XII. SPEECH ANALYSIS*

Prof. M. Halle
G. W. Hughes
A. R. Adolph

A. STUDIES OF PITCH PERIODICITY

In the past a number of devices have been built to extract pitch-period information from speech. These efforts have stemmed mainly from research on communication systems, the immediate goal being bandwidth reduction. For this purpose, some measure of the invariant or nonredundant properties of speech is desired. The fundamental pitch-periodicity or glottal excitation frequency is one such invariant.

Since pitch information was only one aspect of one dimension (speech) of a complex multidimensional code (language), the devices that were evolved en route to bandwidth reduction were based on rather idealized and simplified representations of the speech waveform. The devices did not work, or made errors, over a significant portion of commonly encountered speech sounds. The limited dynamic range of physically realizable systems, improperly utilized, caused errors during transient portions of the speech signals. In tests with a small number of sample inputs, errors of 5 per cent to 20 per cent were reported.

Four distinct approaches to the problem of pitch extraction were made by Dolansky (1), Gruenz and Schott (2), Lerner (3), and Miller and Weibel (4). Their work does not necessarily represent the first or only examples of these approaches but it illustrates the principles involved. Other devices in use are similar to one or more of these four. All of the aforementioned devices work reasonably well for an input consisting of a steady-state, decaying waveform, such as might result when one speaker said the long vowel /e/, as in "gay" in Fig. XII-1a, or /a/, as in "father." However, when a perfectly understandable vowel like /I/ in Fig. XII-2a, or /o/ in Fig. XII-3a occurs, they make errors.

Dolansky's system, which depends on obtaining the envelope of the input speech by asymmetrical detection, would be in error on /I/ in Fig. XII-2a. Figure XII-2b shows the speech half-wave rectified. The effect of the second peaks would be enhanced by integration. A similar effect would occur on /o/ in Fig. XII-3a and b.

Gruenz and Schott use low-pass filtering to extract the fundamental. A problem arises in deciding on a cutoff frequency for the filter. Referring to Fig. XII-2d and e, it is clear that a cutoff high enough to allow both male and female pitches to be extracted (i.e., 500 cps) will be unsuitable. This point is further illustrated in Fig. XII-3d and e, and in Fig. XII-4d and e.

---

* This work was supported in part by the National Science Foundation.

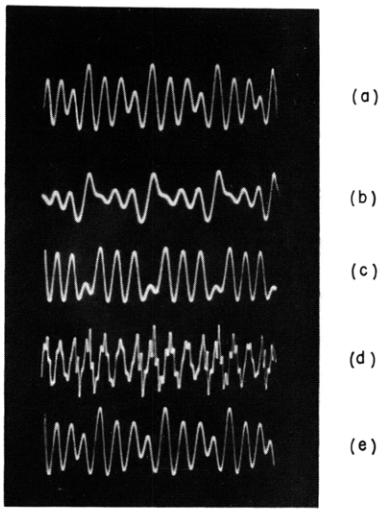

Fig. XII-1. a. Steady-state /e/ sound, as in the word "gay." Male; τ_p = 9 msec.
b. Full-wave rectified, 500 cps low-pass filtered.
c. Half-wave rectified, 500 cps low-pass filtered.
d. Bandpass filtered 600-2500 cps.
e. Bandpass filtered 250-700 cps.

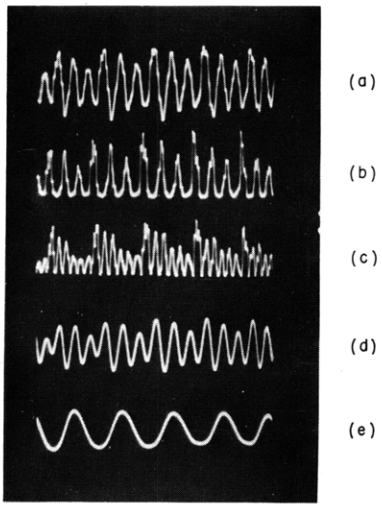

Fig. XII-2. a. Steady-state /I/ sound, as in the word "sill." Male; τ_p = 7.5 msec.
b. Half-wave rectified.
c. Full-wave rectified.
d. 500 cps low-pass filtered 2a.
e. 200 cps low-pass filtered 2a.

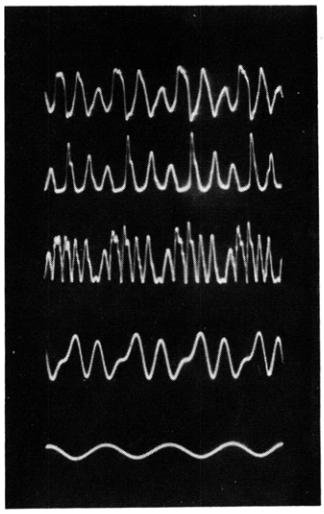

Fig. XII-3. a. Steady-state /o/ sound, as in "doze." Female; τ_p = 6 msec.
b. Half-wave rectified.
c. Full-wave rectified.
d. 500 cps low-pass filtered.
e. 200 cps low-pass filtered.

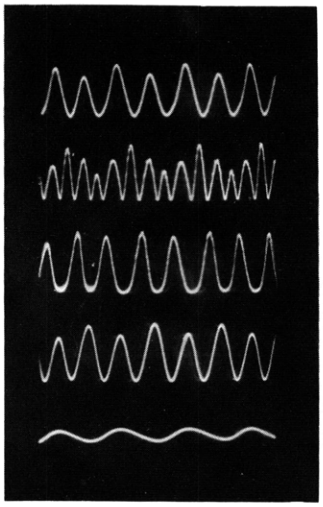

Fig. XII-4. a. Steady-state /u/, as in "spool." Female; τ_p = 6 msec.
b. Full-wave rectified.
c. Half-wave rectified.
d. 500 cps low-pass filtered.
e. 200 cps low-pass filtered.

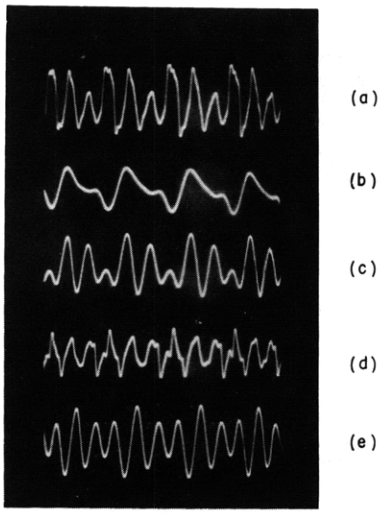

Fig. XII-5. a. Steady-state /o/, as in "doze." Female; τ_p = 6 msec.
b. Full-wave rectified, 500 cps low-pass filtered.
c. Half-wave rectified, 500 cps low-pass filtered.
d. Bandpass filtered 600-2500 cps.
e. Bandpass filtered 250-700 cps.

Lerner represents the speech signal as a rotating vector and depends on a discontinuity in amplitude to indicate the period. Such a system will make an error on vowel examples that exhibit a high first-formant amplitude or correspondingly relatively small decay of their envelopes.

Miller was more cognizant of vowel properties in his approach than were the others. By using a long delay line with a great number of taps which were scanned and whose outputs were subtracted from the undelayed speech, he added a "shape" correlation dimension to the amplitude dimension, and reduced the degree of uncertainty in the pitch determination. Errors caused by transient amplitude variations and formant frequency shifts can occur; however, these would be smaller in number for a true multiplying-integrating type of correlator. Vowels such as the /u/ in Fig. XII-4a illustrate a difficult case for a correlation method of pitch extraction.

One promising method has been examined by the author: full-wave rectification followed by low-pass filtering. The effect of full-wave rectification is to shift the frequencies of energy concentration (formants) upward without changing the pitch frequency. This can best be visualized by thinking of the waveform in terms of its zero-crossing frequency and its fundamental pitch. The zero-crossing frequency is a function of the formant positions. When the waveform is full-wave rectified the effective zero-crossing frequency is doubled, while the fundamental pitch remains unchanged. This is readily seen in Fig. XII-2b and c. Thus a much higher low-pass filter cutoff

frequency can be used, one that is high enough to be effective for both male and female voices, as illustrated in Figs. XII-1b and XII-5b. Another result of the current work is that the second formant energy alone is of little value for pitch determination. See Figs. XII-1d and XII-5d.

A. R. Adolph

References

1. L. O. Dolansky, An instantaneous pitch-period indicator, J. Acoust. Soc. Am. 27, 67-72 (1955).
2. O. O. Gruenz and L. O. Schott, Extraction and portrayal of pitch of speech sounds, J. Acoust. Soc. Am. 21, 487-495 (1949).
3. R. Lerner, Pitch synchronous chopping of speech, Quarterly Progress Report, Research Laboratory of Electronics, M.I.T., Oct. 15, 1956, p. 99.
4. R. L. Miller and E. S. Weibel, Measurement of the fundamental period of speech using a delay line, J. Acoust. Soc. Am. 28, 761 (1956).

B. VARIATIONS OF FORMANT INTENSITY WITH PITCH

The energy contained in a vowel formant may vary with fundamental pitch. It will be near maximum when the ratio of formant frequency to pitch frequency is an integer, x, and near minimum when this ratio equals an odd multiple of 1/2 greater than one.

Calculations were made with a parallel RLC circuit as a simple model of vocal tract resonance. The network was assumed to be driven by periodic unit impulses of current of period $2\pi/\omega_o$. Taking $\log_{10}$ of the sum of $|Z(j\omega)|^2$ at each of the component frequencies of the voltage across the elements will give a measure of the relative power, in decibels, passed by the resonant circuit for various values of the above ratio. Since

$$|Z(j\omega)|^2 = \frac{L/C}{\left(\frac{1}{Q_o}\right)^2 + \left[\frac{\omega_1}{\omega} - \frac{\omega}{\omega_1}\right]^2} \tag{1}$$

the relative total power is

$$\frac{L}{C}\left[\frac{1}{\left(\frac{1}{Q_o}\right)^2 + \left[x - \frac{1}{x}\right]^2} + \frac{1}{\left(\frac{1}{Q_o}\right)^2 + \left[\frac{x}{2} - \frac{2}{x}\right]^2} + \frac{1}{\left(\frac{1}{Q_o}\right)^2 + \left[\frac{x}{3} - \frac{3}{x}\right]^2} + \cdots\right] \tag{2}$$

where $x = \omega_1/\omega_o$, $\omega_1 = LC^{-1/2}$, and $Q_o = \omega_1 RC$.

In Fig. XII-6 the quantity in brackets in Eq. 2 is plotted against x for three values

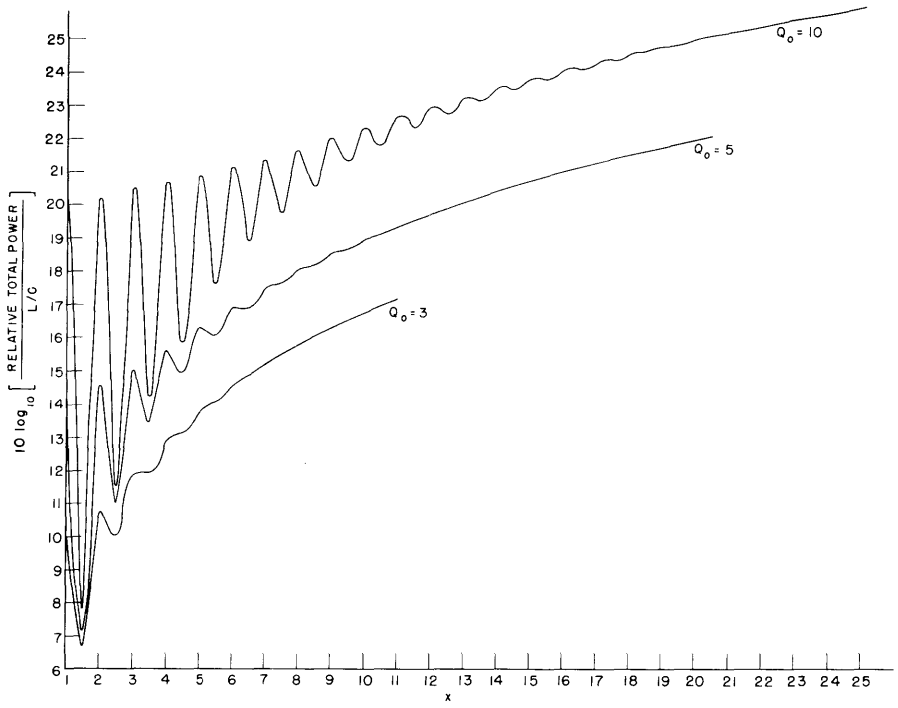

Fig. XII-6. Output of parallel RLC circuit versus ratio (x) of the resonant frequency to the frequency of impulses.

of Q_o. Some idea of how formant intensity in speech varies with pitch is given from the shapes of these curves, although the ordinate numbers themselves mean little. This effect is important for the very low formant frequency position only, since a Q_o of 3 to 5 is appropriate for most vocal tract resonances. Appropriate weighting curves could be applied if the excitation is not a pure impulse.

G. W. Hughes