ARCHIVES OF ACOUSTICS
2, 2, 121-137 (1977)

A MATHEMATICAL MODEL FOR THE DETERMINATION OF SPEECH SIGNAL PARAMETERS WITH THE AID OF INSTANT MEMORY

ZBIGNIEW MARCIN WÓJCIK

Institute of Biocybernetics and Biomedical Technics (Warszawa)

In this paper the concept is presented of a system for the determination of basic parameters of speech signals, including formants and their transients. These parameters are described in stages by means of characteristic functions of the sound features. Values of the characteristic functions for the j-th stage of the sound features are arguments for the characteristic functions of other features of the $(j+1)$-st stage. If the values of these functions equal unity, they are automatically recorded in the memory. The length of the time of storage of these values is determined by physical properties of the sound. Physical properties at the $(j+1)$-st stage are described by means of relations that establish the sequence of appearance of the sound features as a function of time at the output of the j-th stage.

The system enables the analysis of sounds in real time, as well as the automatic segmentation of sound as a function of time. The accuracy of the identification of speech does not depend on the frequency of the larynx tone or on the speech rate, etc. With the system it is possible to use bandpass filters with comparatively broad transmission band widths. A formal description of the system by means of characteristic functions of the features allows direct design and construct of the system by means of generally available integrated circuits.

1. Introduction

Present methods for the identification of speech sounds take advantage of the formant theory of the speech signal. The sounds developed as a result of inducing vibration of the vocal chords (larynx tone), and by air passing through the contracting oral cavity and larynx cavity (noise), are modulated and filtered through the larynx cavity, the oral cavity separated by the tongue, and nasal canal, and have a formant structure. That is to say that one can distinguish frequency ranges in which the spectral energy density assumes locally maximum values, and also frequency ranges in which the spectral energy density decreases strongly to minimum values (antiformants). The changes of geometrical dimensions of the cavities (especially of the oral cavity), observed during speaking, cause the formation of transients, i.e. changes of the frequencies of the formants with time. Formants and their transients are the basic acoustic-phonetic parameters used to describe speech signals and, especially, voiced sounds.

From point of view of essence transmission a huge excess of information in a speech signal occurs. In the system presented in this paper the reduction of this excess is connected with the measurement of basic parameters: formants and their transients. The measurement is effected automatically by the spectral analysis of a speech signal.

2. Assumptions, conceptions and working range

The direction of this paper has been provided by the need to build a system for the automatic identification of speech in real time, which could be constructed with generally obtainable electronic subassemblies, i.e. integrated circuits and medium-size computers.

Proposed and constructed sound identification systems face various difficulties that result from the physical nature of a speech signal, from imperfections of the individual subsystems and from simplifications introduced as assumptions in the solution of technical problems.

In view of the high operating costs of computers with very large memories, it is more advantageous to use a system permitting input of a speech signal into the computer (Fig. 1) with the basic parameters already determined. Such an approach would permit the construction of useful systems within which the computer could be utilized for performing various tasks, its operation being controlled by means of a speech signal. This paper presents the concept of a system intended to determine the basic parameters of a speech signal (block E in Fig. 1).

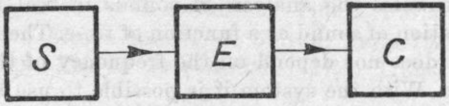

Fig. 1. General block diagram of the system for speech identification
S – source of speech signal, E – system of singling out the basic parameters of the speech sound, C – universal digital machine

To maintain the general character of these considerations no specific technical parameters of individual subsystems have been stated. In the description the possibility has been accepted of the construction of a set of bandpass filters with parameters to enable them to locate (as functions of time and frequency) the formant parameters of a speech sound with sufficient accuracy. These assumptions are substantiated by the results of works cited by SAPOZHKOV [5] and, especially, the results of works obtained by KUBZDELA [4]. The contemporary literature on the subject of the automatic identification of speech points to the necessity of also considering in the analysis the dynamics of changes in the measured parameters of a speech signal. This paper is a proposal to associate the already known parameters with their dynamics and changes in the speech signal (e.g. the definition of a formant is – according to the conception of their detectability stated in items 9 and 12 – very similar to classical definitions). The criteria of the determination of basic parameters of speech signal, accepted simultaneously, are: the magnitude of relative changes of the signals at the output of band filters (i.e. positive and negative amplitude increments), the rate of these changes, and the time intervals between individual and predetermined changes of speech signals. The value of each of determined parameter is stored in the system (in the memory) during a short period of time (the so-called *instant memory time*).

The concept of instant memory and its role in the process of determing the parameters of the speech signal are presented in the following sections. The final verification of the conception presented can unfortunately only be effected experimentally.

The anticipated effects of the conception are stated in the concluding section of the paper.

3. The organization of instant memory

In the proposed system the speech signal is described in stages by means of characteristic functions of the sound features. Each of these functions describes a specific property (feature) of the speech signal. The function assumes a value equal to unity when the shape of the speech signal conforms to the assumed description. Only those properties of the speech signal are described which are of essential significance for the determination of basic speech parameters.

The system is built hierarchically: the values of characteristic functions of the features of the speech sounds of the j-th stage are arguments for the characteristic functions of features of these sounds at the $(j+1)$-st stage of processing. When the characteristic functions assume values equal to "1", an automatic recording of these values into the memory is made.

The length of time t_{pj} of the storage of the value "1" of characteristic functions of the j-th stage of the system are described in the following manner:

$$t_{p(j-1)} \leqslant t_{pj} \approx \tau_{pj}, \tag{1}$$

where $j = 2, 3, \ldots, \tau_{pj}$ is the experimentally determined time of the occurrence of properties of the speech sound, as described at the j-th stage of the system. Thus the duration of time t_{pj} is determined by physical properties of sounds.

It is assumed that the access to the memory (i.e. to the values of the arguments of characteristic functions at each stage of the system) is automatic. The times of reading of the individual values "1" of characteristic functions from the memory, at the transition from the j-th stage to the $(j+1)$-st stage, correspond to the times of their storage at the j-th stage.

Physical properties of sounds at the j-th stage are described by means of relations of partial order in time for the value "1" of the characteristic functions of features of the $(j-1)$-st stage.

Let $D = \{d_1, d_2, \ldots, d_g, \ldots, d_h, \ldots, d_G\}$ be a set of various properties of the sound signals determined at the j-th stage. Any point on the time axis $t_a \in T = \{t_1, t_2, \ldots, t_S\}$ is the time of occurrence of the property d_g of the sound if the value of the characteristic function $Cd_g(t')$ of the feature d_g of the sound satisfies the condition

$$Cd_g(t') = \begin{cases} 0 & \text{for } t' < t_a, \\ 1 & \text{for } t' > t_a. \end{cases} \tag{2}$$

Thus the individual times $\{t_1, t_2, \ldots, t_S\}$ are not known before the analysis of the signal of a speech sound. These times are determined automatically in the course of the analysis as the times at which the characteristic functions have the value "1".

The process of determining the values of the times of the set T can be regarded as a process of an automatic segmentation of the speech signal as a time function.

The function $Cd_g(t')$, which satisfies condition (2), will be denoted by the symbol $Cd_g(t_a)$ or $Cd_g(t)$. The apparent argument t or t_a of the characteristic function will denote the first moment at which this function has the value "1". Let us form the set of functions $\{Cd_g(t_a)\}^{D\times T}$. This is a set of all features determined at the j-th stage during the entire period of analysis (these features may repeat).

Let us select from the set $\{Cd_g(t_a)\}^{D\times T}$ each pair of elements $Cd_g(t_a)$, $Cd_h(t_b)$ $(d_g, d_h \in D; t_a, t_b \in T)$ for which the relation

$$0 < |t_a - t_b| < t_{pj} \tag{3}$$

is satisfied.

For each pair of elements of the set $\{Cd_g(t_a)\}^{D\times T}$, for which relation (3) is satisfied, there is a relation $\mathrel{\varepsilon\!-}$ of partial order in time defined in the following manner:

$$Cd_g(t_a) \mathrel{\varepsilon\!-} Cd_h(t_b) \Leftrightarrow t_a \geqslant t_b. \tag{4}$$

The relation $\mathrel{\varepsilon\!-}$ establishes the sequence of the occurrence of two features d_g and d_h of sounds as a time function; the time between the occurrence of these features is shorter than t_{pj}.

Pairs of elements of the set $\{Cd_g(t_a)^{D\times T}\}$, for which relation (3) is not satisfied, are not arguments for characteristic functions of the speech sound of the $(j+1)$-st stage.

The input data for each stage of the identification are selected in this manner. The reduction of redundant information, and also the extraction of useful information from disturbances, is thus effected at each stage of the operation.

By means of pulses of duration t_{pj} it is possible to detect, with technical facilities available, the features to which these pulses correspond. The duration of the value "1" of the function from the j-th stage is equal to the times of coincidence of the value "1" of the function from the $(j-1)$-st stage, increased by the memory times of the value "1" for the function in the j-th stage.

4. Input system

Investigations of the properties of human hearing show that «with regard to selective properties, the hearing is similar to bandpass filters with a critical band» [5].

In the proposed system the speech sound from the microphone output is transmitted to a set of bandpass filters. At the filter outputs, the signals are rectified and then integrated (Fig. 2). The value of the time constant of the integrating system for each filter should not exceed the reciprocal of the mid-frequency of this filter. In this manner n signals are obtained,

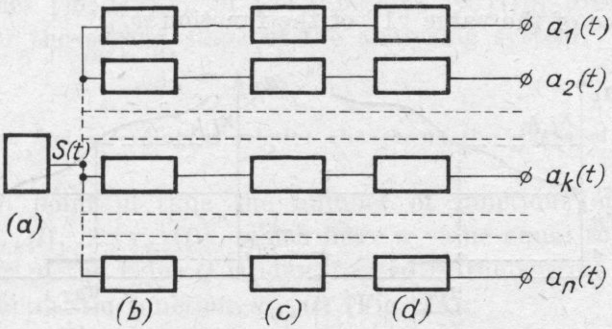

Fig. 2. Diagram of input systems for the analysis of sounds: a) microphone, b) set of bandpass filters, c) rectifying systems, d) integrating systems

$$\{a_k(t)\}_{k\epsilon N}, \qquad N = \{1, 2, \ldots, n\},$$

where k is the number of filters. The signals $\{a_k(t)\}_{k\epsilon N}$ are the arguments for the characteristic functions of the features of sound for the first stage of the system's operation.

5. First stage of system: the determination of the value of characteristic functions of the increase and decrease of signals $\{a_k(t)\}_{k\in N}$

The following properties of the signals output from the analyzer are determined at the first stage of the system: the increase and (independently) the decrease of the signals through set threshold values. It is assumed that for each filter there exists a corresponding set of threshold values $\{x_{1,k}, \ldots \ldots, x_{i,k}, \ldots, x_{q,k}\}$, where k is an index of a band filter and i is an index of the threshold value, $k \in N = \{1, 2, \ldots, n\}$, $i \in E = \{1, 2, \ldots, q\}$.

Each of the elements of the set $\{x_{i,k}\}_{k\in N}$ is denoted by the symbol $x_{i,k}$. Let

$$\{w_{i,k}(t)\}_{k\in N} = \{w_{1,k}(t),\ w_{2,k}(t), \ldots, w_{g,k}(t)\}_{k\in N}$$

and

$$\{m_{i,k}(t)\}_{k\in N} = \{m_{1,k}(t), m_{2,k}(t), \ldots, m_{q,k}(t)\}_{k\in N}$$

be families of sets of characteristic functions which transform signals $\{a_k(t)\}_{k\in N}$ and families of sets of threshold values $\{x_{i,k}\}_{k\in N}$ into sets of values $\big\{\{0, 1\}^E\big\}_{k\in N}$.

Let us consider the function $w_{i,k}(t) \in \{w_{i,k}(t)\}_{k\in N}$. The function $w_{i,k}(t)$ is a characteristic function of the increase of the signal $a_k(t)$ in relation to the preset threshold value $x_{i,k}$ (Fig. 3a),

$$w_{i,k}\big(t, a_k(t), x_{i,k}\big) = \begin{cases} 1 & \text{if } a_k(t) \leqslant x_{i,k} \wedge a_k(t+\Delta t) > x_{i,k}, \\ 0 & \text{otherwise}, \end{cases} \tag{5}$$

where $\big(t, a_k(t), x_{i,k}\big)$ are apparent arguments of the function $w_{i,k}$ and moments of the occurrence of the value "1" of the function $w_{i,k}$.

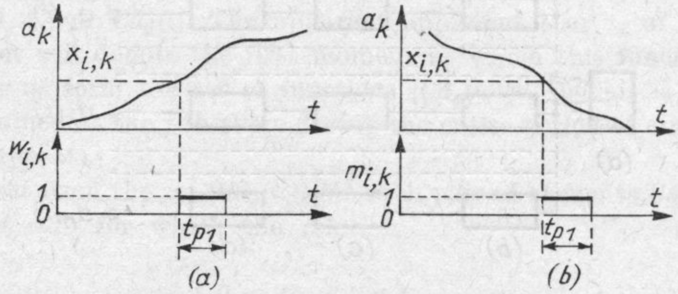

Fig. 3. a) Graphic interpretation of the value "1" of the characteristic function $w_{i,k}(t)$ of the signal increasing $a_k(t)$ through the threshold value $x_{i,k}$, b) physical interpretation of the value "1" of the characteristic function $m_{i,k}(t)$ of the signal decreasing $a_k(t)$ through the threshold value $x_{i,k}$

Let as assume that $m_{i,k}(t) \in \{m_{i,k}(t)\}_{k\in N}$, where $m_{i,k}(t)$ is a characteristic function of the decrease of the signal $a_k(t)$ in relation to the preset threshold value $x_{i,k}$ (Fig. 3b):

$$m_{i,k}[t, a_k(t), x_{i,k}] = \begin{cases} 1 & \text{if } a_k(t) \geqslant x_{i,k} \wedge a_k(t+\Delta t) < x_{i,k}, \\ 0 & \text{otherwise}. \end{cases} \tag{6}$$

Since the human hearing reacts basically to relative change of sound intensities, the values of the thresholds $\{x_{i,k}\}_{k\in N}$ can be distributed in the intervals $\{[x_1, x_q]_k\}_{k\in N}$ according to the logarithmic scale. Such a distribution is justified also by the economics of the system construction (using a smaller number of threshold values).

The values $\{x_{q,1}, x_{q,2}, \ldots, x_{q,k}, \ldots, x_{q,n}\}$ are distributed over a range of corresponding frequencies $\{f_1, f_2, \ldots, f_k, \ldots, f_n\}$, according to the value of the mean speech spectrum.

An essential property of the system is the memorizing of each of the value "1" of the characteristic functions $\{w_{i,k}(t)\}_{k\in N}$ and $\{m_{i,k}(t)\}_{k\in N}$ over short periods of time t_{p1} (Fig. 3), which satisfies the condition

$$t_{p1} \leqslant \tau_{p1}, \tag{7}$$

where τ_{p1} is the smallest duration of an extreme of the sound spectral envelope.

The duration of the maximum of the sound spectral envelope at a frequency f_k is understood as a period during which the output of k-th filter signal exceeds the signals at the $k-1$ and $k+1$ outputs of the filter. More precise determinations of extremes are stated in section 8. The time t_{p1} can be found only by experiment.

The use of the functions $\{w_{i,k}(t)\}_{k\in N}$ and $\{m_{i,k}(t)\}_{k\in N}$ permits further consideration to include only relative changes in the filter output signal stages which satisfy condition (7).

The values $\left\{\{0,1\}^E\right\}_{k\in N}$ of the functions $\{w_{i,k}(t)\}_{k\in N}$ and $\{m_{i,k}(t)\}_{k\in N}$ are arguments for the second stage of the analyzing system.

6. Detection and recording of relative changes in the stages of analyzed signals

For each point in time the number of functions of the set $\{w_{i,k}(t)\} = \{w_{1,k}(t), w_{2,k}(t), \ldots, w_{q,k}(t)\}$, which have a value equal to unity, is totalled. The exceeding of the value Q is identificated in the system by a value of "1" for the characteristic function $w_{Q,k}(t)$ (Fig. 4a):

$$w_{Q,k}[t, \{w_{i,k}(t)\}, Q] = \begin{cases} 1 & \text{if } \left(\operatorname{card}\left\{w_{i,k}(t) \in \{w_{i,k}(t)\} : w_{i,k}(t) = 1\right\}\right) > Q, \\ 0 & \text{otherwise}, \end{cases} \tag{8}$$

where card $\{\ldots\}$ is the number of elements of the set $\{w_{i,k}(t)\}$, which satisfy the condition $w_{i,k}(t) = 1$.

Similar operations are carried out for the set of functions $\{m_{i,k}(t)\} = \{m_{1,k}(t), m_{2,k}(t), \ldots, m_{q,k}(t)\}$:

$$m_{Q,k}[t, \{m_{i,k}(t)\}, Q] = \begin{cases} 1 & \text{if } \left(\operatorname{card}\left\{m_{i,k}(t) \in \{m_{i,k}(t)\}: m_{i,k}(t) = 1\right\}\right) > Q, \\ 0 & \text{otherwise}. \end{cases} \tag{9}$$

A value of "1" for the function $w_{Q,k}(t)$ expresses an increase of the signal $a_k(t)$ in the time interval $\Delta t \leqslant t_{p1}$ by the successive threshold values, with the number of surpassed thresholds being higher than Q (Fig. 4). In a similar manner, a value of "1" for the function $m_{Q,k}(t)$ describes the decrease of the signal $a_k(t)$ in the interval Δt, $\Delta t \leqslant t_{p1}$ by more than Q threshold values. The representation of the set of functions $\{w_{i,k}(t)\}$ gives in effect a large reduction of the input information. The use of the threshold value Q permits a determination of the parameters of the speech signal, which is independent of its stage. The subsequent analysis utilizes only relative changes of the stage of signals output by filters.

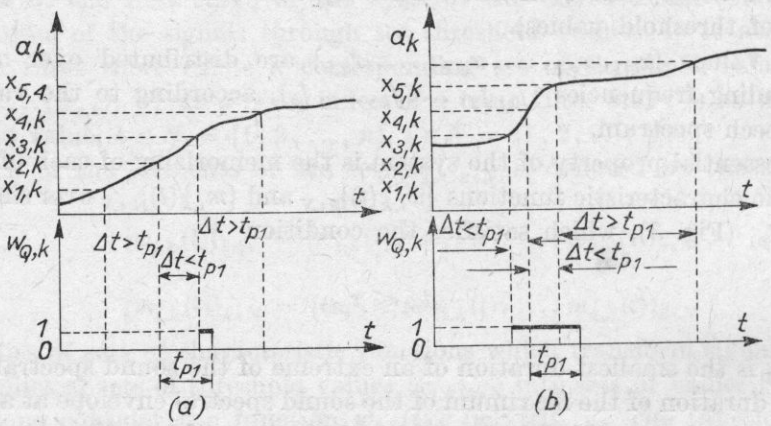

Fig. 4. Physical interpretation of the value "1" of the function $w_{Q,k}(t)$: signals (a) and (b for $Q = 1$ are identified

The longer the time t_{p1} (formula (7)), the smaller the stage changes of the signals $\{a_k(t)\}_{k \in N}$ will be for the values of the functions $\{w_{Q,k}(t)\}_{k \in N}$ and $\{m_{Q,k}(t)\}_{k \in N}$ to assume the value $\{1\}_{k \in N}$. The higher the threshold value Q, the higher the relative increments of signals $\{a_k(t)\}_{k \in N}$ as well as $\{w_{Q,k}(t)\}_{k \in N}$ and $\{m_{Q,k}(t)\}_{k \in N}$ should assume the value $\{1\}_{k \in N}$.

The values $\{0, 1\}_{k \in N}$ of the functions $\{w_{Q,k}(t)\}_{k \in N}$ and $\{m_{Q,k}(t)\}_{k \in N}$ are arguments for the "sharpening" stage of the extremes of the frequency-time envelope of the sound.

7. „Sharpening" of extremes of the frequency-time envelope of the sounds

In order to obtain sufficiently small time constants for the input system it is possible to use bandpass filters with broad, overlapping transmission bands. The entire frequency range to be analyzed can then be covered with a comparatively small number of filters. A single extreme of the sound spectral envelope will then occur at the output of several neighbouring filters.

At this point we will discuss the process of sharpening a single local extreme, independently of whether a formant and its transient will be discovered in the next stages of analysis.

Our system determines automatically at each moment the numbers of filters at the outputs of which the extremes have appeared. The earlier of each two directly adjoining filters is taken. In effect, the local extremes of the spectral envelope are determined.

If an extreme is registered at the output of the k-th filter, followed by a local extreme at the output of the $(k-1)$-st or $(k+1)$-st filter, but not more than a time t_{p3} later (see formula (18)), then the change in the frequency of this extreme will be registered (see item 12). However, if the local extrema occur at directly adjoining (neighbouring) filters after a time interval longer than t_{p3}, these extrema are analyzed separately.

Let us consider the sequence of directly neighbouring filters $k_1, k_2, \ldots, k_m$ at the outputs of which one extreme occurred. In our system the quickest to determine is the value "1", obtained for the function $w_{Q,k}(t)$ amongst all values $\{1\}^M$ of the function $\{w_{Q,k}(t)\}_{k\in M}$, $M = \{k_1, k_2, \ldots, k_m\} \in N$, where $k_1, k_2, \ldots, k_m$ are indices of successive filters at the outputs of which values $\{1\}^M$ of the function $\{w_{Q,k}(t)\}_{k\in M}$ are obtained. The function $w_{Q,k}(t)$, determined in this manner, will be denoted by $w_{r,k}(t)$.

In a similar manner the most quickly determined is value "1", obtained for the function $m_{Q,k}(t)$ amongst all values $\{1\}^M$ of the function $\{m_{Q,k}(t)\}_{k\in M}$. In this case the function $m_{Q,k}(t)$ will be denoted by $m_{r,k}(t)$. Let $g, h \in M, g = h$:

$$w_{Q,g}(t_a), w_{Q,h}(t_b) \in \{C_k(t)\}_{k\in M} \Leftrightarrow 0 < |t_a - t_b| < t_{p2}, \tag{10}$$

$$m_{Q,g}(t_a), m_{Q,h}(t_b) \in \{C_k(t)\}_{k\in M} \Leftrightarrow 0 < |t_a - t_b| < t_{p2}. \tag{11}$$

In the set $\{C_k(t)\}_{k\in M}$ the relation of partial order

$$C_g(t_a) \mathrel{\in\!\!\!-} C_h(t_b) \Leftrightarrow t_a \geqslant t_b \tag{12}$$

is determined.

The relation $\mathrel{\in\!\!\!-}$ arranges the values $\{1\}^M$ of the function $\{w_{Q,k}(t)\}_{k\in M}$ as well as of $\{m_{Q,k}(t)\}_{k\in M}$, that describe one extreme of the frequency-time envelope of the analyzed sequence.

A characteristic property of the system is the memorizing each of the values "1" of the function $\{w_{r,k}(t)\} = \{w_{r,1}(t), w_{r,2}(t), \ldots, w_{r,n}(t)\}$ throughout a short length of time t_{p2}. The length of time t_{p2} is described by the inequality

$$t_{p1} < t_{p2} \leqslant \tau_{p2}, \tag{13}$$

where τ_{p2} is the longest duration of the maxima of the frequency-time envelope of the speech sound (see also the definition of the time τ_{p1}, formula (7)). In analyzing voiced sounds, we obtain the value

$$\tau_{p2} \approx \frac{1}{2\pi f_t}, \tag{14}$$

where f_t corresponds to the lowest frequency of the larynx tone.

If f_t denotes the average value of the frequency of the larynx tone, then an automatic analysis of speech pronounced with a larynx tone frequency smaller than the average will not determine the extremes. Condition (13) results from expression (1).

The duration of the value "1" of the function $w_{r,k}(t)$ is equal to

$$t_{r2} = t_{p2} + \tau_Q, \tag{15}$$

where τ_Q is the time of coincidence of the value "1" of the function $w_{Q,k}(t)$ or $m_{Q,k}(t)$ (see expressions (8) and (5)).

The times of memorizing are not added to the durations of the value "1" of the function

$$\{m_{r,k}(t, m_{Q,k}(t), Q)\} = \{m_{r,1}(t), m_{r,2}(t), \ldots, m_{r,n}(t)\}$$

which are equal to τ_Q.

8. Detection of maxima of the frequency-time envelope of a speech signal

In this paper it is assumed that at the output of the analyzer or, more strictly, at the output of each channel, a rectified signal is obtained with an instant amplitude which changes periodically, according to the frequency of larynx tone (the pulsations with a frequency of the bandpass filter are smoothed out by means of integrating systems with suitable time constants). If in a given channel a formant occurs, then this brings about an increase in the relative changes of the output signal. Generally speaking, the extremes of the signal spectral envelope are determined for each time interval t_{p2} in such a manner that the values of the characteristic functions of the increase $(\{w_{i,k}(t)\})$ and decrease $(\{m_{i,k}(t)\})$ of the signal are determined. The statement that local extremes of the spectral envelope occurred in the k-th channel means that relative changes (increase and decrease) of the signal in this channel were greater and quicker than the corresponding changes in the $(k-1)$-st or $(k+1)$-st channel. However, this assumption makes it possible to overlook the earlier "sharpening" stage of the function $\{m_{r,k}(t)\}$. In general, the characteristic function of the formant of the analyzed sequence is determined in the following manner:

$$wm_k[t, w_{r,k}(t_1), m_{r,k}(t_2)] = \begin{cases} 1 & \text{if } w_{r,k}(t_1) \mathrel{\varepsilon\!\!-} m_{r,k}(t_2), \\ 0 & \text{otherwise}. \end{cases} \tag{16}$$

An essential property of the system is the storage of each value "1" of the function $\{wm_k(t)\}$ for an instant of time (Fig. 5). The length of time t_{p3} is determined by the inequality

$$t_{p2} < t_{p3} \leqslant \tau_{p3}, \tag{17}$$

where τ_{p3} is somewhat shorter than the duration of the shortest voiced sound.

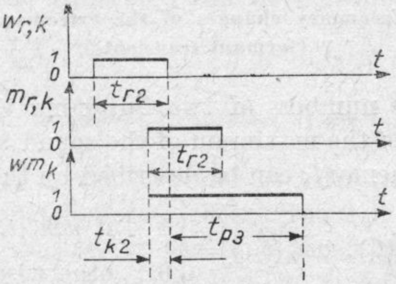

Fig. 5. Graphical interpretation of the function described by formula (16)

The duration of the value "1" of the function $wm_k(t)$ is

$$t_{r3} = t_{p3} + t_{k2}, \tag{18}$$

where t_{k2} is the time of coincidence of the value "1" of the function $w_{r,k}(t)$ (Fig. 5).

9. Determination of the position in the spectrum of formants of voiced sounds

The function of the voiced formant $\varphi_k(t)$ is described by formula

$$\varphi_k[t, wm_k(t), t_{r3}, c, t_{p3}] = \begin{cases} 1 & \text{if } t_{r3} > ct_{p3}, \\ 0 & \text{otherwise}, \end{cases} \tag{19}$$

where c is a constant, chosen experimentally. The value of the constant c should vary within a range from 1 to 3.

From equation (15) it results that in order to detect a formant of frequency f_k, the maxima of the spectral envelope have to occur in the k-th filter several times with the frequency of larynx tone (see expressions (17), (13) and (1)).

The values "1" of the function $\{\varphi_k(t)\} = \{\varphi_1(t), \varphi_2(t), \ldots, \varphi_n(t)\}$ are memorized in the system over period of time t_{p4}.

The value of t_{p4} can be described by the inequality

$$t_{p3} < t_{p4} \leqslant \tau_{p4}, \tag{20}$$

where τ_{p4} equals at least twice the duration of an average sound. Such an assumption is accepted in view of the need to ensure the coincidence of pulses when measuring formant transients (see item 11).

The duration of the value "1" of the function $\varphi_k(t)$ or of the function $a_k(t)$ is

$$t_{r4} = t_{p4} + t_{k3}, \tag{21}$$

where t_{k3} is the time of coincidence of the value "1" of the function $wm_k(t)$ or of the function $mw_k(t)$.

10. Measurements of frequency changes of the extremes of the sound envelope (formant transients)

Let g and h be the numbers of two adjoining channels of the system; $g, h \in N$. The transition of the maximum of the sound signal envelope from the frequency f_g to the frequency f_h can be described by the expression

$$M_{g,h}[t, wm_g(t_1), wm_h(t_2)] = \begin{cases} 1 & \text{if } wm_g(t_1) \leftarrowtail wm_h(t_2), \\ 0 & \text{otherwise}. \end{cases} \tag{22}$$

According to expression (22), the characteristic function $M_{g,h}(t)$ takes the value "1" when a maximum of the envelope in the channel g, and then a maximum of the envelope in the neighbouring channel h, there are identified in succession: in the time interval t_{p3}.

We notice that conditions for the increase and decrease of signals are satisfied even for quick changes of the frequency of formants. However, only the changes slower than the time constants of bandpass filters (see item 4) are detected, i.e. we must have

$$|t_1 - t_2| > \tau_k, \tag{23}$$

where τ_k is the time constant of the filter of the input system.

To satisfy inequality (23) it is necessary to use filters with suitably broad bands.

11. Detection of formant transients of voiced sounds

Let g and h be the voiced numbers of two neighbouring channels of the system $g, h \in N$. The transient formant beginning with a frequency f_g and ending up with a frequency f_h can be detected by means of the characteristic function $Y_{g,h}(t)$:

$$Y_{g,h}[t, \varphi_h(t_2), wm_g(t_1)] = \begin{cases} 1 & \text{if } \varphi_h(t_2) \leftarrowtail wm_g(t_1), \\ 0 & \text{otherwise}. \end{cases} \tag{24}$$

The formant transient, beginning with a frequency f_h and ending up with a frequency f_g, is described by the characteristic function $T_{g,h}(t)$:

$$T_{g,h}[t, wm_g(t_2), \varphi_h(t_1)] = \begin{cases} 1 & \text{if } wm_q(t_2) \leftarrowtail \varphi_h(t_1), \\ 0 & \text{otherwise}. \end{cases} \tag{25}$$

The value "1" of the characteristic functions

$$\{Y_{g,h}(t)\} = \{Y_{1,2}(t), Y_{2,3}(t), \ldots, Y_{n-1,n}(t_1)\}$$

and

$$\{T_{g,h}(t)\} = \{T_{1,2}(t), T_{2,3}(t), \ldots, T_{n-1,n}(t_1)\}$$

are memorized in the system for a short period of time t_{p4} (see inequality (20)). Current frequencies of the transients can be determined from relations (30) and (31).

12. Determination of current frequencies of the formant transients

The considerations of sections 12 and 13 refer to the case where, in the frequency range of two neighbouring filters, the k-th and $(k+1)$-st, there is at most one maximum of the spectral envelope, the energy of which considerably exceeds the energy of other maxima within the range of these two filters. We assume that the bands of the k-th and $(k+1)$-st filters overlap.

The solution presented hitherto permits determination of positions of the transients with an accuracy given by the interval $f_{k+1}-f_k$. A method of determinating current frequencies of the transients with an accuracy of $(f_{k+1}-f_k)/q$ (with the assumption of a linear distribution of threshold values $\{x_{1,k}, x_{2,k}, \ldots, x_{q,k}\}$ in the interval $[x_{1,k}, x_{q,k}]$) will now be presented.

We denote the known coefficients of the attenuation of the bandpass filter of channels k and $k+1$ by $B_k(f)$ and $B_{k+1}(f)$, respectively. For values of the signal $A(t,f)$ with a frequency f, we obtain the following relations:

$$a_k(t) = B_k(f)A(t,f), \tag{26}$$

$$a_{k+1}(t) = B_{k+1}(f)A(t,f). \tag{27}$$

Let us divide the members of equation (26) by equation (27):

$$\frac{a_k(t)}{a_{k+1}(t)} = \frac{B_k(f)}{B_{k+1}(f)} = B(f). \tag{28}$$

For the current frequency of the signal $A(t,f)$ we obtain the following expression:

$$f = B^{-1}\left[\frac{a_k(t)}{a_{k+1}(t)}\right]. \tag{29}$$

Values of the signals $\{a_k(t)\}_{k\in N}$ are also represented in the system by values of elements of the sets

$$\{w_{i,k}(t) \in \{w_{i,k}(t)\} : w_{i,k}(t) = 1\}_{k\in N}$$

and

$$\{m_{i,k}(t) \in \{m_{i,k}(t)\} : m_{i,k}(t) = 1\}_{k\in N},$$

while values of the signals $\{a_{k+1}(t)\}_{k\in N}$ are represented by values of elements of the sets

$$\big\{w_{i,k+1}(t) \in \{w_{i,k+1}(t)\} : w_{i,k+1}(t) = 1\big\}_{k\in N}$$

and

$$\big\{m_{i,k+1}(t) \in \{m_{i,k+1}(t)\} : m_{i,k+1}(t) = 1\big\}_{k\in N}.$$

The current frequency of the maximum that is shifting from the frequency f_k to the frequency f_{k+1} is determined by the following expression:

$$w_f = B^{-1}\left(\frac{\operatorname{card}\big\{m_{i,k}(t) \in \{m_{i,k}(t)\} : m_{i,k}(t) = 1\big\}}{\operatorname{card}\big\{w_{i,k+1}(t) \in \{w_{i,k+1}(t)\} : w_{i,k+1}(t) = 1\big\}}\right), \tag{30}$$

while the current frequency of the maximum that is shifted from the frequency f_{k+1} to the frequency f_k is determined by the relation

$$m_f = B^{-1}\left(\frac{\operatorname{card}\big\{w_{i,k}(t) \in \{w_{i,k}(t)\} : w_{i,k}(t) = 1\big\}}{\operatorname{card}\big\{m_{i,k+1}(t) \in \{m_{i,k+1}(t)\} : m_{i,k+1}(t) = 1\big\}}\right). \tag{31}$$

13. Determination of initial frequencies and of final transients

The frequency at which the transient begins to shift in the direction of smaller frequencies is evaluated by means of the expression

$$LM \approx B^{-1}\left(\frac{\min\limits_i \{w_{i,k}(t)\}}{\max\limits_i \{m_{i,k+1}(t)\}}\right), \tag{32a}$$

where

$$\min_i \{w_{i-1,k}(t)\} = \sim w_{i,k}(t_1) \mathrel{-\!\!\!\ni} w_{i,k}(t_2) \mathrel{-\!\!\!\ni} T_{k+1,k}(t_3), \tag{32b}$$

$$\max_i \{m_{i,k+1}(t)\} = \sim m_{i+1,k+1}(t_1) \mathrel{-\!\!\!\ni} m_{i,k+1}(t_2) \mathrel{-\!\!\!\ni} T_{k+1,k}(t). \tag{32c}$$

From (32b) it results that the threshold value $x_{i,k}$ has been exceeded by the signal a_k at the moment t_2 and the threshold value $x_{i-1,k}$ has not been exceeded by the signal a_k in the time interval $[t_2 - t_{pi}, t_2]$, and the formant transient that shifts from the filter $k+1$ towards the filter k is recorded in the time interval $[t_2, t_2 + t_{pi}]$.

The initial frequency of the transient which shifts towards higher frequencies can be determined from the relation

$$LW \approx B^{-1}\left(\frac{\max\limits_i \{m_{i,k}(t)\}}{\min\limits_i \{w_{i,k+1}(t)\}}\right), \tag{33a}$$

where

$$\max_{i}\{m_{i,k}(t)\} = \sim m_{i+1,k}(t_1) \ni m_{i,k}(t_2) \ni T_{k,k+1}(t_3), \tag{33b}$$

$$\min_{i}\{w_{i,k+1}(t)\} = \sim w_{i-1,k+1}(t_1) \ni w_{i,k+1}(t_2) \ni T_{k,k+1}(t_3). \tag{33c}$$

The frequency at which the transient shifting towards the higher frequencies finishes is determined by means of the expression

$$WL \approx B^{-1}\left(\frac{\min\limits_{i}\{m_{i,k}(t)\}}{\max\limits_{i}\{w_{i,k+1}(t)\}}\right), \tag{34a}$$

where

$$\min_{i}\{m_{i,k}(t)\} = \sim m_{i-1,k}(t_3) \in m_{i,k}(t_2) \in Y_{k,k+1}(t_1), \tag{34b}$$

$$\max_{i}\{w_{i,k+1}(t)\} = \sim w_{i+1,k+1}(t_3) \in w_{i,k+1}(t_2) \in Y_{k,k+1}(t_1). \tag{34c}$$

The frequency at which the transient shifting towards decreasing frequencies finishes can be calculated from the relation

$$ML \approx B^{-1}\left(\frac{\max\limits_{i}\{w_{i,k}(t)\}}{\min\limits_{i}\{m_{i,k+1}(t)\}}\right), \tag{35a}$$

where

$$\max_{i}\{w_{i,k}(t)\} = \sim w_{i+1,k}(t_3) \in w_{i,k}(t_2) \in Y_{k+1,k}(t_1), \tag{35b}$$

$$\min_{i}\{m_{i,k+1}(t)\} = \sim m_{i-1,k+1}(t_3) \in m_{i,k+1}(t_2) \in Y_{k+1,k}(t_1). \tag{35c}$$

14. Modifications of times of the instant memory

For the correct operation of the system, the constancy of times t_{pj} at the j-th stage of the identification for all channels is very important:

$$\underset{j}{\forall}\,\underset{k}{\forall}(t_{pj} = \text{const}). \tag{36}$$

In order to give consideration to the technical problems connected with the satisfaction of condition (36) and also to the construction of band filters with identical time constants, it is possible to modify inequality (3) to the condition

$$t_{p0} < |t_a - t_b| < t_{pj}, \tag{37}$$

where t_{p0}, chosen experimentally, permits consideration to be given to the above-mentioned difficulties.

If condition (36) is unsatisfied, then it causes that the frequency changes of the extremes are detected despite their absence in the real sound signal.

15. Advantages of the system

The use of instant memory at each stage of the identification permits analysis of sounds in the real time, and also permits automatic segmentation of sounds in the time domain. This segmentation consists in the division of the sound sequences analyzed into uneven lengths of time, with lengths not shorter than the times $\{t_{pj}\}$. The normalization of lengths of the time $\{t_{pj}\}$ up to magnitudes that correspond to the extreme values of the occurrence of certain features in the analyzed sequences permits the results of the identification of the speech to be made independent of the speeking speed, the frequency of the larynx tone, etc. The use of suitable times $\{t_{pj}\}$ at various stages of the system allows for a careful consideration of physical properties of speech sounds and the determination of basic parameters of speech signals amongst external disturbances and excess information. This selection takes place at each of the $\{j\}$ stages of the system.

In the system it is possible to use bandpass filters with comparatively broad and overlapping transmission bands. Owing to this it is possible to obtain filters with sufficiently small time constants.

The system proposed permits the determination, at any moment, of local times of the extremes of the speech signal (section 7). Their number can theoretically reach the value $n/2+1$.

The system permits a determination of parameters of the speech signal independent of the absolute stage of the speech signal (section 6), e.g. an increase of the signals $\{a_k(t)\}_{t\in N}$ from various initial values, and a decrease of these signals to various final values.

The pulsations of signals with the frequency of the larynx tone at the output of a set of filters are utilized in the system to determine the feature of the voiced ability of the analyzed signals (sections 9 and 11). The effects of phase incompatibility at the outputs of the filters are eliminated by the application of an instant memory (memory times $\{t_{pj}\}$).

Initial and final frequencies of the transients (section 13) should provide the basis for hypotheses about the position of locuses through the system for the identification of sounds (block (c) in Fig. 1).

The description of the system by means of characteristic functions of the features of sounds permits the system to be designed and built directly with the aid of integrated circuits [6].

The individual definition of the function of increase and decrease permits the automatic determination of directions of the course of transients in the frequency domain, the initial or final routes of transients, the detection of periodicity (soundability) and also a description of duration of these features of speech sounds.

In the case of an absence of periodicity the values of the function $\varphi_k(t)$ are equal to zero for each channel.

Acknowledgments. I am sincerely indebted to Dr R. Gubrynowicz and M. Sobolewski, M. Sc. for precious hints and assistance rendered in the elaboration of this concept.

References

[1] J. L. Flanagan, *Speech analysis, synthesis and perception*, Springer-Verlag, Berlin 1971.

[2] R. Jacobson, C. Fant, M. Halle, *Preliminaries to speech analysis. The distinctive features and their correlates*, MIT Press, Cambridge 1964.

[3] J. L. Kulikowski, *Cybernetic identification systems*, PWN, Warszawa 1972 [in Polish].

[4] H. Kubzdela, *Automatic extraction of the frequency of larynx tone as well as of first formants of speech signal*, IFTR Reports, PAN, Warszawa 1973.

[5] M. A. Sapożkov, *Speech signal in telecommunications and cybernetics*, PWN, Warszawa 1965 [in Polish].

[6] Z. M. Wójcik, *Conception of instant memory in the identification of sounds*, Works by IBIB-PAN, Warszawa 1975 [in Polish].

Received on 5th November 1976