Determination of the instant of glottal closure from the speech wave

Hans Werner Strube

Drittes Physikalisches Institut, Universität Göttingen, Germany

(Received 21 May 1974; revised 11 July 1974)

For vowels excited by vigorous glottal vibrations, the instant of glottal closure is tentatively identified with the moment of strongest excitation (at not too low frequencies) and worst linear predictability. Some predictor methods for its determination are reviewed, which do not always yield reliable and unequivocal results. Then Sobakin's method using the determinant of the autocovariance matrix is examined critically and reinterpreted such that the determinant is maximum if the beginning of the interval on which the autocovariance matrix is calculated coincides with the glottal closure. This hypothesis is tested by comparison with the predictor methods and by looking at the inversely filtered waveforms and the formants obtained from predictors determined on a shifted interval. The determinant method seems to be very reliable even for otherwise difficult cases, such as the vowel /u/.

Subject Classification: 70.40, 70.20.

INTRODUCTION

Linear prediction of the sound pressure as a function of time is a useful means for extracting various physical parameters of the speech signal for coding[1-3] or analyses. For these purposes, the transfer function of the vocal tract and the shape of the glottal volume velocity pulses are assumed to be approximately representable by an all-pole filter excited by a sequence of delta impulses. The prediction error is usually minimized over a speech segment of, for example, 10 – 30-msec duration or over a fundamental period. During such a time interval, however, for voiced sounds the formant frequencies and bandwidths vary, even with constant articulator positions, due to the periodically variable glottal impedance. Moreover, the glottal pulse is not actually representable by an all-pole filter of low order. Both difficulties can be avoided by calculating the predictor (from the autocovariance matrix) only for intervals where the glottis is closed; furthermore the applicability of the usual hard-glottis model of the vocal tract is better then.

I. PITCH-PULSE DETERMINATION BY PREDICTION

The acoustical determination of the instant of glottal closure may be based on the assumption that, at not too low frequencies, the vocal tract is excited most strongly at this instant (the glottis closes abruptly and opens more slowly)[4,5] and that, subsequently, a time interval of free oscillation follows. Thus the instant of glottal closure will lie close to the strongest increase of amplitude of the (high-frequency emphasized) sound pressure, and the prediction error will be great there, since for good predictability the signal must be representable as the free oscillations of an all-pole filter after disappearance of the input signal. Atal[6] determined "excitation impulses" from the maxima of the prediction error. It may be, however, that these maxima do not coincide with the instants of closure, because Atal represents the glottal pulse shape by two poles of the transfer function and assumes delta-impulse excitation. Moreover it is not clear for what intervals the predictor is calculated here *before* the position of these impulses is known. If a 30-msec segment with Hamming weighting is chosen, the prediction error shows clear peaks at the points of strong excitation (Fig. 1), which, however, are not always uniquely defined. Multiple peaks of different polarity occur. Sometimes (e.g., with /u/) the peaks do not stand out clearly; furthermore the polarity of the signal must be known. Markel[7,8] employs only the autocorrelation function of the prediction error for determining the period and thus loses all information about the absolute position of the periods.

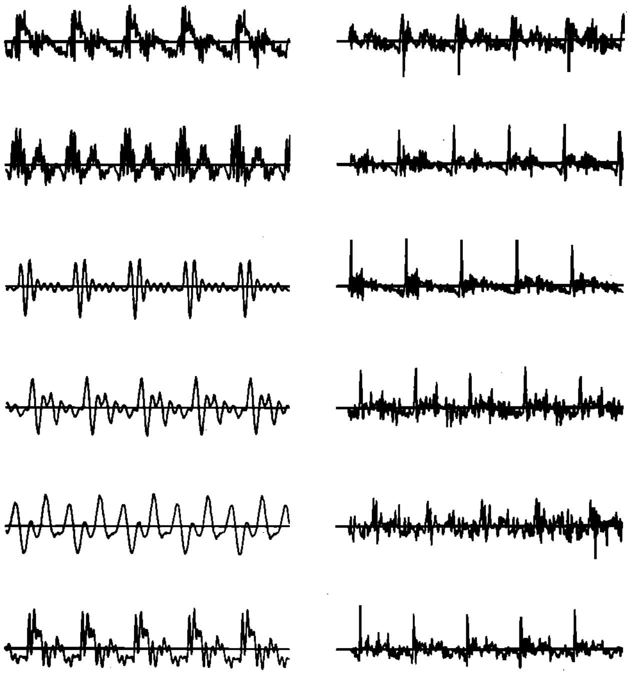

FIG. 1. Differentiated sound pressure (left) and prediction error (right) for vowels /i, e, a, o, u, y/. Length of segment shown 34 msec, sampling frequency 8.825 kHz, degree of predictor $p = 12$, predictor determined for a Hamming-weighted 30-msec segment.

If the predictors are calculated from the autocovariance matrices for overlapping segments, short compared with a period (e.g., 3 msec), one may assume that, in a segment containing the instant at which the glottis closes, the mean-squared prediction error normalized by the mean-squared signal is maximum. In subsequent segments it rapidly decreases; a predictor calculated there will be relatively good and the prediction error will show sharp peaks at the instant of glottal closure (similar to Fig. 7). An automatic strategy based hereon at first seemed to work reliably but failed for a larger sample of speech recordings. Multiple peaks of the prediction error occurred, or the behavior of the rms prediction error was not clear. A more specific investigation of the rms error as a function of the segment position (Fig. 2) in most cases does show an abrupt rise (sometimes with a high peak) and fall when the supposed point of glottal closure enters and leaves the segment, respectively, but the overall behavior is often rather irregular, especially for /u/, and difficult to evaluate by an automatic procedure.

Several computationally extremely fast adaptive prediction methods are given by Maksym[9]; from the prediction error a sequence of "excitation impulses" is derived by pulse shaping. My own experiments, however, yielded a relatively large prediction error from which the pulses did not stand out any more clearly than in the other methods described (particularly for /u/). Also the correlation of the pulse position with the instant of glottal closure is theoretically at least as uncertain.

II. PITCH-PULSE DETERMINATION USING THE AUTOCOVARIANCE DETERMINANT

A different approach is indicated by Sobakin.[10] Prediction without error means linear dependence of certain signal segments: Let s_k be samples of the signal, $a_1, \ldots, a_p$ the predictor coefficients. If for $k=K_1, \ldots, K_2$

$$s_k=\sum_{n=1}^{p} a_n s_{k-n}\;, \qquad (1)$$

then the vectors $\mathbf{s}_n \equiv (s_{K_1-n}, s_{K_1+1-n}, \ldots, s_{K_2-n})^T$ (superscript T denotes "transpose"), $n=0, \ldots, p$, are linearly dependent, i.e.,

$$\mathbf{s}_0=\sum_{n=1}^{p} a_n \mathbf{s}_n\;. \qquad (2)$$

In this case the (nonnegative) Gram determinant det $(\mathbf{s}_n^T\mathbf{s}_m)$ vanishes. However,

$$\mathbf{s}_n^T\mathbf{s}_m=\sum_{k=K_1}^{K_2} s_{k-n}s_{k-m}\;. \qquad (3)$$

Thus the Gram determinant is the determinant of the $(p+1)\times(p+1)$ autocovariance matrix of the signal segment from K_1 to K_2. Sobakin concludes from this: When the glottis is closed during the whole segment, the determinant will be small, because prediction and thus the linear dependence will be especially good. Further, the value of the determinant as a function of the temporal position of the segment $\mu(t)$ should qualitatively correspond to the glottal pulses. In view of his Fig. 3(c), however, this conclusion is doubtful. The minima of the determinant lie too far to the right relative to the supposed position of the closure intervals. A more detailed investigation is obviously necessary. There are three objections to the original interpretation: (1) Even with nonvanishing input signal, prediction may be good as long as the considered signal segment can be interpreted as a free decaying process of an all-pole filter (however, the formants calculated from the roots of the predictor polynomial then become more or less incorrect). (2) It must be clarified more in detail which point of the segment is most suitable as a time reference point for definition of the function $\mu(t)$; Sobakin has chosen the center point. (3) The value of the determinant is not only a measure for the linear dependence, but is strongly influenced by the signal amplitude, namely, is proportional to its $2(p+1)$th power.

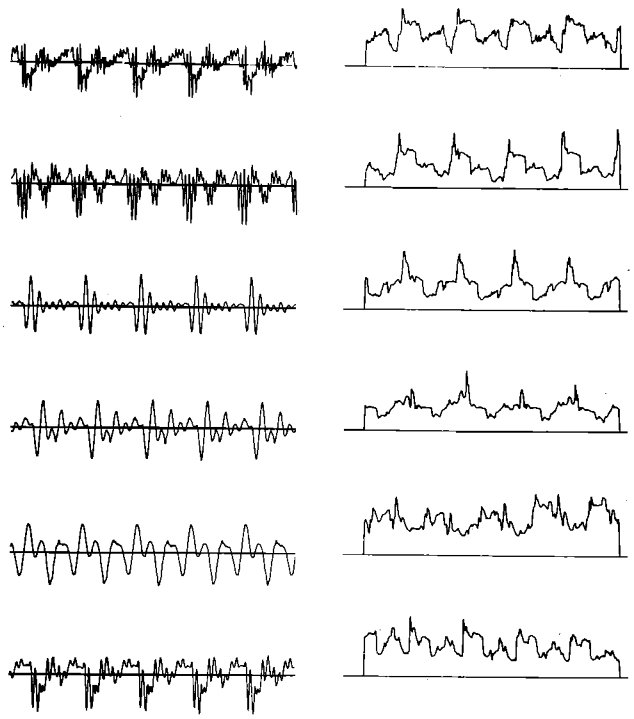

FIG. 2. Right: rms prediction error divided by signal rms value on an interval of 21 samples (2.38 msec) as a function of the interval position (abscissa: center of interval).

Let us consider a periodically excited filter whose input signal is nonzero only during a small part of the period, and whose output signal is linearly predictable only with nonzero error. The analysis interval $[K_1, K_2]$ is shifted in time, its length being constant and small compared to the signal period. The order of prediction p is fixed. In intervals when the input signal is zero, let the prediction error be on the average proportional to the signal amplitude, and the determinant μ proportional to a power of the error. Assuming that the output signal decays approximately exponentially, the logarithm of the determinant is a linearly decreasing function of the interval position (Fig. 3, t_1 to t_2). But if an interval of nonzero excitation enters the shifted analysis interval, the determinant will grow strongly. The reason is the increase of the prediction error and also of

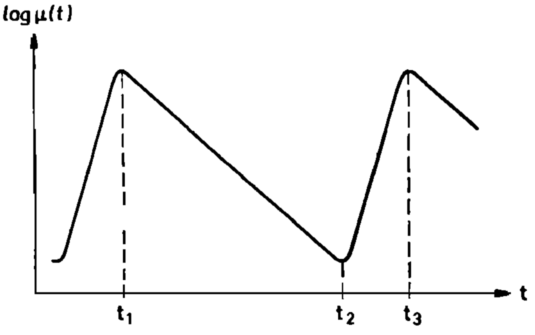

FIG. 3. Schematical curve of the logarithm of the determinant of the autocovariance matrix as a function of the interval position.

the mean amplitude due to the excitation (t_2 to t_3). The maximum of μ is reached when the point of excitation coincides with the beginning of the interval. On further shift of the analysis interval, the determinant decreases again (after t_3), since prediction improves. Thus the overall curve of $\log\mu(t)$ should exhibit a sawtooth-like shape. $t_3 - t_2$ is the width of the shifted interval plus width of the small interval in which the excitation occurs. Thus, if the beginning of the shifted analysis interval is chosen as reference point for its position, the maximum of $\mu(t)$ indicates the end of the excitation.

In spite of the simplicity of the assumptions made here the curve of $\log\mu(t)$ for speech generally shows great similarity with Fig. 3; see Fig. 4. The sound pressure was differentiated by an RC circuit, the sampling frequency is 8.825 kHz. The differentiation strongly emphasizes the point of glottal closure, so that the assumption of a short excitation interval is less artificial than it appears at first. The curve does not depend very much on the rank $p+1$ of the covariance matrix and on the length of the shifted interval, but becomes smoother when these quantities increase. For $p=12$, $\mu(t)$ ranges over up to 13 decades, so in a linear plot rather narrow pulses would be obtained. The shifted interval has a width of 15 to 25 samples (1.7 to 2.9 msec). Even for usually difficult cases such as /u/, clear main maxima result, although the total range of values is smaller.

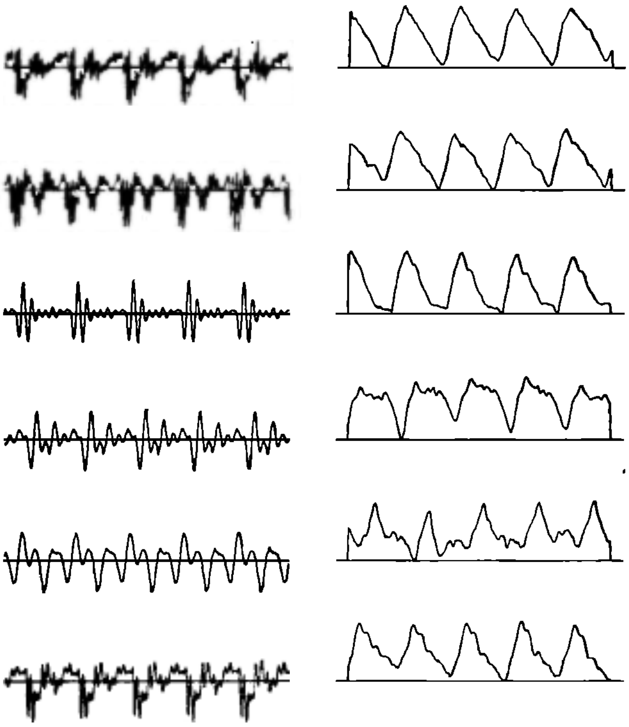

FIG. 4. Right: Logarithm of the determinant of the autocovariance matrix as a function of the position of the interval beginning. $p=12$, interval length 16 samples = 1.81 msec.

The assumption that the maxima of the $\mu(t)$ curves indicate the end of excitation, and thus the instant of glottal closure, cannot really be proved without synchronous observation of the vocal cord vibrations by optical or electrical means. However, a comparison with the prediction error methods outlined above shows that the results agree whenever they are unequivocal. The peaks in Fig. 1 and the abrupt rms-error decrease in Fig. 2 lie at the same points as the maxima of $\mu(t)$ (Fig. 5). For /u/, on the contrary, the maxima of μ lie at points not clearly discernible in Fig. 2 but quite possible as main excitation points, according to Fig. 1 and according to the visual impression of the differentiated sound pressure itself. A further possibility to test our hypothesis

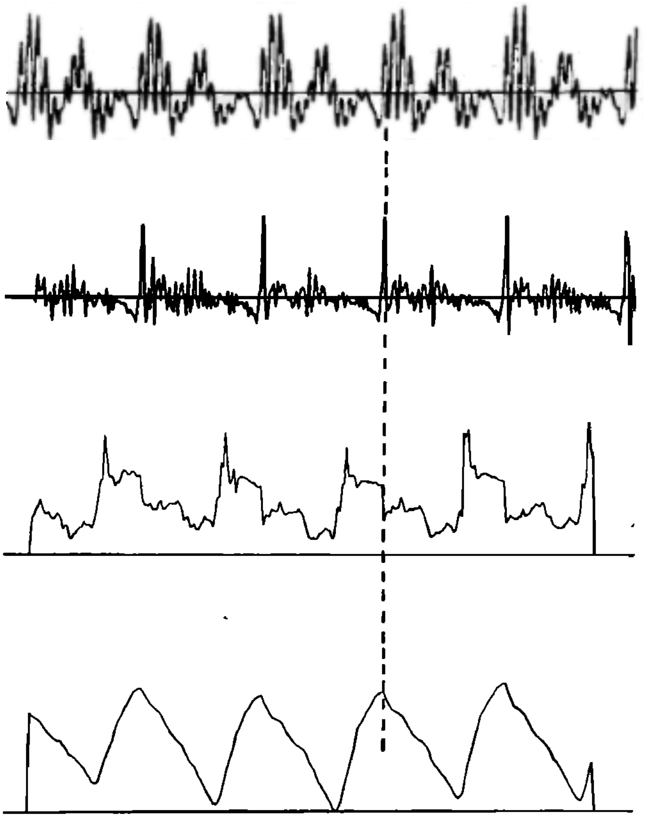

FIG. 5. Comparison of sound pressure, prediction error after Fig. 1, normalized rms prediction error after Fig. 2, and $\log\mu(t)$ for /e/. For the two latter curves the abscissa denotes the interval beginning. Interval length 21 samples = 2.38 msec, $p=12$.

about the meaning of the maxima of μ is the inverse filtering of the speech with the predictor polynomial as transfer function, i.e., calculation of the prediction error, and comparison of the "glottal pulses" obtained with the "theoretical" shape. This shape is not really known but should exhibit a rising, falling, and closure phase like the synthetic pulses in Eq. 4 (Fig. 6). The pulse shape should come out bad if the predictor has been determined for a segment in which the glottis was just closing; for a shift of the segment into the closure interval, the pulse shape should suddenly assume its best appearance. The prediction error here corresponds to the second derivative of the glottal volume velocity: For a hard-walled vocal tract, the transfer function of volume velocity has only poles (apart from a strongly damped zero due to the real part of the radiation admittance, whose influence is negligible at low frequencies), its denominator is the best predictor polynomial; the sound pressure in the radiation field is proportional to the derivative of the volume velocity in the mouth opening, further, it is once more differentiated when being recorded. When the tube wall is compliant, however, at least one more pole-zero pair appears in the transfer function. Its influence was investigated with the help of synthetic vowels where the wall impedance was crudely simulated by an RL shunt at the tube input end[11] ($R = 25$ g cm^{-4}sec^{-1}, $L = 0.02$ g cm^{-4}). The excitation pulses were of the shape

$$u(t) = \begin{cases} \sin^2 \dfrac{\pi t}{2T_1}\,, & (0 \le t \le T_1)\,; \\ \cos \dfrac{\pi(t - T_1)}{2(T_2 - T_1)}\,, & (T_1 \le t \le T_2)\,; \\ 0\,, & (T_2 \le t \le T)\,; \end{cases} \tag{4}$$

with $T_1 = 2.89$ msec, $T_2 - T_1 = 1.16$ msec, period $T = 7.25$ msec. The pulses were bandlimited to 4.4 kHz by a linear-phase low-pass filter. Figure 6 shows the result of the inverse filtering with the predictor obtained in the interval with $u(t) = 0$ and of the following twofold integration. The result is apparently rather good. Figure 7 shows the same for real speech. Superimposed remnants of power-line hum stand out strongly due to the integration causing a low-frequency emphasis. The closure interval does not always show up clearly: At its beginning irregularities are found or it goes over into the next pulse smoothly. The curves shown, however, are the optimal ones that can be obtained by shifting the segment on which the predictor was determined; already a shift of few samples makes the similarity with the expected shape decrease considerably. The segment was shifted in a "dialogue" (interactive computing) procedure making use of a visual display; the best result agrees quite closely with that of the determinant method. An additional check is provided by the formants determined from the roots of the predictor polynomial, which look more or less reasonable, dependent on the position of the shifted segment.

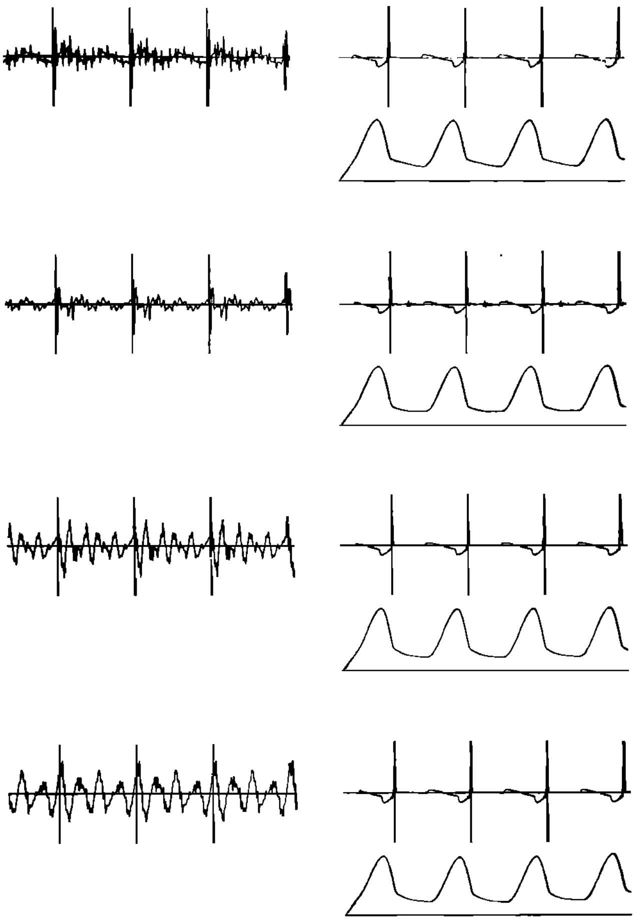

FIG. 6. Differentiated sound pressure (left), prediction error and reconstructed glottal wave (right) for synthetic vowels. The predictor was determined in the closed-glottis interval. (The vertical lines denote the moments of closure.)

Pursuing further Sobakin's interpretation of the determinant as a measure of linear dependence, I tried to eliminate the influence of the signal amplitude by various normalizations. The most obvious is to replace the vectors $\mathbf{s}_n$ by the corresponding unit vectors $(\mathbf{s}_n/|\mathbf{s}_n|)$, i.e., to replace the autocovariance matrix by the autocorrelation matrix. This is equivalent to dividing the determinant by $\prod_{n=0}^{p} |\mathbf{s}_n|^2$. The result shows an irregular behavior not clearly enough marking out the point of excitation. Other normalization attempts were based on division by

$$\left(\sum_{n=0}^{p} |\mathbf{s}_n|^2\right)^{p+1}, \quad |\mathbf{s}_0|^{2(p+1)}, \quad |\mathbf{s}_0|^{2p}, \quad \left(\sum_{i=K_1-p}^{K_1-1} s_i^2\right)^{p},$$
$$\left[\sum_{m,n} (\mathbf{s}_m^{\mathrm{T}}\mathbf{s}_n)^2\right]^{(p+1)/2}.$$

In no case did a curve result that was as smooth and as easy to evaluate automatically as the unnormalized determinant. Indeed, by reason of the above objection (1), it cannot be necessarily expected that a mere measure of linear dependence show a clear relation to the input signal. By the way, $\mu(t)$ behaves much more irregularly for synthetic vowels than for natural ones, because the synthetic signal contains portions where the determinant actually approaches zero.

III. REMARKS

A disadvantage of the method is the relatively great expense of computation time if a fully automatic pro-

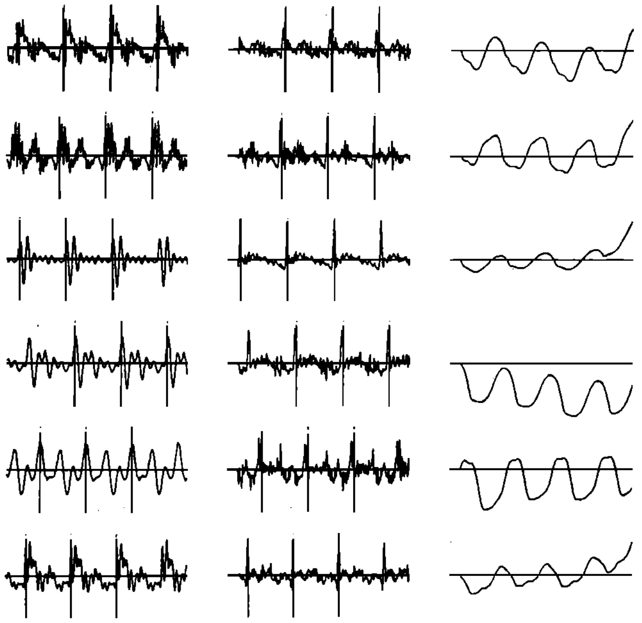

FIG. 7. Differentiated sound pressure (left), prediction error (middle) and reconstructed glottal wave (right). The predictor was determined in an interval following a maximum of $\mu(t)$. The vertical lines denote the position of the maxima of μ.

cedure is desired, because for a signal segment comprising at least one period the function μ must be calculated; so many covariance matrices and their determinants must be formed. Fortunately, for each new matrix only one row has to be computed anew, and the determinant is obtained relatively simply without destroying the matrix by the Choleski method[12] (see Appendix). If desired, a second closure point may be determined by searching for the maximum over an interval of at least one but less than two periods, starting from the first closure point (so input of a crudely estimated period value is necessary). The search for the nth point is made over a smaller interval located about a period after the $(n-1)$th. The method has been tested with stationary male vowels only; discontinuous period changes by a factor of two obviously cannot be recognized like this. Furthermore, according to the underlying assumptions, only vigorous vocal cord vibrations with sharp glottal closure ("chest register") can be considered. Generally, the value of the method will apparently consist not in the formation of a sequence of excitation pulses for transmission and synthesis of longer speech segments, but in the determination of glottal closure for special investigations of single or a few periods. For transmission and synthesis of longer segments, other methods, e.g., Maksym's, are more suitable.

ACKNOWLEDGMENT

I am grateful to Professor M. R. Schroeder and Dr. V. Mellert for critically reading the manuscript.

APPENDIX

The determinant of a positive definite symmetrical matrix is easily computed the following way. For any such $n\times n$ matrix $\mathrm{A}=(a_{ij})$, there is a lower triangular matrix $\mathbf{V}=(v_{ji})$, $v_{ji}=0$ for $i>j$, so that

$$\mathbf{A}=\mathbf{V}\mathbf{V}^{\mathrm{T}}\,;\qquad a_{ij}=a_{ji}=\sum_{k=1}^{i} v_{ik}v_{jk}\,;\quad i\leq j\,. \tag{A1}$$

By solving Eq. A1 for the term with $k=i$, for the v_{ji} one immediately obtains the recursion formulas

$$v_{ii}^2=a_{ii}-\sum_{k=1}^{i-1} v_{ik}^2\,,\qquad \left[v_{ji}=\left(a_{ij}-\sum_{k=1}^{i-1} v_{ik}v_{jk}\right)\Big/ v_{ii}\,,\ j=i+1\,,\,\ldots\,,\,n\right];\qquad i=1\,,\,\ldots\,,\,n\,. \tag{A2}$$

For the determinant

$$\det\mathbf{A}=(\det\mathbf{V})^2=\prod_{i=1}^{n} v_{ii}^2\,. \tag{A3}$$

To avoid floating-point overflow, it is advisable to compute log det **A** as the sum $\sum_i \log v_{ii}^2$ rather than the logarithm of the product in Eq. A3.

The v_{ji} with $i<j$ may be stored in one half of the matrix **A**, for the v_{ii}^2 only a single storage word is needed. Thus all the a_{ij} are preserved without requirement of additional storage.

[1] B. S. Atal, M. R. Schroeder, Proc. 1967 Conf. on Speech Commun. and Processing (M. I. T., Cambridge, Mass., 1967), pp. 360–361.

[2] B. S. Atal, M. R. Schroeder, Rep. 6th Int. Cong. Acoust. (Tokyo, 1968), pp. C-13–C-16.

[3] B. S. Atal, M. R. Schroeder, Bell Syst. Tech. J. 49, 1973–1986 (1970).

[4] R. L. Miller, J. Acoust. Soc. Am. 31, 667–677 (1959).

[5] K. Ishizaka, J. L. Flanagan, Bell Syst. Tech. J. 51, 1233–1268 (1972).

[6] B. S. Atal, S. L. Hanauer, J. Acoust. Soc. Am. 50, 637–655 (1971).

[7] J. D. Markel, IEEE Trans. Audio Electroacoust. AU-20, 367–377 (1972).

[8] J. D. Markel, IEEE Trans. Audio Electroacoust. AU-21, 154–160 (1973).

[9] J. N. Maksym, IEEE Trans. Audio Electroacoust. AU-21, 149–154 (1973).

[10] A. N. Sobakin, Sov. Phys. Acoust. 18, 84–90 (1972).

[11] G. Fant, Speech Trans. Lab.-Q. Prog. Stat. Rep. (Royal Institute of Technology, Stockholm), No. 2–3, 28–52 (1972).

[12] R. A. Buckingham, *Numerical Methods* (Pitsman & Sons, London, 1962), pp. 351–353, 365.