Zero-Crossing Measurements for Analysis and Recognition of Speech Sounds

MABO ROBERT ITO, Member, IEEE
Radio and Electrical Engineering Division
National Research Council of Canada
Ottawa, Ont., Canada

ROBERT W. DONALDSON
Department of Electrical Engineering
University of British Columbia
Vancouver, B. C., Canada

Abstract

Existing mathematical relations between the power spectral density and the mean zero-crossing rate of an ergodic random process are used to derive relations between spectral measurements and mean zero-crossing rates of a speech signal and its derivatives. Of particular significance is the equation relating zero-crossing rates to the formant parameters of vowels and vowel-like sounds. Reasonably close agreement between measured zero-crossing rates and those calculated from spectral measurements were observed for virtually all phonemes in a variety of contextual environments.

Measured zero-crossing rates and corresponding calculated values for speech and its first derivative are presented for vowels, unvoiced fricatives, and unvoiced stops, all in many different contextual environments. Unvoiced fricatives /s/, /ʃ/, and /f/ are shown to be distinguishable from each other solely on the basis of the zero-crossing rate of the derivative signal. Some vowels are shown to be differentiable from other vowels, although measurements other than zero-crossing rates of filtered, unfiltered, and differentiated speech are shown to be necessary for complete vowel separation. For unvoiced stop consonants, the zero-crossing rate of either the signal or its derivative is shown to be useful for classification, provided some information concerning the contextual environment is available.

Manuscript received May 11, 1970; revised February 16, 1971.

This work was supported by the National Research Council of Canada under Grant NRC 3308 and by The Defence Research Board of Canada under Grant DRB 2801-26.

I. Introduction

During the past few years zero-crossing rates (ZCR) of speech signals have proved useful for analysis, segmentation, and recognition of speech sounds. ZCR measurements are well suited to digital processing, virtually independent of talker volume, and apparently less speaker dependent than spectral data [1]. Peterson [2] used the ZCR to estimate the resonant frequency of a periodic signal operated on by an RLC circuit. Reddy [3]–[5] used the amplitude and ZCR of speech signals for segmentation, for preliminary classification of the segmented sounds, and as an aid in final classification. Reddy and Vincens [6] used the amplitude and ZCR of low-pass and high-pass filtered speech for segmentation. Bezdel and Chandler [7] used the ZCR of speech to classify five different vowels. Bezdel and Bridle [8] later used the ZCR of low-pass and high-pass filtered speech to recognize the digits 1 to 9 with 90 percent accuracy. Scarr [9] achieved some success in relating the ZCR of vowels to vowel formant frequencies by using an approximate method to estimate the number of times a deterministic periodic signal crosses zero in a given time interval.

The research described in this paper was conducted in order to further explore the use of ZCR of unfiltered, filtered, and differentiated speech waveforms for analysis and recognition, and to express ZCR in terms of the more commonly used spectral characterizations. Expressing ZCR in terms of spectral data is accomplished by modeling any segment of a speech waveform as a sample function from an ergodic random process. Such a model is realistic and avoids difficulties previously encountered in attempting to relate ZCR for vowels and vowel-like sounds to spectral measurements [9].

Measured ZCR for all phonemes in a variety of contextual environments were observed to be in reasonable agreement with ZCR calculated from spectral data. Measured ZCR and corresponding calculated values are shown below for vowels, unvoiced fricative consonants, and unvoiced stop consonants. Unvoiced fricatives /s/, /ʃ/, and /f/ were found to be differentiable from each other solely on the basis of the ZCR of the derivative signal. Some vowel separation was observed, although measurements other than ZCR for filtered, unfiltered, and differentiated speech appear necessary for complete vowel separation. For unvoiced stop consonants, the ZCR of either the signal or its derivative is shown to be useful for classification, provided some information concerning the contextual environment is available.

II. Zero-Crossing Rates of Random Processes

Let $x(t)$ be a stationary random process with autocorrelation function $R_x(\tau)=\mathrm{E}[x(t)x(t-\tau)]$ and power density spectrum

$$S_x(f) = \int_{-\infty}^{\infty} R_x(\tau)e^{-j2\pi f\tau}d\tau,$$

where expectation E is an average over the ensemble of waveforms in the process. Chang, Pihl, and Essigmann [10] have shown that the expected zero-crossing rate z_0 of $x(t)$ is given by the following equation.

$$z_0 = 2\phi_0\left[\int_{-\infty}^{\infty} f^2 S_x(f)df \Big/ \int_{-\infty}^{\infty} S_x(f)df\right]^{1/2} \quad (1)$$

$$\phi_0 = \left[\int_{-\infty}^{\infty} |v/\xi_1|\, p(u/\xi_0, v/\xi_1)d(v/\xi_1)\right]_{u=0}. \quad (2)$$

In (2), $p(u, v)$ is the joint amplitude probability density of $x(t)$ and its first derivative, and ξ_0 and ξ_1 are, respectively, the variances of $x(t)$ and dx/dt.

Let $x^n(t)$ be the nth derivative of $x(t)$ and let z_n be the expected zero-crossing rate of $x^n(t)$. The power spectral density of $x^n(t)$ is $f^{2n}S_x(f)$, which is an even function of f. It follows from (1) and (2) that

$$z_n = 2\phi_n\left[\int_0^{\infty} f^{2n+2} S_x(f)df \Big/ \int_0^{\infty} f^{2n} S_x(f)df\right]^{1/2}. \quad (3)$$

In (3), ϕ_n is given by (2), with $p(u, v)$ now defined to be the joint amplitude probability density of $x^n(t)$ and $x^{n+1}(t)$ and ξ_0 and ξ_1 defined, respectively, as the variances of $x^n(t)$ and $x^{n+1}(t)$.

For stationary Gaussian processes, $\phi_n = 1$ [10], [11]. When $x(t) = A\sin(2\pi\alpha[t+\theta])$, where A is any constant and θ is a random variable uniformly distributed between 0 and $1/\alpha$, direct calculation shows that $z_n = 2\alpha$, and that

$$2\left[\int_0^{\infty} f^{2n+2} S_x(f)df \Big/ \int_0^{\infty} f^{2n} S_x(f)df\right]^{1/2} = 2\alpha. \quad (4)$$

Thus, $\phi_n = 1$ for this sinusoidal random process, which suggests that ϕ_n may be close to unity for some non-Gaussian random processes.

In many cases of interest $x(t)$ can be considered, over a small time interval Δ, to be a portion of a sample function from an ergodic random process having power spectral density $S_x(f)$. In this case time and ensemble averages are equal, with the result that z_n in (3) approximates the mean zero-crossing rate of $x^n(t)$ during time interval Δ.

III. Data Base and Data Processing

Words used in our study were from lists used by Hughes and Halle [12] and Halle, Hughes, and Radley [13] to study the static spectral properties of fricative and stop consonants. All words used were spoken by a 29-year-old male having a western Canadian accent. The words were recorded in a soundproof room using a high-quality audio tape-recording system consisting of an Altec–Lansing-type 681A microphone and a Tandberg-type 64X tape recorder. During recording, care was taken to obtain natural, but well articulated words. All word-final stop consonants, for example, were required to be released. Our data were found to be typical in that its spectral details were similar to those of others [12]–[17].

Processing of the recorded words was done by a PDP-9 digital computer having 16K 18-bit words of core memory, 3 Dectape transports, a display and light pen, a 32-channel multiplexer, and D/A and A/D converters. The display and light pen, in conjunction with audio monitoring, proved very useful for on-line selection and display of desired portions of speech waveforms, and for display of the results of the various processing operations. Digital recording of analog speech waveforms followed perfiltering, sampling at 16 kHz, and 9-bit uniform quantization. Prefilters used included low-pass filters of 8- and 1-kHz bandwidth and a band-pass filter having a 1–8-kHz passband. For convenience, 8-kHz low-pass filtered speech will be referred to as normal speech. The low-pass and bandpass filters were Khronkite #3342 filters having a 48-dB/decade cutoff rate. Also used was a 32-channel filter-bank spectrum analyzer with center frequencies logarithmically spaced from 300 Hz to 7.6 kHz. Each channel consisted of a third-order Chebychev bandpass filter, a half-wave rectifier, and an RC smoothing filter with a 10-ms time constant.

As a first step in processing, the digitized computer-input signal was divided into adjacent 10-ms time intervals. The amplitude A assigned to any 10-ms interval was obtained by averaging the amplitudes of all 160 samples in the interval. Zero-crossing rates $z_0 = z$ and $d = z_1$ for the interval were obtained from the following equations, where $N = 160$ and x_k is the amplitude of sample k in the interval.

$$z = \sum_{k=1}^{N} [1 - \operatorname{sgn}(x_{k+1})\operatorname{sgn}(x_k)]/2 \quad (5)$$

$$d = \sum_{k=1}^{N} [1 - \operatorname{sgn}(\Delta_{k+1})\operatorname{sgn}(\Delta_k)]/2 \quad (6a)$$

$$\Delta_k = [x_{k+1} - x_k]. \quad (6b)$$

If Δt is the time between samples x_k and x_{k+1}, then $\Delta_k/2\Delta t$ is an estimate of dx/dt at the time corresponding to sample x_k. In our work $\Delta t = (1/16)$ms. Fig. 1 shows amplitude $A(t)$ and zero-crossing rates $z(t)$ and $d(t)$, quantized to 10-ms time intervals, for the normal version of the word "sit."

IV. Zero-Crossing Analysis of Vowels and Vowel-Like Sounds

Most speech production models of vowels, semivowels, glides, and nasals consist of a linear filter excited by a periodic signal $p(t)$ having the following form [18], [19].

$$p(t) = \sum_{k=0}^{\infty} [p_k \cos 2\pi k\alpha(t+\theta) + \hat{p}_k \sin 2\pi k\alpha(t+\theta)]. \quad (7)$$

In (7) α is the fundamental voicing frequency, p_k and $\hat{p}_k$ are constants, and θ is a random variable uniformly distributed between 0 and $1/\alpha$. This same model has also been used in virtually all speech synthesizers [20]–[26]. The acoustic output signal has the same form as (7), except that the coefficients p_k and $\hat{p}_k$ are changed by the filter. The assumption that θ in (7) is a uniformly distributed random

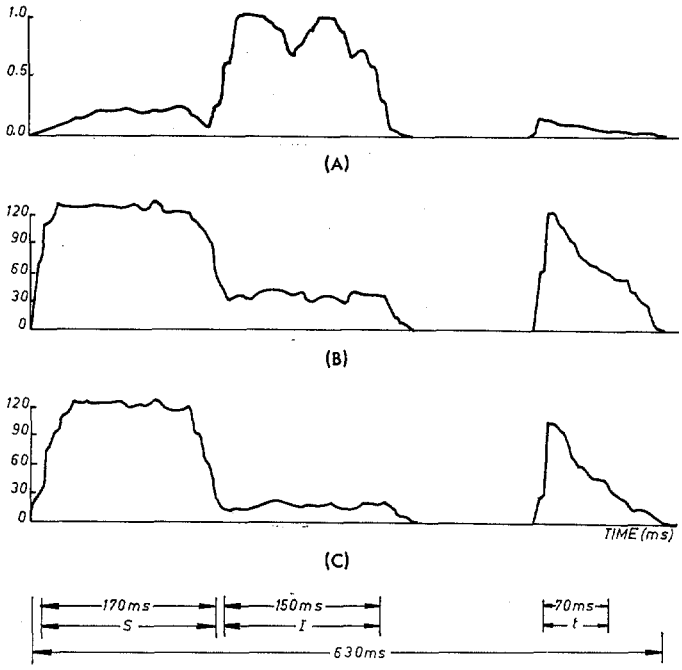

Fig. 1. Experimental data for "sit." Phonemes /s/, /I/, and /t/ resulting from visual segmentation are indicated. (A) Amplitude. (B) Zero-crossing rate d of derivative of normal speech. (C) Zero-crossing rate z of normal speech.

variable is needed to enable stable regions[1] of the acoustic output signal to be regarded as segments from sample functions of ergodic random processes. The assumption is realistic because phase angle θ is not essential for listener identification.

If the transfer function of the filter over a stable region of the phoneme is $H(f)$ then, the power spectral density of the acoustic output signal over this stable region is

$$S_x(f) = \sum_{k=-\infty}^{\infty} |H(f)|^2 P_k^2 \delta(f - k\alpha) \tag{8a}$$

where $\delta(f)$ is the unit impulse at $f=0$ and

$$P_k^2 = \begin{cases} [p_{|k|}^2 + \hat{p}_{|k|}^2]/4, & k \neq 0 \\ p_0^2, & k = 0. \end{cases} \tag{8b}$$

An appropriate choice for $H(f)$ is [18], [19]

$$H(f) = \sum_{i=1}^{N} K_i \left\{ \frac{\sigma_i^2 + (2\pi f_i)^2}{[\sigma_i + j2\pi(f + f_i)][\sigma_i + j2\pi(f - f_i)]} \right\} \tag{9}$$

where positive real numbers f_i, σ_i, and K_i are, respectively, the center frequency, bandwidth, and gain parameter of the vocal tract's ith resonance. Substitution of (8a) into (3) yields the following equation for z_n in terms of the harmonic amplitudes $|H(\alpha k)|^2 P_k^2$ of the vowel or vowel-like sound.

[1] A stable region of a random process is one where the statistics of interest are virtually constant over the region.

$$(z_n/2\phi_n)^2 = \sum_{k=0}^{\infty} (k\alpha)^{2n+2} |H(k\alpha)|^2 P_k^2 \Big/ \sum_{k=0}^{\infty} (k\alpha)^n |H(k\alpha)|^2 P_k^2. \tag{10}$$

To express z_n in terms of formant parameters, $|H(f)|^2 = H(f)H(-f)$ is first expanded in partial fractions, with use being made of the quadratic symmetry of the poles of $|H(f)|^2$, as follows (* denotes complex conjugate).

$$|H(f)|^2 = \sum_{i=1}^{N} \left[\frac{B_i}{\sigma_i + j2\pi(f - f_i)} + \frac{B_i^*}{\sigma_i + j2\pi(f + f_i)} + \frac{B_i^*}{\sigma_i - j2\pi(f - f_i)} + \frac{B_i}{\sigma_i - j2\pi(f + f_i)} \right] \tag{11a}$$

where

$$B_i = \{[\sigma_i + j2\pi(f - f_i)]\,|H(f)|^2\}_{f = \left(f_i + \frac{j\sigma_i}{2\pi}\right)} \tag{11b}$$

Let

$$S_x(f) = \sum_{i=1}^{N} S_{x_i}(f) \tag{12a}$$

where

$$S_{x_i}(f) = \sum_{k=0}^{\infty} P_k^2 \delta(f - k\alpha) \left[\frac{B_i}{\sigma_i + j2\pi(f - f_i)} + \frac{B_i^*}{\sigma_i + j2\pi(f + f_i)} + \frac{B_i^*}{\sigma_i - j2\pi(f - f_i)} + \frac{B_i}{\sigma_i - j2\pi(f + f_i)} \right]. \tag{12b}$$

For all vowels and vowel-like sounds, $\sigma_i \ll f_i$, with the result that B_i and $S_{x_i}(f)$ for $f>0$ are closely approximated as follows.

$$B_i \simeq \sigma_i |H(f_i)|^2 \tag{11c}$$

$$S_{x_i}(f) \simeq \sum_{k=0}^{\infty} 2P_k^2 \delta(f - k\alpha)\sigma_i^2 |H(f_i)|^2 \cdot \left[\frac{1}{\sigma_i^2 + 4\pi^2(f - f_i)^2} \right], \qquad f > 0. \tag{12c}$$

Equation (12c) now yields the following equation.

$$\int_0^{\infty} f^q S_{x_i}(f)df \simeq \sum_{k=0}^{\infty} 2P_k^2 (k\alpha)^q \sigma_i^2 |H(f_i)|^2 \cdot \left[\frac{1}{\sigma_i^2 + 4\pi^2(k\alpha - f_i)^2} \right]. \tag{13}$$

The significant terms in (13) are those involving values of k in the region near f_i/α. Because the bandwidth σ_i of the ith resonance is small in comparison with the fundamental vowel frequency, P_k^2 in (13) can be approximated by P_{β_i}

where β_i is the integer nearest to f_i/σ. The term $(k\alpha)^q$ can be approximated by $f_i{}^q$, since its effect on the sum in (13) will be small in comparison with that of $[\sigma_i{}^2+4\pi^2(f-f_i)^2]^{-1}$. Approximation of (13) by an integral and subsequent use of (12a) yields the following equations.

$$\int_0^\infty f^q S_{x_i}(f)df \simeq f_i{}^q P_{\beta_i}{}^2 B_i/\alpha \tag{14}$$

$$\int_0^\infty f^q S_x(f)df \simeq \sum_{i=1}^{N} f_i{}^q \sigma_i A_i{}^2/4\alpha \tag{15}$$

where

$$A_i = 2\,|P_{\beta_i}|\,\sqrt{B_i/\sigma_i}. \tag{16}$$

Note that A_i closely approximates the amplitude of the harmonic of the acoustic signal near $f=f_i$, and that f_i closely approximates the frequency of the ith formant. From (15), (16), and (3) one now obtains the following equation for z_n.

$$(z_n/2\phi_n) = \sum_{i=1}^{N} f_i{}^{(2n+2)} A_i{}^2\sigma_i \Big/ \sum_{i=1}^{N} f_i{}^{2n} A_i{}^2\sigma_i. \tag{17}$$

Specifically, for $N=3$

$$(z/\phi_0)^2 = 4f_1{}^2[1 + R_2{}^2(f_2/f_1)^2 + R_3(f_3/f_1)^2] \cdot/[1 + R_2{}^2 + R_3{}^2] \tag{18}$$

$$(d/\phi_1)^2 = 4f_1{}^2[1 + R_2{}^2(f_2/f_1)^4 + R_3{}^2(f_3/f_1)^4] \cdot/[1 + R_2{}^2(f_2/f_1)^2 + R_3{}^2(f_3/f_1)^2] \tag{19}$$

where $z=z_0$, $d=z_1$, and $R_i=A_i\sqrt{\sigma_i}/A_1\sqrt{\sigma_1}$. From (17)–(19) it follows that zero-crossing rates are not simply and directly related to formant frequencies alone. However, if the ith formant is dominant, in the sense that $A_i\sqrt{\sigma_i}$ is significantly larger for this formant than for any other, then $(z_n/\phi_n)\simeq 2f_i$, as one would expect.

As an example illustrating the use of (17)–(19), the probable locations in the d versus z plane for vowels /i/, /Λ/ and /u/ in normal speech will be determined. Because vowel formant amplitudes and bandwidths are less stable than formant frequencies [16], [27], [28], we consider the effect of variations in R_2 and R_3 on z_0 and z_1. From the spectral data of ourselves and others [14]–[16] we obtained the following average values for the first three formant frequencies.

/i/: $f_1 = 350$ Hz; $f_2 = 2350$ Hz; $f_3 = 3480$ Hz

/Λ/: $f_1 = 650$ Hz; $f_2 = 1400$ Hz; $f_3 = 2546$ Hz

/u/: $f_1 = 350$ Hz; $f_2 = 925$ Hz; $f_3 = 2400$ Hz.

Fig. 2 shows the expected regions for /i/, /Λ/, and /u/ as obtained from (18) and (19). The values of R_2 and R_3 shown represent somewhat larger ranges typical of data of ourselves and others. Fig. 2 suggests, among other things, that good separation of /i/, /Λ/, and /u/ on the basis of d and z is possible for normal speech.

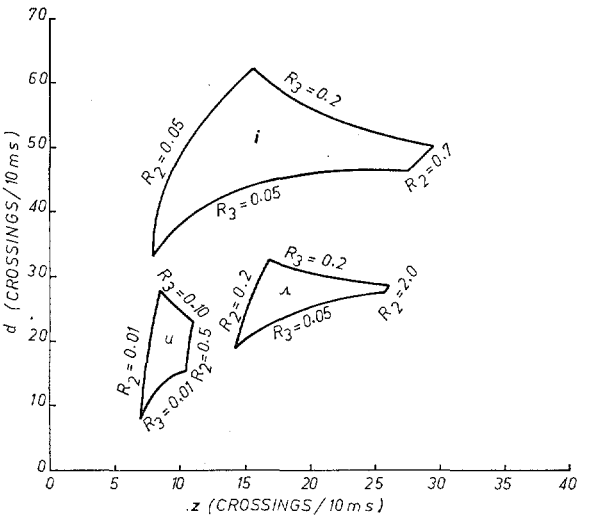

Fig. 2. Probable locations of vowels /i/, /Λ/, and /u/ from normal speech.

V. Zero-Crossing Measurements for Vowels

Zero-crossing data for vowels in various contexts appear in Figs. 3 and 4 and in Table I. The point of measurement was selected manually as a 10-ms segment in the stable part of the vowel nucleus. Fig. 3 shows d versus z for normal speech. Fig. 4 shows z for speech prefiltered by the 1–8-kHz band-pass filter on the vertical axis versus z for speech low-passed at 1 kHz on the horizontal axis. Table I compares measured values of d and z with corresponding values calculated using (18) and (19) with $\phi_0=\phi_1=1$. Formant frequencies, bandwidths, and amplitudes for the middle third of the vowel were obtained from the spectrum analyzer described in Section III.

The locations of /i/, /Λ/, and /u/ in Fig. 3 agree generally with those predicted by Fig. 2. For a vowel like /i/, large values of f_2/f_1 and f_3/f_1 cause the higher formant terms to dominate the expression for d and to have a fairly strong effect on z. The dispersion pattern for /i/ is, therefore, mainly horizontal. For /u/, the smaller values of f_2/f_1 and f_3/f_1 combined with the relatively weak amplitudes of the higher formants causes the higher formants to mainly affect d and not z. Consequently, the dispersion for /u/ tends to be vertical. A similar analysis confirms the horizontal dispersion pattern of /Λ/.

Table I shows reasonably good agreement between measured and calculated values of z and d. The most noticeable discrepancy is for the vowel /i/. The major source of error here is due to measurement error in f_1 caused by the first formant falling below the lowest frequency channel (280–310 Hz) of our spectrum analyzer. Other sources of error for /i/ and the other vowels result from approximations involved in developing (17)–(19) and errors in measuring formant parameters.

Fig. 4 is similar to plots of f_2 versus f_1 [14]–[16] for

TABLE I

Comparison of z and d as Measured for Vowels from Normal Speech with Corresponding Values Calculated Using (18) and (19)

Utterance	z (/10 ms)		d (/10 ms)	
	Measured	Calculated	Measured	Calculated
g i l	8	11	55	57
d i l	12	14	50	51
b i l	9	16	50	51
t i l	17	20	53	56
n i s	17	21	49	50
k i l	13	24	49	60
p i l	24	24	53	58
f i ə̂	13	26	49	58
l i ʃ	15	26	50	52
r i f	17	29	53	60
f I t	15	14	31	36
s I d	13	15	31	35
s I p	13	16	32	34
w I s t	12	15	31	33
w I s k	12	15	28	31
v I l	13	17	31	37
s I t	17	19	35	39
s I g	19	20	42	41
s I k	15	21	36	37
s I b	19	21	39	40
w I s p	14	21	37	38
z ɛ l	14	16	30	24
s ɛ k t	14	16	25	30
h æ z	19	17	32	30
s æ v	21	24	35	36
s æ v	18	23	28	34
ʃ æ k	26	26	38	34
b ʌ n	23	20	31	25
g ʌ n	23	21	34	30
p ʌ n	24	21	33	27
l ʌ ʃ	20	21	31	26
d ʌ n	25	22	31	29
t ʌ n	24	22	32	28
k ʌ f	25	22	30	27
b ʌ s	22	23	31	27
v ɔ n	17	16	22	22
f ɔ l	17	16	21	19
l ɔ g	20	17	26	24
l ɔ d	21	17	24	21
l ɔ b	21	17	24	20
l ɔ k	23	19	24	21
l ɔ t	23	21	24	23
b U ʃ	12	14	16	18
ʃ u ə̂	8	8	12	12
t u l	9	9	13	14
k u l	9	10	16	15
p u l	9	10	16	16
d u l	9	9	18	18
d u l	9	9	18	18
r u ə̂	6	10	20	23
l u z	10	10	32	28
f u d	8	11	16	22
z u m	9	11	20	26
g u l	8	12	14	23

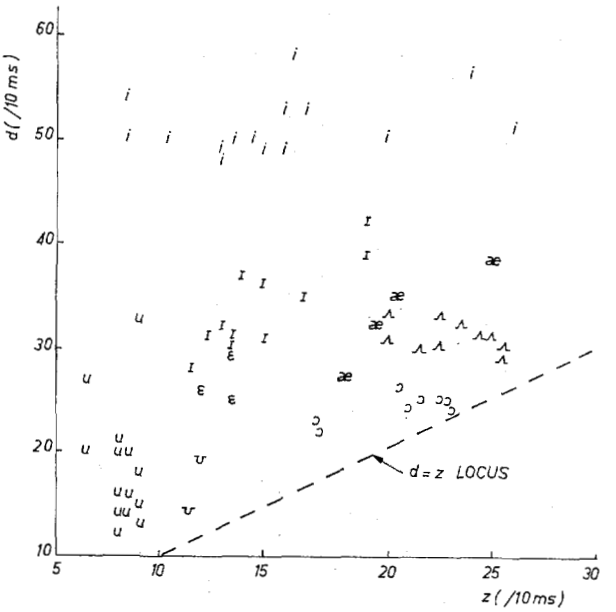

Fig. 3. d versus z for vowels measured in various contexts in normal speech.

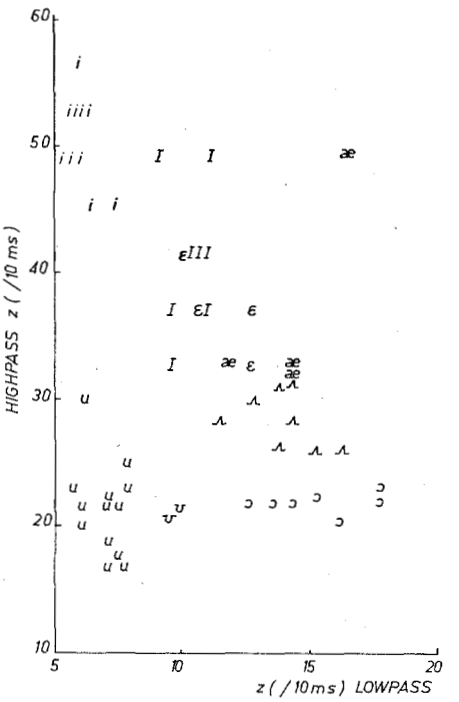

Fig. 4. Zero-crossing rates for vowels low-pass and high-pass filtered at 1 kHz.

normal speech. This is to be expected, since the 1 kHz low-pass filter usually passes only the first formant, while the 1–8-kHz band-pass filter passes the second and third formant.

Fig. 3 indicates that z and/or d permit partial separation of vowels. Fig. 4 shows more compact clustering of similar vowels and better separation of different vowels than does Fig. 3. The clustering and separation evident in Fig. 4 is comparable to that in graphs of f_2 versus f_1 [14]–[16] and indicates partial, but not complete, vowel separation.

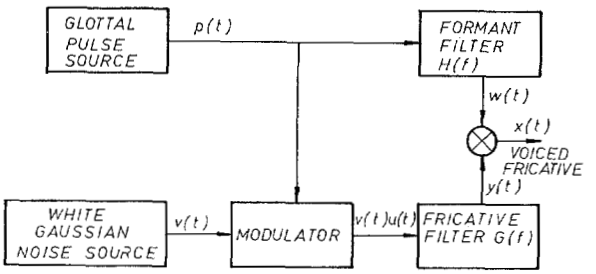

Fig. 5. Model for production and synthesis of voiced fricatives.

VI. Zero-Crossing Analysis of Fricative and Stop Consonants

Unvoiced fricatives (uvf) have been modeled and synthesized by exciting linear filters with Gaussian noise [18]–[26]. Since the filter output signal is Gaussian if the input is Gaussian, it is reasonable to model stable regions of uvf's as portions of sample functions from an ergodic Gaussian process, in which case (3) applies with $x(t)$ as the uvf acoustic waveform and $\phi_n = 1$.

The frictional noise region of unvoiced stop consonants has been modeled [18], [19], with some success, by a source-filter system like that used for unvoiced fricatives. One would therefore expect z_n in the frictional noise region to be approximated by (3) with $\phi_n = 1$. Because good models for some regions of unvoiced stop consonants are lacking,[2] precise quantitative behavior of z_n in such regions is difficult to predict.

Fig. 5 shows a circuit model that has been used in both speech production [4], [12], [15] and speech synthesis [20], [21], [24], [26] studies of voiced fricatives. The model is reasonably realistic in that voicing modulation of the frictional noise is considered. Excitation $v(t)$ is a sample function from a zero-mean Gaussian random process which is uncorrelated with voicing signal $p(t)$. Periodic signal $u(t)$ is synchronized with $p(t)$ and is usually considered to be a rectangular wave. Over those regions where $H(f)$ and $G(f)$ can be considered time-invariant, $S_x(f) = S_w(f) + S_y(f)$. From arguments similar to those leading to (8) it follows that $S_w(f)$ equals $S_x(f)$ in (8) and that

$$S_y(f) = |G(f)|^2 \sum_{k=-\infty}^{\infty} U_k^2 S_v(f - k\alpha) \tag{20a}$$

where

$$U_k^2 = \begin{cases} u_0^2, & k = 0 \\ (u_{|k|}^2 + \hat{u}_{|k|}^2)/4, & k \neq 0 \end{cases} \tag{20b}$$

and u_k and $\hat{u}_k$ are the Fourier coefficients of $u(t)$ [see (7) and (8)]. From (3) it follows that

$$\left(\frac{z_n}{2\phi_n}\right)^2 = \frac{\sum_{k=0}^{\infty}\left[P_k^2 |H(k\alpha)|^2 (k\alpha)^{2n+2} + U_k^2 \int_0^{\infty} f^{2n+2} |G(f)|^2 S_v(f - k\alpha)df\right]}{\sum_{k=0}^{\infty}\left[P_k^2 H(k\alpha)(k\alpha)^{2n} + U_k^2 \int_0^{\infty} f^{2n} |G(f)|^2 S_v(f - k\alpha)df\right]}. \tag{21}$$

Power spectrum $S_v(f)$ can usually be assumed constant over the audio range, in which case (21) simplifies somewhat.

The case $U_k^2 \ll P_k^2$ in (21) corresponds to very weak frictional noise, or equivalently, to strong voicing dominance, with the result that (21) reduces to a form similar to (10) for vowel-like sounds. The case $U_k^2 = 0$ for $k > 0$ and $P_k^2 = 0$ for all k occurs when voicing is absent, in which case (21) reduces to the relation for unvoiced fricatives.

The complexity of voiced fricatives in real speech arises from several factors including variation of ratios U_k^2/P_k^2 with time, temporal, and contextual dependence of U_k^2/P_k^2, and slow variation of $G(f)$ and $H(f)$ with time. These temporal variations cause the durations of stable regions in voiced fricatives to be relatively short, and result in voiced fricatives being considerably more complex than vowels, vowel-like sounds, and unvoiced fricatives.

Regarding voiced stops, the model in Fig. 5 has been used, with limited success, for speech production studies [18] and for synthesis [21]–[26]. However, the extremely short duration (on the order of 20 ms) of voiced stops suggests that (21) and the assumptions on which (21) is based will be, at best, reasonable approximations to reality.

VII. Zero-Crossing Measurements for Unvoiced Fricatives

Fig. 6 shows d versus z for /s/, /ʃ/, and /f/ in word-initial and word-final position. As with vowels, the 10 ms segment selected for measurement was manually selected to be in a stable region of the uvf nucleus. Separation of /s/, /ʃ/, and /f/ solely on the basis of d is seen to be possible, while z can be used to separate /s/ from /ʃ/. The dispersion pattern for /f/ is larger than that of /s/ and /ʃ/, which indicates that the acoustic properties of /f/ are more variable than those of /s/ and /ʃ/. The data also suggests

[2] For example, behavior immediately following the release of the stop is not well understood.

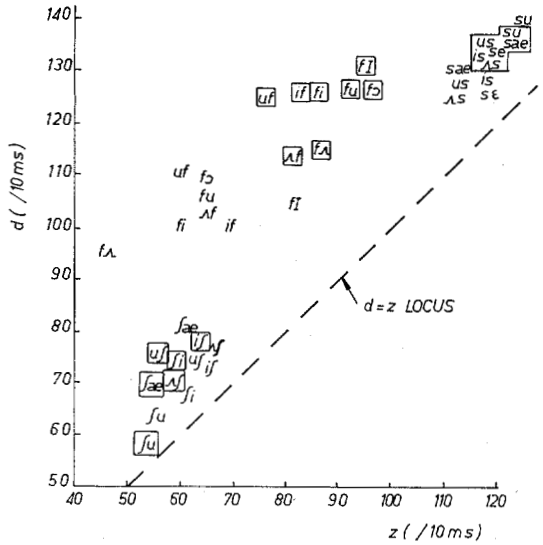

Fig. 6. d versus z for fricative consonants measured in various contexts. Fricatives are in either initial or final position in a word. Squares denote z and d as calculated from spectral data.

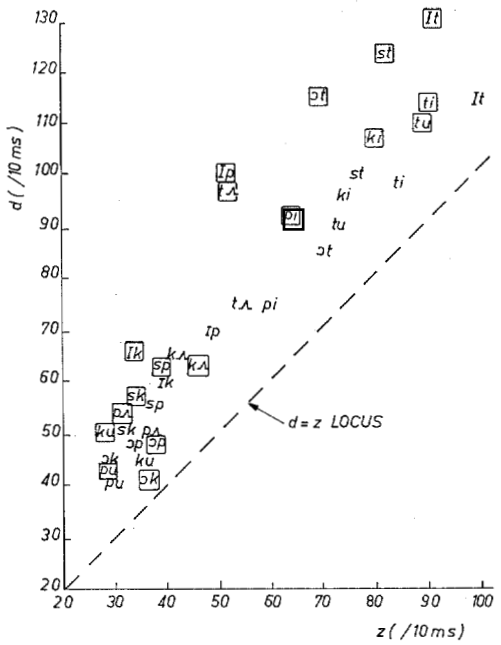

Fig. 7. d versus z for unvoiced stop consonants measured in various contexts. All stops are in either word-initial or word-final position. Squares indicate d and z as calculated from spectral data.

that initial /f/ is more variable than final /f/. This latter phenomenon is attributed to the fact that initial /f/ were shorter in duration and more plosive-like than final /f/, as also observed by Reddy [4], [5].

Values of z and d calculated from spectral measurements are shown enclosed in squares in Fig. 6 and are in reasonable agreement with directly measured zero-crossing rates for /s/ and /ʃ/. The displacement of the calculated points for /f/ is attributable to two factors. First, /f/ is a very weak sound, with the result that the signal-to-noise ratio for the filter-bank output signal is low. Second, the relation between z and d contains a high-frequency preemphasis factor, with the result that any noise in the high-frequency channels tends to bias z and d to higher than normal values, as in Fig. 6.

VIII. Zero-Crossing Measurements for Unvoiced Stops

Fig. 7 shows d versus z for unvoiced stops, most of which are in either word-initial or word-final position. The point of measurement was selected manually to be that 10-ms region of the stop for which z and d were largest. Thus, the values of z and d shown correspond to the frictional noise region if this region is present and dominant; otherwise they correspond to the aspiration region.

The points in Fig. 7 enclosed by squares represent estimates of z and d obtained using (3) with $\phi_n = 1$. The point of spectral measurement was manually selected to be in the temporal region immediately following stop release. Comparison of measured and estimated values shows reasonably good agreement for /p/ and /k/. For /t/, however, the calculated values are all noticeably higher than the corresponding measured values. A possible reason for this discrepancy is as follows. The frictional noise region for /t/ is short in duration, and generally overlaps the following aspiration region. The 10-ms time window over which z and

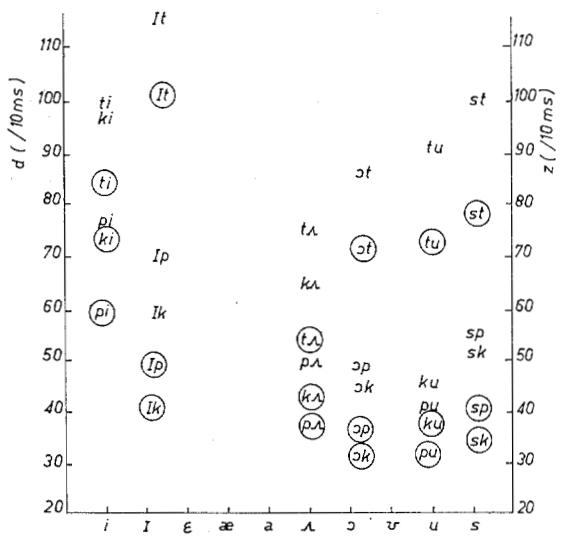

Fig. 8. d versus z for stop consonants in Fig. 7 plotted versus contextual environment. Circles indicate z while absence of circles indicates d.

d were measured included the initial part of the aspiration region, whereas the spectral values were averaged over a 5-ms time window which included only frictional noise. Since the z and d values of the frictional noise region are higher than those of the aspiration region, averaging z and d over 10 ms results in values lower than their peak values. Another possible source of discrepancy between measured and estimated values results from the high-frequency noise enhancement properties inherent in (3).

Fig. 8 shows the measured values of d and z of the stops

in Fig. 7 plotted versus the preceding or the following vowel. The contextual dependence of *z* and *d* is seen to be so great that identification of the stop solely on the basis of *z* and *d* is not possible, in general. Some information on the contextual environment is needed, as others have noted [13], [16], [29].

References

[1] J. Licklider and I. Pollack, "Effects of differentiation, integration, and infinite peak clipping on the intelligibility of speech," *J. Acoust. Soc. Amer.*, vol. 20, Jan. 1948, pp. 42–51.

[2] G. E. Peterson and J. R. Hayne, "Examination of two different formant estimation techniques," *J. Acoust. Soc. Amer.*, vol. 38, 1962, pp. 1865–1875.

[3] D. R. Reddy, "Segmentation of speech sounds," *J. Acoust. Soc. Amer.*, vol. 40, 1966, pp. 307–312.

[4] D. R. Reddy, "Phoneme grouping for speech recognition," *J. Acoust. Soc. Amer.*, vol. 41, 1966, pp. 1295–1300.

[5] D. R. Reddy, "Computer recognition of connected speech," *J. Acoust. Soc. Amer.*, vol. 42, 1967, pp. 329–347.

[6] D. R. Reddy and P. J. Vincens, "A procedure for the segmentation of connected speech," *Proc. 34th Convention Audio Eng. Soc.*, vol. 16, Oct. 1968, pp. 404–411.

[7] W. Bezdel and H. J. Chandler, "Results of an analysis and recognition of vowels by computer using zero-crossing data," *Proc. Inst. Elec. Eng.*, vol. 112, Nov. 1965, pp. 2060–2066.

[8] W. Bezdel and J. S. Bridle, "Speech recognition using zero-crossing measurements and sequence information," *Proc. Inst. Elec. Eng.*, vol. 116, Apr. 1969, pp. 613–617.

[9] R. W. A. Scarr, "Zero crossings as a means of obtaining spectral information in speech analysis," *IEEE Trans. Audio Electroacoust.*, vol. AU-16, June 1968, pp. 247–255.

[10] S.-H. Chang, G. E. Pihl, and M. W. Essigmann, "Representations of speech sounds and some of their statistical properties," *Proc. IEEE*, vol. 39, Feb. 1951, pp. 147–153.

[11] A. Papoulis, *Probability, Random Variables and Stochastic Processes.* New York: McGraw-Hill, 1965, pp. 485–495.

[12] G. W. Hughes and M. Halle, "Spectral properties of fricative consonants," *J. Acoust. Soc. Amer.*, vol. 28, 1956, pp. 303–310.

[13] M. Halle, G. W. Hughes, and J. P. A. Radley, "Acoustic properties of stop consonants," *J. Acoust. Soc. Amer.*, vol. 29, Jan. 1957, pp. 107–116.

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

[15] G. W. Hughes and J. F. Hemdal, "Speech analysis," Purdue Research Foundation, Lafayette, Ind., Final Rep., July 1, 1965.

[16] K. N. Stevens and A. S. House, "Perturbation of vowel articulations by consonantal context: An acoustical study," *J. Speech Hear. Res.*, vol. 6, June 1963, pp. 111–128.

[17] J. M. Heinz and K. N. Stevens, "On the properties of voiceless fricative consonants," *J. Acoust. Soc. Amer.*, vol. 33, May 1961, pp. 589–594.

[18] G. Fant, *Acoustic Theory of Speech Production.* New York: Humanities Press, Inc., 1960, ch. 2.

[19] J. L. Flanagan, *Speech Analysis, Synthesis and Perception.* New York: Academic Press, 1965, chs. 3 and 5.

[20] G. Rosen, "Dynamic analog speech synthesizer," Mass. Inst. Tech. Electron. Res. Lab., Cambridge, Tech. Rep. 353, 1962.

[21] L. R. Rabiner, "A model for synthesizing speech by rule," *IEEE Trans. Audio Electroacoust.*, vol. AU-17, Mar. 1969, pp. 7–13.

[22] S. E. Estes, H. R. Kerby, H. D. Maxey, and R. M. Walker, "Speech synthesis from stored date," *IBM J. Res. Develop.*, vol. 8, Jan. 1964, pp. 2–12.

[23] N. R. Dixon and H. D. Maxey, "Terminal analog synthesis of continuous speech using the diphone method of segment assembly," *IEEE Trans. Audio Electroacoust.*, vol. AU-16, Mar. 1968, pp. 40–50.

[24] J. C. W. A. Liljencrants, "The OVE III speech synthesizer," *IEEE Trans. Audio Electroacoust.*, vol. AU-16, Mar. 1968, pp. 137–140.

[25] M. P. Haggard and I. S. Mattingly, "A simple program for synthesizing British English," *IEEE Trans. Audio Electroacoust.*, vol. AU-16, Mar. 1968, pp. 95–99.

[26] R. S. Tomlinson, "SPASS—An improved terminal analog speech synthesizer," Mass. Inst. Tech., Electron. Res. Lab., Cambridge, Quart. Progr. Rep. 80, Jan. 1967, pp. 198–205.

[27] I. Lehiste and G. B. Peterson, "Transitions, glides and diphthongs," *J. Acoust. Soc. Amer.*, vol. 33, 1961, pp. 268–277.

[28] G. E. Peterson, "Parameters of vowel quality," *J. Speech Hear. Res.*, vol. 4, 1961, pp. 10–29.

[29] K. N. Stevens, A. S. House, and A. P. Paul, "Acoustical description of syllabic nuclei: An interpretation in terms of a dynamic model of articulation," *J. Acoust. Soc. Amer.*, vol. 40, 1966, pp. 123–132.