# XX. SPEECH COMMUNICATION*

## Academic and Research Staff

| | | |
|---|---|---|
| Prof. K. N. Stevens | Prof. A. V. Oppenheim | Dr. J. Suzuki† |
| Prof. M. Halle | Dr. Margaret Bullowa | C.-W. Kim |
| Prof. W. L. Henke | Dr. Paula Menyuk | N. Benhaim |
| Prof. D. H. Klatt | | J. S. Perkell |

## Graduate Students

| | | |
|---|---|---|
| J. K. Frediani | L. R. Rabiner | M. Y. Weidner |
| A. J. Goldberg | R. S. Tomlinson | J. J. Wolf |

## A. ON THE MECHANISM OF GLOTTAL VIBRATION FOR VOWELS AND CONSONANTS

In the study of speech it has been assumed that glottal vibration is independent of the configurations of the supraglottal cavity or, at least, that the effects of different configurations are negligible. In connection with the word saving, for example, it is usually stated that the entire utterance is voiced except for the initial consonant; no distinction is made between the kind of voicing that occurs during the vowels and during the voiced consonant v.

There is some evidence to suggest, however, that the positioning of the vocal cords and the manner in which they vibrate may be quite different, the difference depending upon whether the supraglottal tract is unconstricted, as in a vowel, or is constricted to some degree, as in stops, fricatives, and certain semivowels. This evidence comes in part from an analysis of the mechanism of larynx vibration, and in part from measurements of air-flow events during speech.

Analysis of the mechanism of vocal-cord vibration has shown that the process involves the interaction of several forces. In the beginning of the time interval in which the vocal folds are moving apart, the principal force tending to separate the folds is due to the positive subglottal pressure. During the time interval in which the glottis is closing, on the other hand, the principal force is the Bernoulli force, which is a consequence of the rapid air flow in the glottis.

High-speed motion pictures and other kinds of data have indicated that during vowel production the glottal opening as a function of time is roughly triangular in form, and the glottis is closed over part of the cycle. A typical open time for a male voice might be approximately 4 msec, with a closed interval of comparable duration. The subglottal pressure for normal stressed vowels is typically ~10 cm $H_2O$, and each glottal pulse

*This work was supported principally by the U.S. Air Force (Electronic Systems Division) under Contract AF19(628)-5661; and in part by the National Institutes of Health (Grant 5 RO1 NB-04332-04).

†On leave from Radio Research Laboratories, Tokyo, Japan.

generates a flow of the order of 1 $cm^3$.

Let us examine now what would happen if the vocal tract becomes constricted at some point along its length, and let us suppose, for the moment, that the vocal cords continue to vibrate as before, generating the same volume flow for each cycle. It is possible to make an approximate calculation of the sound pressure immediately above the glottis by convolving the triangular volume-velocity pulse with the impulse response of the vocal tract as seen from the glottis. When the first-formant frequency is low, the response is approximately that of a simple resonant circuit tuned to the frequency of the first formant. This calculation shows that, for a resonant frequency of approximately 250 cps, and an assumed bandwidth of ~100 cps, a 1-$cm^3$ air pulse of 4-msec duration would give rise to a peak sound pressure above the glottis of the order of 10 cm $H_2O$. For higher first-formant frequencies, i.e., for first-formant frequencies that are normally found in vowels, the sound pressure would be considerably less, but for resonant frequencies lower than 250 cps, as found in consonants, this peak sound pressure would be substantially greater. Under the latter conditions, the peak supraglottal sound pressure would be of the order of the steady subglottal pressure, and the pressure drop across the glottis would undergo extreme fluctuations. Flow conditions through the glottis would be greatly modified, and the vocal-cord vibrations would become highly erratic. Regular vocal-cord vibration could be maintained with a constricted vocal tract only under one or both of two conditions: the width of the glottal pulse must increase considerably, and the damping of the first formant must be increased substantially, probably by creating a larger average glottal opening.

In addition to this effect of increased driving-point impedance of the vocal tract at low resonant frequencies, a second consequence of the constricted vocal tract associated with consonantal articulation is that the resistance to air flow at the constriction may become appreciable, thereby causing a rise in average pressure in the mouth. As a result of the heightened mouth pressure, there will be an increase in the net pressure force tending to separate the vocal cords. Furthermore, the air flow through the open glottis will decrease as a consequence of the reduced pressure across the glottis (under the assumption that the subglottal pressure does not change appreciably), thereby causing a reduction of the Bernoulli force. Thus there will be a tendency for the pattern of vibration to change, with the open time being longer, and possibly with the vocal cords remaining separated during the entire cycle. In fact, if vocal-cord vibration is to continue through the constricted consonant, it is reasonable to suppose that an overt adjustment in vocal-cord position toward a more open state is made in order to accommodate vibration with reduced pressure across the glottis. (It is known that this kind of adjustment is made when a vocal-cord vibration is to be maintained during a vowel that is generated with low subglottal pressure.[1-3]

From this simple analysis it is evident that there are rather drastic adjustments in

vocal-cord positioning and in the manner in which the vocal cords vibrate when voicing is to be maintained during certain consonants. Further evidence for this change in laryngeal operation comes from studies of air flow for vowels and for voiced fricative consonants.[4] During a voiced fricative, the air flow tends to be somewhat greater than during a vowel. At first glance, this increased flow is contrary to expectation; one might expect that the supraglottal consonantal constriction would cause an increased resistance to flow and hence a reduction in flow, under the assumption that the subglottal pressure does not change appreciably. The fact that flow is greater during the fricative suggests that the glottal resistance must be less in the consonant than in the vowel, and thus indicates that the open time of the glottis is increased or, more likely, that the vocal cords are separated so that the glottis remains open during the entire vibratory cycle. A similar adjustment of the vocal cords would probably also be made for the production of a voiced-stop consonant in English. In such a consonant, vocal-cord vibration often occurs in the stopped interval before the release of the consonant, and the effects of increased supraglottal pressure and decreased resonant frequency of the supraglottal cavity are even more pronounced than for a voiced fricative.

During the production of a voiceless consonant, the vocal cords are also separated, of course, so that the resistance to air flow at the glottis is small and, under normal circumstances, there is little or no glottal vibration. The vocal-cord separation is much greater for a voiceless consonant than for a voiced consonant. In the case of a voiceless consonant the mouth pressure tends to be higher or to rise more rapidly than during a voiced consonant; this increased mouth pressure would presumably act on the vocal cords to produce a still greater separation.

In the air-flow trace for a vowel preceding a voiced or voiceless consonant there tends to be an increase in air flow in the latter part of the vowel in anticipation of the consonant.[5] It appears that 100 msec or more before the time when the consonantal constriction is achieved an adjustment of vocal-cord positioning is initiated in preparation for the production of the constricted consonant. Likewise, there is a similar time interval following a voiceless consonant before the air flow drops to a value appropriate for a steady vowel. These observations would suggest, therefore, that the rate at which vocal-cord positioning can be achieved is relatively slow, and that a talker must compensate for this slow response in timing his commands to the larynx musculature. In the case of a voiceless consonant following a vowel, it might be argued that the necessary wide separation of the vocal cords can be achieved more rapidly than the more finely adjusted smaller separation for a voiced consonant. Furthermore, the abduction maneuver for a voiceless consonant can be assisted or speeded up by the increased mouth pressure that occurs in such a consonant. This line of reasoning would lead, then, to a logical explanation for the difference in vowel length in English before voiced and voiceless consonants. The longer laryngeal adjustment time required for a voiced

consonant would necessitate an increased duration of the preceding vowel; the consonantal constriction cannot be effected before the vocal cords are positioned in a way that will guarantee uninterrupted vocal-cord vibration during the constricted interval.

Table XX-1. Average vowel durations for symmetrical consonant-vowel-consonant syllables in English. Data for each consonant represent averages over 36 utterances (12 vowels, 3 speakers).

| Consonantal Environment | Vowel Duration (msec) |
|---|---|
| b | 270 |
| m | 240 |
| d | 310 |
| n | 260 |

Partial support for this explanation is found by comparing vowel length before voiced stops and nasals in English. Before a nasal consonant, which presumably does not require the special vocal-cord adjustment, a vowel is found to be consistently shorter than before a voiced-stop consonant. Table XX-1 summarizes measurements of vowel length in symmetrical CVC syllables, averaged over three talkers.

It is evident, then, that during a consonant the vocal cords may be positioned with various degrees of spacing between them. Adjustment of vocal-cord spacing is, however, only one parameter that determines whether or not the vocal cords will vibrate during the consonantal interval. It is important that an adequate glottal air flow be provided if vibration is to be maintained. The amount of air flow is determined both by the subglottal pressure and by the configuration of the supraglottal articulators. The supraglottal tract must have an adequate constriction through which air can flow (as in voiced fricatives, liquids, glides, and nasals) or, in the case of a stopped configuration, continuously increasing volume of the cavities must be provided (as in voiced-stop consonants in English).[6] Thus a talker has under his control a variety of independent or quasi-independent parameters that he can use to generate voiced and voiceless consonants of various types. Any inventory of phonetic features which purports to describe the various stop and fricative consonants in different languages must specify these parameters, which include the subglottal pressure, vocal-cord positioning, and the means by which air is allowed to flow into the supraglottal cavities.

M. Halle, K. N. Stevens

References

1. J. L. Flanagan, Speech Analysis, Synthesis and Perception (Academic Press, New York, 1965), p. 44.
2. A. Bouhuys, D. F. Proctor, and J. Mead, "Kinetic Aspects of Singing," J. Appl. Physiol. 21, 483 (1966).
3. A. Bouhuys, J. Mead, D. F. Proctor, and K. N. Stevens, "Pressure-Flow Events during Singing," Ann. N. Y. Acad. Sci. (in press).
4. D. H. Klatt, K. N. Stevens, and J. Mead, "Studies of Articulatory Activity and Air Flow during Speech," Ann. N. Y. Acad. Sci. (in press).
5. D. H. Klatt, "Articulatory Activity and Air Flow during the Production of Fricative Consonants," Quarterly Progress Report No. 84, Research Laboratory of Electronics, M. I. T., January 15, 1967, pp. 257-260.
6. J. Perkell, "Studies of the Dynamics of Speech Production," Quarterly Progress Report No. 76, Research Laboratory of Electronics, M. I. T., January 15, 1956, pp. 253-257.

## B. HOMOMORPHIC SPEECH PROCESSING*

It is generally accepted that an approximate model for the speech waveform consists of a cascade of a system G whose impulse response g(t) has the form of a single glottal pulse, a system V whose impulse response v(t) corresponds to the response of the vocal tract, and a system R representing the radiation characteristics. The excitation to this system is an impulse train whose spacing corresponds to the fundamental frequency of the resulting waveform.

Both for speech bandwidth compression and basic studies of the nature of the speech wave, it is desirable to attempt to isolate the effects of each of these systems. If we consider that, on a short-time basis, each of these systems is representable as a linear, time-invariant system, then we may represent s(t), the speech waveform, as

$$s(t) = p(t) \otimes w(t) \tag{1}$$

with

$$w(t) = g(t) \otimes v(t) \otimes r(t),$$

where $\otimes$ denotes convolution. The work reported here represents the application of a previously proposed technique for the separation of convolved signals[1] to separation of the effects of p(t), which represents the fine structure in the spectrum of s(t), and w(t), which represents the spectral envelope of s(t). The motivation for this study lies both in the notion of what could conveniently be termed homomorphic deconvolution, and the

*This work was supported in part by Lincoln Laboratory, a center for research operated by the Massachusetts Institute of Technology, with support of the U.S. Air Force.

success of cepstral pitch detection as considered by Noll.[2]

Since the processing illustrated here was simulated on a digital computer, it is convenient to rewrite Eq. 1 in terms of a set of equally spaced samples of s(t). Under the assumption that the pitch period is an integer multiple of the sampling period, it follows from Eq. 1 that

$$s(k) = p(k) \otimes w(k), \tag{2}$$

where s(k), p(k), and w(k) are the $k^{th}$ samples of s(t), p(t), and w(t), respectively, and $\otimes$ now denotes a discrete convolution. To separate p(k) and w(k) from s(k), we wish to operate on s(k) with a transformation D defined by

$$z[D(s(k))] = \log [z(s(k))],$$

where z denotes the Z-transform of the sequence. After linear filtering, the result is then transformed by the inverse of the transformation D. If p(k) contains N unit samples spaced by r and starting at k = 0, then it can be shown that

$$D[p(k)] = \frac{r}{k} \sum_{m=1}^{\infty} \delta(k-mr) - \frac{Nr}{k} \sum_{m=1}^{\infty} \delta(k-mNr) \qquad k \neq 0$$

where

$$\delta(k-j) = 1 \qquad k = j$$

$$= 0 \qquad k \neq j.$$

Under certain conditions, we may write w(k) approximately as

$$w(k) = v(k) \otimes h(k),$$

where v(k) are samples of v(t), and h(k) are samples of g(t) $\otimes$ r(t). If we represent v(t) by the impulse response of a simple acoustic cavity with small loss, so that the transfer function of v(t) is

$$|V(j\omega)| = \left| \frac{1}{\cosh \left[ a + j\frac{\omega}{c} \right] \ell} \right|,$$

then

$$v(k) = \sum_{m=0}^{\infty} (-1)^m e^{-2a\ell m} \delta\left(k - \frac{2m\ell}{c}\right) \qquad k \neq 0$$

and

$$D[v(k)] = \frac{2\ell}{kc} \sum_{m=1}^{\infty} (-1)^m e^{-2a\ell m} \delta\left(k - \frac{2\ell m}{c}\right). \qquad k \neq 0$$

Under the idealization that g(t) is a symmetrical triangular pulse of duration $T\Delta$, where $\Delta$ is the sampling rate, and that r(t) can be represented by a differentiator,

$$D[h(k)] = \frac{1}{k} - \frac{T}{k} \sum_{m=0}^{\infty} \delta\left(k - \frac{mT}{2}\right). \qquad k \neq 0$$

We observe, in this case, that D[h(k)] and D[v(k)] provide their primary contribution for small values of k, as compared with D[p(k)]. By multiplying the sequence D[s(k)] by unity up to some value of k, and zero thereafter, we should, to some approximation, have retained only values relevant to h(k).

The processing described above was carried out on the Lincoln Laboratory TX-2 computer. To determine the response of the system D, the inverse transform of the logarithm of the discrete Fourier transform was computed. The use of the discrete transform, rather than the Z-transform, introduces an aliasing effect into the expressions derived previously, although it is usually possible to minimize this effect. For the present studies, phase information was not included; this means working with the autocorrelation function of the speech sample, rather than with the speech sample itself. In this case, the sequence $D[\phi_{ss}(k)]$, with $\phi_{ss}(k)$ denoting the autocorrelation function of s(k), is equivalent to the cepstrum.

To illustrate the result of this processing, consider the sample of the vowel "ah" in father as shown in Fig. XX-1, representing 41 msec sampled at 25 kHz. The spectrum

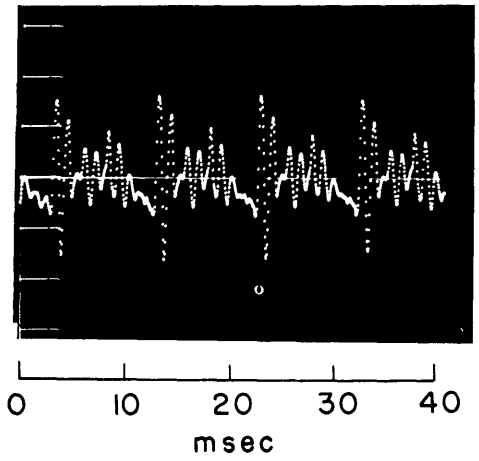


Fig. XX-1. 41 msec of "ah" spoken in isolation sampled at 25 kHz.

of this sample was computed by weighting the speech sample with a raised cosine window, 41 msec long, followed by 41 msec of zero. The logarithm of the magnitude of

the spectrum obtained in this way is shown in Fig. XX-2.

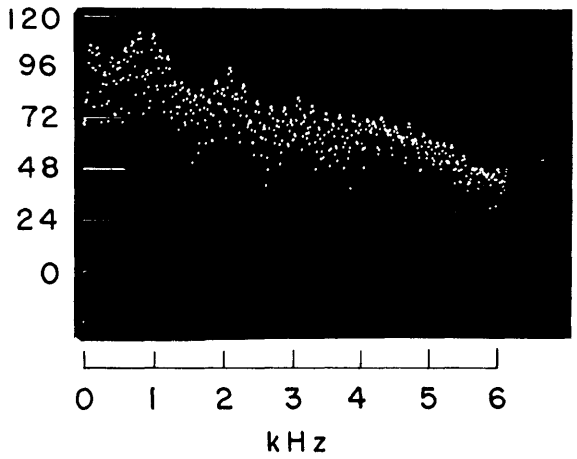


Fig. XX-2. Log magnitude of the spectrum of Fig. XX-1 after weighting with a raised cosine window and terminating in 41 msec of zero.

Figure XX-3 represents the inverse transform of the spectrum of Fig. XX-2, with a pronounced peak occurring at a time corresponding to the pitch period.

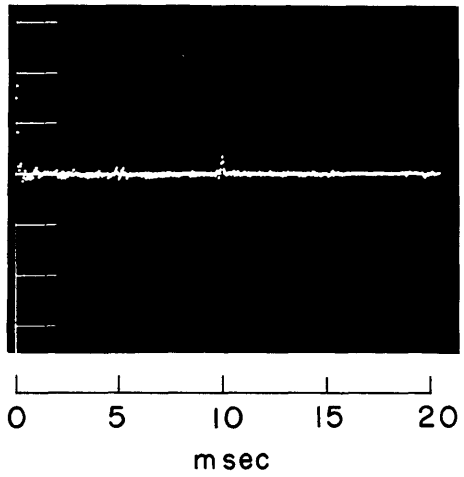


Fig. XX-3. Inverse transform of the spectrum of Fig. XX-2.

As described above, filtering of the logarithmic spectrum is carried out by multiplying the series of Fig. XX-3 by unity until some time $\tau$ and zero thereafter. Figures XX-4, XX-5, and XX-6 represent the result of filtering and transforming for $\tau$ = 7.7 msec,

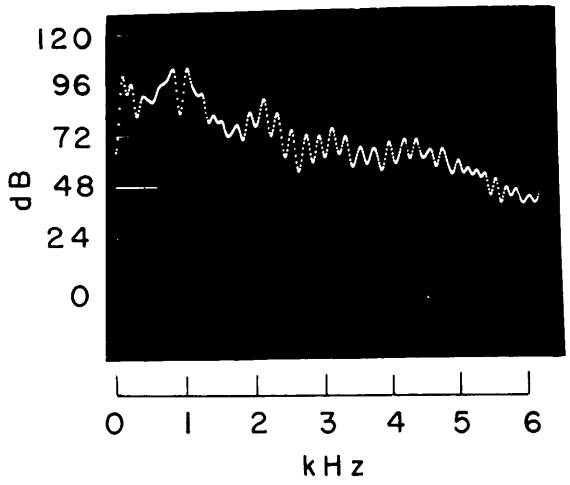


Fig. XX-4. Low-time filtered log magnitude spectrum with filter cutoff time, $\tau$ = 7.7 msec.

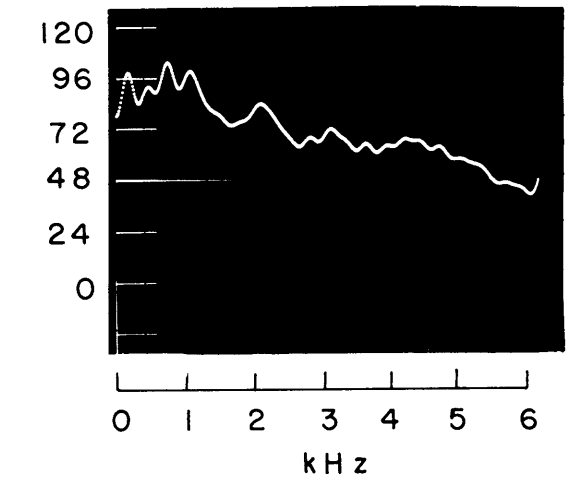


Fig. XX-5. Low-time filtered log magnitude spectrum with filter cutoff time, $\tau$ = 3.8 msec.

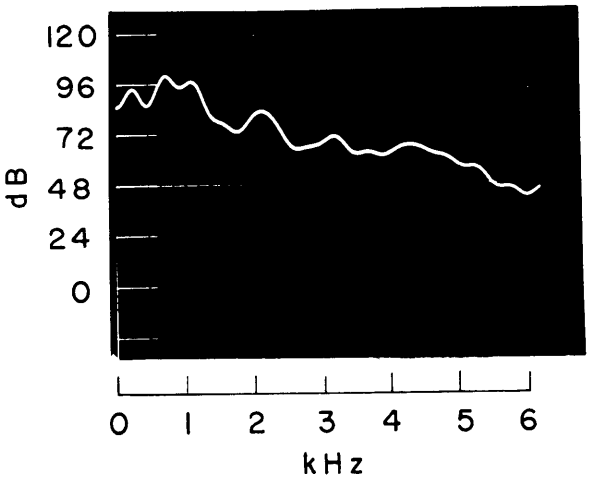


Fig. XX-6. Low-time filtered log magnitude spectrum with filter cutoff time, $\tau$ = 2.6 msec.

$\tau$ = 3.8 msec, and $\tau$ = 2.6 msec, respectively. Figure XX-7 shows the superposition of the spectra of Fig. XX-2 and Fig. XX-5.

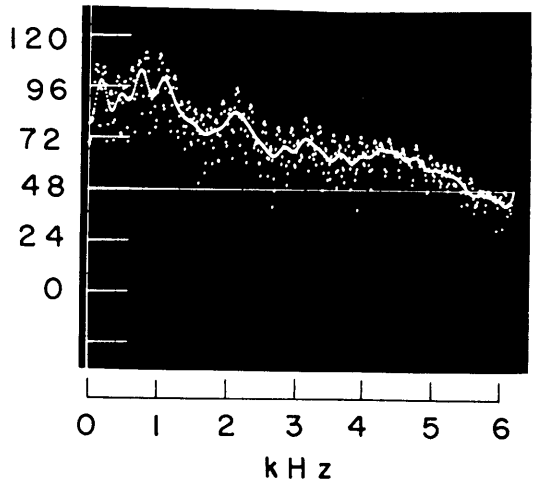


Fig. XX-7. Series of Fig. XX-2 and Fig. XX-5 superimposed.

To implement the inverse of the system D, the filtered spectrum will, in general, be operated on by an exponential transformation, and then the inverse Fourier transform will be taken. The result of an exponential transformation on the spectrum of Fig. XX-6 is shown in Fig. XX-8.

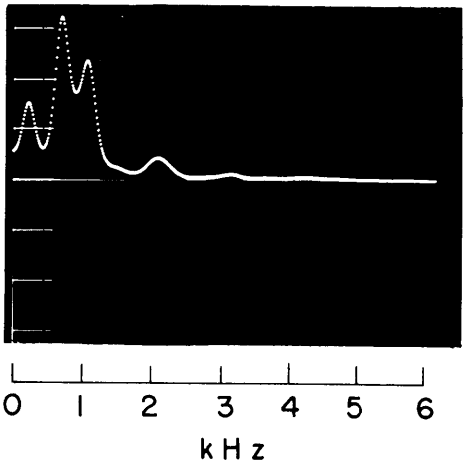


Fig. XX-8. Spectrum of Fig. XX-5 after an exponential transformation. Vertical scale is linear.

In a case for which high-frequency emphasis of this smoothed spectrum is desired, the series of Fig. XX-3 can be "band-time filtered" rather than "low-time filtered." The result of introducing high-frequency emphasis for the spectrum of Fig. XX-8 is shown in Fig. XX-9.

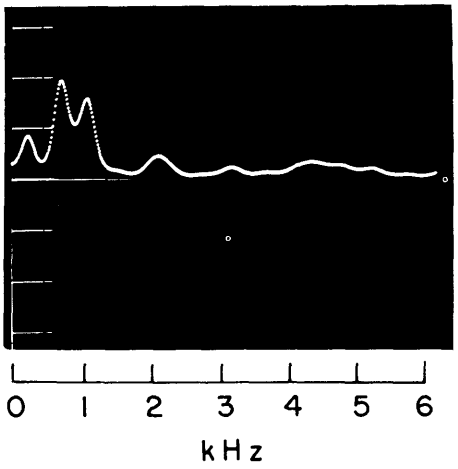


Fig. XX-9. Spectrum of Fig. XX-5 after an exponential transformation with high-frequency emphasis obtained by band-time filtering. Vertical scale is linear.

In comparing the spectra of Figs. XX-4, XX-5, and XX-6, we observe that for values of $\tau$ slightly less than the pitch period, a ripple is apparent in the spectrum. As $\tau$ decreases the ripple disappears, bringing more into evidence the individual formants, while apparently broadening the resonance bandwidths. The ripple for larger values of $\tau$ is apparently contributed by the terms that are due to the glottal pulse, the spacing of the peaks being related to glottal pulse duration.

In continuing this study, the objective is to isolate the effects of pitch, glottal pulse, and vocal-tract impulse response. Although for constant pitch, the pitch period can be measured directly from the series of Fig. XX-3, the contribution from pitch is generally more pronounced after "long-time" filtering and operating on the resulting series with the inverse of the system, D. To recover the glottal pulse, the log spectrum corresponding to a series of resonators will be subtracted from the smoothed log spectrum of Fig. XX-4 and the result operated on with the inverse of the system, D. In order to recover the glottal pulse, rather than its minimum phase counterpart, phase information must be included in the determination of the series corresponding to Fig. XX-3. This procedure for recovering the glottal pulse is similar to inverse filtering; the parameters of the inverse filter are obtained by a spectral matching procedure on the smoothed log spectrum.

A. V. Oppenheim

## References

1. A. V. Oppenheim, "Nonlinear Filtering of Convolved Signals," Quarterly Progress Report No. 80, Research Laboratory of Electronics, M.I.T., January 15, 1966, pp. 168-175.

2. A. M. Noll, "Short-time Spectrum and Cepstrum Techniques for Vocal-Pitch Detection," J. Acoust. Soc. Am. 36, 296-302 (1964).