V, in bits. Thus, for example, 8 bits corresponds to a ratio of 256, while $8\frac{1}{2}$ bits is $\sqrt{512}$. Listeners agreed that this configuration corresponded most closely with the threshold of audible noise. Speaking rather loosely, if we allow a reasonable tolerance for problems such as transients caused by formant changes, it would seem that a computer with an 18-bit register length would satisfy fidelity requirements on a digital formant synthesizer.

We should keep in mind that the numbers obtained hold for a 5-pole 10-kHz system. If the number of poles is increased, the situation worsens. More noise is generated, and the problem of maintaining fairly uniform dynamic range becomes more difficult. If the sampling rate is increased, the situation also worsens, since, then, the poles come closer to the unit circle, so that the gain of the system increases. Again, this means that uniformly distributing register lengths becomes more difficult, although the effect on the signal-to-noise ratio is not clear.

By contrast with the configuration of Fig. 23, where gains were judiciously adjusted, Fig. 24 and Table VI show the result of a rather arbitrary arrangement of formants. Notice that, although the register lengths need to be larger in this case, comparable signal-to-noise ratio results. Thus, we see that some care in the ordering of the elements results in a more efficient system, and may make the difference between a successful and nonsuccessful run on an 18-bit computer.

TABLE V

Register Lengths for 540321 Synthesizer Configuration

Vowel	Node										
	1	2	3	4	5	6	7	8	9	10	SNR (bits)
IY	10	12	13	11	12	14	13	14	13	8	4.5
I	10	12	13	11	12	13	12	12	12	8	4
E	10	12	13	11	12	12	12	12	12	9	4.5
AE	10	12	13	11	11	12	11	11	11	9	4
UH	10	12	13	11	11	12	11	10	11	8	5
A	10	12	13	11	11	12	11	10	11	9	4.5
OW	10	12	13	11	11	12	11	9	12	9	4
U	10	12	13	11	11	12	11	10	11	7	4
OO	10	12	13	11	11	12	11	9	12	7	4
ER	10	12	13	11	11	11	11	10	12	8	4
Maximum over all vowels	10	12	13	11	12	14	13	14	13	9	

For each digital resonator, there are three noise sources, corresponding to the three multipliers. We discussed earlier a method of reducing the number of multipliers by one, for the fixed resonators, by removing the numerator multiplier. However, this method cannot be used for the variable formants because the numerator contains terms which depend on the frequency of the resonator. A method for reducing the number of multipliers to two per formant for both variable and fixed formants has been suggested by C. H. Coker.[8] Fig. 25 shows this method of realizing a digital formant. Differences of the input signal and delayed versions of the output signal are the multiplier inputs, thereby eliminating the output multiplier.

One would expect the noise variance at the output of the formant network of Fig. 25 to be about two-thirds the noise variance of Fig. 3. This is not the case, however. The noise variance at the output node of Fig. 25 is identical to the noise variance at the output of the summer of Fig. 3 since, in both cases, the comparable noises go through identical loops. However, the noise of Fig. 3 is then multiplied by the numerator coefficient which, for frequencies less than 1667 Hz, is less than 1 in magnitude. Hence, the noise of Fig. 3 can be less than the noise of Fig. 25 by an appreciable amount.

The formant network of Fig. 25 was used in Fig. 23 to replace formants 1, 2, and 3 (the low-gain formants), and signal-to-noise ratios were measured and compared with those using the network of Fig. 3. The results are presented in Table VII. Column I shows the signal-to-noise ratio (in bits) using the network of Fig. 25, and column II shows signal-to-noise ratios for the network

[7] The registers containing $y(nT-T)$ and $y(nT-2T)$ will be of the same length as the register containing $y(nT)$.

[8] Private communication.

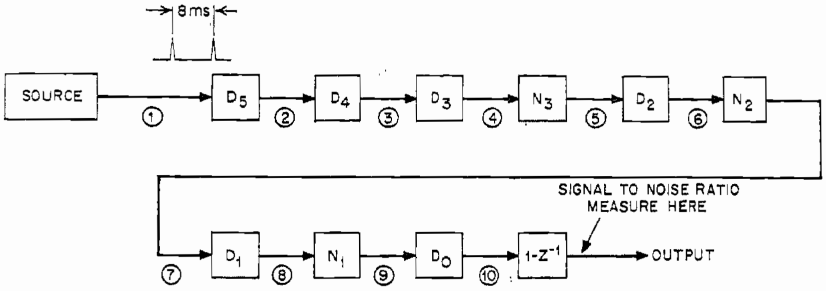


Fig. 24. 543210 sequence of digital formants.

TABLE VI

Register Lengths for 543210 Synthesizer Configuration

Vowel	Node										
	1	2	3	4	5	6	7	8	9	10	SNR (bits)
IY	12	14	15	15	16	16	17	15	10	12	5
I	12	14	15	14	15	15	15	14	10	12	5
E	12	14	15	14	15	14	15	14	10	12	5
AE	12	14	15	14	15	14	14	13	11	12	5
UH	12	14	15	14	15	14	13	12	9	12	5½
A	12	14	15	14	15	14	13	12	10	13	5
OW	12	14	15	14	15	13	12	12	9	12	4
U	12	14	15	14	15	13	12	12	8	12	4½
OO	12	14	15	14	15	13	11	12	7	12	4½
ER	12	14	15	13	13	13	12	13	9	12	5
Maximum over all vowels	12	14	15	15	16	16	17	15	10	13	

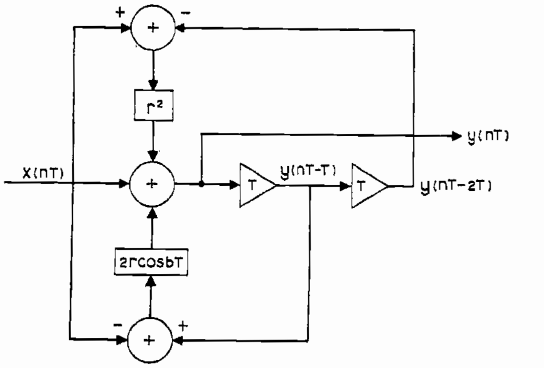


Fig. 25. Digital formant with two multipliers.

TABLE VII

Comparison Between Two Formant Networks

Vowel	I SNR (bits)	II SNR (bits)
IY	2	4½
I	2½	4
E	3	4½
AE	3½	4
UH	3½	5
A	3½	4½
OW	2½	4
U	2	4
OO	1	4
ER	3	4

TABLE VIII

Noise Variance in Bits as a Function of Input Level for Synthesizer of Fig. 23

Input Level (bits)	Vowel									
	IY	I	E	AE	UH	A	OW	U	OO	ER
9	2	2	3	3	2	4	5	3	1	3
10	1	3	2	3	2	3	3	2	1	3
11	2	2	2	3	3	5	3	3	2	2
12	2	3	4	4	4	6	6	3	3	4
13	2	3	4	3	4	5	4	3	2	4
15	1	2	2	2	3	4	3	2	2	4
16	2	3	3	3	4	5	5	3	3	3
17	2	2	2	4	4	5	4	2	2	4
18	2	2	2	3	3	5	4	4	2	3
19	2	4	3	4	3	6	4	4	4	4

of Fig. 3. The signal-to-noise ratios are from ½ bit to 3 bits lower using the network of Fig. 25. Even for the high-gain formants (F_4 and F_5), the network of Fig. 25 provides no advantages over the network of Fig. 3. This is because we do not have to use the high-gain numerator multiplier for these fixed formants. Therefore, the internal noise generated by both networks is identical. However, the network of Fig. 25 automatically includes the high-gain multiplier; therefore, the noise at the input to the network (as well as the signal) will be amplified. This is an undesirable feature when trying to keep register lengths uniform.

Experimental study of the noise generated by a digital formant synthesizer showed that this noise was correlated both with the pitch and the vowel; so much so, that one could detect by eye the pitch period from the noise waveform, and hear the vowel when listening to the noise.

The dependence of the noise variance upon input level was investigated quantitatively using the synthesizer of Fig. 23. The results of this investigation are presented in Table VIII. For any one vowel, the noise variance depends upon the input level, but not in a smooth, continuous way. The peak variation in noise variance (in bits) for any one vowel was 3 bits. Table VIII indicates a fairly significant variation of noise variance with signal level. This variation is greater than would have been expected from Table IV. This is possibly due to the low noise levels of the data of Table VIII. The agreement between theory and experiment may be better when a significant amount of noise is generated, as is the case for the data of Table IV.

References

[1] J. F. Kaiser and F. Kuo, Eds., *System Analysis by Digital Computer*. New York: Wiley, 1966.

[2] C. M. Rader and B. Gold, "Digital filter design techniques in the frequency domain," *Proc. IEEE*, vol. 55, pp. 149–171, February 1967.

[3] R. M. Golden and J. F. Kaiser, "Design of wideband sampled-data filters," *Bell Sys. Tech. J.*, vol. 43, pp. 1533–1546, July 1964.

[4] J. L. Flanagan, C. H. Coker, and C. M. Bird, "Digital computer simulation of a formant-vocoder speech synthesizer," 15th Ann. Meeting of the Audio Engrg. Soc., 1963.

[5] L. R. Rabiner, "Speech synthesis by rule: an acoustic domain approach." Ph.D. dissertation, Dept. of Elec. Engrg., M.I.T., Cambridge, Mass., May 1967.

[6] J. L. Flanagan, *Speech Analysis Synthesis and Perception.* New York: Academic Press, 1965.

[7] G. Fant and J. Martony, "Speech synthesis," Speech Transmission Lab., University of Stockholm, Quart. Prog. Rept., July 1962.

[8] R. S. Tomlinson, "SPASS—An improved terminal-analog speech synthesizer," *J. Acoust. Soc. Am.*, vol. 38, p. 940(A), 1965.

[9] J. F. Kaiser, "Some practical considerations in the realization of linear digital filters," *1965 Proc. 3rd Allerton Conf.*, pp. 621–633.

[10] J. B. Knowles and R. Edwards, "Effect of a finite-word-length computer in a sampled-data feedback system," *Proc. IEE* (London), vol. 112, pp. 1197–1207, June 1965.

[11] B. Gold and C. M. Rader, "Effects of quantization noise in digital filters," *1966 Proc. Spring Joint Computer Conf.*, pp. 213–219.

[12] G. Peterson and H. Barney, "Control methods used in a study of the vowels," *J. Acoust. Soc. Am.*, vol. 24, pp. 175–184, 1952.

[13] H. Dunn, "Methods of measuring vowel formant bandwidths," *J. Acoust. Soc. Am.*, vol. 33, pp. 1737–1746, 1961.

[14] G. Fant, *Acoustic Theory of Speech Production.* 's-Gravenhage: Mouton & Co., 1960.

[15] W. R. Bennett, "Spectra of quantized signals," *Bell Sys. Tech. J.*, vol. 27, pp. 446–472, July 1948.

Bernard Gold (M'49) was born in New York, N. Y., on March 31, 1923. He was graduated from the City College of New York, New York, N. Y., in 1944 with the B.S.E.E. degree, and from the Brooklyn Polytechnic Institute of Brooklyn, Brooklyn, N. Y., in 1948 with the Ph.D. degree in electrical engineering.

He has worked at the Avion Instrument Company, Hughes Aircraft Company, and, since 1953, at the M.I.T. Lincoln Laboratory, Lexington, Mass., on problems of radar, noise theory, pattern recognition, and speech communications. In 1965–1966, he was on leave from Lincoln Laboratory as a Visiting Professor of Electrical Engineering at the Massachusetts Institute of Technology, Cambridge.

Dr. Gold is a member of the Acoustical Society of America.

Lawrence R. Rabiner (S'62–M'67) was born in Brooklyn, N. Y., on September 28, 1943. He received the S.B. and S.M. degrees simultaneously in June, 1964, and the Ph.D. degree in June, 1967, in electrical engineering, all from the Massachusetts Institute of Technology, Cambridge.

From 1962 through 1964, he participated in the cooperative plan in electrical engineering at the Bell Telephone Laboratories, Inc., in Whippany, and Murray Hill, N. J. He worked on digital circuitry, military communications problems, and problems in binaural hearing. Presently he is engaged in research on speech communications at the Bell Telephone Laboratories.

Dr. Rabiner is a member of Eta Kappa Nu, Sigma Xi, Tau Beta Pi, and the Acoustical Society of America.