ARCHIVES OF ACOUSTICS
5, 1, 31-46 (1980)

AUTOMATIC DETERMINATION OF THE FREQUENCY BEHAVIOUR OF THE LARYNX TONE BY THE METHOD OF LINEAR PREDICTION

ANDRZEJ DZIURNIKOWSKI

ASW (02-656 Warszawa)

This paper presents a method of automatically tracing the frequency of the larynx tone, based on a numerical implementation of the method of linear prediction on a digital computer. It also discusses the algorithm for the practical implementation of this method, derived by the author, and presents the results obtained from the analysis of a continuous speech signal with a duration of approximately 2 seconds.

1. Introduction

The frequency of the larynx tone F_0, also called the *fundamental frequency*, is one of the basic accoustic parameters in the investigation of a speech signal. It is generally known that the intonation of continuous speech is connected with the parameter F_0. If the speaker is in the state of emotional excitement, the frequency F_0 of his larynx tone usually increases. From the acoustical point of view it is connected with changes in the duration of the vibration of the vocal chords which depends on the air pressure in the larynx and the physiological excitation of the chords [6]. The frequency F_0 can be measured by different methods: being determined by the analysis of a speech signal in the time domain in real time using the PPD method [8] or by the analysis of the parameters of a speech signal in the frequency domain, e.g. by the calculation of the cepstrum [1]. The objective of the present paper is to present a method for automatic determination of the parameter F_0 in a continuous speech signal, based on the method of linear prediction and using an electronic digital computer.

Developed in the 1950's, the technique of linear prediction was first used to analyse and synthesize speech by SAITO and ITAKURA [9] in the 1970's. It is based on a linear model of speech production which was created by FANT in the late 1950's [4, 6]. This model reflects, with certain simplifications, the most essential phenomena occurring during the course of the process of speech signal generation by man. This process can be roughly described in the following way. An acoustic speech wave consists of a variable acoustic pressure whose source is a positive pressure in the lungs. During articulation the air flows through the larynx and the pharyngo-oral cavity. The following parameters affect the characteristics of the acoustic wave at the outlet from the vocal channel that is formed by the lips and, in particular, its pressure: the pressure of the air exhaled from the lungs, the velocity of vibration of the vocal chords (depending on their mass, elasticity etc., which is a characteristic feature of the speaker), and the anatomical structure of the vocal tract, particularly those of its elements which participate in the dynamic changes of the parameters of the tract (e.g., lips, tongue, soft palate) [6, 11]. During generation of non-nasal sounds, the soft palate closes the nasal cavity, separating it from the other part of the vocal tract. Plosive sounds in turn occur when the mouth is violently opened after previous accumulation of air in the vocal tract. Unvoiced sounds are created when the vocal chords remain open and the air flows freely through the larynx. From the point of view of the present considerations we are, however, interested first of all in generation of such speech sounds as are characterized by the quasiperiodicity of the vocal wave signal (sonorous, voiced fricatives), i.e. those whose generation involves the vocal chords.

The vibration frequency of vocal chords can be measured by the acoustical method as the inverse of the time interval of the wave periods observed in the acoustic pressure at the outlet of the vocal tract. This frequency is defined as the F_0 frequency. The linear model of speech generation given by Fant (Fig. 1) permits voiced sounds, voiced fricatives and other sounds of quasiperiodic structure to be represented under the assumption of the vocal tract being excited by an impulse function or by white noise [6]. It is thus possible to estimate the F_0 frequency in the speech signal based on this model.

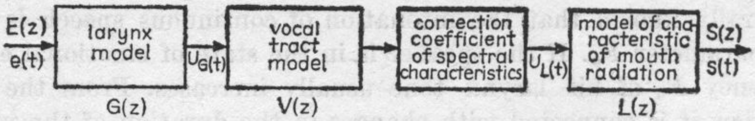
E(z)
e(t)
larynx model
G(z)
$U_G(t)$
vocal tract model
V(z)
correction coefficient of spectral characteristics
$U_L(t)$
model of characteristic of mouth radiation
L(z)
S(z)
s(t)

Fig. 1. A model of the generation of speech sounds

Excitation with pulses of amplitude σ, separated in time by $T_0 = IT$, where I is a positive whole number and T – a standardized sampling interval (for the sake of simplification it can be assumed that $T = 1$), can be expressed in the z-domain by the formula

$$E(z) = \sigma \sum_{n=0}^{\infty} (z^{-I})^n = \frac{\delta}{1-z^{-I}} \quad \text{for } |z| > 1. \tag{1}$$

In the case of quasiperiodic sounds this excitation is applied at the outlet of the larynx model $G(z)$, represented in Fant's model by a low-pass bipolar filter. The signal $U_G\,(t)$, obtained at the outlet of the filter $G(z)$, is in turn supplied to the model of the vocal tract $V(z)$ which has the form of step-like resonators connected with each other. The number of resonators is connected with the number of the resonance frequencies of formants assumed in a given implementation of the model [6]. In practical research the number of formants assumed is relatively small, resulting in an imprecise representation of the spectral characteristic over the low frequency range. In order to equalize the imprecision of the representation of the spectrum over this range, Fant introduced in his model the so-called *correlation factor of the spectral characteristic* [6] which, however, in numerical representations of this model is not usually taken into consideration [7]. The signal $U_L(t)$ which is formed at the output of the model is supplied to the system $L(z)$ representing the frequency characteristic of the mouth radiation. It is the last element of the linear model of speech sound generation. The model given by Fant can be written as a z-transform using the equation

$$X(z) = E(z)G(z)V(z)L(z), \tag{2}$$

where $X(z)$ is the z-transform of the discrete signal $X(nt)$. On account of the fact that in the present model excitation may be assummed to be with either periodic pulses or noise and that a limited fixed number of formant frequencies and formant bandwidths are assumed, Markel [6] suggests that this model may be directly used for the representation of vowel and fricative sounds. Based on the analytical model (Fant's model), the investigation of quasiperiodic speech segments can be performed in order to define successive periods of larynx excitation and, at the same time, of successive values of the frequency of the larynx tone.

With simplification, the model of speech sound analysis can be written as

$$E(z) = X(z)A(z), \tag{3}$$

where

$$A(z) = \frac{1}{G(z)V(z)L(z)} \tag{4}$$

is the inverse filter. It can be expressed, with simplification, by the following equation in the z-domain:

$$A(z) = \sum_{i=0}^{M} a_i z^{-i}, \quad a_0 = 1, \tag{5}$$

where a_i are the coefficients of the filter $A(z)$, which in addition to the order of the filter M and the signal $X(z)$ define completely the analytical model given by formula (3). In the (discrete) time domain equation (3) can be written as

$$e(n) = \sum_{i=0}^{M} a_i x(n-i) = x(n) + \sum_{i=1}^{M} a_i x(n-i), \tag{6}$$

which means that the M-th coefficient of the inverse filter, otherwise known as the *coefficient of prediction* of the sample $x(n)$, requires a linear combination of the previous M samples. In order to define the coefficients of the inverse filter (i.e. the prediction coefficients), in order to minimize $e(n)$, M linear equations must be solved [5, 6]. In practice, this is achieved by the use of covariance or autocorrelation methods of analysis of the speech signal samples. This means that on the basis of the model of linear prediction, the parameters of the model of speech analysis can be determined in a relatively easy manner which is essential for the further considerations.

2. Determination of the F_0 parameter using the method of linear prediction

The model of speech sound analysis, based on the technique of linear prediction, is a model which can, in a natural manner, due to its relatively easy numerical implementation, be used in speech analysis performed with an electronic digital computer. The methods for the estimation of the frequency of the larynx tone which have been used so far, used the idea of flattening (smoothing) of the spectrum of the signal analyzed, in which the phases of the individual harmonic components of the spectrum were reduced to zero. The signal obtained through the inverse transformation had a structure of the form of pulses repeating at intervals equal to a period of the larynx tone (in the case of a voiced signal) or had a random character (in the case of an unvoiced signal) [10].

The technique of linear prediction and, in particular, the method of analysis by which we obtain the sequence $\{e(n)\}$ with parameters described by formula (6), can give similar effects. The sequence $\{e(n)\}$, obtained in the analysis by the method of linear prediction, called in this method the *error signal*, is a signal with distinct pulses occurring at successive moments of larynx tone generation. The distance between them is the period of this tone (if a voiced quasiperiodic signal is analysed). The signal $\{e(n)\}$ has a random character for all other signal classes. This is in agreement with the assumption of the linear model of speech generation, which is the basis for linear prediction. There is, however, a certain "fault" in the characteristics of the error signal obtained for a quasiperiodic signal, which consists in a rather significant decrease in the amplitude of pulses and changes of sign within the sounds analyzed or in transitions from phoneme to phoneme. Fig. 2 shows the error signal obtained with the present model for a fragment of a word "pokoju" contained in the phrase "w pokoju paliła się

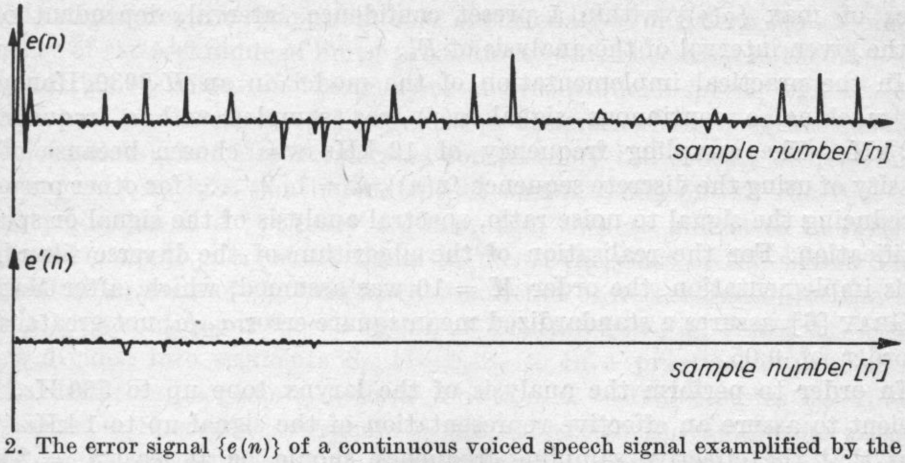

Fig. 2. The error signal $\{e(n)\}$ of a continuous voiced speech signal examplified by the word "pokoju" ($\delta = 1$)

słaba żarówka". This fault can be overcome by adequate operation on the coefficient σ, for example by making it dependent on the amplitude of the real signal. In spite of the above remarks, it can be stated that this approach provides a basis for the development of an effective method for estimation and extraction of the parameter F_0 in a signal of continuous speech, using the technique of linear prediction. A general functional diagram for implementing the estimation of the frequency of the larynx tone is shown in Fig 3. At the

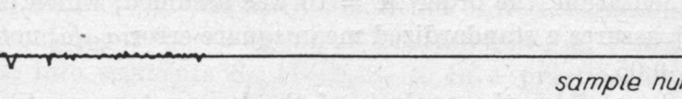

Fig. 3. A functional model of the estimation of the frequency of the larynx tone

output of the model in Fig. 3 we obtain the sequence $\{F_{0i}\}$ of values of the frequency of the larynx tone $F_0 = 1/T_0$ corresponding to successive periods of larynx excitation in a quasiperiodic signal of voiced human speech. An analysis of the length of the period of the larynx tone T_0 can be performed using an autocorrelation analysis of the sequence of the error signal $\{e(n)\}$, described by the following expression:

$$\varrho(j) = \sum_{n=0}^{N-1-j} e(n)e(n+j). \tag{7}$$

In the estimation of F_0 the sequence of values obtained for the autocorrelation $\{\varrho(j)\}$ is in practice calculated for $j \leqslant N/2$. The choice of the values of N depends on the sampling frequency of the analogue signal f_p and affects the effect of averaging the values of the sequence $\{F_{0i}\}$. The sequence $\{\varrho(j)\}$ obtained serves directly for determination of duration of the period T_0 by finding the values of max $\{\varrho(j)\}$ within a preset confidence interval, dependent on f_p and the given interval of the analysis of F_0.

In the practical implementation of the model on an H-6030 Honeywell Bull machine, a continuous signal $x(f)$ was sampled with a frequency $f_p = 12$ kHz. The sampling frequency of 12 kHz was chosen because of the necessity of using the discrete sequence $\{x(n)\}$, $n = 1, 2, \ldots,$ for other purposes, e.g. reducing the signal to noise ratio, spectral analysis of the signal or speaker identification. For the realisation of the algorithm of the inverse filter $A(z)$ in this implementation, the order $M = 10$ was assumed, which, after MARKEL and GRAY [6], assures a standardized mean square error a_M/a_0 not greater than the order of 0.05.

In order to perform the analysis of the larynx tone up to 500 Hz, it is sufficient to assure an effective representation of the signal up to 1 kHz. This means that the effective sampling frequency should be at least $f_p = 2$ kHz. To achieve this, for a signal sampled at $f_p = 12$ kHz, it is sufficient to use the procedure of initial filtering consisting in the selection of samples with numbers increasing by a constant value K assumed in the relation $f_p = f_p' K$. This approach additionally ensures a considerable reduction in the number of elements of the sequence $\{x(n)\}$ to be analyzed, which significantly affects the speed of signal processing. In the present situation $K = 6$ was assumed. 480 samples of the primary signal $\{x(n)\}$ were included in each analytical step, which permitted an autocorrelation analysis, defined by formula (7), of a signal of duration $N = 80$. Considering formula (7) and the expression for the constant j occurring in this formula, $j_{\max} = 39$ was assumed which permits, in practice, an analysis of the larynx tone for frequencies above $F_d = 51$ Hz. The upper frequency limit of the analysis is fixed by practical considerations of searching for the absolute maximum in the sequence $\{\varrho(i)\}$.

In other words, we search for the absolute maximum value of the sequence $\{\varrho(i)\}$ in a certain preset interval Q, subsequenly called the *confidence interval.* In the algorithm implemented here the lower boundary of the confidence interval was taken as $l = 6$, and the upper boundary as $k = 37$, which permits an estimation of the frequency of the larynx tone up to $F_g = 333$ Hz and involves the assumption of such a value of i that the following condition would be satisfied within the voiced signal:

$$\max_{l \leqslant i \leqslant k} \{\varrho(i)\} > \varrho(l). \tag{8}$$

3. Algorithm for determining F_0 by the method of linear prediction

The algorithm presented for using the method of linear prediction served for its numerical implementation in the form of a programme in Fortran.

It was assumed from the beginning that the analysis would be performed only within the class of voiced signals of continuous speech. This assumption not only eliminates the "imperfections" of the analysis of speech signals inherent in the use of the technique of linear predictions, which occurred in the implementation of this technique in the analysis of a continuous speech signal without separation of signal classes [6] (e.g. within a signal which does not belong to the voiced class), but also accelerates considerably the process of determining the parameter F_0 in a continuous speech signal. Thus one of the first steps in this algorithm is a "primary segmentation" whose aim is to define, with the precision up to a constant signal segment, the class of the signal. This is an essential element of the schematic algorithm of determination of the parameter F_0 in a continuous speech signal, as illustrated in Fig. 4. A discrete signal $x(n)$ is divided into segments S_l, $l = 1, 2, \ldots$ In a practical implementation of the algorithm, segments with a duration $T = 128$ samples (~ 10.6 ms) were assumed. Successive segments S_l are subject to the amplitude criterion. The segments are considered as silent if the maximum amplitude of a signal within a given segment does not exceed β times the threshold value of β_1. The threshold value β_1 is determined for each segment S_l separately as 15% of the maximum amplitude of the signal $x(n)$ from 8 successively analyzed segments S_l; but it cannot take a value less than 10 (with 128 levels of quantization of the amplitude of the input signal).

Let $a_{1,l}$ denote the factor by which the threshold β_1 is exceeded, where

$$a_{\beta_{1,l}} = n\{X_i\colon\ X_i \geqslant \beta_1 \wedge X_i \in S_l\}, \tag{9}$$

$X_i \in S_l$ being the values of the l-th segment of duration T (the values of the signal sampled and quantized), while $n\{\cdot\}$ is the size of the set of the values of the amplitude of the samples of the signal exceeding the threshold value β_1 in the segment S_l. The value of $a_{\beta_{1,l}}$ is the basis for determining the class of the signal represented by the segment S_l. A segment is assigned to the class of "no signal" if the following condition is satisfied:

$$a_{\beta_{1,l}} < \beta. \tag{10}$$

$\beta = 30$ was assumed for this algorithm. Only those segments S_l for which $a_{\beta_{1,l}} \geqslant \beta$ were subject to further analysis. Segments with a quasiperiodic character are recognized as all those l-th segments of duration T for which $a_{\beta_{2,l}} \geqslant \beta_0$, where β_2 is the threshold value of the amplitude, and $\beta_0 = 40$. The parameter β_2, similarly to β_1, is dynamically determined as 80% of the value of the maximum amplitude (its value cannot be less than 42). All the remaining segments undergo further analysis based on the frequency criterion. They are initially assigned to the class of "noise" signals, to the class of "voiced quasiperiodic" signals or to the class of signals conditionally recognized as "noise" signals ("conditional noise"). The frequency criterion is based on the calculation of the zero-crossing parameter. The segments for which the number of zero-crossings is greater than $\beta_3 = 42$ (which for $f_p = 12$ kHz corresponds to a frequency of about 2 kHz) are recognized as "noise" signal. However, the segments with a number of zero-crossings over $\beta_4 = 25$ (~ 1.2 kHz) and a maximum amplitude satisfying the condition

$$\max\{X_i\} < \beta_2, \qquad X_i \in S_l, \tag{11}$$

are considered as "conditional noise". The segments of "conditional noise" can be considered as segments of voiced signal or as "noise". If in the direct neighbourhood of the segment recognized as a "noise" segment, a group of

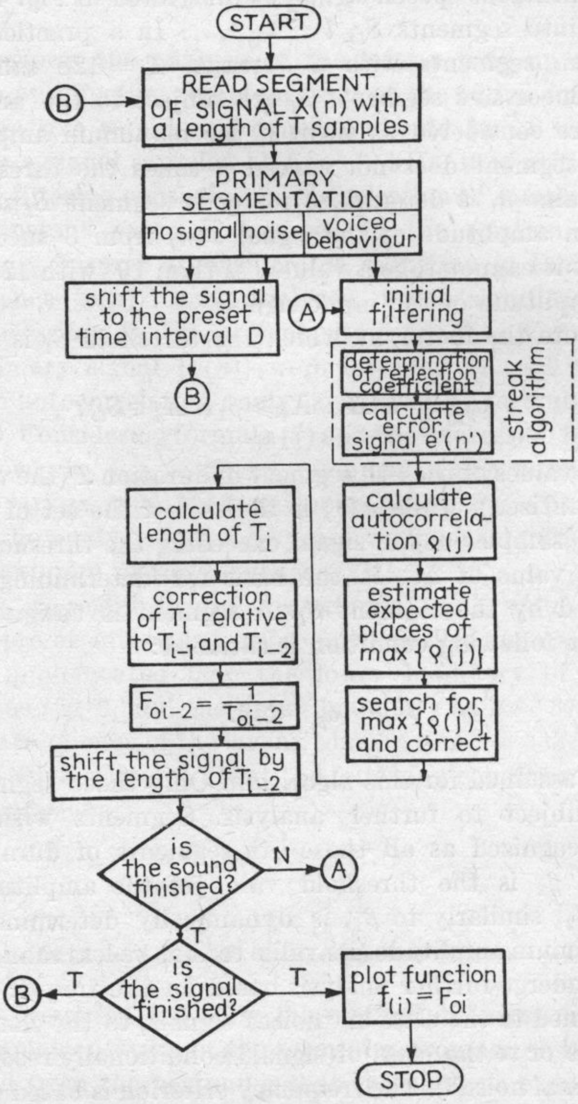
START
B
READ SEGMENT OF SIGNAL X(n) with a length of T samples
PRIMARY SEGMENTATION
no signal
noise
voiced behaviour
shift the signal to the preset time interval D
B
A
initial filtering
determination of reflection coefficient
calculate error signal e'(n')
streak algorithm
calculate length of T_i
correction of T_i relative to T_{i-1} and T_{i-2}
$F_{oi-2} = \frac{1}{T_{oi-2}}$
shift the signal by the length of T_{i-2}
calculate autocorrelation $\varrho(j)$
estimate expected places of $\max\{\varrho(j)\}$
search for $\max\{\varrho(j)\}$ and correct
is the sound finished?
N
A
T
is the signal finished?
T
B
T
plot function $f_{(i)} = F_{oi}$
STOP

Fig. 4. A schematic diagram of the algorithm for determination of the parameter F_0 in a continuous speech signal; N – no, T – yes

"conditional noise" segments occurs, the whole group is in its entirety considered to be "noise". In the opposite case the "conditional noise" segments are assigned to the class of voiced segments.

A fragment of the algorithm for the determination of the parameter F_0 in a continuous speech signal, concerning the primary segmentation, is shown in Fig. 5. Sections of the speech signal with a duration of about 85.3 ms were

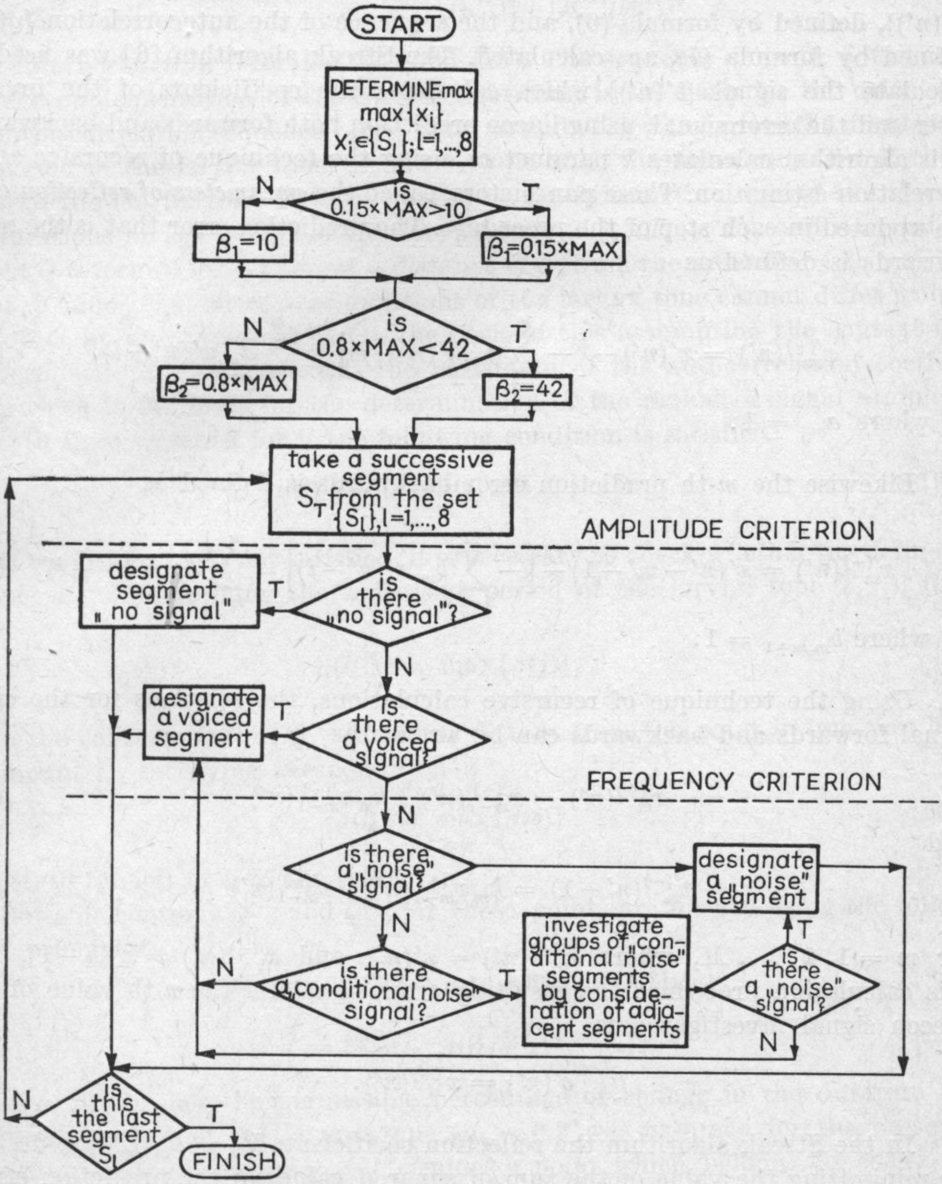
START
DETERMINEmax
max {x_i}
x_i∈{S_l}; l=1,...,8
N
is 0.15×MAX>10
T
β_1=10
β_1=0.15×MAX
N
is 0.8×MAX<42
T
β_2=0.8×MAX
β_2=42
take a successive segment S_T from the set {S_l}, l=1,...,8
AMPLITUDE CRITERION
designate segment „no signal"
T
is there „no signal"?
N
designate a voiced segment
T
is there a voiced signal?
FREQUENCY CRITERION
N
is there a „noise" signal?
designate a „noise" segment
N
N
is there a „conditional noise" signal?
T
investigate groups of „conditional noise" segments by consideration of adjacent segments
is there a „noise" signal?
T
N
N
is this the last segment S_l
T
FINISH

Fig. 5. The algorithm of the primary segmentation for a group of eight segments S_l
N – no, T – yes

subject to a single primary segmentation which corresponded to two tape blocks with 512 samples of the signal sampled at $f_p = 12\,\mathrm{kHz}$. Within the segments assigned to the voiced class, further denoted by $\{x(n)\}_D$, the initial filtering (described in Section 2) was conducted, as a result of which we obtained the signal $\{x'(n')\}_D \subset \{x(n)\}_D$, where $n' = Ki$, $i = 1, 2, \ldots$ and $n = 1, 2, \ldots$ According to the foregoing discussion, the size of the set $\{x'(n')\}_D$, taken for analysis in one step of the algorithm, is 80. For this signal the error signal $\{e'(n')\}$, defined by formula (6), and the sequence of the autocorrelation $\{\varrho(j)\}$, defined by formula (7), are calculated. The Streak algorithm [6] was used to calculate the signal $\{e'(n')\}$ which calculates the coefficients of the inverse filter and the error signal, using linear prediction both forwards and backwards. This algorithm calculates k parameters, using the technique of recursive autocorrelation estimation. These parameters, called the ***parameters of reflection*** of k, are updated in each step of the procedure. The prediction error that is the m-th forwards is defined as

$$x'^{(+)}_m(n') = x'(n') - \Big[-\sum_{i=1}^{m} a_{m_i} x'(n'-i)\Big] = \sum_{i=c}^{m} a_{m_i} x'(n'-i), \tag{12}$$

where $a_{m_0} = 1$.

Likewise the m-th prediction error backwards is defined as

$$x'^{(-)}_m(n') = x'(n'-m'-1) - \Big[-\sum_{i=1}^{m} b_{m_i} x'(n'-i)\Big] = \sum_{i=1}^{m+1} b_{m_i} x'(n'-i), \tag{13}$$

where $b_{m,m+1} = 1$.

Using the technique of recursive calculations, the relations for the error signal forwards and backwards can be written as

$$x'^{(+)}_m(n') = x'^{(+)}_{m-1}(n') + k_m x'^{(-)}_{m-1}(n') \tag{14}$$

and

$$x'^{(-)}_m(n'+1) = k_m x'^{(+)}_{m-1}(n') + x'^{(-)}_{m-1}(n') \tag{15}$$

for $m = 1, 2, \ldots, M$, where $x'^{(+)}_0(n') = x'(n')$ and $x'^{(-)}_0(n') = x'(n-1)$. For this calculation procedure the resulting error signal for the n-th value of the speech signal investigated is

$$e'(n') = x'^{(+)}_M(n'). \tag{16}$$

In the Streak algorithm the reflection coefficients $k_m = k_m(n')$ are defined by minimizing the value of the sum of squared values of the prediction errors forwards and backwards:

$$[x'^{(+)}_{m}(n')]^2+[x'^{(-)}_{m}(n'+1)]^2 \\ = 2k_m x'^{(+)}_{m-1}(n')\,x'^{(-)}_{m-1}(n')+(1+k_m^2)\{[x'^{(+)}_{m-1}(n')]^2+[x'^{(-)}_{m-1}(n')]^2\}. \tag{17}$$

This minimization gives, in effect, the expression

$$k_m = k_m(n') = \frac{-2x'^{(+)}_{m-1}(n')\,x'^{(-)}_{m-1}(n')}{[x'^{(+)}_{m-1}(n')]^2+[x'^{-1}_{m-1}(n')]^2}, \tag{18}$$

taken into account in formulae (14) and (15). The sequence $\{e'(n')\}$, obtained from the implementation of the Streak algorithm, is the basis for the calculation of the correlation sequence $\{\varrho(j)\}$ using formula (7), which gives the duration of a period of the larynx tone. In order to do this, the mean duration of the period of the larynx tone T_s is calculated as the arithmetic mean of the three last durations of the period of the larynx tone T_{i-2}, T_{i-1}, T_i, and a certain context Q is formed for a point at a distance of T_s from the preceding $j_{\max}\{\varrho(j)\}$. It was assumed that successive durations of the larynx tone cannot differ from each other by more than 30 %. On the basis of this assumption the context Q is determined. In the context Q the maximum of the autocorrelation coefficient, which is the basis for the determination of the period of signal samples of length l_s, is searched for if the following condition is satisfied:

$$\bigvee_{l\in Q}\,[\varrho(l) = \max\{\varrho(j)\};\ 6\leqslant j\leqslant 37]. \tag{19}$$

If condition (19) is not satisfied, it is necessary to check whether the defined distance l_{i+1} (determining the successive period of the larynx tone T_{i+1}), for which

$$\varrho(l_{i+1}) = \max_{j}\{\varrho(j)\},$$

lies in the contexts $Q^{(+)}$ or $Q^{(-)}$ of points distant from the point in question by an amount l_s, satisfying the condition

$$\varrho(l_s) = \max_{j\in Q}\{\varrho(j)\} \tag{20}$$

by $0.5l_s$ to the left or l_s to the right.

Suitable contexts $Q^{(-)}$ and $Q^{(+)}$ for these points are formed using the following principles:

$$Q^{(-)} \equiv [0.5l_s(1-\mu_1),\ 0.5l_s(1+\mu_1)], \tag{21}$$

$$Q^{(+)} \equiv [2l_s(1-\mu_1),\ 2l_s(1+\mu_1)], \tag{22}$$

where μ_1 establishes the permissible percentage of change in the duration of successive periods of the larynx tone ($\mu_1 = 0.2$ was assumed for the present algorithm). If the distance l_{i+1} determines a point which belongs to either of the contexts $Q^{(-)}$ or $Q^{(+)}$, then the distance l_s is accepted only as the duration of a successive period of the larynx tone, recognizing that an error of a double decrease or increase of the distance occurred in the course of the analysis. Examples of these type of errors are reflected in Figs. 6 and 7. In both cases

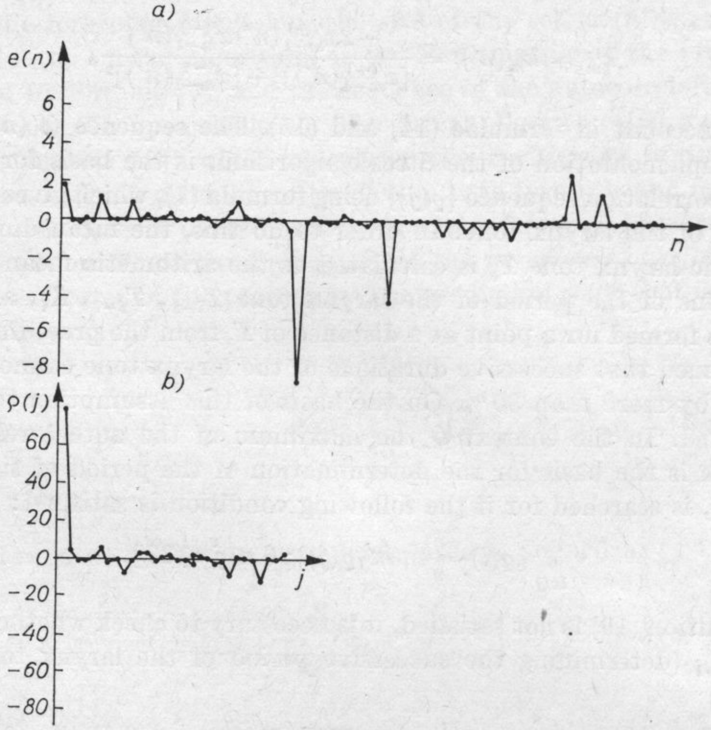

Fig. 6. The error signal $\{e(n)\}$ of the final phase of the duration of the sound /u/ in the word "pokoju" – (a), and the corresponding sequence of the values of the autocorrelation $\{\varrho(j)\}$ – (b)

they apply to points in the real signal $\{x(n)\}$, where the values of the signal samples are lower than the others in a specific segment of the signal, e.g. in the final phase of the duration of a sound in the word "pokoju" in the phrase whose signal is presented in Fig. 8 (Fig. 6), or when the amplitude of the signal is decreased and its structure is changed, e.g. in the transition between sounds (a) and (l) in the word "paliła" represented by the signal in Fig. 8 (Fig. 7). If, however, the distance determined does not belong to the contexts $Q^{(-)}$ or $Q^{(+)}$, we have

$$\varrho(l_s) \geqslant u\,\varrho(T_{i+1}), \tag{23}$$

where u is the threshold coefficient, which in the present algorithm took the value $\mu = 0.75$. If condition (23) is satisfied, then l_s is assumed as the duration of a successive period of the larynx tone. If not, the duration of this period is established as l_{i+1}.

Thus, errors in the definition of the larynx tone which were due to a large variation in the values of the sequence $\{e'(n')\}$ for successive periods, and hence in the sequence $\{\varrho(j)\}$, were eliminated (cf. Figs. 6 and 7).

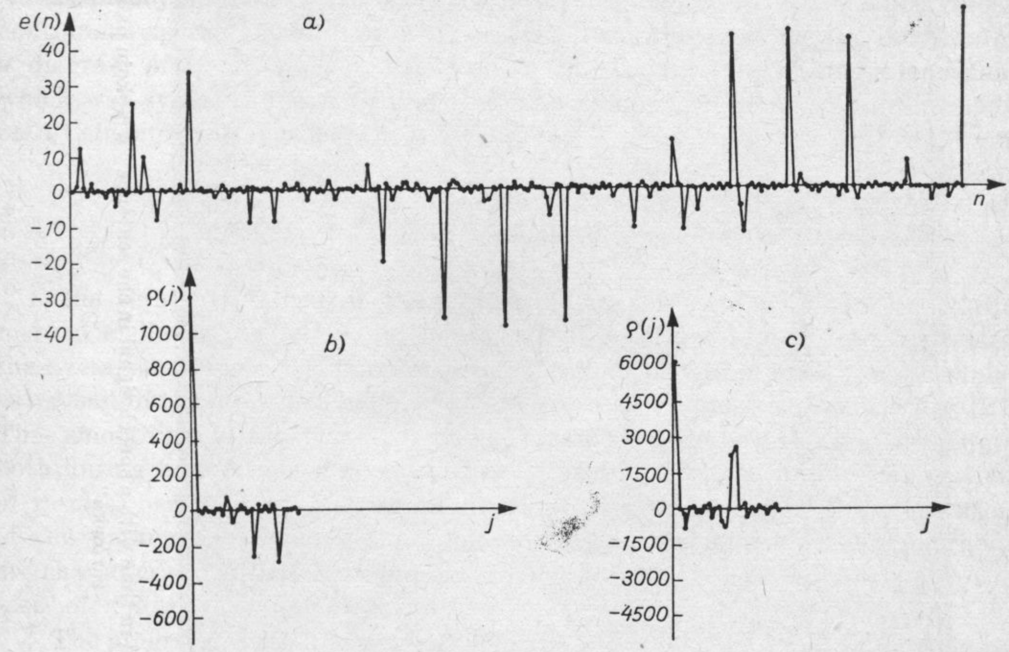

Fig. 7. The error signal $\{e(n)\}$ calculated for the transition between the sounds /a/ and /l/ in the word "wypaliła" $-(a)$, and synchronised therewith in terms of the origin of the sequence of values of the autocorrelation $\{\varrho(j)\}-(b)$ and (c)

The calculated duration of the period of the larynx tone T_{i-1} and T_i is subsequently corrected in order to eliminate errors resulting from large changes in the amplitude of the signal $e(n)$ (for $\delta = \text{const.}$ see formula (2)). In the correction stage the three successive values of the length of the period of the larynx tone T_{i-1}, T_i and T_{i+1}, determined in the preceding steps of the algorithm, are considered. Taking the trend of these values and approximating them by linear extrapolation (such extrapolation was assumed for the present algorithm after the extrapolation given by Markel [6]), new values for the duration of the larynx tone are determined, which effectively "corrects" those values which deviated from the extrapolated ones and were caused by the errors mentioned earlier.

In this way, in the i-th step of the algorithm the final frequency of the larynx tone, determined in the $(i-2)$-nd step, is calculated from the period T_{i-2}.

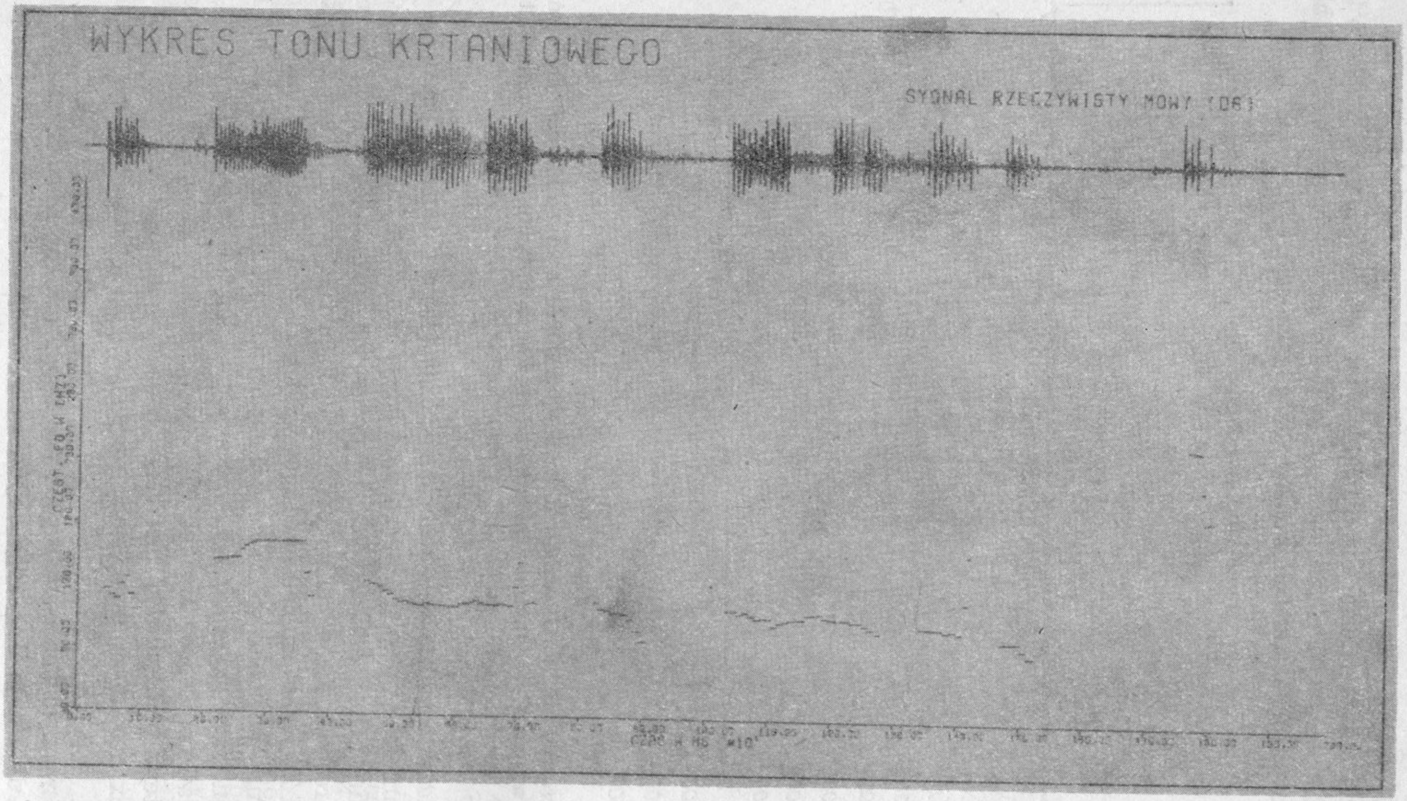

Fig. 8. A diagram of the larynx tone synchronised in time with a speech signal in the sentence "w pokoju paliła się słaba żarówka"

In the present algorithm this frequency is calculated according to the formula

$$F_{0i-2} = \frac{12000}{(T_{i-2}-1)\cdot 6} \tag{24}$$

and is one of the values of the sequence $\{F_{0i}\}$ corresponding to the values of the frequency of the larynx tone dynamically calculated for the class of voiced continuous speech signal. The sequence $\{F_{0i}\}$ was the basis for the creation of a diagram of the frequency variation in the larynx tone during phonation, which was synchronised with real time. An example of such a diagram made on a Calcomp plotter is shown in Fig. 8.

4. Conclusions

The results obtained of the estimation of the parameter F_0, using the method of linear prediction, in the algorithm described above, fully represent the averaged (smoothed) signal of the real values of F_0 obtained, for example, as a result of the implementation of an algorithm of "primary segmentation" [2]. This smoothing is the result of the averaging of the values of parameters both during calculation of the autocorrelation function and the approximation of results, permitting a partial elimination of disturbances in the estimation of the parameters. Thus the parameters $\{F_{0i}\}$, determined by this method, are characterised with a considerably smaller scatter of their values than in the case of "primary segmentation".

The program, written in Fortran 6000 for a H-6030 Honeywell Bull machine, developed by the author, was not optimized either in terms of memory occupied or processing time, and in the present form occupies about 15 K words of memory (together with Fortran subprogrammes and programmes for plotter operation). Thuś as an example the implementation time of a programme for a 2-second continuous speech signal (recorded in 50 tape blocks, 512 words-signal samples each) was $\simeq$ 57.9 s (processor time). It is a relatively long time and excludes the implementation of investigations in real time. It should, however, be noted that the H-6030 machine was designed for data processing and not for scientific and technical calculation, and with a calculation speed of the order of about 200000-300000 operations per second naturally offered no such possibilities.

The mean number of wrongly estimated values of the sequence $\{F_{0i}\}$, using the method of linear prediction, calculated for a continuous speech signal with a duration of 80 s, did not exceed 2.5% of the total number of values estimated. The errors occurred particularly in those places of the signal where plosives appeared, which could be classified as voiced quasiperiodic signals. These can be eliminated by the introduction of a more precise way of determining the boundaries of the segments of voiced sounds in the stage of signal classification. In spite of those imperfections, the results obtained can be used in further stages of human speech analysis. While the results obtained from "primary segmentation" [2] served for a speech recognition oriented spectral analysis synchronised with the larynx tone [3], the results obtained from the implementation of the programme, based on the method of linear prediction, are not suitable for this type of analysis. They may, however, still be used in statistical investigations of the parameter F_0 with a view to identification or verification of a speaker on the basis of the statistical characteristics of the distribution of the frequency F_0.

References

[1] J. W. Bayless, S. J. Campanella, A. J. Goldberg, *A survey of speech digitization techniques*, Proceedings of Carnahan conference on electronic crime countermeasures, 80-81 (1972).

[2] A. Dziurnikowski, *Microphonemes as fundamental segments of a speech wave, Primary segmentation – automatic searching for microphonemes*, Advence Papers of the IV IJCAI – 75, Vol. 2, Tbilisi 1975.

[3] A. Dziurnikowski, *Primary sequentation of speech sound signals in the SUSY system* (in Polish), Reports JJ UW, **52** (1976).

[4] G. C. M. Fant, *Acoustic theory of speech production*, Mouton and Co., s' – Gravenhage, The Netherlands, 1960.

[5] K. Jaszczak, *Digital modelling of speech signal using linear prediction*, JJ UW, **62** (1977).

[6] J. D. Markel, A. M. Gray, *Linear prediction of speech*, Springer Verlag, Berlin – Heidelberg – New York 1976/1977, 132-134, 190-206.

[7] L. R. Rabiner, *Digital-formant synthesizer for speech synthesis*, JASA, **43**, 822-828 (1968).

[8] D. R. Reddy, *Pitch period determination of speech sound*, Communications of the ACM, **10**, 6, 343-348 (1967).

[9] S. Saito, F. Itakura, *The theoretical consideration of statistically optimum methods for speech spectral density*, Report No. 3107, Electr. Commun. Lab., N.T.T., Tokyo 1966.

[10] M. M. Sondhi, *New methods of pitch extraction*, IEEE Trans., AU-16, 262-266 (1968).

[11] B. Wierzchowska, *Polish pronunciation*, PZWS, Warszawa 1971.

Received on February 23, 1979.