A STATISTICAL APPROACH TO SPEECH ANALYSIS BY WAVEFORM

PHILIPP CHRISTOV*

Let us take as model for the speech source a linear dynamic system characterized by a memory function which, driven by a forcing function, gives the speech signal at its output.[1] The forcing function is a quasiperiodic train of laryngeal excitations which follow one after another, by $t > 0$, in intervals equal to the pitch period[2] and have so short a duration that at the end of each pitch period equal zero.[3] As the memory function seems to have such characteristics that at the end of each pitch period the response of each single glottal pulse also equals zero,[4] the latter can be expressed with the 'identical' functions[5] or *elementary waves*

$$x_k(\tau;\, a_k) = \begin{cases} x_k(\tau; a_k) & \text{if} \quad t_k - T_{ok} \leqq \tau < t_k \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where **a** is an aggregate of waveform parameters, T_{ok} is the value of the pitch 'period' in the k[th] elementary wave and k is chosen arbitrarily.

From the speculation about the nature of the speech signal, made above, it follows that each single elementary wave contains no periodic components and is statistically independent. Hence each vocal segment of the speech signal, defined as a *phone*[6] (Fig. 1), can be regarded as a non-stationary *random function* $X(\tau;\, \mathbf{a})$[7] or *random speech wave* presented by the set of its trials

$$x_1(\tau;\, a_1),\ x_2(\tau;\, a_2),\ x_3(\tau; a_3) \ldots x_k(\tau;\ \ a_k),\ \ldots x_h(\tau{:}a_h) \tag{2}$$

* Bulgarian Language Institute, Sofia, Bulgaria.

[1] Huggins, W. H., "A Note in Autocorrelation Analysis of Speech Sounds", *JASA*, **26**, 790—794 (1954).

[2] Fant, G., "Acoustic Theory of Speech Production", пепевод с английского под редакцией В. С. Григорьева, из-во Наука, Москва (1964), стр. 29.

[3] Miller, R. L., "Nature of the Vocal Cord Wave", *JASA*, **31**, 667—677 (1959).

[4] Drew, R. O., and Kellog, E. W., "Starting Characteristics of Speech Sounds", *JASA*, **12**, 95—103 (1940), p. 103.

[5]) Поливанов, К. М., ,,Теоретические основы электротехники", част I, из-во Энергия, Москва, Ленинград (1965), стр. 374.

[6] Bloch, B., "Studies in Colloquial Japanese", *Language*, **26**, 86 (1950), p. 89.

[7] Middleton, D., "An Introduction to the Statistical Communication Theory", перевод с английского, том I, из-во Советское радио, Москва (1961), стр. 676.

which occur in the fixed interval $(\tau_0, \tau_0 + T_0)$, where $T_0 = C^{te} \geqq T_{0k}$ and whereas k is chosen arbitrarily, it may be chosen in an orderly manner as well.

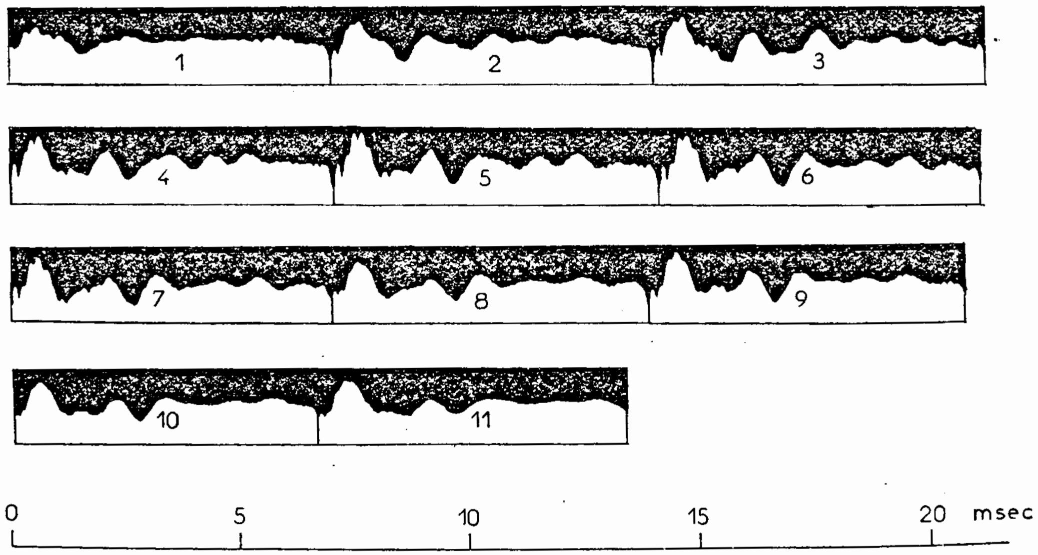

Fig. 1. High speed oscillogram of the vowel [*á*] as in [*dáp*]. Speaker A. M. The oscillogram is obtained by photographical enlargement of bilateral motion picture sound recording.

Let us take samples $x_k(\tau_m; a_k) \equiv x_{km}$ of the k^{th} elementary wave in its successive phases of development $\tau_m[m = 0, 1, 2, 3, \dots n)$ $(k)]$ at constant sampling intervals $\Delta\tau = \tau_m - \tau_{m-1} = 1/2W_x$, where W_x is the signal band width, which are sufficiently long to ensure the independence of the samples.[8] Then each elementary wave is presented with the set of samples

$$x_k^*(\tau; a_k) \equiv x_{k0}, x_{k1}, x_{k2}, x_{k3}, \dots x_{km}, \dots x_{kn(k)}, \dots x_{kn} \tag{3}$$

where the samples x_{k0} in the points of zero-crossing and the samples with numbers of sampling between $n(k) = T_{0k}/\Delta\tau$ and $n = T_0/\Delta\tau$ (See Eq. 1) equals zero.

If the sections of the random speech wave $X(\tau; \mathbf{a})$ in the fixed phases of development τ_m are grouped to form *ensembles*

$$X(\tau_m; \mathbf{a}) = x_{1m}, x_{2m}, x_{3m}, \dots x_{km}, \dots x_{hm} \tag{4}$$

then the ensembles or random variables $X(\tau_m; \mathbf{a})$ form the system of random variables $X^*(\tau; \mathbf{a})$

$$X^*(\tau; \mathbf{a}) \equiv [X(\tau_1; \mathbf{a}), X(\tau_2; \mathbf{a}), X(\tau_3; \mathbf{a}), \dots X(\tau_m; \mathbf{a}), \dots X(\tau_n; \mathbf{a})] \tag{5}$$

[8] Lee, Y. W., "Statistical Theory of Communication", Fourth printing, John Willey & Sons, Inc., New York, London, Sydney (1946), p. 278.

Let us assume that the *storage rule* for the information about the system of random variables $X^*(\tau; \mathbf{a})$ is given by the rectangular matrix χ

$$\chi = \| x_{km} \| \; (k = 1, 2, 3, 4, \dots h) \; (m = 1, 2, 3, 4, \dots n\,(k), \dots n) \tag{6}$$

Then we can consider as a portrayal of the system of random variables the discrete vector field created in the measurement space as a result of the storage operation, i.e., as a discrete vector variable $\mathbf{X}^* = f\chi(\mathbf{Q})$, where the operator, $f\chi$ designates that $\mathbf{X}^*$ takes its values in the points $\mathbf{Q}$ of the measurement space, according to the rules given by the matrix χ.

Assuming a measurement space determined by amplitude, phase and time, $\mathbf{Q}(x, \tau, t)$, we can decompose the vector variable χ^* in the components belonging to the subspaces of the measurement space (Fig. 2):

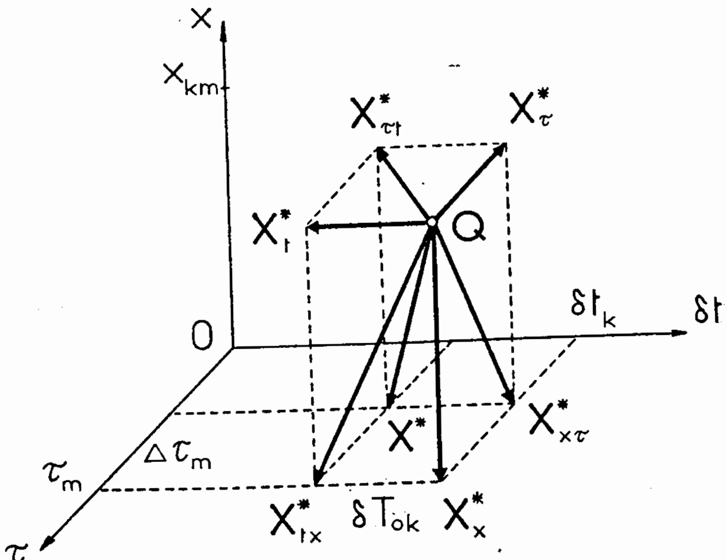

Fig. 2. The discrete vector variable $\mathbf{X}^* = f\chi(\mathbf{Q})$ and its components in the point (x_{km}, τ_m, t_k) of the measurement space, determined by amplitude x, phase τ and time t ($\delta = C^{te}$ is scale reduction coefficient).

1. *Unidimensional Substances.* $\mathbf{X}^* = \mathbf{X}_x^* + \mathbf{X}_\tau^* + \mathbf{X}_t^*$ where $\mathbf{X}_x^* = -x_{km}x_x^0$ $\mathbf{X}_\tau^* = -\Delta\tau x_\tau^0$ and $X_t^* = -\delta T_{0k} x_t^0$ and where $\delta = C^{te}$ is scale reduction coefficient which takes values between zero and unity, $0 < \delta \leqq 1$. The unidimensional components form sets of measurements which generates the statistical parameters (mean, variance, etc.) of the distributions of amplitude and pitch 'period', and the duration of the speech sound. The *space filling properties* of the portrayal $\mathbf{X}^*$ are given by the mixed product $V_{\mathbf{X}^*} = \mathbf{X}_x^* \cdot (\mathbf{X}_\tau^* \times \mathbf{X}_t^*)$. The unit vector $\mathbf{x}^0 = \mathbf{X}^*/|\mathbf{X}^*|$ of $\mathbf{X}^*$ gives its *directional properties*.

2. *Two Dimensional Subspaces.* $\mathbf{X}^* = \mathbf{X}_{x\tau}^* + \mathbf{X}_{\tau t}^* + \mathbf{X}_{tx}^*$, which generate a set of subportrayals: the plot $\bar{x}_p(t)$ of mean peak amplitude ($\bar{x}_p = x_{p-p}/2$) *vs* time or mean

amplitude envelope, the plot $T_{0(t)}$ of pitch 'period' *vs* time and the family of waveforms of the trials of the random function.

It is known,[9] that if we let the total number of samplings per speech sound tend to discontinuity then the system of random variables (Eq. 5), presented by the pattern **X***, becomes equivalent to the random speech wave $X(\tau, \mathbf{a})$. Hence any visual representation of speech based on this principle[10,11] should be considered as reliable (Fk 3).

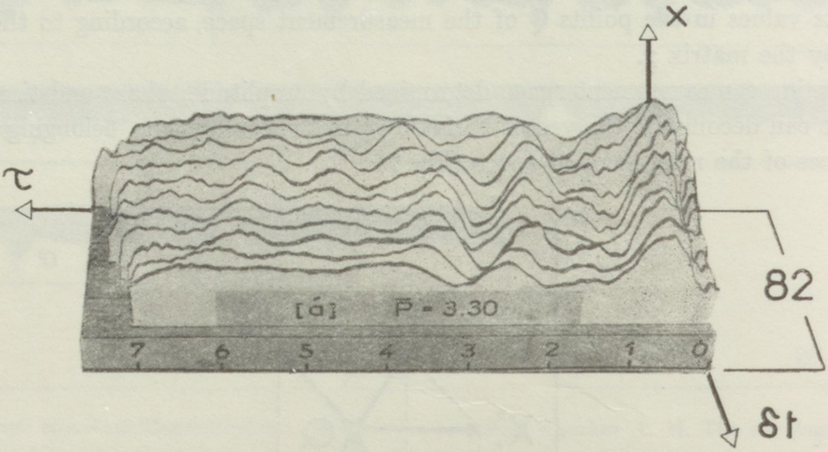

Fig. 3. Waveform (WF) Portrayal of the vowel [*á*] as in [*dáp*] build from the high speed oscillogram shown in Fig. 1. Total number of elementary waves h = 11. Recorder scale [*sec.* 10^{-3}]. Duration 82.10^{-3} *sec* ($\delta = 0{,}032$).

The pitch 'period' normalized portrayal **B*** of the random function generates its mean, variance, correlation function and power spectra. Assuming that each trial $x_k(\tau; a_k)$ exists between two successive positive-going zero-crossings, the portrayal **B*** can be obtained after prearranging of the time-domain samples x_{km} in a new phase-angle basis according to the rule given by the matrix $\boldsymbol{B}$

$$\boldsymbol{B} = \|\boldsymbol{B}_{km}\| \quad (k = 1, 2, 3 \ldots h)\,(M = 1, 2, 3 \ldots N)$$

where $\boldsymbol{B}_{km} = \| x_{km} \ldots x_{k(m+\vartheta_k-1)} \|$

and where $\pm\vartheta_k \approx \pm\, n(k)/2\pi$ is an entire digit.

But we may choose another way for analysis of the waveform portrayal by which no pitch normalization is needed; we can use parameters for direct evaluation of the successive waveforms like the *crest factor* C and the *form factor* F. These parameters, together with the *slope of the pulse front* S = Peak/Rise time,

---

[9]) Пугачев, В. С., ,,Теория случайных функции", Физматгиз, Москва (1960), p. 204.

[10] Christov, P., "New Methods for Investigation of Speech Sounds", *Technika*, **11** 215—218 (1962) (In Bulgarian).

[11] Grützmacher, M., "Demonstration eines Tonhöhenschreibers", *Phonetica*, **13**, 3—17 (1965).

can be applyied also in regard to the mean envelope $\overline{x}_p(t)$ and to the plots of the another waveform parameters *vs* time: $T_0(t)$, $C(t)$, $F(t)$, etc.

Some preliminary results of the application of the methods described in this paper to the practical problems of speech analysis can be considered as encouraging. Five stressed Bulgarian vowels, uttered by speaker A. M., in nonsense syllables, was subjected to speech analysis by the waveform. Sampling with $\Delta\tau = 0{,}0001$ *sec* was

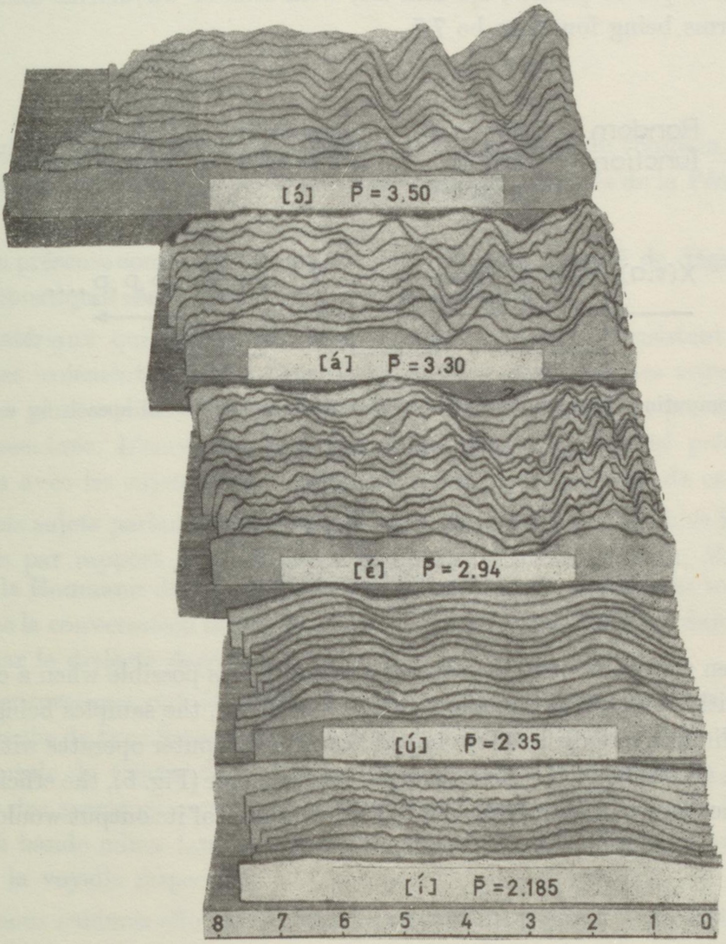

Fig. 4. WF Portrayals of five vowels uttered by speaker A. M., in the syllables: [*dík*], [*pút*], [*tʃét*], [*dáp*] and [*tɔ́n*] (Phonetic designations according to I.P.A.). Recorder scale [sec. 10^{-3}].

acomplished manually from their Waveform (WF) Portrayals shown in Fig. 4. The mean product-factor of the waveform $\overline{P} = \overline{C} \,.\, \overline{F}$, presented in the right—hand side of Fig. 4., seems to match the complexity of the waveform and, since it is non-dimensional, it appears that it is closely correlated to the phonemic value of these vowels. The observation that the sudden change in voice effort during the stress results in corresponding changes in the waveform[12] makes it reasonable to suggest that the product $(C_{P(t)} \cdot C_{\bar{x}_p(t)})$ would be effective by digital evaluation of stress. The envelope form of IrI[13] can be evaluated by the product factor of the envelope $P_{\bar{x}_p(t)}$, the ratio between its value in the modulated and non-mudulated segments of the carrier vowel sound being greater or equal to 1,55. It is sugested that the most important acoustic cue of IjI is the slope $S_{T_0(t)}$ of the plot of pitch vs time,[14] the ratio between the slopes of phones, IjI and IiI, with similar waveforms and amplitude envelope forms being found to be 7,5.

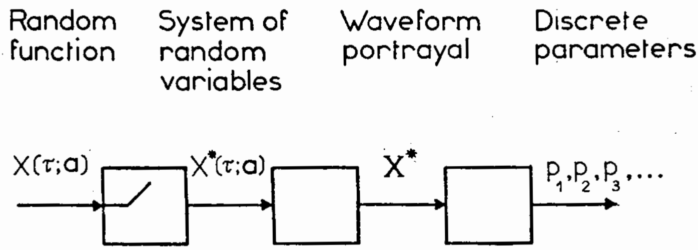

Fig. 5. Computing circuit of the process of statistical analysis of speech by waveform.

$$Q(x_{\mathrm{km}},\ \tau_{\mathrm{m}},\ t_{\mathrm{k}})$$

$$\mathbf{X}^* = -x_{\mathrm{km}}\mathbf{x}_{\mathrm{x}}^0$$

$$\mathbf{X}_\tau^* = -\frac{\tau m}{m}\,\mathbf{x}_\tau^0$$

$$\mathbf{X}_{\mathrm{t}}^* = \delta(t_{\mathrm{k-1}} - t)\,\mathbf{x}_{\mathrm{t}}^0$$

It has been shown recently that speech recognition is possible when a computer is presented with short samples of the acoustic waveform, the samples being processed without preliminary analysis.[15] It is hoped that if a computer operates with the input speech wave according to the ideas set forth in this paper (Fig. 5), the efficiency of the process of mechanical recognition as well as the quality of its output would be improved.

---

[12] Christov, P., "Some Peculiarities of the Bulgarian Vowels", *Bulgarian Language*, **13**, 22—27 (1963), p. 26 (In Bulgarian. French abstract in Bull., Anyl., Litt., Sci., Bulg., A. 1/1963, No 21).

[13] Christov, P., "Experiment for Changing the Envelope Form of Vowels", 5e C.I.A., *Reports*, **1a**, Liège (1965), A11.

[14] Christov, P., "Investigation Upon the Sonants in the Bulgarian Language", *Bulgarian Language*, **14**, 32—39 (1964), p. 37 (In Bulgarian. French abstract in Bull., Anal., Litt., Sci., Bulg., A. 1/1964, No 8).

[15] Reddy, D. R., "Segmentation of Speech Sounds", *JASA*, **40**, 307—312 (1966).