POLSKA AKADEMIA NAUK
KOMITET ELEKTRONIKI I TELEKOMUNIKACJI

KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJI

ELECTRONICS AND TELECOMMUNICATIONS QUARTERLY

TOM 46 — ZESZYT 4

WYDAWNICTWO NAUKOWE PWN
WARSZAWA 2000

RADA REDAKCYJNA

Przewodniczący
Prof. dr hab. inż. STEFAN HAHN
czł. koresp. PAN

Członkowie
prof. dr hab. inż. DANIEL JÓZEF BEM — czł. koresp. PAN, prof. dr hab. inż. MICHAŁ BIAŁKO — czł. rzecz. PAN, prof. dr hab. inż. JAN CHOJCAN, prof. dr hab. inż. MAREK DOMAŃSKI, prof. dr hab. inż. ANDRZEJ HAŁAS, prof. dr inż. WŁADYSŁAW MAJEWSKI, prof. dr hab. inż. JÓZEF MODELSKI, prof. dr inż. JERZY OSIOWSKI, prof. dr hab. inż. EDWARD SĘDEK, prof. dr hab. inż. MICHAŁ TADEUSIEWICZ, prof. dr hab. inż. WIESŁAW WOLIŃSKI — czł. koresp. PAN, prof. dr inż. MARIAN ZIENTALSKI

REDAKCJA
Redaktor Naczelny
prof. dr hab. inż. WIESŁAW WOLIŃSKI

Zastępca Redaktora Naczelnego
doc. dr inż. KRYSTYN PLEWKO

Sekretarz Odpowiedzialny
mgr ELŻBIETA SZCZEPANIAK

ADRES REDAKCJI
00-665 Warszawa, ul. Nowowiejska 15/19 Politechnika, pok. 470
Instytut Telekomunikacji, Gmach im. prof. JANUSZA GROSZKOWSKIEGO

Dyżury Redakcji: środy i piątki, godz. 14–16
tel/fax (022) 660 77 37

Telefony domowe: Redaktora Naczelnego: 812 17 65
Zast. Red. Naczelnego: 826 83 41
Sekretarza Odpowiedzialnego: 633 92 52

WYDAWNICTWO NAUKOWE PWN SA
Warszawa, ul. Miodowa 10

Ark. wyd. 17 Ark. druk. 13,00	Podpisano do druku w styczniu 2001 r.
Papier offset. kl. III 80 g. B-1	Druk ukończono w lutym 2001 r.

SKŁAD I ŁAMANIE: PHOTOTEXT, Warszawa, ul. Chmielna 120
DRUK I OPRAWA: Drukarnia Braci Grodzickich, Piaseczno, ul. Geodetów 47a

ŁKO — czł.
dr hab. inż.
SKI, prof. dr
USIEWICZ,
TALSKI

Szanowni Autorzy

„Kwartalnik Elektroniki i Telekomunikacji" — Electronics and Telecommunications Quartery jest kontynuatorem tradycji powstałego 46 lat temu kwartalnika p.t. „Rozprawy Elektrotechniczne".

Kwartalnik jest czasopismem Komitetu Elektroniki i Telekomunikacji Polskiej Akademii Nauk. Wydawany jest przez Wydawnictwo Naukowe PWN SA w Warszawie. Kwartalnik jest czasopismem naukowym, na którego łamach są publikowane artykuły i komunikaty prezentujące wyniki oryginalnych prac teoretycznych i doświadczalnych, a także przeglądowych. Związane są one z szeroko rozumianymi dziedzinami współczesnej elektroniki, telekomunikacji, mikroelektroniki, optoelektroniki, radiotechniki i elektroniki medycznej.

Autorami publikacji są wybitni naukowcy, znani specjaliści o wieloletnim doświadczeniu, a także młodzi badacze — głównie doktoranci.

Artykuły charakteryzują się oryginalnym ujęciem zagadnienia, interesującymi wynikami badań, krytyczną oceną teorii lub metod, omówieniem aktualnego stanu, lub postępu danej gałęzi techniki oraz omówieniem perspektyw rozwojowych. Sposób pisania matematycznej części artykułów zgodny jest z wytycznymi IEC (International Electronics Commision) oraz ISO (International Organization of Standarization).

Wszystkie publikowane w Kwartalniku artykuły są recenzowane przez znanych krajowych specjalitów, co zapewnia że publikacje te są uznawane jako autorski dorobek naukowy. Opublikowanie w kwartalniku wyników prac naukowych zrealizowanych w ramach „GRANTów" Komitetu Badań Naukowych spełnia więc jeden z wymogów stawianych tym pracom.

Czasopismo dociera do wszystkich zajmujących się elektroniką i telekomunikacją krajowych ośrodków naukowych oraz technicznych, a także szeregu instytucji zagranicznych. Jest ponadto prenumerowane przez liczne grono specjalistów i biblioteki.

Każdy Autor otrzymuje bezpłatnie 20 egzemplarzy nadbitek swojego artykułu, co ułatwia przesłanie go do indywidualnych wybranych przez Autora osób i instytucji w kraju lub za granicą. Ułatwia to dodatkowo fakt, że w Kwartalniku są publikowane artykuły w języku angielskim.

Nadesłane do redakcji artykuły są publikowane w terminie około pół roku, w przypadku sprawnej współpracy Autora z Redakcją. Wytyczne dla Autorów dotyczące formy publikacji są zamieszczone w zeszytach Kwartalnika, można je także otrzymać w siedzibie Redakcji.

Artykuły można dostarczać osobiście, lub pocztą pod adresem zamieszczanym na stronie redakcyjnej w każdym zeszycie.

Redakcja

KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJI, 2000, 46, z. 4, ss. 461–477

Analysis of stability of multiple-feedback Delta-Sigma modulators with leakage

KRZYSZTOF OLEJARCZYK

Instytut Systemów Elektronicznych, Politechnika Warszawska

Otrzymano 2000.03.23
Autoryzowano 2000.09.14

In this paper a new method of proving stability of Delta-Sigma modulators in the way of bounding outputs of integrators is presented. The proposed approach is theoretical as well as computational. The method concerns multiple-feedback modulator topologies with single quantiser (and can be used for higher-order systems also). Input signal is assumed to be constant. The method has been implemented as a computer program. Exhaustive test have been made for second-order modulators, and presented in the paper.

Key words: sigma-delta, noise-shaping, chaos, nonlinear analysis.

1. INTRODUCTION

In recent years $\Delta\Sigma$ modulation has become very popular. Although $\Delta\Sigma$ modulators are successfully designed and manufactured, there is a lack of solid theoretical foundation. Methods of design are based on empiricial knowledge obtained mainly in the way of great amount of simulation. Stability, which is the basic parameter of modulators, can be obtained only by simulation in higher-order case. Obviously, even exhaustive simulation cannot prove in strict sense that a modulator is stable. It follows that it is impossible to find the optimal design of $\Delta\Sigma$ modulator in certain class of solutions, in the strict sense.

Due to the fact that $\Delta\Sigma$ modulator contains highly nonlinear element, the quantiser, in a feedback loop, the analysis of the behaviour of $\Delta\Sigma$ modulators is extremely difficult. Only some topologies have been analysed exactly [1]. First-order modulator has been described in many works, for example in [2]. Standard second-order modulators are the subject of theoretical work of Wang [3] as well as Hein and Zakhor [4]. The last work yields the method of determining convex trapping region with parabolic edges in the state space [4]. A computational approach to determining convex trapping region

is presented by Schreier *et al.* [5]. This method produces tighter bounds then those of Hein and Zakhor [4]. The trapping region is a convex polygon (in the 2-D case; in 3-D case the region is a convex polyhedron). The advantage of this approach is that the method can be applied to higher-order modulators. The fully theoretical approach of Farrell and Feely [6] takes account of the jagged form of the outer boundary of the trapping region (which is no longer constrained to be convex). This method yields better results for standard second-order modulators then those of Schreier *et al.* [5].

The presented method is based on the similar idea as the approach of Farrell and Feely [6]. The method of Farrell and Feely is dedicated to standard second-order ΔΣ modulators (without any local feedback loops), and there was a need to find out more general method that can be applied to wider class of modulators. In this paper the new method will be described with an example of multiple-feedback second-order ΔΣ modulator (with single-bit quantiser) shown in Figure 1. The presented approach can be applied to higher-order modulators also.

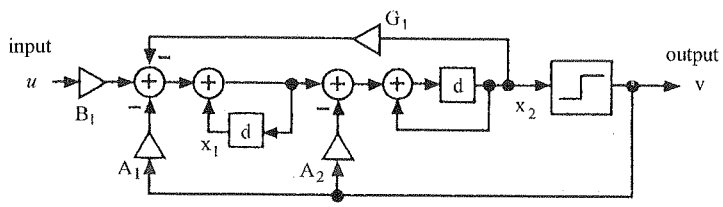


Fig. 1. Multiple-feedback second-order ΔΣ modulator with single-bit quantiser; d - delay.

Finding a positively invariant set (PIS) in state space of a ΔΣ modulator is the simplest way of proving (in the strict sense) that the modulator is stable. The bounds of the PIS give us the information about stability of the modulator. The presented method enables calculation these bounds for certain class of modulators. The starting point of the method is the observation that the sequence of samples of the output signal of the modulator (denoted $v_{(n)}$, has certain important property: there is finite (and relatively small) number of possible subsequences of arbitrary lenght. The best introduction to the method will be an example, described in the next section.

2. EXAMPLE OF MULTIPLE-FEEDBACK SECOND-ORDER ΔΣ MODULATOR

The modulator is assumed to be single-bit in this paper. Consider the signal at the quantiser output. The signal is a sequence of +1s and −1s. (There is a convention that the quantiser outputs are normalised). Let all consecutive +1s and −1s be grouped and denoted as N^+ and N^- (these symbols will be called *numbers of positive and negative iterations*, respectively, or shortly: *iteration numbers*). For instance: (−1, +1, +1, +1, +1, +1, +1, +1) = (1^-, 7^+). This notation means also, that the modulator had to have at least one positive iteration before the sequence (1^-, 7^+), and there is at least one negative iteration after it.

In this paper there is assumption that the input signal u is constant. The modulator model is shown in Figure 1 (it is one of the most common topologies of second-order $\Delta\Sigma$ modulators with single-bit quantiser). Let the input signal be equal $u = 0.696$, and coefficients of the modulator model equal as follows: $A_1 = 0.215$, $A_2 = 0.559$, $B_1 = 0.215$, $G_1 = 320356/1603602025 \approx 0.00019977$.

Long simulations of the model has shown that only few sequences of numbers of positive and negative iterations may appear. Two-element sequences obtained during a simulation are listed in Table 1. (The initial state of modulator equalled $\mathbf{x}(0) = (0, 0)^T$, simulation consisted from 60,000 time steps, but 10,000 initial time steps were discarded). Table 2 shows four-element sequences for comparison. One can notice that not all concatenations of two-element sequences from Table 1 appear in Table 2. It suggests that there exist rules of succession of iteration numbers.

Table 1
Two-element sequences (N_1, N_2) of consecutive negative and positive iterations

N_1^-	N_2^+	How many times the sequence (N_1^-, N_2^+) appeared
1	4	697
1	5	2929
1	6	2947
1	7	1038

Table 2
Four-element sequences $(N_2^-, N_2^+, N_3^-, N_4^+)$ of consecutive iterations

N_1^-	N_2^+	N_3^-	N_4^+	How many times the sequence $(N_1^-, N_2^+, N_3^-, N_4^-)$ appeared
1	4	1	6	105
1	4	1	7	592
1	5	1	5	638
1	5	1	6	1845
1	5	1	7	446
1	6	1	4	178
1	6	1	5	1782
1	6	1	6	987
1	7	1	4	520
1	7	1	5	509
1	7	1	6	9

The modulator model (Figure 1) can be described with following difference equation (given in matrix-vector notation):

$$\mathbf{x}(n+1) = f(\mathbf{x}(n)) = \begin{bmatrix} 1 & -G_1 \\ 1 & 1-G_1 \end{bmatrix} \cdot \mathbf{x}(n) + \begin{bmatrix} B_1 \cdot u \\ B_1 \cdot u \end{bmatrix} + \begin{bmatrix} -A_1 \\ -A_1 - A_2 \end{bmatrix} \cdot \operatorname{sgn}(x_2(n)), \qquad (1)$$

where $\operatorname{sgn}(\cdot) = \begin{cases} 1 & \text{for argument} \geq 0 \\ -1 & \text{for argument} < 0 \end{cases}$, and $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$.

Let Γ_P and Γ_N be the positive and negative half-planes: $\Gamma_P = \{\mathbf{x}: x_2 \geq 0\}$, $\Gamma_N = \{\mathbf{x}: x_2 < 0\}$. It can be seen that the mapping f is piecewise-affine, and there exist such two affine mappings

$f_P: \mathbf{R}^2 \to \mathbf{R}^2$ and $f_N: \mathbf{R}^2 \to \mathbf{R}^2$, that

$$f(\mathbf{x}) = \begin{cases} f_P(\mathbf{x}) & \text{for } \mathbf{x} \in \Gamma_P \\ f_N(\mathbf{x}) & \text{for } \mathbf{x} \in \Gamma_N \end{cases}. \qquad (2)$$

The symbol $\mathbf{R}$ denotes real numbers set.

Consider two-element sequences of numbers of iterations. One of the sequences which have appeared during the simulation is $(1^-, 7^+)$. It will be shown, that after the modulator has performed the sequence $(1^-, 7^+)$, its state has to enter the quadrangle that is shown in Figure 2C and D (Figure 2D shows the zoomed-in central part of Figure 2C). The set α shown in Figure 2, is defined as follows: $\alpha = \{\mathbf{x}: x_2 = 0\}$. In other words α is simply the horizontal axis. Before the sequence $(1^-, 7^+)$, the modulator state had to be inside the positive half-plane. After the negative iteration (the first iteration of the sequence), the trajectory must enter the sector $\Gamma_N \cap f_P(\Gamma_P)$, shown in Figure 2A (as the darkest region). Next, after the first positive iteration the modulator state enters the set $\Gamma_P \cap f_N(\Gamma_N) \cap f_N \circ f_P(\Gamma_P)$, that is shown in Figure 2B (as a triangle and the darkest region). Finally, when the trajectory has performed the sequence $(1^-, 7^+)$ and additionally one negative iteration, the modulator state has to enter the set:

$$\eta_1 = \Gamma_N \cap f_P(\Gamma_P) \cap f_P^7 \circ f_N(\Gamma_N) \cap f_P^7 \circ f_N \circ f_P(\Gamma_P), \qquad (3)$$

The set η_1 can be seen in Figure 2C and D (actually, the set η_1 is a quadrangle). (The symbol f^n denotes the n-fold composition of f with itself). It is obvious that if a trajectory with an arbitrary initial conditions contains the sequence $(1^-, 7^+)$, then the set η_1 is not empty. In this case it is true, due to the fact, that the sequence was obtained as a result of simulation (in this case, rational arithmetic was employed in simulation, so rouding errors couldn't appear). The set η_1 obtained as an intersection of a several half-planes is convex. The way of determining regions generated by sequences of the form (N_1^-, N_2^+) will be formulated for the general case in the next section.

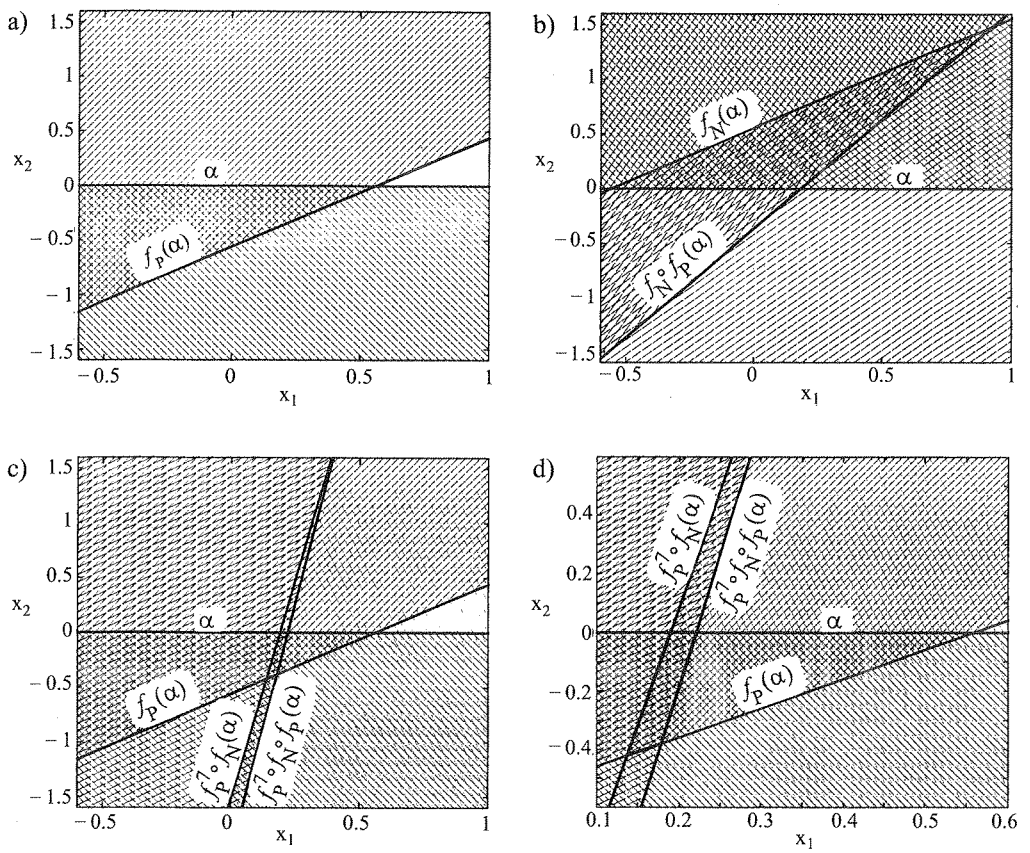


Fig. 2. Determining a region in state space corresponding to the sequence of iterations $(1^-, 7^+)$

Consider the inverse implication: what sequences of the form (N_1^-, N_2^+) may be generated, if trajectories start from the set η_1? It is easier to answer this question, if the set η_1 is close. The closure of η_1 will be denoted $\overline{\eta}_1$. Obviously, $\eta_1 \subset \overline{\eta}_1$. All the possible sequences of iterations can be found in the way of simulation. The polygon $\overline{\eta}_1$ is iterated with the mapping f. Because $\overline{\eta}_1$ lays in the negative half-plane, the whole polygon $\overline{\eta}_1$ will be transformed by affine mapping f_N (Figure 3A). Affine mapping preserves structure of a set (straight lines maps to straight lines and the interior or a polygon maps to the interior of the polygon's image), so it is sufficient to map only the tops, in order to obtain the image of the polygon. During the iteration of the mapping, a polygon may cross the splitting line (*i.e.* α). In such a case the polygon is cut into pieces and the pieces are mapped further individually (Figure 3). In this way new tops appear, *i.e.* the split points (points where the line α intersect polygon edges). The iteration finishes when every pieces of the polygon $\overline{\eta}_1$ have crossed the splitting line α twice (because two-element sequences of iterations are considered).

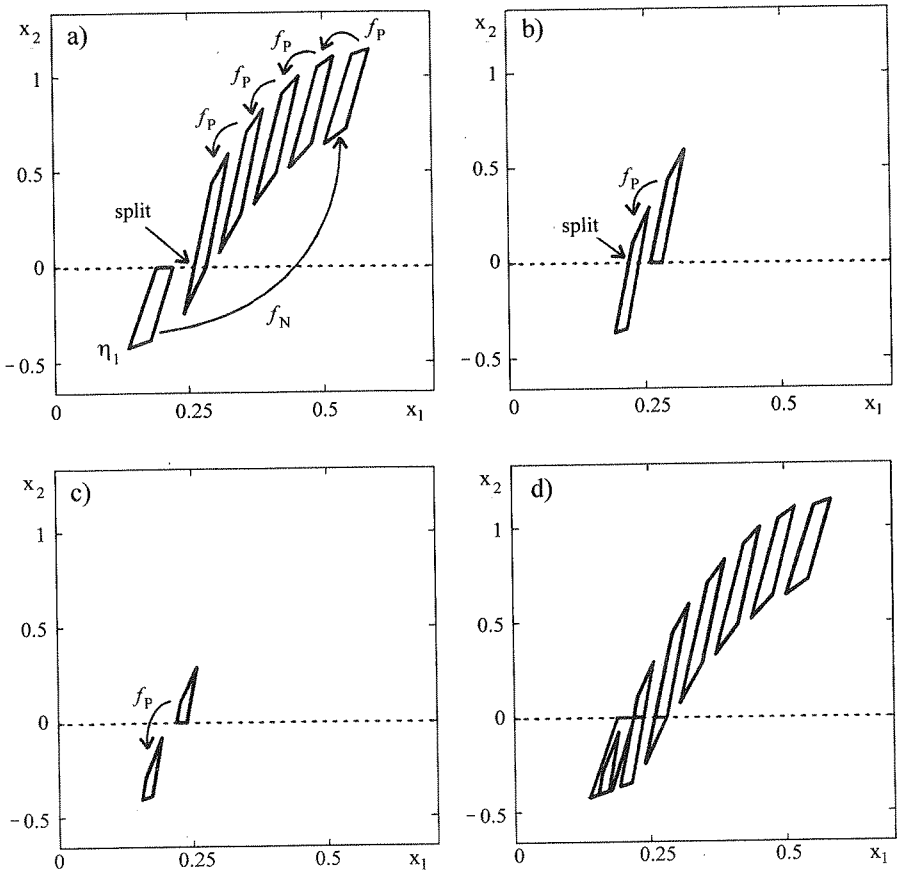


Fig. 3. Polygons obtained during the process of determining all possible sequences of iterations which correspond to the polygon η_1. A) The polygon η_1 and its images before the first split. Obtained result: $(1^-, 4^+)$. B) The new polygon and its image. Obtained result: $(1^-, 5^+)$. C) The second new polygon and its image. Obtained result: $(1^-, 6^+)$. D) All polygons obtained during the process

An algorithm that derives automatically implications like this described above has been originally developed by the author. The algorithm is implemented in C++ language and uses rational arithmetic (in order to avoid rounding errors). For the case mentioned above following sequences have been obtained: $(1^-, 4^+)$, $(1^-, 5^+)$ and $(1^-, 6^+)$. It means, that after the sequence of iterations $(1^-, 7^+)$, one of these three above-mentioned sequences must follow. It doesn't mean however that *all* sequences obtained as a result concatenation *i.e.* $(1^-, 7^+, 1^-, 4^+)$, $(1^-, 7^+, 1^-, 5^+)$ and $(1^-, 7^+, 1^-, 6^+)$ must be acceptable, it means only that a sequence which follows the sequence $(1^-, 7^+)$ has to be a member of the set $\{(1^-, 4^+), (1^-, 5^+), (1^-, 6^+)\}$. Such a relationship will be called *sequence relationship*, and will be denoted symbolically as follows:

$$(1^-, 7^+) \rightarrow \{((1^-, 4^+), (1^-, 5^+), (1^-, 6^+)\}.$$

It would be useful to find a set of sequences which may follow each new (just obtained) sequence, *i.e.* $(1^-, 4^+)$, $(1^-, 5^+)$ and $(1^-, 6^+)$. For the case mentioned above following sequence relationships have been obtained:

$(1^-, 7^+) \rightarrow \{(1^-, 6^+), (1^-, 4^+), (1^-, 5^+)\}$,

$(1^-, 6^+) \rightarrow \{(1^-, 6^+), (1^-, 4^+), (1^-, 5^+), (1^-, 7^+)\}$,

$(1^-, 4^+) \rightarrow \{(1^-, 6^+), (1^-, 4^+), (1^-, 5^+), (1^-, 7^+), (1^-, 3^+)\}$,

$(1^-, 5^+) \rightarrow \{(1^-, 6^+), (1^-, 4^+), (1^-, 5^+), (1^-, 7^+)\}$,

$(1^-, 3^+) \rightarrow \{(1^-, 6^+), (1^-, 4^+), (1^-, 5^+), (1^-, 7^+), (1^-, 3^+), (1^-, 8^+)\}$,

$(1^-, 8^+) \rightarrow \{(1^-, 6^+), (1^-, 4^+), (1^-, 5^+)\}$.

All polygons obtained during the process of determining the above-mentioned sequence relationships are shown in Figure 4. It can be noticed, that the sum of these polygons is an positively invariant set.

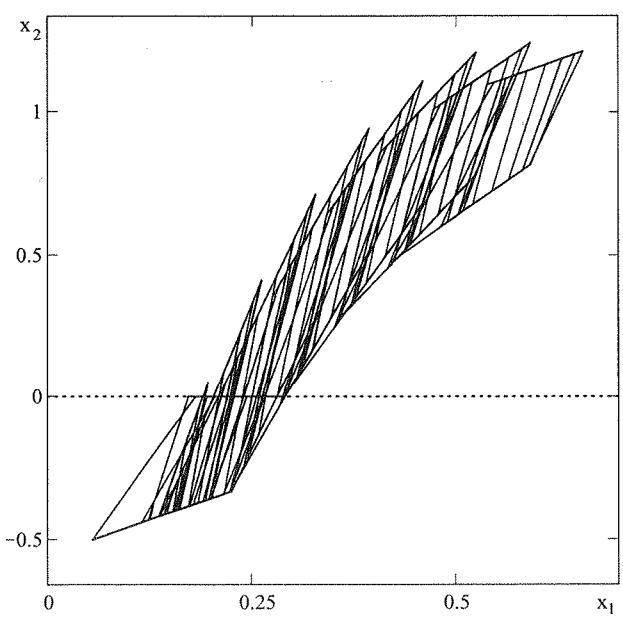


Fig. 4. All polygons obtained during the process of determining a trapping set of sequences of iterations. The sum of the polygons yields an positively invariant set in state spece.

Let define a set of sequences:

$$\Omega = \{(1^-, 3^+), (1^-, 4^+), (1^-, 5^+), (1^-, 6^+), (1^-, 7^+), (1^-, 8^+)\}. \tag{4}$$

Consider sequence of iterations $(N_1^-, N_2^+, N_3^-, N_4^+)$. It can be seen that if $(N_1^-, N_2^+) \in \Omega$, than $(N_3^-, N_4^+) \in \Omega$. It follows that the set Ω is a trapping set of sequences. Obviously, if Ω is finite, and its elements contain finite numbers of iterations, then the modulator is stable (for the given value of the constant input signal). It follows that stability of the modulator is proven for the above-mentioned example.

3. GENERAL FORMULA OF ALGORITHM

In this section general formula of the relationship between sequence of numbers of iterations and a region in state space will be presented. The algorithm has been originally developed by the author. Consider a D-order modulator, $D \geq 2$, which can be described with a mapping f: $\mathbf{R}^D \to \mathbf{R}^D$. Let define Γ_P and Γ_N as follows: $\Gamma_P = \{\mathbf{x}: x_D \geq 0\}$, $\Gamma_N = \{\mathbf{x}: x_D < 0\}$. There is an assumption that f is piecewise-affine, so that it can be expressed as:

$$f(\mathbf{x}) = \begin{cases} f_P(\mathbf{x}) & \text{for } \mathbf{x} \in \Gamma_P \\ f_N(\mathbf{x}) & \text{for } \mathbf{x} \in \Gamma_N \end{cases}. \tag{5}$$

where the mappings f_P and f_N are affine. The majority of commonly used $\Delta\Sigma$ modulators with single-bit quantiser satisfies these above-mentioned assumptions.

Let S be the initial state of the quantiser, $S \in \{-1, +1\}$. Without loss of generality it can be assumed that the zero time instant corresponds to the change from the quantiser state $-S$ to S, that is the initial quantiser outputs form the following sequence:

$$v_0, v_1, \ldots, v_{\xi_2 + \ldots + \xi_M - 1} = \underbrace{S, \ldots S}_{\xi_1}, \underbrace{-S, \ldots -S}_{\xi_2}, \underbrace{S, \ldots S}_{\xi_3}, \ldots, \underbrace{-S, \ldots -S}_{\xi_M}, \tag{6}$$

where M is even, $v_{-1} = -S$, $v_M = S$. The sequence of quantiser outputs (6) corresponds to the M-element sequence of iterations $(\xi_1, \xi_2, \ldots \xi_M)$. (In this notation of sequence of iterations super-script-signs have been omitted, because they depend on S). Let define a set η in the state space, in the following way:

$$\mu_1 = \begin{cases} f_N(\Gamma_N) \cap f_N^{\xi_1} \circ f_P(\Gamma_P) & \text{for } S = -1 \\ f_P(\Gamma_P) \cap f_P^{\xi_1} \circ f_N(\Gamma_N) & \text{for } S = +1 \end{cases} \tag{7}$$

$$\mu_{i+1} = \begin{cases} f_P^{\xi_{i+1}}(\mu_i) \cap f_P(\Gamma_P) & \text{if } (S = -1 \wedge \text{i-odd}) \vee (S = +1 \wedge \text{i-even}) \\ f_N^{\xi_{i+1}}(\mu_i) \cap f_N(\Gamma_N) & \text{if } (S = -1 \wedge \text{i-even}) \vee (S = +1 \wedge \text{i-odd}) \end{cases} \tag{8}$$

for $i = 1,\ldots M-1$.

$$\eta = \begin{cases} \mu_M \cap \Gamma_P & \text{if } (S=-1 \wedge \text{ M-odd}) \vee (S=+1 \wedge \text{ M-even}) \\ \mu_M \cap \Gamma_N & \text{if } (S=-1 \wedge \text{ M-even}) \vee (S=+1 \wedge \text{ M-odd}) \end{cases} \tag{9}$$

It can be seen that the modulator state after the above-mentioned M-element sequence of iterations, $\mathbf{x}(\xi_1+\xi_2+\ldots+\xi_M)$, belongs to the set η. In order to apply the method of determining a region in state space corresponding to a sequence of iterations, a closed-form expressions for f_P^n (the n-fold composition of f_P with itself) and f_N^n, for any $n \geq 0$, are needed. For multiple-feedback higher-order modulators (with single quantiser) such expressions can be obtained in the way described in [7]. For example following formulas correspond to the difference equation (1) (as well as the model shown in Figure 1):

$$f_{P,N}^n(\mathbf{x}) = \mathbf{M} \cdot \left(\begin{bmatrix} \cos(n \cdot \alpha) & -\sin(n \cdot \alpha) \\ \sin(n \cdot \alpha) & \cos(n \cdot \alpha) \end{bmatrix} \cdot (\mathbf{M}^{-1} \cdot \mathbf{x} + \mathbf{R}_{P,N}) - \mathbf{R}_{P,N} \right), \tag{10}$$

where

$$\alpha = \text{a sin}\left(-\tfrac{1}{2} \cdot \sqrt{G_1 \cdot (4-G_1)}\right), \tag{11}$$

$$\mathbf{M} = \begin{bmatrix} \frac{\sqrt{2}}{2} \cdot G_1 & \frac{\sqrt{2}}{2} \cdot \sqrt{G_1 \cdot (4-G_1)} \\ \sqrt{2} & 0 \end{bmatrix}, \tag{12}$$

$$\mathbf{R}_P = \begin{bmatrix} \frac{-\sqrt{2}}{2} \cdot \frac{B_1 \cdot u - A_1}{G_1} \\ \frac{\sqrt{2}}{2} \cdot \frac{B_1 \cdot u - A_1 - 2 \cdot A_2}{\sqrt{G_1 \cdot (4-G_1)}} \end{bmatrix}, \tag{13}$$

$$\mathbf{R}_N = \begin{bmatrix} \frac{-\sqrt{2}}{2} \cdot \frac{B_1 \cdot u + A_1}{G_1} \\ \frac{\sqrt{2}}{2} \cdot \frac{B_1 \cdot u + A_1 + 2 \cdot A_2}{\sqrt{G_1 \cdot (4-G_1)}} \end{bmatrix}. \tag{14}$$

Notation $\mathbf{M}^{-1}$ means the inverse matrix of $\mathbf{M}$. It can be seen that the trajectory of the modulator corresponds to a circle movement in certain basis. Examples of formulas corresponding to other topologies can be found in [7].

In the general case determining all possible sequences of iterations which correspond to a convex polytope (the D-dimensional analogue of 2-dimensional polygon and 3-dimensional polyhedron) is very similar to the exemplary 2-dimensional case described above: it is sufficient to calculate images of the tops to obtain an image of the polytope. In D-dimensional state space the splitting line α (defined above for 2-dimensional case) has to be redefined as a hyperplane: $\alpha = \{\mathbf{x}: x_D = 0\}$. When a

D-dimensional convex polytope crosses the hyperplane α, many new tops may appear (they would be points where the hyperplane α intersect the polytope edges). Obviously, calculating these split points is more difficult in the higher-order case.

4. LEAKY SYSTEMS

One of typical analogue nonidealities is integrator leakage. There is an assumption that the leakage of all integrators is equal (for the sake of simplicity) and is described with a parameter L. Introducing the leakage modifies the circuit shown in Figure 1 in the way shown in Figure 5. The leaky system can be described with the following difference equation:

$$\mathbf{x}(n+1) = f_{\text{Leaky}}(\mathbf{x}(n)) = \mathbf{L}\cdot\left(\begin{bmatrix}1 & -G\\ 1 & 1-G\end{bmatrix}\cdot\mathbf{x}(n)+\begin{bmatrix}B_1\cdot u\\ B_1\cdot u\end{bmatrix}+\begin{bmatrix}-A_1\\ -A_1-A_2\end{bmatrix}\cdot\operatorname{sgn}(x_2(n))\right). \tag{15}$$

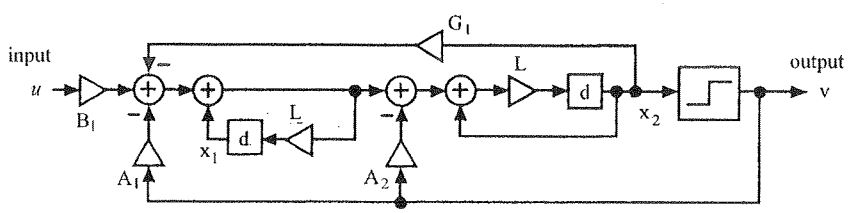


Fig. 5. Leaky multiple-feedback second-order $\Delta\Sigma$ modulator with single-bit quantiser; d - delay. (Parameter L equals 1 for the ideal system. The value slightly less than 1 corresponds to the „natural'' leakage, and the value slightly greater than 1 appears when the leakage is over-compensated.)

Comparison of equations (1) and (15) shows that the leakage introduces slight difference into the equation but it can change rapidly the behaviour of the modulator. Changes of the parameter L may destabilise a stable system or, on the other hand, stabilise an unstable one.

Trajectories of the leaky systems belong to the same class as the ideal ones, *i.e.* the trajectories can be expressed as a movement along spirals or ellipses. The close-form equations for leaky systems can be expressed as follows:

$$f^{n}_{P,N}(\mathbf{x}) = \mathbf{M}\cdot\left(\mathbf{L}^{n}\cdot\begin{bmatrix}\cos(n\cdot\alpha) & -\sin(n\cdot\alpha)\\ \sin(n\cdot\alpha) & \cos(n\cdot\alpha)\end{bmatrix}\cdot\mathbf{M}^{-1}\cdot(\mathbf{x}-\mathbf{x}^{*}_{P,N})+\mathbf{M}^{-1}\cdot\mathbf{x}^{*}_{P,N}\right) \tag{16}$$

where α and $\mathbf{M}$ are expressed by equations (11) and (12) respectively, $\mathbf{x}^{*}_{P}$ is a fix point of the map f_P, *i.e.* $\mathbf{x}^{*}_{P}$ is a solution of the following equation:

$$\mathbf{x}_P^* = L \cdot \left(\begin{bmatrix} 1 & -G_1 \\ 1 & 1-G_1 \end{bmatrix} \cdot \mathbf{x}_P^* + \begin{bmatrix} B_1 \cdot u \\ B_1 \cdot u \end{bmatrix} + \begin{bmatrix} -A_1 \\ -A_1 - A_2 \end{bmatrix} \right) \tag{17}$$

and $\mathbf{x}_N^*$ is a fix point of the map f_N, and satisfies the following equation:

$$\mathbf{x}_N^* = L \cdot \left(\begin{bmatrix} 1 & -G_1 \\ 1 & 1-G_1 \end{bmatrix} \cdot \mathbf{x}_N^* + \begin{bmatrix} B_1 \cdot u \\ B_1 \cdot u \end{bmatrix} - \begin{bmatrix} -A_1 \\ -A_1 - A_2 \end{bmatrix} \right). \tag{18}$$

The equation (16) has been derived in the analogous way as the equation (10) [7]. The derivation is based on the fact that:

$$\left(L \cdot \begin{bmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{bmatrix} \right)^n \equiv L^n \cdot \begin{bmatrix} \cos(n \cdot \alpha) & -\sin(n \cdot \alpha) \\ \sin(n \cdot \alpha) & \cos(n \cdot \alpha) \end{bmatrix} \text{ for any integer n.} \tag{19}$$

The final formula for fix points is expressed as follows:

$$\mathbf{x}_{P,N}^* = \frac{\mathbf{1}}{(\mathbf{L}^2 - \mathbf{2} \cdot \mathbf{L} + \mathbf{L} \cdot \mathbf{G}_1 + \mathbf{1})} \cdot \begin{bmatrix} (-B_1 \cdot u \pm A_1 \pm G_1 \cdot A_2) \cdot L^2 + (B_1 \cdot u \mp A_1) \cdot L \\ \pm L^2 \cdot A_2 + (B_1 \cdot u \mp A_1 \mp A_2) \cdot L \end{bmatrix}. \tag{20}$$

5. RESULTS OF TESTS OF ALGORITHM

Results of tests of the same exemplary modulator as this described in section 2 will be shown in this section. The presented algorithm was tested for different values of the lenght of sequence of iterations M, and for different values of constant input signal u. For M = 4 and u = 0.696, following sequence relationships have been obtained:

$\chi_1 \to \{\chi_1, \chi_2, \chi_3, \chi_4, \chi_5, \chi_6, \chi_7\}$, $\chi_2 \to \{\chi_1, \chi_2, \chi_3, \chi_5, \chi_8\}$, $\chi_3 \to \{\chi_1, \chi_4, \chi_6, \chi_7, \chi_9, \chi_{10}\}$,

$\chi_4 \to \{\chi_1, \chi_2, \chi_3, \chi_8\}$, $\chi_5 \to \{\chi_7, \chi_9, \chi_{10}, \chi_{11}\}$, $\chi_6 \to \{\chi_1, \chi_4, \chi_6, \chi_7, \chi_9, \chi_{10}\}$,

$\chi_7 \to \{\chi_1, \chi_2, \chi_3, \chi_4, \chi_5, \chi_6\}$, $\chi_8 \to \{\chi_1, \chi_2, \chi_3, \chi_4, \chi_5, \chi_6, \chi_7\}$, $\chi_9 \to \{\chi_7, \chi_9, \chi_{10}, \chi_{11}\}$,

$\chi_{10} \to \{\chi_1, \chi_4, \chi_5, \chi_6, \chi_7, \chi_9\}$, $\chi_{11} \to \{\chi_1, \chi_2, \chi_3, \chi_5\}$, where $\chi_1 = (1^-, 5^+, 1^-, 6^+)$,

$\chi_2 = (1^-, 4^+, 1^-, 7^+)$, $\chi_3 = (1^-, 5^+, 1^-, 5^+)$, $\chi_4 = (1^-, 5^+, 1^-, 7^+)$, $\chi_5 = (1^-, 6^+, 1^-, 4^+)$,

$\chi_6 = (1^-, 6^+, 1^-, 5^+)$, $\chi_7 = (1^-, 6^+, 1^-, 6^+)$, $\chi_8 = (1^-, 4^+, 1^-, 6^+)$, $\chi_9 = (1^-, 7^+, 1^-, 4^+)$,

$\chi_{10} = (1^-, 7^+, 1^-, 5^+)$, $\chi_{11} = (1^-, 7^+, 1^-, 6^+)$.

All polygons obtained during the process of derivation the above-mentioned relationships are shown in Figure 6. The sum of these polygons is a positively invariant set in

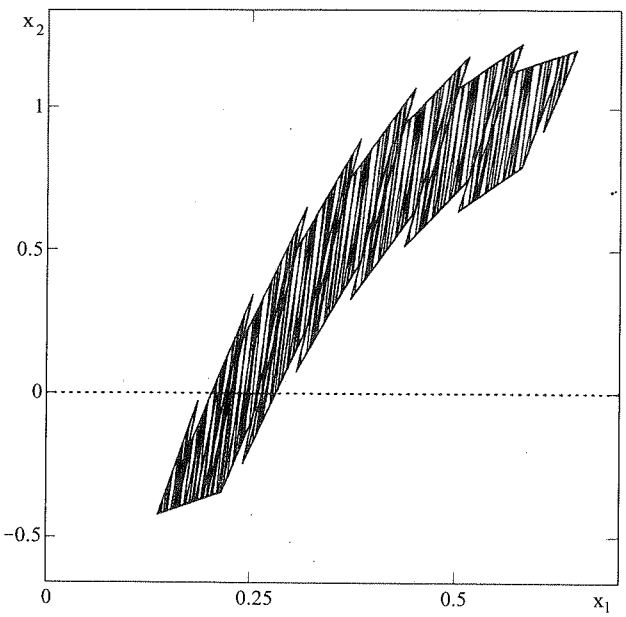


Fig. 6. All polygons obtained during the process of determining a trapping set of sequences of iterations for parameter M = 4. The sum of the polygons is an positively invariant set.

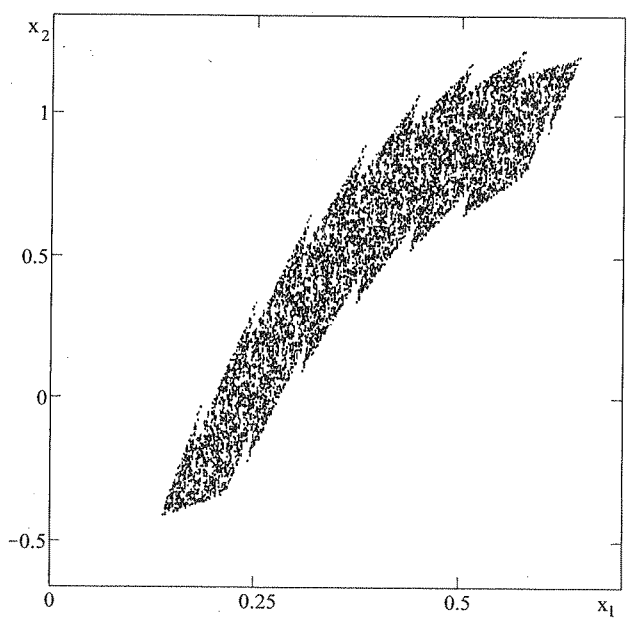


Fig. 7. Trajectory of the modulator obtained in the way of the ''ordinary'' simulation.

state space. It can be noticed that the PIS obtained for M = 2 includes the PIS obtained for M = 4 (compare Figures 4 and 6). There is exemplary trajectory shown in Figure 7. The trajectory has been obtain as a result of an "ordinary" simulation (with rounding errors). It contains 10,000 time steps, the initial 1000 time steps have been discarded. Initial condition $\mathbf{x} = [0.001\ 0.001]^T$ was used. Comparison of Figures 6 and 7 suggests that the PIS obtained for M = 4 is very tight and close to the smallest one.

The computer program that realises the above-mentioned algorithm calculates minimal and maximal values of co-ordinates x_1 and x_2 for each positively invariant set that has been obtained. It enables an analysis of the bounds of the PIS versus the input signal value u. The results of the computations are shown in Figures 8 and 9. For parameter M equal two the algorithm of determining PIS has been run 132 times for different values of u ranging from 0 to 0.9. For M = 4 the algorithm has been run for 76 constant input values equally spaced between 0 and 0.9. The simulation has been performed 451 times with input values equally spaced between 0 and 0.9 also. Each simulation contained 1179648 time steps and initial 65536 time steps were discarded.

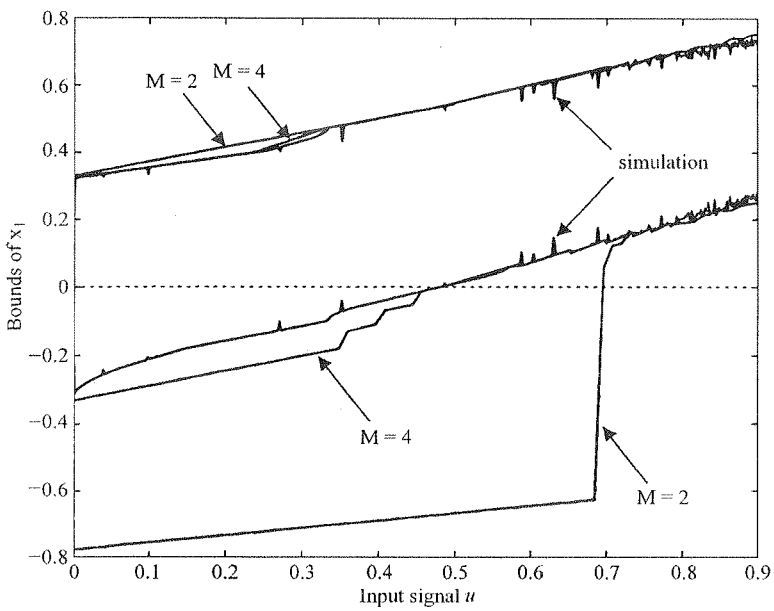


Fig. 8. Bounds (the maximal and minimal values) of the signal x_1. The outer bounds have been obtained by the algorithm for M = 2, the tighter bounds have been achieved for M = 4. Bounds calculated as a result of long simulations are denoted as "S".

Leaky ΔΣ systems have been tested by the author less extensively than the ideal systems - only some exemplary modulators have been analysed. For example a modulator with parameters L = 1.006 and u = 0.9 has been analysed (all other parameters of the modulator were the same as before). A positively invariant set corresponding to the modulator is shown in Figures 10-12.

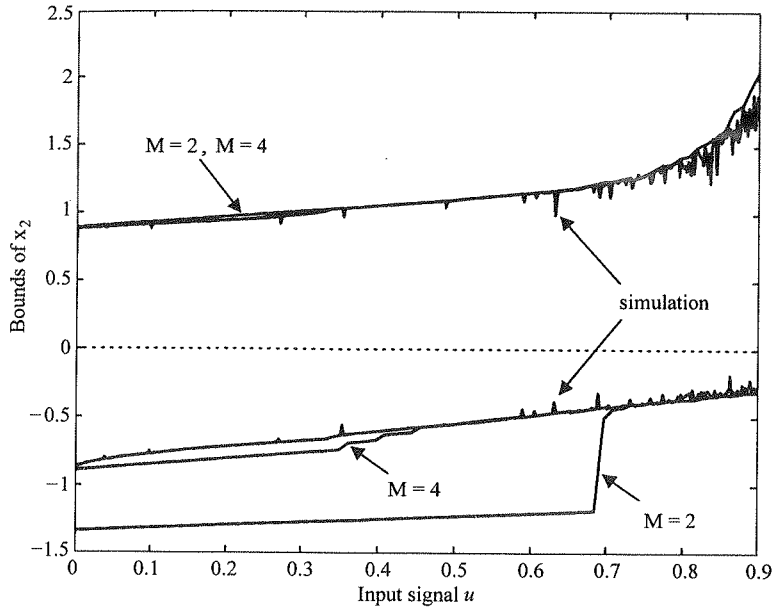


Fig. 9. Bounds of the signal x_2. The outer bounds have been obtained by the algorithm for M = 2, the tighter bounds have been calculated for M = 4. Bounds obtained as a result of long simulations are denoted as "S".

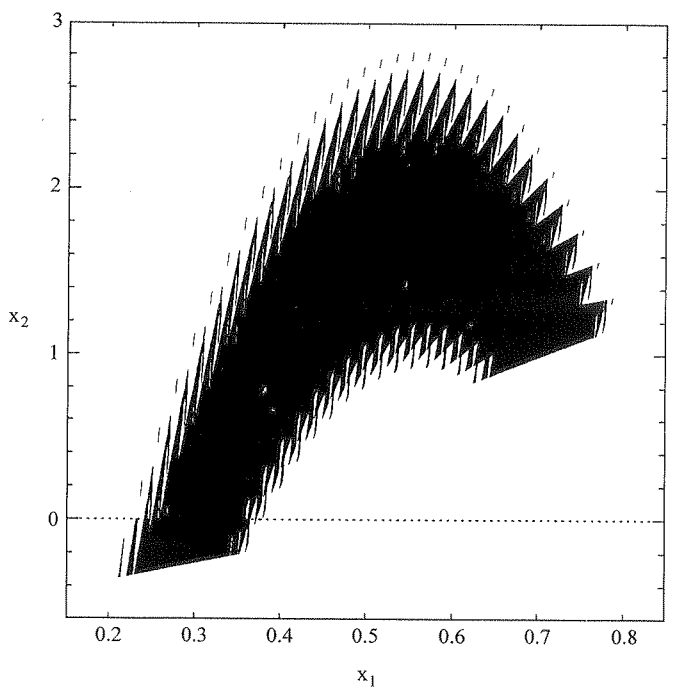


Fig. 10. The positively invariant set in state space obtained for a leaky ΔΣ modulator.

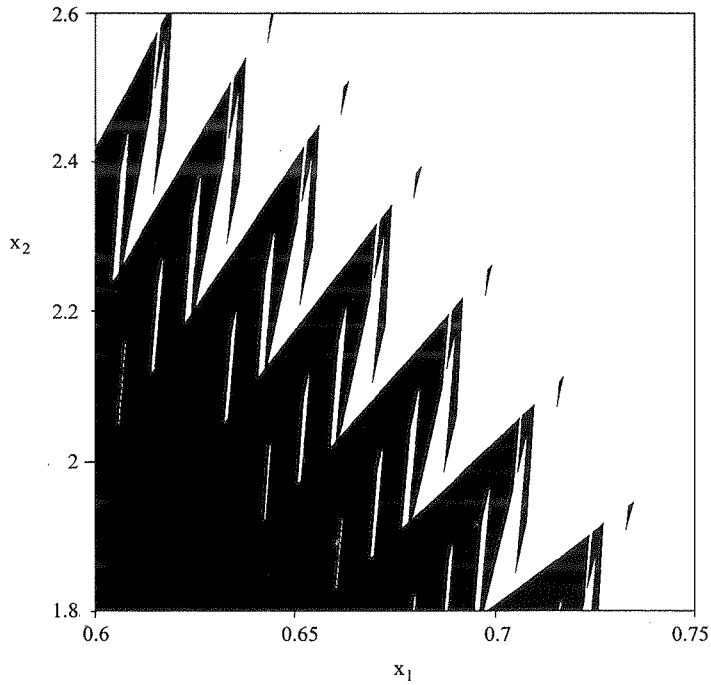


Fig. 11. A zoomed-in fragment of the positively invariant set shown in Figure 10.

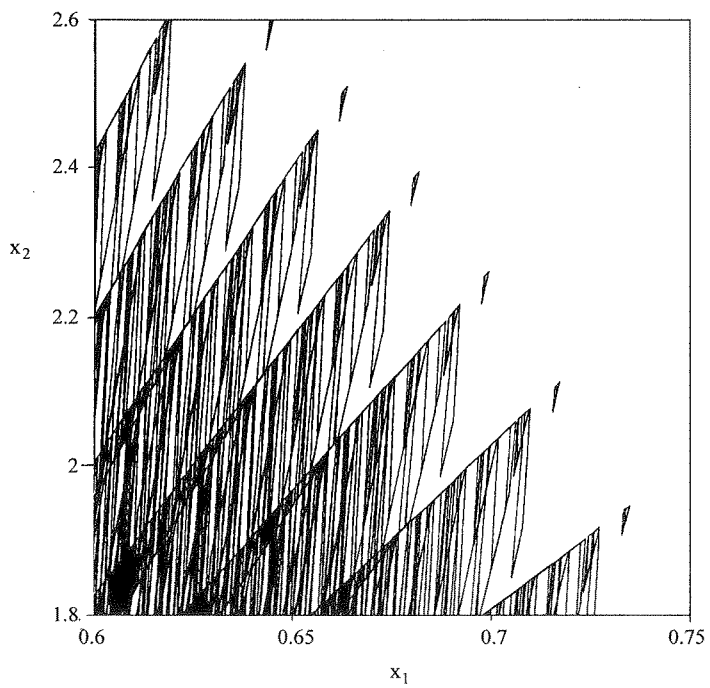


Fig. 12. Polygons which have created the fragment of PIS shown in Figure 11.

The values of parameters L and u have been chosen in such a way that the modulator is very close to instability. In this case a very intricate form of the boundary of the PIS can be seen (Figure 10 and 11). One can notice that the PIS includes discontinuities (Figure 11) and the PIS consists of really great amount of polygons, due to the fact that the modulator is close to instability (Figure 12).

6. CONCLUSIONS

The new method of determining positively invariant set in state space has been originally developed. There is a similarity between the presented method and the approach of Farrell and Feely [6]: they both emphasise the fact that there exists relationship between sequences of numbers of iterations and the position of the modulator state.

The presented method enables calculation of bounds of integrators outputs and leads to proving stability of modulator for certain value of constant input signal. The method has been implemented as a computer program for second-order $\Delta\Sigma$ modulators only. The advantage of the algorithm is that it can be applied to multiple-feedback higher-order $\Delta\Sigma$ modulators and the bounds obtained for $M = 4$ are tight. The disadvantage of the method is that the algorithm is timeconsuming especially for values of input signal close to $+1$ or -1. Tests for greater values of parameter M has been made also. The result is that the bounds of signals x_1 and x_2 obtained for $M \geq 6$ are almost the same as those obtained for $M = 4$ (but the speed of calculations falls down). Results of exemplary analysis of leaky $\Delta\Sigma$ modulator has been described also in order to show flexibility of the presented method.

REFERENCES

1. S. Norsworthy, R. Schreier, G. Temes: *Delta-Sigma Data converters.* IEEE Press, 1997.
2. O. Feely, L. Cha: *Nonlinear dynamics of a class of analogue-to-digital converters.* International Journal of Bifurcations and Chaos, vol. 2, no. 2, 1992.
3. H. Wang: *A geometric view of $\Sigma\Delta$ modulators.* IEEE Transactions on Circuits and Systems II, vol. 39, no. 6, June 1992.
4. S. Hein, A. Zakhor: *On the stability of Sigma-Delta modulators.* IEEE Transactions on Signal Processing, vol. 41, July 1993.
5. R. Schreier, M. Goodson, B. Zhang: *An algorithm for computing convex positively invariant sets for Delta-Sigma modulators.* IEEE Transactions on Circuits Systems I, vol. 44, no. 1, pp. 38-44, January 1997.
6. R. Farrell, O. Felly: *Bounding the integrator outputs of second-order Sigma-Delta modulators.* IEEE Transactions on Circuits and Systems II, vol. 45, no. 6, pp. 691-702, June 1998.
7. K. Olejarczyk: *A geometric view of multiple feedback $\Sigma\Delta$ modulator, in Proceedings of Third International Conference on Advanced A/D and D/A Conversion Techniques and their Applications.* Glasgow, pp. 66-69, July 1999. This paper is available also from WWW site http://www.ise.pw.edu.pl/~kpole.

K. OLEJARCZYK

Analiza stabilności modulatorów Delta-Sigma z lokalnymi pętlami sprzężenia zwrotnego

Streszczenie

W artykule przedstawiono nową metodę dowodzenia stabilności modulatorów Delta-Sigma poprzez wyznaczenie granic sygnałów na wyjściach układów całkujących. Zaproponowana metoda w równym stopniu bazuje na analizie teoretycznej, jak i na zastosowaniu algorytmów obliczeniowych. Zaproponowany sposób badania stabilności może być zastosowany do modulatorów z wieloma lokalnymi pętlami sprzężenia zwrotnego, ale z jednym tylko kwantyzatorem (metoda może być zastosowana do modulatorów wyższych rzędów). Założono, że sygnał wejściowy musi być stały. W oparciu o prezentowaną metodę napisano program komputerowy. Przeprowadzono wyczerpujące testy algorytmu dla modulatorów drugiego rzędu a wyniki przedstawiono w artykule.

Słowa kluczowe: sigma-delta, kształtowanie szumu, chaos, analiza nieliniowa.

to ju
a cir
theor
softw
a qua
This
nece

Key

Very
implemen
chips wi
microproc
a VLSI
semi-cust
(MPGAs)
several n

KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJI, 2000, 46, z. 4, ss. 479–491

An efficiency measure of FPGA based logic synthesis tools

HENRY SELVARAJ

Departament of Electrical and Computer Engineering
University of Nevada, Las Vegas

PIOTR SAPIECHA

Warsaw University of Technology

Otrzymano 2000.06.28
Autoryzowano 2000.09.18

In FPGA-based designs, the number of Logic Cells (LCs) needed is an important criterion to judge whether a desing is good or not. But the total number of LCs needed to implement a circuit differs vastly from tool to tool. Normally, vendor software use more LCs than the theoretical maximum needed by functional decomposition to implement a circuit. Academic software uses less number of LCs. So far, we are not aware of any technique that would give a quantitative measure to judge the comparable silicon area efficiency of a logic synthesis tool. This paper presents a technique to calculate the minmax number of logic cells (Q) which are necedssary to implement a logic circuit.

Key words: logic synthesis, FPGA, functional decomposition

1. INTRODUCTION

Very Large Scale Integration Technology (VLSI) has opened the door to the implementation of powerful digital circuits at low cost. It has become possible to build chips with more than a million transistors, as exemplified by the state-of-the-art microprocessors. Such chips are realized using full-custom approach, where all parts of a VLSI circuit are carefully tailored to meet a set of specific requirements and the semi-custom approaches such as Standard Cells and Mask-Programmed Gate Arrays (MPGAs). Both techniques, however, require extensive manufacturing effort, taking several months from beginning to end. In electronics industry it is vital to reach the

market with new products in the shortest possible time. Furthermore, it is important that the financial risk incurred in the development of a new product be limited so those more new ideas can be prototyped. Field Programmable Gates Arrays (FPGAs) have emerged as the ultimate solution to these time-to-market and risk problems because they provide instant manufacturing and very low cost prototyping.

A designer who wants to make good use of FPGAs must have access to an efficient CAD system. The traditional steps involved in a typical system for implementing a circuit using FPGA is given in Fig 1. The starting point for the desing process is the initial logic entry of the circuit that is to be implemented, like schematic capture, VHDL description or Boolean expressions. This is followed by translating the description into a standard from, which traditionally has been the Boolean expression. The Boolean expressions are then processed by logic optimization tool, which manipulates the expressions. The optimized Boolean expressions are then transformed into a circuit of logic blocks. This is done by a technology mapping program. Having mapped the circuit into logic blocks, it is necessary to decide where to place each block in the FPGA's array. A placement program is used to solve this problem. The final step in the CAD system is performed by the routing software, which assigns the FPGA's wire segments and chooses programmable switches to establish the required connections among logic blocks.

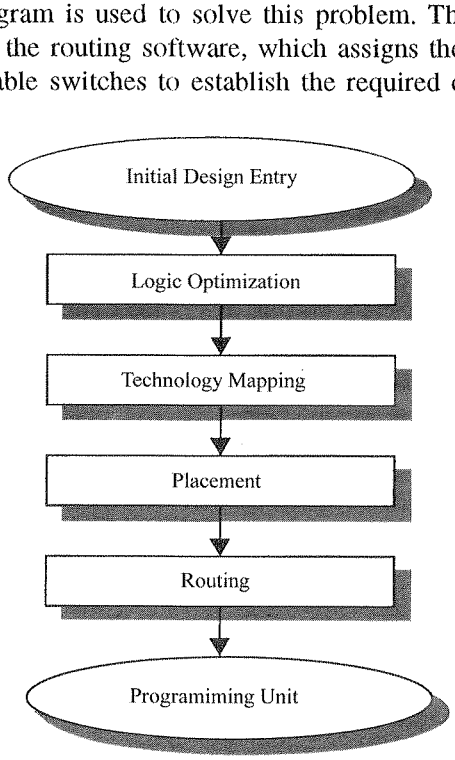


Fig. 1. Procedure of A Typical CAD System

FPGAs produced by different vendors have different logic cell structure. Generally, a logic cell can be viewed as a functional unit capable of implementing any function of a fixed number of variables. In the desing process, the number of the logic cells needed

is an important criterion to judge whether a design is good or not. In this paper, it is shown that we can calculate the „*minmax*" number (Q) of LCs needed to implement a combinational circuit. Q is the maximum number of logic cells needed to implement a circuit without redundancy. In other words, the circuit can be implemented using a maximum of number of LCs. However, most vendor software uses more LCs than the theoretical maximum (Q). The *minmax* value Q is obtained using functional decomposition [Lub95, Lub95a, Joz95, Sas93].

2. MAXIMUM DECOMPOSITION THEOREM (MDT)

A Boolean function F of n binary variables X_1, X_2 ... X_n, can be expressed as:

$$F: D^n \rightarrow \{0,1,-\}^m$$

where '-' is called a 'don't care', X is the input set and Y is the outpout set.

There are two main decomposition strategies: parallel and serial decomposition [Sel96]. The first one, parallel decomposition, relies on dividing the outputs into two disjoint subsets so that each subset of outputs depend on lesser ot the same number of variables as the given function F:

$$Y = Y_{G1} \cup Y_{G2}$$

The respective support sets of Y_{G1} and $Y_{G2} \subseteq Y$, X_{G1} and $X_{G2} \subseteq X$. It is shown in Fig. 2.

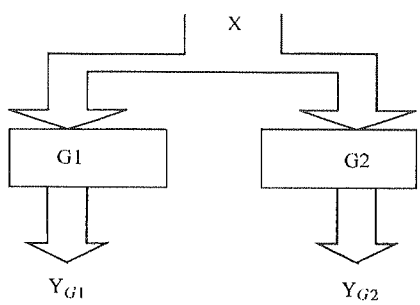


Fig. 2. Parallel decomposition

The second one, called serial decomposition, relies on partitioning the input variables in such a way as to obtain a two-level functional dependency $F = H(A,G(B \cup C))$. It is shown in Fig. 3.

These two strategies are used in the multilevel decomposition method developed by Selvaraj et. al. [Lub95, Sel96].

Theorem: Let F(X) denote a Boolean function F with a set of input variables X. Let $X = A \cup B$ and $A \cap B = \emptyset$, where B is the set of bound variables (G block inputs) and A is the set of free variables (direct inputs to the H block). For every function

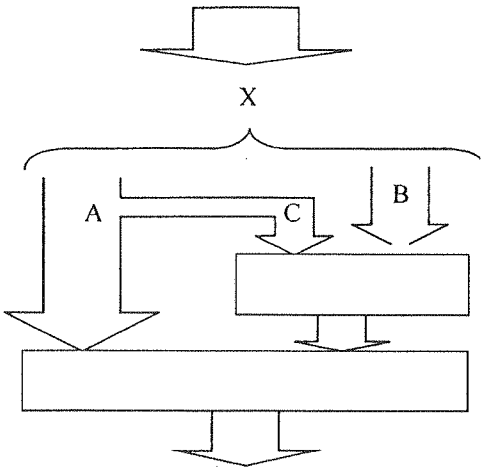


Fig. 3. Serial decomposition

F with |X| inputs, |Y| outputs, |A| free variables and |B| bound variables, if $|X| > 2|Y|$ and $|B| > |Y|2^{|A|}$, there exists a maximal disjoint serial decomposition such that:

$$F = H(G(B),A)$$

where the number of G block outputs = $|Y|2^{|A|}$.

We apply this theorem to prove that functional decomposition can be use to decomposition any given function.

We use this theorem to prove that functional decomposition can be used to decompose any given function.

3. MAXIMUM DECOMPOSITION STRUCTURE

Using the MDT, we can construct the Maximum Decomposition Structure (MDS) for any unitary function. The *minmax* value Q for the design is calculated as given below. Let us assume that a combinational circuit with N inputs and M outputs is implemented using Logic Cells with G inputs and 1 output. The *minmax* value Q is given by the equation:

$$Q = M(2^{(N-G+1)} - 1) \tag{1}$$

If the LCs have G inputs and two outputs, Q is expressed as:

$$Q = |M/2|(2^{N-G+1} - 1) + K*2^{N-G} (G > 4,\ N > 5) \tag{2}$$

where: K = 1, if M is odd and K = 0, if M is even. If G is 4, there is a special case:

$$Q = \lfloor M/2 \rfloor (5*2^{n-5} - 1) + K*5*2^{N-5} (G = 4,\ N > 5)$$
$$Q = \lfloor M/2 \rfloor *4 + K*2 \qquad (G = 4,\ N = 5) \tag{3}$$

where: K = 1, if M is odd and K = 0, if M is even.

Proof of formula (1):

Let us assume we use 3:1 LCs (Logic Cells with 3 inputs and 1 output) to implement N:1 circuit (a combinational circuit with N inputs and 1 output).

$$\text{If } N = 3, \text{ then } Q = 2^1 - 1 = 1.$$

For N = 4, the Maximum Decomposition Structure can be constructed as given in Fig. 4.

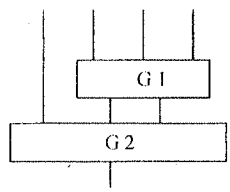


Fig. 4. Using 3:1 LCs to implement a 4:1 circuit

And part G1 can be decomposed into two 3:1 LCs using parallel decomposed an in Fig. 5.

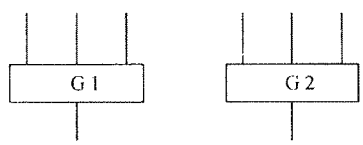


Fig. 5. Using 3:1 LCs to implement a 3:2 circuit

Therefore, $Q = 2^2 - 1 = 2*1 + 1 = 3$. Fig. 6 gives the decomposition structure for N = 5.

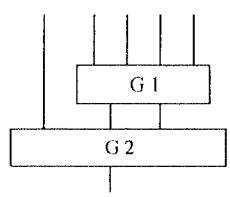


Fig. 6. Using 3:1 LCs to implement a 5:1 circuit

And part G1 can be decomposed into two 4:1 LCs using parallel decomposition.
So, $Q = 2*3 + 1 = 7 = 2^3 - 1$.

Using the some procedure, we can write: for N = 6, Q = 2*7+1 = 15 = 2^4-1, and for N = 7, Q = 2*15+1 = 31 = 2^5-1;

Therefore, we can generalize for N as:

$$Q = 2^{N-2}-1 \ (N>=3) \tag{4}$$

Suppose we use 4:1 LCs to implement N:1 combinational circuit, Q is the maximum number of LCs needed according to the MDT.

If N = 4, then Q = 1.

For N = 5, the Maximum Decomposition Structure is in Fig. 7.

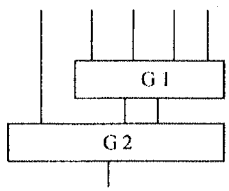


Fig. 7. Using 4:1 LCs to implement a 5:1 circuit.

And the part G1 can be decomposed into two 4:1 LCs using parallel decomposition.

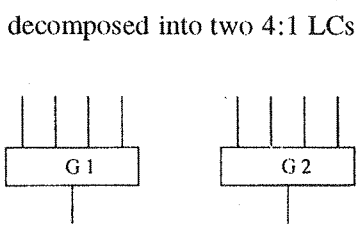


Fig. 8. Using 4:1 LCs to implement a 4:2 circuit

So, Q = 2*1+1 = 3 = 2^2-1. Let N = 6, The decomposed structure is in Fig. 9.

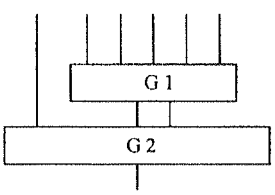


Fig. 9. Using 4:1 LCs to implement a 6:1 circuit

And part G1 can be decomposed into two 4:1 LCs using parallel decomposition. So, Q = 2*3+1 = 7 = 2^3-1. Using the same procedure, we can write: for N = 7, Q = 2*7+7+1 = 15 = 2^4-1; and for N = 8, Q = 2*15+1 = 31 = 2^5-1;

– 1, and for

(4)

e maximum

composition.

g. 9.

position. So,
for N = 7,

Therefore, we can generalize for N as:

$$Q = 2^{N-3} - 1 \ (N >= 4) \tag{5}$$

Further, we can derive that: if we use 5:1 LCs to implement N:1 combinational circuit, then:

$$Q = 2^{N-4} - 1 \ (N >= 5) \tag{6}$$

and if we use 6:1 LCs to implement N:1 combinational circuit, then:

$$Q = 2^{N-5} - 1 \ (N >= 6) \tag{7}$$

We generalize that: if we use G:1 LCs to implement a N:1 combinational circuit, then:

$$Q = 2^{(N-G+1)} - 1 \ (N >= G) \tag{8}$$

If the number of outputs of the combinational circuit is not 1, but M , then it can first be decomposed into M N:1 sub-circuits using parallel decomposition, then by applying the procedure discussed above, we can calculate the Q value. Therofore, we can state that combinational circuit with N inputs, M outputs can be implemented by using basic Logic Cells with G inputs and 1 output. The *minmax* number of LCs, value Q can be calculated in the folloning way:

$$Q = M(2^{(N-G+1)} - 1)$$

Proof of formula (2):

To derive the formula, we have to calculate Q for two conditions:

- Using G:2 LCs to implement a N:2 circuit.
- Using G:2 LCs to implement a N:1 circuit.

1. Using G:2 LCs to implement a N:2 circuit.

Suppose we use 5:2 LCs to implement N:2 combinational circuit, Q is the number of Cs needed.

If N = 5, then Q = 1. For N = 6, the decomposition structure is in Fig. 10.

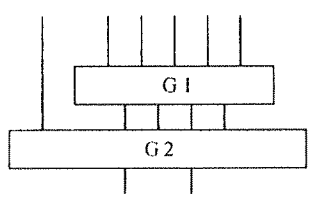


Fig. 10. Using 4:2 LCs to implement a 6:2 circuit

And the part G1 can be decomposed into two 5:2 LCs using parallel decomposition. So, $Q = 3 = 2^2 - 1$. The decomposition structure for N = 7is in Fig. 11.

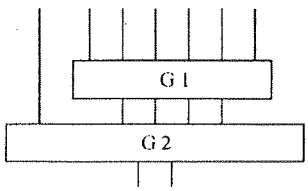


Fig. 11. Using 4:2 LCs to implement a 7:2 circuit

And the part G1 can also decomposed into two 5:2 LCs using parallel decomposition. Them $Q = 2 * 3 + 1 = 7 = 2^3 - 1$. We can get: N:2 circuit using 5:2 LCs to implement $Q = 2^{N-4} - 1$. Suppose we use 6:2 LCs to implement N:2 combinational circuit, If N = 6, then Q = 1. For N = 7, the decomposition proces is shown in Fig. 12.

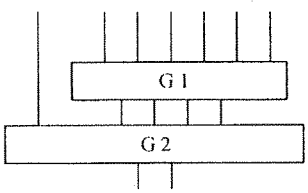


Fig. 12. Using 6:2 LCs to implement a 7:2 circuit

And the part G1 can be decomposed into two 6:2 LCs using parallel decomposition. Then, $Q = 3 = 2^2 - 1$. For N = 8, the decomposition structure is in Fig. 13.

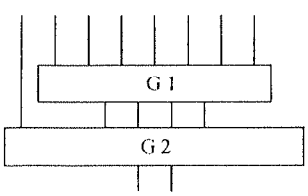


Fig. 13. Using 6:2 LCs to implement a 8:2 circuit

And the part G1 can also be decomposed into two 6:2 LCs using parallel decomposition.

Therefor, $Q = 2*3+1 = 7 = 2^3-1$. In this case, we can get: using 6:2 LCs to implement a N:2 circuit,

$$Q = 2^{N-5}-1 \tag{9}$$

Applying the same procedure, we can get: using 7:2 LCs to implement a N:2 circuit,

$$Q = 2^{N-6}-1 \tag{10}$$

Using 8:2 LCs to implement a N:2 circuit,

$$Q = 2^{N-7}-1 \tag{11}$$

Them consequently using G:2 LCs to implement a N:2 circuit,

$$Q = 2^{N-G+1}-1 \tag{12}$$

2. Then cousequently, using G:2 LCs to implement a N:1 circuit.

Suppose we use 5:2 LCs to implement a N:1 circuit. If N = 5, then Q = 1. For N = 6, the decomposition structure is show in Fig. 14.

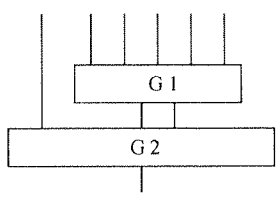


Fig. 14. Using 5:2 LCs to implement a 6:1 circuit

Let, Q = 2. For N = 7, the decomposition structure is shown in Fig. 15.

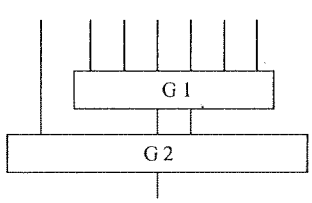


Fig. 15. Using 5:2 LCs to implement a 7:1 circuit

From formula (8), we will use 3 5:2 LCs to implement a 6:2 circuit (part G1), so $Q = 3+1 = 4 = 2^2$. For N = 8, the decomposition structure is shown in Fig 16.

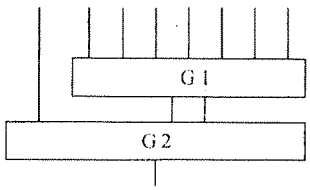


Fig. 16. Using 5:2 LCs to implement a 8:1 circuit

From formula (8), we will use 7 5:2 LCs to implement a 7:2 circuit (part G1), so $Q = 7+1 = 2^3$;

Therefore, we can generalize for N as:

$$Q = 2^{N-5} \tag{13}$$

Applying the some method, we can derive that use 6:2 to implement a N:1 circuit:

$$Q = 2^{N-6} \tag{14}$$

Use 7:2 to implement a N:1 circuit

$$Q = 2^{N-7} \tag{15}$$

Adequately using G:2 to implement a N:1 circuit, the number of the LCs needed is:

$$Q = 2^{N-G} \tag{16}$$

So, suppose we want to use G:2 to implement a circuit W with N input, M output.

1. If M is even, we first decompose in into M/2 N:2 sub-circuits, then we can use G:2 LCs to implement it using the method mentioned above.
2. If M is odd, we first decompose in into M-1/2 N:2 and a N:1 sub-circuits, then we can use G:2 LCs to implement it using the method mentioned above.

So, we get formula (2).

Proof of formula (3):

A special case: Using 4:2 LCs to implement N:1 or N:2 circuits. Suppose we use 4:2 LCs to implement a N:2 circuit, if N = 4, then Q = 1 and if N = 5, the decomposition structure is shown in Fig. 17.

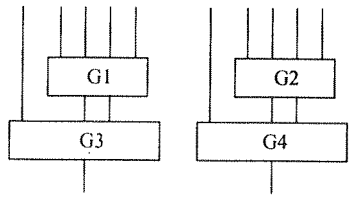


Fig. 17 Using 4:1 LCs to implement a 5:2 circuit

Let, Q = 4. We have to do such kind of decomposition just because if we don't use parallel decomposition first, instead, we use serial decomposition, from the MDT, we can get the decomposition shown in Fig. 18.

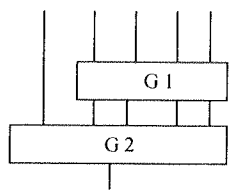


Fig. 18. Using 4:1 LCs to implement a 5:1 circuit

Part G1 has 4 input and 4 output, it is meaningless. And this is also why we have to mention 4:2 LCs as a special case. Using the deduction procedure mentioned before, we can get, if N = 6, Q = 2 * 4 + 1 = 9 and if N = 7, Q= 2 * 9 + 1 = 19.

Therefore, we can generalize for N as:

$$Q = 5*2^{N-5} - 1 (N >= 5)$$

Also, we can get, if we want to use 4:2 LCs to implement N:1 circuits,

$$Q = 5*2^{N-5} (N >= 6)$$

$$Q = 2 (N = 5)$$

So, suppose we want to use 4:2 to implement a circuit W with N input, M output.

1. If M is even, we first decompose in into M/2 N:2 sub-circuits, then we can use 4:2 LCs to implement it using the method mentioned above.
2. If M is odd, we first decompose in into M-1/2 N:2 and a N:1 sub-circuits, then we can use 4:2 LCs to implement it using the method mentioned above.

Fiually, we proved formule (3).

4. RESULT AND CONCLUSION

We calculated the Q value for MCNC benchmarks and compared them with actual LCs needed to implement the benchmarks using an academic software DECMAIN and Alterra's Max +Plus II system. The results are presented in Table 1. The Q parameter provides a reference value to compare the two software. The chosen MCNC benchmarks are small enough so that placement and routing problems don't influence the cell count.

Table 1

Comparision of our method and other method

Output: Input	Name	Max+Plus II	DECMAIN	Q
9:1	9sym	98	11	63
9:5	Clip	404	61	315
7:3	Rd73	98	10	45
8:5	Root	113	57	155
10:4	SAO	125	53	508
7:4	Z4	88	7	60
8:4	Rd84	147	10	124

5. CONCLUSIONS

In this paper was presented a technique to calculate the numeber of logic cells necessary to implement logic circuit in FPGA – based technology. This number would to be the important measure, which help to judge on quality of design, whether a design is good or bad.

REFERENCES

1. [Joz95] L. Józwiak: *General decomposition and its use in digital circuit synthesis.* VLSI *Desing*, Vol. 3, Nos. 3–4, pp. 225–248, OPA, Amsterdam, 1995.
2. [Lub95] T. Łuba, H. Selvaraj: *A general approach to Boolean function decomposition and its applications in FPGA based-synthesis.* VLSI Design, Vol. 3, Nos. 3–4, pp. 289–300, OPA, Amsterdam, 1995.
3. [Lub95a] T. Łuba, H. Selvaraj, M. Nowicka and A. Kraśniewski: *Balanced multilevel decomposition and its applications in FPGA-based synthesis* in G. Saucier, A. Mignotte (ed.), „Logic and Architecture Synthesis", Chapman & Hall, 1995.
4. [Sap00] P. Sapiecha, H. Salvaraj, M. Pleban: *Decomposition of Boolean relations and functions in logic synthesis and data analysis.* RSCTC, Canada 2000.
5. [Sas93] T. Sasao: *FPGA design by generalized functional decomosition, in logic synthesis and optimization.* Sasao T. (Ed.), Kluwer Academic Publishers, 1993, pp. 233–258.
6. [Sel96] H. Selvaraj: *FPGA-based logic synthesis, PhD. dissertation*, Warsaw University of Technology, Warsaw, Poland, 1996.

H. SELVARAJ, P. SAPIECHA

MIARY JAKOŚCI SYNTEZY LOGICZNEJ W STRUKTURACH FPGA

Streszczenie

W projektach wykonanych w technologii FPGA liczba komórek logicznych (KL) jest ważną miarą oceny jakości, pozwala ona udzielić odpowiedzi na pytanie: czy jest to dobry projekt, czy też nie. Liczba KL potrzebnych do realizacji układu logicznego jest istotnie zależna od używanego do projektowania narzędzia CAD. Oprogramowanie komercyjne wykorzystuje więcej KL niż tyle ile wynika z rozważań teoretycznych, dotyczących dekompozycji funkcjonalnej układów logicznych. Oprogramowanie akademickie używa mniejszej liczby KL. Dotychczas, nie znaliśmy żadnej techniki, która pozwalałaby ocenić wymaganą powierzchnię krzemu potrzebną do realizacji projektu w technice FPGA. W artykule prezentujemy metodę pozwalającą wyznaczyć ograniczenie górne liczby KL. W pracy wykażemy poprawność tego oszacowania.

Słowa kluczowe: synteza logiczna, FPGA, dekompozycja funkcjonalna.

jes
ws
op
wy
błę

Sło

Mo
widma
przyrzą
stosowa
płaską,
zwierci
w poło
źródeł
optyczn
niem te
w połąc

W
technik
z mono

KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJI, 2000, 46, z. 4, ss. 493–510

Homodynowy układ detekcji słabych sygnałów optycznych do współpracy z monochromatorami

JERZY SIUZDAK, ROMAN NOWAK, TOMASZ CZARNECKI

Instytut Telekomunikacji, Politechnika Warszawska

Otrzymano 2000.06.12
Autoryzowano 2000.09.21

W pracy opisano homodynowy układ pomiaru mocy słabych sygnałów optycznych, który jest przeznaczony do współpracy z monochromatorem. Zamieszczono również krótki opis współpracującego z nim oprogramowania. Najważniejszym problemem projektowym była optymalizacja szumowa fotoodbiornika. Z przeprowadzonej analizy błędów pomiarowych wynika, że oprócz błędów związanych z monochromatorem najistotniejszą rolę odgrywają błędy związane z niedokładnością określenia zależności czułości fotodiody od długości fali.

Słowa kluczowe: detekcja homodynowa, monochromator, widmo optyczne

1. WSTĘP

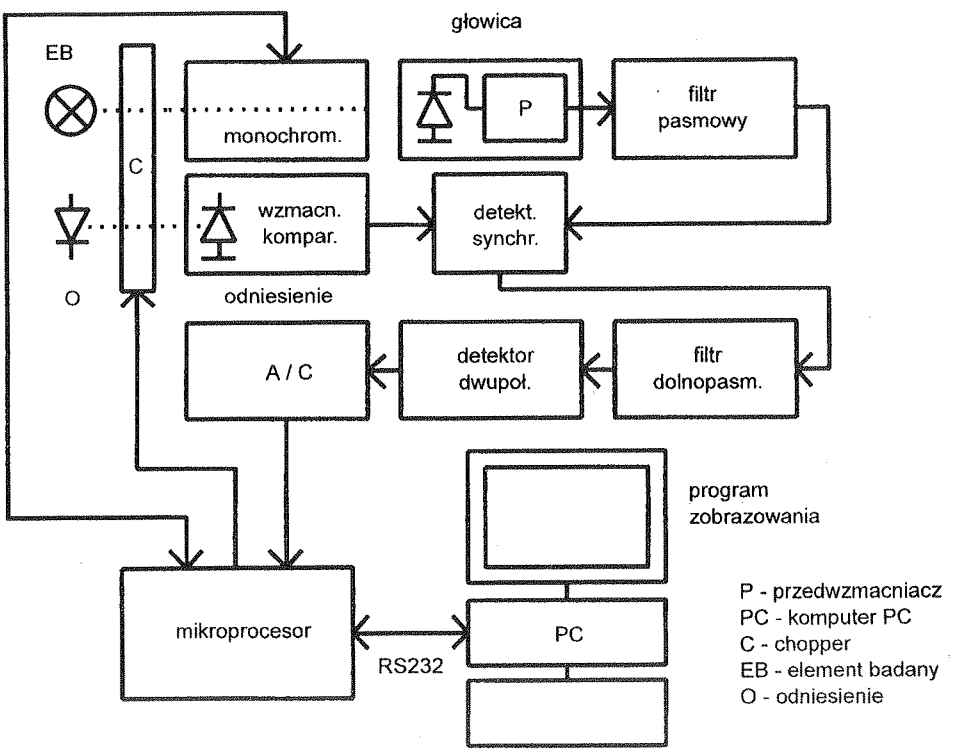
Monochromator jest jednym z najważniejszych urządzeń służących do pomiaru widma sygnału optycznego i jako taki jest powszechnie stosowany w nowoczesnych przyrządach pomiarowych, na przykład w optycznych analizatorach widma. W często stosowanej konfiguracji (Czerny-Turner [1]) widmo źródła światła jest formowane przez płaską, obrotową siatkę dyfrakcyjną typu odbiciowego, używaną w połączeniu z dwoma zwierciadłami sferycznymi. Dana długość fali jest wyodrębniona przez wąską szczelinę w położonym za nią fotodetektorem. Częstym problemem, zwłaszcza przy badaniu źródeł światła o małej mocy i rozproszonym widmie, jest zbyt niski poziom sygnału optycznego w porównaniu z poziomem szumów własnych fotoodbiornika. Rozwiązaniem tego problemu jest modulacja amplitudowa mocy badanego strumienia świetlnego w połączeniu z wykorzystaniem detekcji homodynowej w odbiorniku.

W niniejszej pracy opisano układ opracowany w Instytucie Telekomunikacji Politechniki Warszawskiej, który działa w oparciu o powyższą zasadę i współpracuje z monochromatorem. Artykuł skonstruowany jest w następujący sposób. W paragrafie 2

zamieszczono krótki opis działania układu, jak również współpracującego z nim oprogramowania. W paragrafie 3 opisano problemy występujące przy projektowaniu, a dotyczące optymalizacji szumowej układu. W następnym punkcie zanalizowano błędy pomiarowe. W paragrafie 5 podano przykłady uzyskanych wyników pomiarowych. Wreszcie ostatni punkt poświęcony jest podsumowaniu.

2. OGÓLNY OPIS DZIAŁANIA

Schemat blokowy układu pokazano na rys. 1. Zasada działania jest niezwykle prosta. Moc sygnału optycznego z badanego źródła jest modulowana przez obracającą się tarczę z otworami (chopper). Synchronicznie, przez ten sam chopper, modulowana jest moc w kanale odniesienia składającym się z transportera wraz z układem formującym (komparatorem). Jak zostanie to wykazane w następnym paragrafie, uniezależnia to działanie układu do ewentualnych wahań w prędkości obrotowej choppera. Badany sygnał jest podawany na monochromator, przy czym aktualne położenie monochromatora (transmitowana długość fali) jest ustawiana programowo z komputera, który poprzez złącze RS 232 i mikroprocesor steruje jego silnikiem krokowym. Sygnał optyczny wychodzący z monochromatora jest odbierany przez szerokopasmową fotodiodę p-i-n



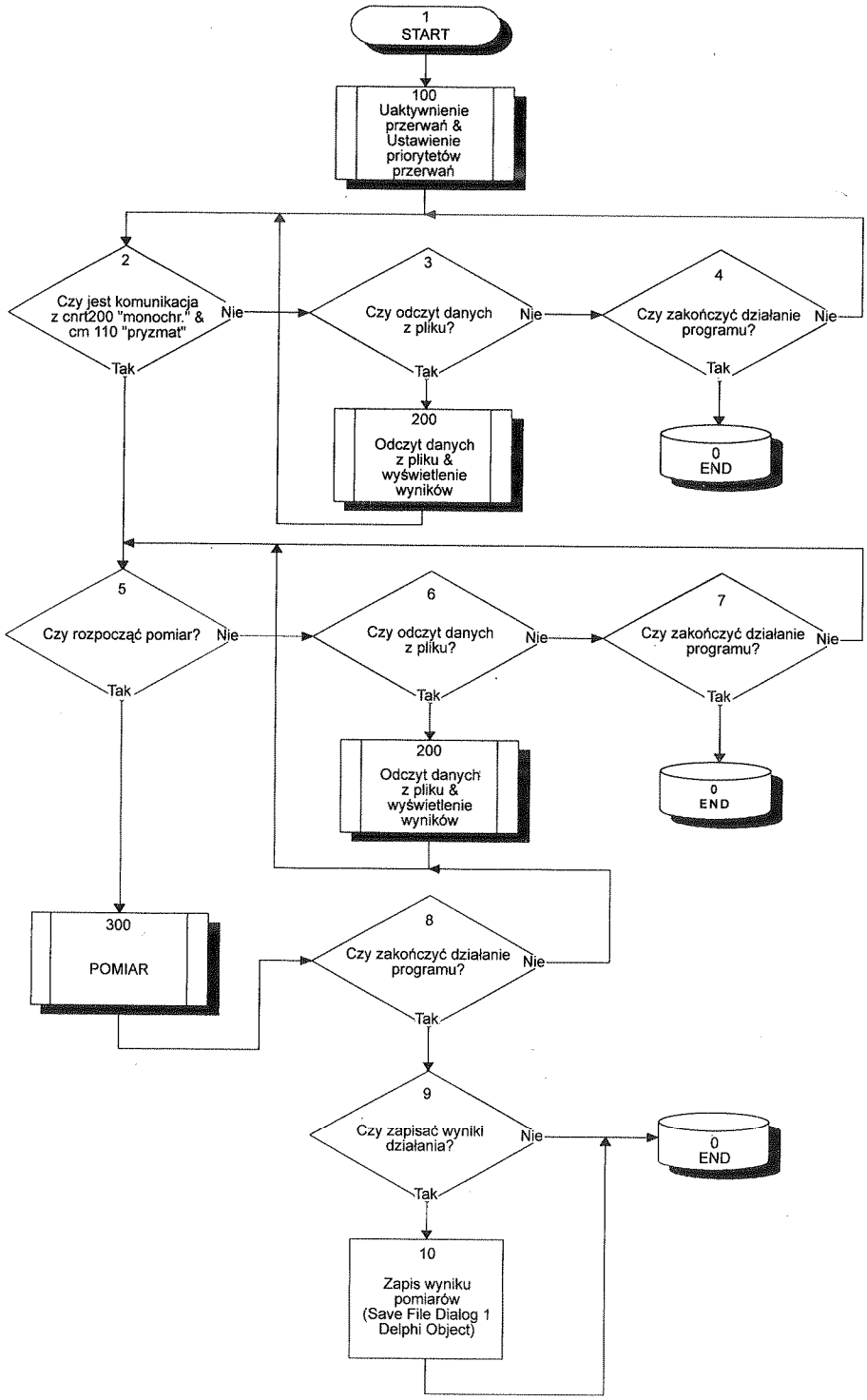
Rys. 1. Schemat blokowy układu pomiarowego

cego z nim
ojektowaniu,
owano błędy
omiarowych.

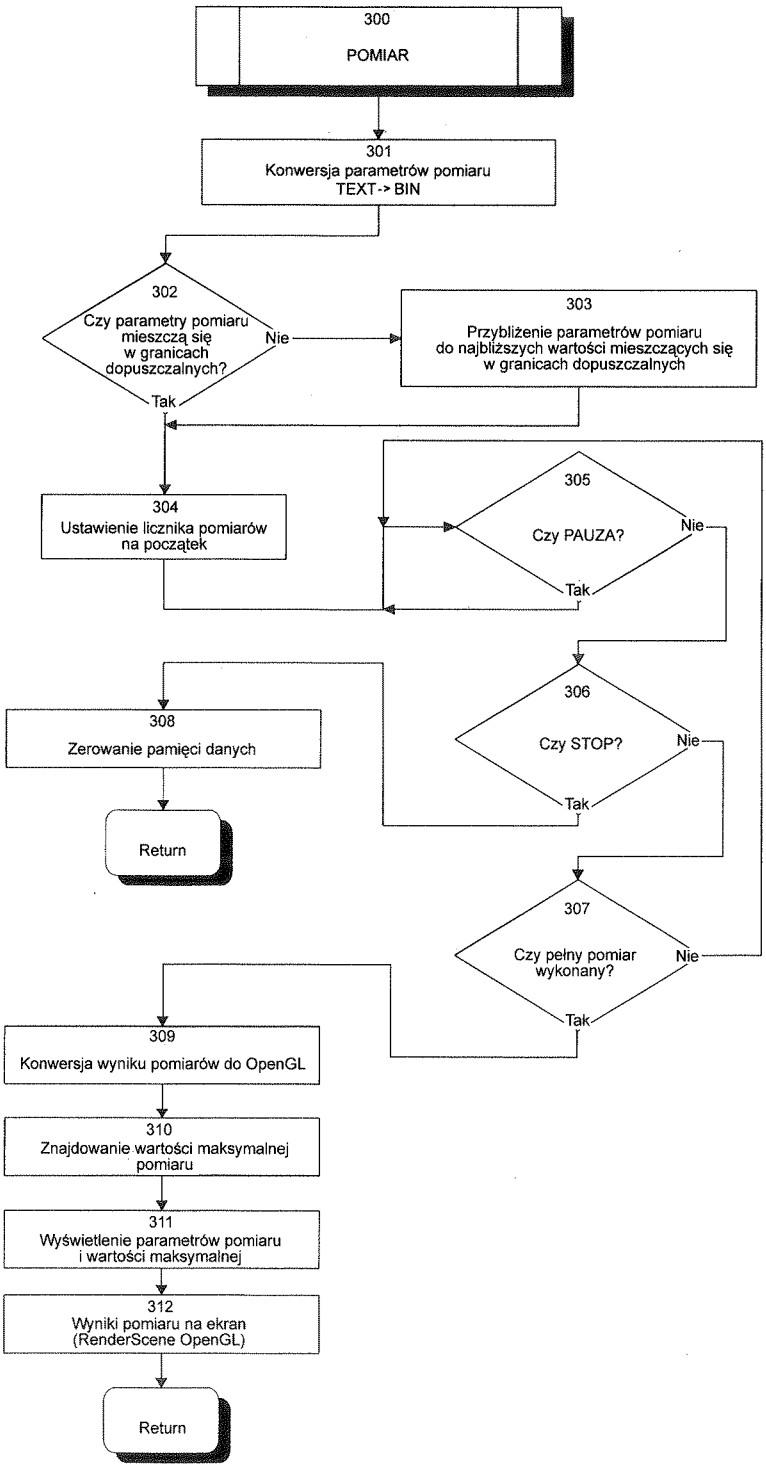
wykle prosta.
cą się tarczę
na jest moc
formującym
ezależnia to
era. Badany
nochromato-
tóry poprzez
ał optyczny
odiodę p-i-n

cniacz
PC

adany



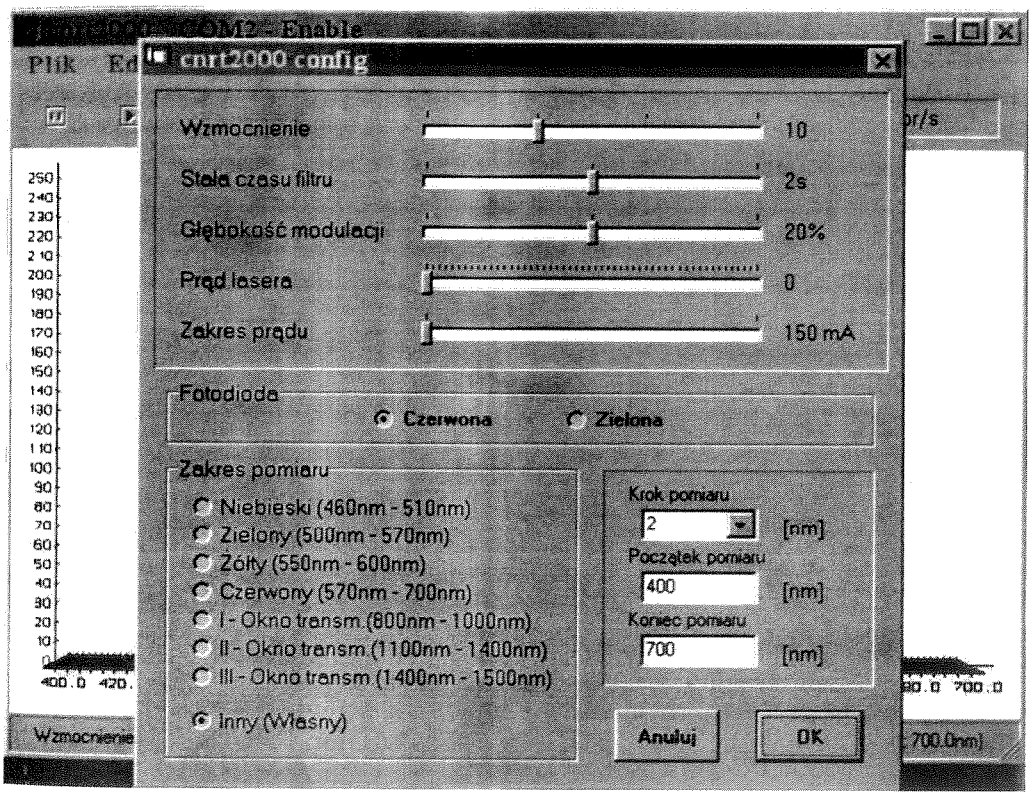
Rys. 2. Sieć działań pętli głównej



Rys. 3. Sieć działań pętli pomiarowej

połączoną z przedwzmacniaczem. Obydwa te elementy znajdują się w wymiennej głowicy. Poprzez filtr pasmowy odebrany sygnał doprowadzany jest do detektora synchronicznego sterowanego przebiegiem ze wspomnianego na początku kanału odniesienia. Sygnał wyjściowy z detektora synchronicznego (proporcjonalny do mocy odbieranego sygnału optycznego) jest poprzez filtr dolnopasmowy i detektor podawany na przetwornik analogowo-cyfrowy. Detektor dwupołówkowy nie ma wpływu na błędy pomiarowe, a został zastosowany ze względu na konieczność ustalenia polaryzacji na wejściu przetwornika analogowo-cyfrowego. Dane dotyczące mocy odbieranego sygnału są poprzez mikroprocesor i złącze RS 232 podawane do komputera, gdzie służą do uzyskania zobrazowania charakterystyki widmowej badanego elementu. Uzyskana charakterystyka jest korygowana zgodnie z zależnością zmian czułości fotodiody od długości fali.

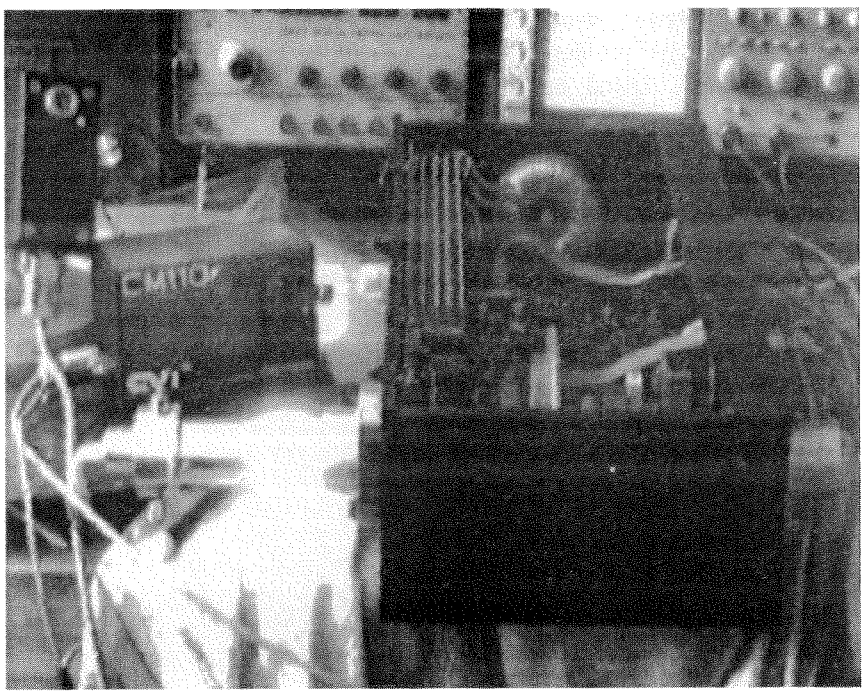
Napisany w tym celu program komputerowy pracuje pod systemami operacyjnymi Win95/ Win/98 WinNT4.0/ Win2000 i umożliwia sterowanie modułu głowicy pomiarowej (kontrola i odczyt pomiarów), modułu silnika krokowego oraz wyświetlania otrzymanych wyników. Został napisany w języku obiektowym Delphi z importem



Rys. 4. Wygląd okna ustawiania parametrów programu

szybkich procedur asemblera x86. Do wyświetlania szybkiej grafiki trójwymiarowej zastosowano technikę OpenGL. Dzięki takiemu podejściu do zagadnienia tworzenia oprogramowania, wynikowy kod programu nie przekracza 400kB, poprawnie działa już na komputerach klasy PC-i486 z zegarem 100MHz i pamięcią operacyjną 32MB. Aby w pełni wykorzystać efekty trójwymiarowej wizualizacji wyników pomiarów zalecany jest komputer z kartą graficzną VGA minimum 255 kolorów i sprzętowym akceleratorem obsługującym procedury języka OpenGL.

Oprogramowanie całego urządzenia składa się z części programu uruchamianego z komputera klasy PC i części programu znajdującego się w module pomiarowym głowicy pomiarowej napisanego dla procesora Intel MCS51, który kontroluje również moduł silnika krokowego. Ogólny schemat działania programu ilustruje sieć działań pętli głównej (rys. 2), oraz sieć działań pętli pomiarowej (rys. 3).



Rys. 5. Widok urządzenia

Interfejs użytkownika programu jest zgodny ze współczesnymi standardami. Wygląd okna programu i okna konfiguracyjnego przedstawiono na rys. 4, natomiast na rys. 5 pokazano wygląd samego urządzenia.

3. OPTYMALIZACJA SZUMOWA STOPNIA WEJŚCIOWEGO

3.1. WIADOMOŚCI OGÓLNE

Ze względu na mechanizm powstawania, szumy w interesującym nas układzie można podzielić na [2,3]:

— szumy cieplne,
— szumy śrutowe,
— szumy niskoczęstotliwościowe typu 1/f,
— szumy wybuchowe.

Szum cieplny powstaje wskutek przypadkowego ruchu cieplnego nośników ładunku w przewodnikach. Określa on minimalny poziom szumu w elemencie rezystywnym. Średniokwadratowy prąd szumów w elemencie o rezystancji r wyraża się zależnością

$$i_T^2 = \frac{4kTB}{r}, \tag{1}$$

gdzie k jest stałą Boltzmanna ($1{,}38 \times 10^{-23}$ J/K), T – bezwzględną temperaturą przewodnika, zaś B – pasmem układu. Szum cieplny jest szumem białym, tzn. jego gęstość widmowa nie zależy od częstotliwości.

Szum śrutowy jest związany z przepływem prądu przez barierę potencjału, jaka występuje np. na złączu p-n elementu półprzewodnikowego (fotodiody, tranzystora) [2,3]. Wynika on stąd, że prąd składa się z olbrzymiej liczby impulsów prądowych związanych z przepływem pojedynczych nośników ładunku. Podobnie jak szum cieplny, szum śrutowy jest szumem białym, a średniokwadratowy prąd szumów wyraża się wzorem

$$i_s^2 = 2qIB. \tag{2}$$

Tutaj q jest ładunkiem elektronu ($1{,}6 \times 10^{-19}$ C), zaś I prądem stałym płynącym poprzez dany element.

Główną przyczyną występowania szumu 1/f są zjawiska występujące na powierzchni materiałów (generacja/rekombinacja nośników) oraz przepływ prądu przez ośrodek nieciągły. Widmowa gęstość mocy tych szumów daje się opisać zależnością typu $1/f^a$, gdzie a jest zazwyczaj bliskie jedności [2].

Szum wybuchowy pojawia się w niektórych elementach półprzewodnikowych. Jego źródłem są defekty materiałowe, zaś jego gęstość widmowa jest wyrażana przez zależność typu $1/f^a$, gdzie a jest zazwyczaj zbliżone do dwóch. Przy prawidłowej pracy wpływ tego szumu jest możliwy do pominięcia [2].

Zakładając, że szumy niskoczęstotliwościowe mają gęstość widmową odwrotnie proporcjonalną do częstotliwości, można gęstość widmową szumów dla niskich częstotliwości wyrazić wzorem

$$N(f) = A(1 + f_d/f), \tag{3}$$

gdzie A jest pewną stałą, zaś f_d jest częstotliwością, dla której gęstość widmowa szumów wzrasta dwukrotnie w porównaniu z wartością asymptotyczną. Z danych katalogowych różnych elementów wynika, że f_d jest rzędu 10 ... 1000 Hz [4,5]. Jest oczywiste, że częstotliwość pracy urządzenia powinna znacznie przekraczać f_d,. Nie może ona jednak być ona zbyt duża ze względu na wzrost współczynnika szumów spowodowany pojemnością bocznikującą. Zostanie to dokładniej przedyskutowane w następnym punkcie.

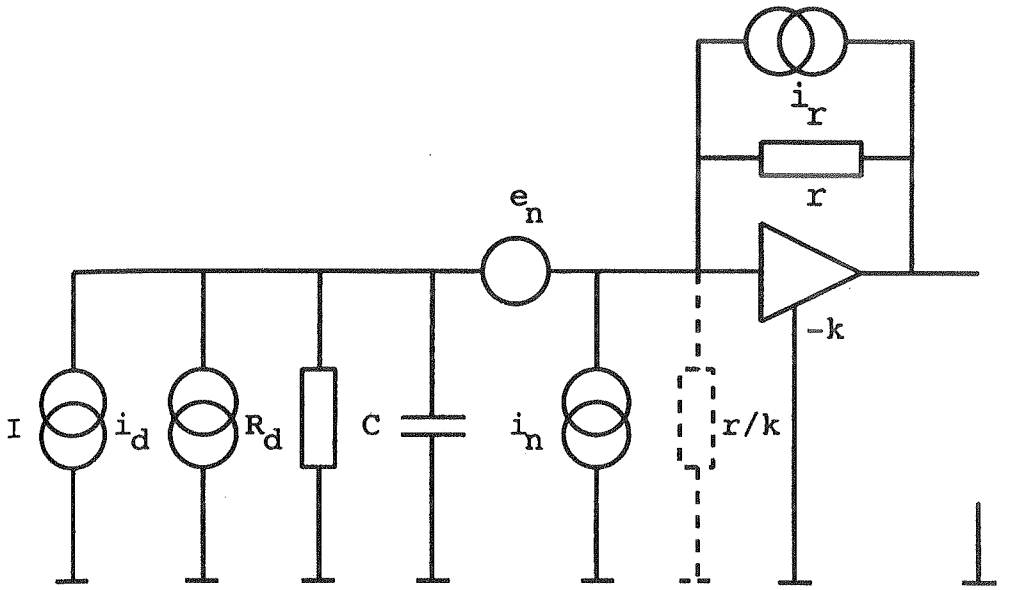
3.2. WYBÓR KONFIGURACJI ODBIORNIKA

W omawianym przypadku można w praktyce rozpatrywać jedynie dwie konfiguracje odbiornika: odbiornik o wysokiej impedencji oraz odbiornik transimpedancyjny [6]. Wybór konfiguracji jest podyktowany parametrami fotodetektora. Ponieważ wymagany był szeroki zakres długości fal (0,4...1,7 μm) oraz duża powierzchnia (> 1 mm^2) wybrano fotodiodę p-i-n, gdyż nie dysponowano fotodiodami lawinowymi o żądanych parametrach. W zakresie podczerwieni użyto fotodiody p-i-n InGaAs typu C30665 o średnicy 3 mm, natomiast w zakresie widzialnym zastosowano fotodiodę krzemową. Obydwa fotodetektory wraz z podwzmacniaczami umieszczono w wymiennych głowicach. Ze względu na niewystarczające pasmo odbiornika o dużej impedencji, do realizacji wybrano konfigurację transimpedancyjną, która jest szeroko stosowana w różnego rodzaju odbiornikach optycznych np. systemów światłowodowych [3,6]. Dostępne są również realizacje scalone takich odbiorników. Schemat odbiornika transimpedancyjnego wraz ze źródłami szumów pokazano na rys. 6. Pasmo B takiego odbiornika w przypadku b. dużej impedencji wejściowej i dużego wzmocnienia k jest określone przez efekt Millera i wynosi

$$B = \frac{k+1}{2\pi rC}. \tag{4}$$

W omawianym przypadku pasmo jest rzędu kilkudziesięciu kHz, co jest zupełnie wystarczające. Ze względu na trudności realizacyjne ze stabilnością wzmacniaczy zbudowanych z elementów dyskretnych zdecydowano się na realizację scaloną odbiornika w postaci wzmacniacza operacyjnego. O jakości odbioru, oprócz dostarczanego pasma, decyduje współczynnik szumów stopnia wejściowego.

Na rys. 6 przyjęto następujące oznaczenia gęstości widmowych szumów:
– $i_r^2 = 4kT/r$ — prąd szumów rezystora sprzężenia zwrotnego, np. dla r = 10 MΩ — $i_1 = 0{,}4$ pA/Hz $^{1/2}$,
– e_n, i_n — odpowiednio napięcie i prąd szumów wzmacniacza,
– $i_d = 2qI_c + 4kT/R_d$ — prąd szumów fotodiody (I_c jest prądem ciemnym), dla danego elementu i_d = 0,04..0,2 pA/$Hz^{1/2}$; w projektowaniu przyjęto i_d = 0,14 pA/$Hz^{1/2}$.



Rys. 6. Szumowy schemat zastępczy transimpedancyjnego stopnia wejściowego

Przy pominięciu szumów 1/f współczynnik szumów F wyraża się zależnością

$$F = \frac{i_d^2 + i_r^2 + i_n^2 + e_n^2(1/R_d^2 + 4\pi^2 f^2 C^2)}{i_d^2} \tag{5}$$

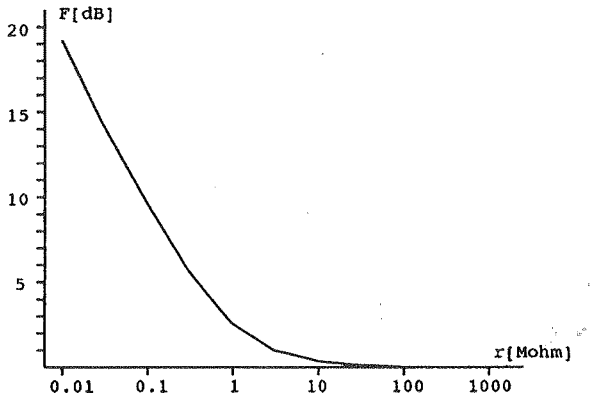
Ze zrozumiałych względów współczynnik szumów powinien być jak najbliższy jedności. Wartości parametrów szumowych wzmacniacza zależą przede wszystkim od użytej technologii. Typowe wartości [4,5] dla wzmacniaczy bipolarnych to e_n równe kilka $nV/Hz^{1/2}$, zaś i_n równe kilka $pA/Hz^{1/2}$; dla wzmacniaczy FET e_n wynosi od kilku do kilkudziesięciu $nV/Hz^{1/2}$, zaś i_n – 0,1...10 $fA/Hz^{1/2}$.

Ze wzoru (5) wynika, że wzrost częstotliwości pracy (w badanym przypadku powyżej 1 kHz) nie jest korzystny ze względu na wzrost szumów związanych z e_n^2. Jeśli częstotliwość pracy $f < 1$ kHz, to wówczas ze względu na dużą wartość R_d dominującym źródłem szumów wzmacniacza jest i_n^2. Dlatego też w układach optymalizowanych szumowo należy użyć wzmacniacza FET, który ma prąd szumów o kilka rzędów wielkości niższy aniżeli wzmacniacz bipolarny. Dodatkową zaletą wzmacniacza FET jest bardzo duża rezystancja wejściowa $r_{we} > 10^{12}$ Ω oraz znacznie niższa częstotliwość graniczna szumów 1/f, zwłaszcza w odniesieniu do prądu szumów.

W wypadku wzmacniacza FET i optymalnej częstotliwości pracy współczynnik szumów można przybliżyć zależnością

$$F \approx 1 + \frac{i_r^2}{i_d^2} \tag{6}$$

Zależność współczynnika szumów od rezystancji r pokazano na rys. 7. Przy rezystancji r > 10 MΩ wzrost poziomu szumów jest minimalny.



Rys. 7. Zależność współczynnika szumów od rezystancji r

Innym ważnym czynnikiem, który należy wziąć pod uwagę przy projektowaniu jest maksymalny poziom sygnału, jaki może być odebrany bez zniekształceń. Im większa transimpedancja r, tym ten poziom jest mniejszy. Wniosek jest prosty: fotoodbiornik optymalizowany szumowo nie nadaje się do odbioru silnych sygnałów optycznych ze względu na możliwość przesterowania. Istnieje kilka rozwiązań problemu ograniczenia dynamiki optymalizowanych szumowo wzmacniaczy transimpedancyjnych. Wymienimy je kolejno [6].

1. Zastosowanie w pętli sprzężenia zwrotnego zamiast rezystora – tranzystora FET, pracującego jako rezystancja sterowana napięciem z pętli ARW. Wadą tego rozwiązania jest to, że przy wzroście napięcia wyjściowego tranzystor FET zaczyna się zachowywać jako źródło prądowe, a nie sterowana rezystancja.

2. Użycie aktywnego sprzężenia zwrotnego w postaci pojemności sterowanej z wzmacniacza całkującego, która zachowuje się jako rezystor. Wadą tego rozwiązania jest jego złożoność i problemy z zapewnieniem stabilności układu. Ponadto obydwa wymienione do tej pory rozwiązania jedynie nieznacznie poprawiają dynamikę odbiornika.

3. Zastosowanie na wejściu tranzystora FET bocznikującego fotodiodę i pracującego jako rezystancja sterowania napięciowo z pętli ARW. Zaletą tego rozwiązania jest to, że napięcie dren- źródło jest zawsze bliskie zera (ze względu na punkt masy pozornej), co pozwala na prawidłową pracę. Natomiast zasadniczą wadą jest to, że taki tranzystor wprowadza na wejście swoje szumy pogarszając współczynnik szumów odbiornika.

Wspólną wadą wszystkich powyższych rozwiązań jest po pierwsze brak realizacji scalonych omówionych układów, a po drugie trudność z realizacją pętli ARW, która ze względu na długie czasy skanowania monochromatora musiałaby mieć złożoną strukturę analogowo-cyfrową.

Z tych przyczyn zdecydowano się na zastosowanie wymiennych głowic, zmienianych w zależności od zakresu długości fal i poziomu mocy badanego sygnału.

3.3. WYBÓR KSZTAŁTU APERTURY CHOPPERA

Ponieważ dla zwiększenia czułości urządzenia strumień świetlny pochodzący z badanego elementu jest modulowany zachodzi natychmiast pytanie, jaki powinien być kształt apertury urządzenia modulującego (choppera). Musi on być oczywiście tego rodzaju, aby zapewnić maksymalną zawartość podstawowej harmonicznej w zmodulowanym sygnale f(t). Moc maksymalna sygnału jest ograniczona właściwościami źródła światła. Stąd należy rozwiązać zagadnienie maksymalizacji zawartości składowej podstawowej przy ograniczeniu mocy sygnału tzn.

$$\int_0^{\pi} f(\alpha)\sin\alpha\, d\alpha = \max, \qquad f(t) < P_{\max} \tag{7}$$

Nietrudno jest wykazać, że rozwiązaniem tego zagadnienia jest sygnał prostokątny o wypełnieniu 1/2 (a nie sinusoida, jak możnaby się spodziewać na pierwszy rzut oka). Zysk z zastosowania sygnału prostokątnego jest dosyć duży. Jego moc (elektryczna — w odbiorniku) jest 2,1 dB większa aniżeli sygnału sinusoidalnego i np. aż o 3,9 dB większa aniżeli sygnału trójkątnego. Aby uzyskać sygnał maksymalnie zbliżony do sygnału prostokątnego, należy wyeliminować strefy przejściowe przy przechodzeniu od pełnego otwarcia przesłony modulatora do jej pełnego zamknięcia. Stąd wniosek, że średnica modulowanej wiązki powinna być jak najmniejsza w porównaniu z wymiarami otworów i przesłon choppera. W granicznym przypadku chopper powinien mieć kształt półkola osadzonego na osi. W praktyce jednak liczba przesłon choppera jest ograniczona od dołu maksymalną prędkością obrotową silnika sterującego.

4. BŁĘDY POMIAROWE

4.1. DOKŁADNOŚĆ POMIARU DŁUGOŚCI FALI

Dokładność pomiaru długości fali zależy od użytego monochromatora. Do współpracy z omawianym urządzeniem użyto monochromatora Digikrom CM110 [1]. Błąd określenia wartości długości fali podany przez producenta urządzenia wynosi ±0,2 nm, rozdzielczość jest lepsza od 0,2 nm, zaś pasmo przepustowe jest mniejsze od 1 mm przy szczelinie 0,3 mm. Należy przy tym pamiętać, że długość fali jest związana z kątem obrotu w sposób nieliniowy poprzez zależność:

$$\lambda = \frac{2d\cos\phi\,\sin\Theta}{n}, \tag{8}$$

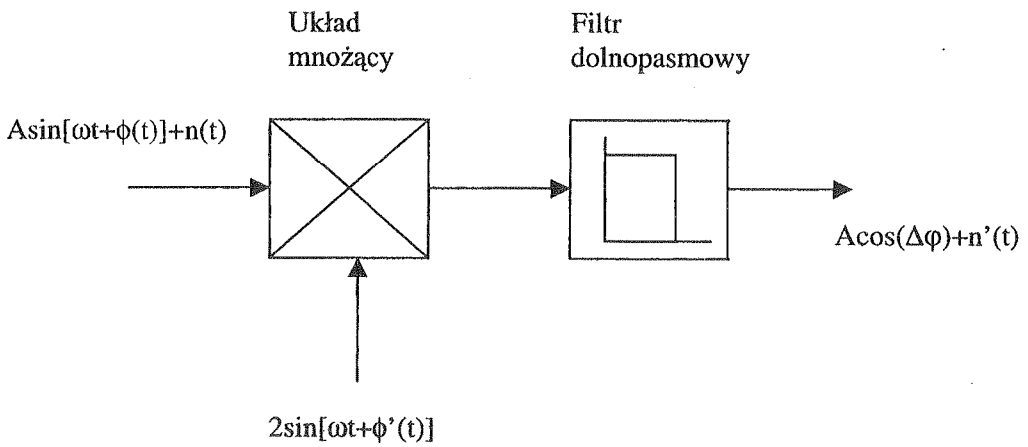
gdzie d — odstęp między rysami w siatce dyfrakcyjnej w m, ϕ — tzw. kąt Eberta, określony przez wzajemne pozycje siatki i obydwu luster i równy w opisywanym przypadku 12.7 stopnia, Θ — kąt obrotu siatki mierzony od punktu odbicia zwierciadlanego światła bbiałego, zaś n jest rzędem dyfrakcji (n = −/+1). Powyższa zależność musi być uwzględniona przy obróbce wyników pomiaru, gdyż przy różnych kątach obrotu obrót o kąt dyskretyzacji (najmniejszy kąt) odpowiada różnej zmianie długości fali. Przy użytej siatce (1200 rys/mm) maksymalny przyrost długości fali dla kroku kątowego wynosi 0.0076 nm.

4.2. DOKŁADNOŚĆ POMIARU GĘSTOŚCI WIDMOWEJ

Dla określenia błędów związanych z pomiarami gęstości widmowej badanego źródła światła należy najpierw przedyskutować zasadę pomiaru homodynowego, która jest wykorzystana w opracowanym urządzeniu. Jak wiadomo pomiary homodynowe są stosowane w przypadku pomiarów słabych sygnałów dodatkowo zakłóconych przez szumy. Zasadę pomiaru przedstawiono na rys. 8. W rozpatrywanym przypadku pomiarowi podlega amplituda napięcia A. Sygnał wejściowy jest mnożony w układzie mnożącym przez niezaszumiony sygnał odniesienia. Produkt mnożenia jest przepuszczany przez filtr dolnopasmowy w celu eliminacji szumów i drugiej harmonicznej. Błąd przypadkowy pomiaru zależy od gęstości widmowej szumów i pasma filtru dolnoprzepustowego. Względny błąd przypadkowy wynosi

$$\sigma = \frac{\sqrt{BN'}}{A} \tag{9}$$

gdzie B jest pasmem szumowym filtru, zaś N′ — gęstością widmową szumu na wejściu filtru dolnoprzepustowego. Z zależności (9) wynika, że błędy przypadkowe rosną wraz ze wzrostem szerokości pasma filtru dolnoprzepustowego B.



Rys. 8. Zasada pomiaru homodynowego

Błąd systematyczny pomiaru zależy od rozfazowania (różnicy faz) sygnału mierzonego i sygnału odniesienia:

$$\Delta\phi = \phi(t) - \phi'(t) \tag{10}$$

W przypadku spotykanych praktycznie układów pomiarowych różnica faz obydwu sygnałów jest niewielka ($\Delta\phi << 1$) i ma charakter przypadkowy. Wówczas systematyczny błąd pomiaru (względny) można przybliżyć zależnością

$$\delta A_w = \frac{E\{A\cos\Delta\phi) - A}{A} = E\{\cos\Delta\phi)\} - 1 \approx E\{(\Delta\phi)^2\}/2 = \delta^2/2 \tag{11}$$

gdzie δ jest odchyleniem standardowym różnicy faz wyrażonym w radianach. Wynika stąd, że systematyczny błąd pomiaru jest w zasadniczym stopniu wyznaczony przez fluktuacje fazowe pomiędzy sygnałem mierzonym a sygnałem odniesienia.

Sygnał odniesienia jest najczęściej uzyskiwany w pętli synchronizacji fazowej [7]. Średniokwadratowy błąd śledzenia takiej pętli, równoważny wariancji omawianej różnicy faz δ^2 może być wyrażony wzorem [7]

$$\delta^2 = \delta_\phi^2 + \delta_n^2 \tag{12}$$

Pierwszy składnik po prawej stronie oznacza składnik błędu związany z niestabilnością fazową śledzonego sygnału, zaś drugi- jest spowodowany szumami addytywnymi w odbieranym sygnale. W powyższym wzorze przyjęto brak modulacji sygnału oraz brak przesunięcia Dopplera w odbieranym sygnale. Jeśli funkcja przenoszenia zamkniętej pętli synchronizacji fazowej jest wyrażana przez H(f), to wówczas obydwie składowe błędu śledzenia wyrażają się zależnościami [7]:

$$\delta_\phi^2 = \int_{-\infty}^{+\infty} |1 - H(f)|^2 S_\phi(f)\, df$$

$$\delta_n^2 = \int_{-\infty}^{+\infty} |H(f)|^2 \frac{S_n(f)}{A^2}\, df \tag{13}$$

W powyższych wzorach przez S z odpowiednim indeksem wyrażone są gęstości widmowe szumu fazowego i szumu addytywnego.

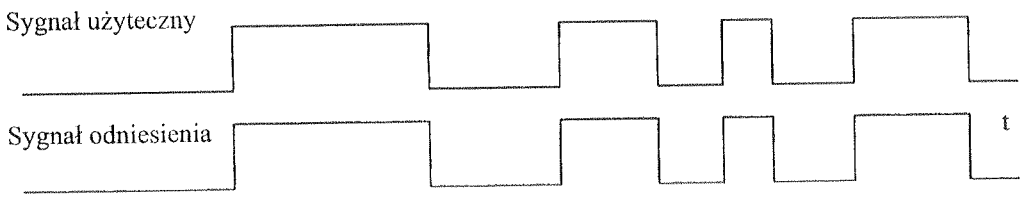
Wymagania na jednoczesną minimalizację obydwu błędów są sprzeczne. Im jest mniejsze pasmo pętli W, określone zależnością

$$W = \int_{-\infty}^{+\infty} |H(f)|^2 df \tag{14}$$

tym mniejszy jest błąd wywołany szumami, ale tym większy błąd spowodowany niestabilnością fazową. Dla pętli fazowej drugiego rzędu można pokazać [7], że

$\delta_\phi^2 \sim 1/W$ zaś $\delta_n^2 \sim W$, a zatem znając widma obydwu szumów można dobrać optymalne pasmo pętli.

W przypadku skonstruowanego układu sygnał odniesienia jest uzyskiwany poprzez modulację amplitudową sygnału dodatkowego transoptora tą samą tarczą obrotową, która jest stosowana do modulacji amplitudy sygnału użytecznego. Kiedy obydwie grupy otworów w tarczy (tzn. otwory dla modulacji sygnału użytecznego i transoptora) zajmują identyczne pozycje kątowe na tarczy i mają podobny kształt (tzn. kształt uzyskiwanych impulsów jest identyczny) wówczas w pierwszym przybliżeniu pomiar nie jest wrażliwy na ewentualne fluktuacje prędkości obrotowej wirującej tarczy. Dzieje się tak pomimo tego, że wspomniane fluktuacje prowadzą do poszerzenia widma każdego z tych sygnałów. Jest to oczywiste, gdyż obydwie grupy otworów są ze sobą sztywno związane, a zatem jakiekolwiek zmiany fazy sygnału użytecznego pociągają za sobą identyczne zmiany sygnału odniesienia. Pokazano to na rysunku 9.



Rys. 9. Przebieg sygnału użytecznego i odniesienia w przypadku fluktuacji częstotliwości obrotowej tarczy (choppera)

W praktyce jednak istnieją pewne tolerancje wykonania mechanicznego otworów; otwory nie zajmują również dokładnie identycznych pozycji kątowych na tarczy. Ten drugi fakt jest przyczyną tego, że w pewnym, minimalnym stopniu zmiany częstotliwości prędkości obrotowej są przenoszone na fluktuacje różnicy faz sygnału odniesienia i sygnału użytecznego. Jednakże zmierzona wartość średniokwadratowa tych fluktuacji, powodowana niedokładnością wykonania mechanicznego tarczy oraz niestabilnością częstotliwości obrotowej nie przekraczała 1% kąta pełnego (0,06 rad), co daje względny systematyczny błąd pomiaru amplitudy na poziomie 0,2%.

4.3. BŁĘDY ZWIĄZANE Z NIEDOKŁADNOŚCIĄ OKREŚLENIA CZUŁOŚCI WIDMOWEJ FOTODETEKTORA

Wskazania przyrządu były korygowane zgodnie z katalogową zależnością czułości widmowej fotodetektora $R(\lambda)$, gdyż brak było możliwości sprawdzenia rzeczywistego przebiegu czułości. Dla określenia wpływu błędu określenia przebiegu czułości fotodetektora na dokładność pomiaru częstotliwości środkowej widma przyjmijmy, że widmo gęstości mocy widmowej można w pobliżu maksimum λ_0 przybliżyć zależnością

$$S(\lambda) = 1 - \alpha(\lambda - \lambda_0)^2 \tag{15}$$

Natomiast zależność błędu czułości fotodetektora od długości fali można wyrazić przez

$$\Delta R = 1 + \beta(\lambda - \lambda_0) \tag{16}$$

Tutaj β określa błąd nachylenia, przy czym ograniczono się jedynie do członu pierwszego rzędu. Wypadkowa (obserwowana na ekranie) charakterystyka widmowa będzie miała kształt

$$S_w(\lambda) = \Delta R \cdot S(\lambda) = [1 - \beta(\lambda - \lambda_0)][1 - \alpha(\lambda - \lambda_0)^2] \tag{17}$$

Błąd określenia środkowej długości fali będzie różnicą między maksimum powyższej funkcji a dokładną wartością tej długości fali, λ_0. Szukany błąd łatwo jest wyliczyć znajdując maksimum funkcji $S_w(\lambda)$ poprzez jej zróżniczkowanie względem długości fali i przyrównanie do zera. Uwzględniając to, że parametr α jest związany z szerokością linii widmowej (FWHM) B zależnością $\alpha = 2/B^2$ otrzymujemy szukane wyrażenia na błąd pomiaru $\Delta\lambda$ w postaci

$$\Delta\lambda = \frac{\beta}{2\alpha} = \beta\frac{B^2}{4} \tag{18}$$

Szacując niedokładność określenia czułości β w zakresie pomiarowym 500 nm na 20% (β = 0,0004/nm), możemy obliczyć poszukiwany błąd. Jak wynika z powyższego wzoru błąd bezpośrednio zależy od szerokości linii widmowej mierzonego elementu. W przypadku lasera wielomodowego o szerokości widmowej 5 nm błąd wynosi 0,01 nm, zaś w przypadku diody elektroluminescencyjnej o szerokości linii widmowej 50 nm błąd wynosi 1 nm.

Przyjęcie podobnych do powyższych założeń przy obliczaniu błędu wyznaczenia szerokości linii widmowej prowadzi do wniosku, że omawiany błąd jest znacznie mniejszy. Istotnie, błędy wypadkowej charakterystyki na krańcach pasma przepustowego wynoszą odpowiednio $-\beta B/4$ oraz $+\beta B/4$. Pochodna funkcji Sw(λ) wyraża się wzorem

$$S_w'(\lambda) = -3\alpha/\beta x^2 - 2\alpha x + \beta,$$

gdzie (19)

$$x = \lambda - \lambda_0.$$

Stąd nachylenie charakterystyki na krańcach pasma wynosi odpowiednio

$$\begin{aligned} S_w'(\lambda_0 + B/2) &= -3\alpha\beta x^2 - 2\alpha x + \beta = -2/B - \beta/2 \\ S_w'(\lambda_0 - B/2) &= -3\alpha\beta x^2 + 2\alpha x + \beta = +2/B - \beta/2 \end{aligned} \tag{20}$$

Dlatego szukany błąd określenia szerokości linii widmowej ΔB wyraża się zależnością

$$\Delta B = \frac{\frac{\beta B}{4}}{2/B-\beta/2} - \frac{\frac{\beta B}{4}}{2/B+\beta/2} = \frac{\beta B^2}{8}\left(\frac{1}{1-\beta B/4} - \frac{1}{1+\beta B/4}\right) \approx \beta^2 B^3/16 \qquad (21)$$

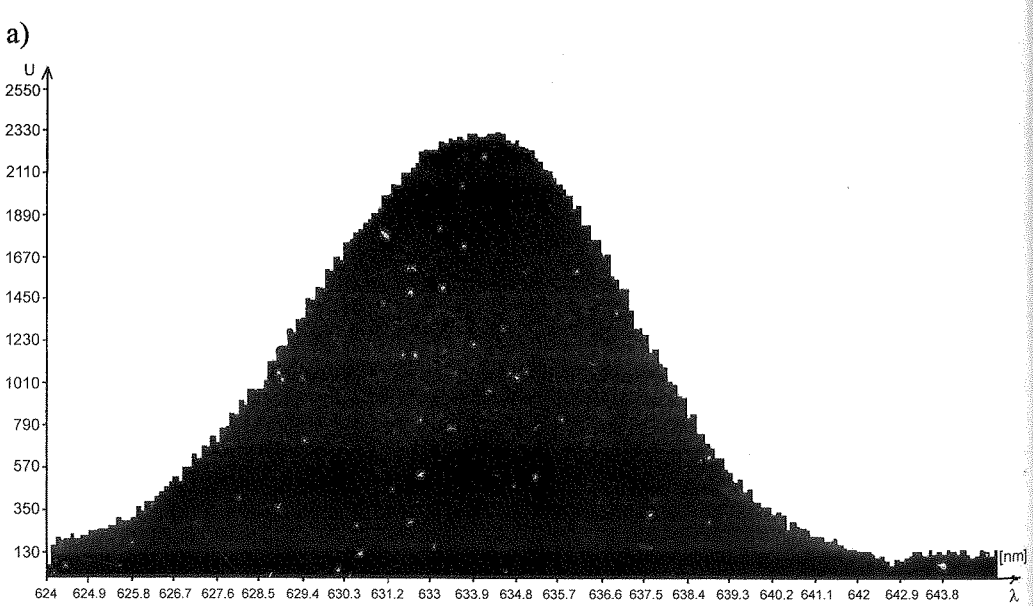
Założono tu, że $\beta B/4 << 1$.

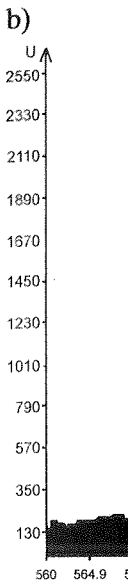
Wykonując obliczenia dla poprzednich danych (B = 5 i 50 nm) otrzymujemy, że szukane błędy wyznaczenia szerokości pasma są rzędu odpowiednio jednej milionowej i jednej tysięcznej nanometra.

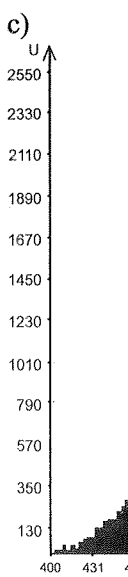
Wynika stąd, że niedokładność w określeniu czułości fotodetektora nieznacznie przesuwa charakterystykę widmową na osi długości fali, zaś w mniejszym stopniu oddziaływuje na jej szerokość.

5. PRZYKŁADOWE WYNIKI POMIARÓW

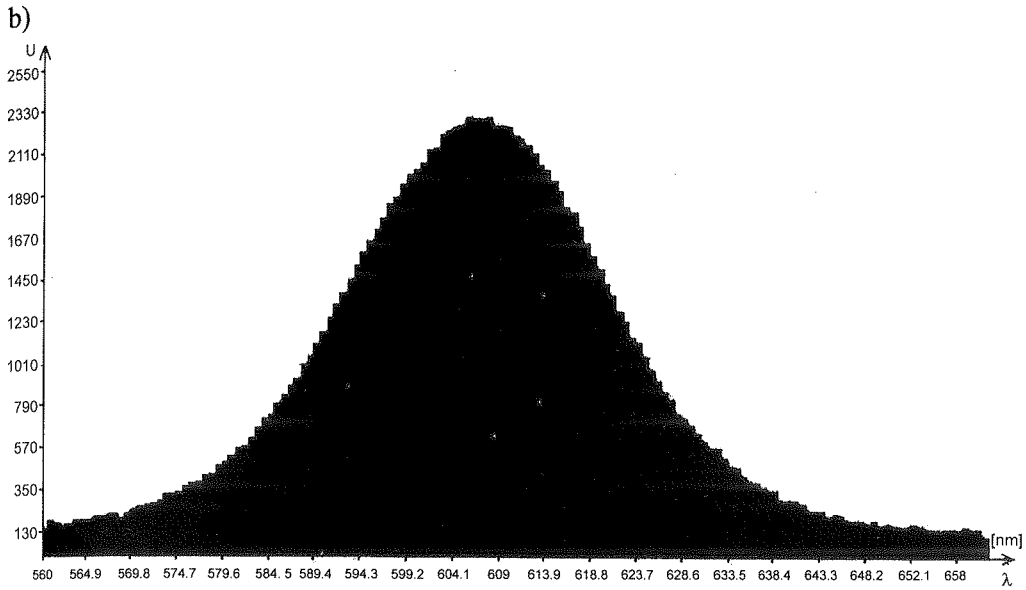
Przykładowe wyniki pomiarów różnych źródeł światła omawianym urządzeniem pokazano na rys. 10.

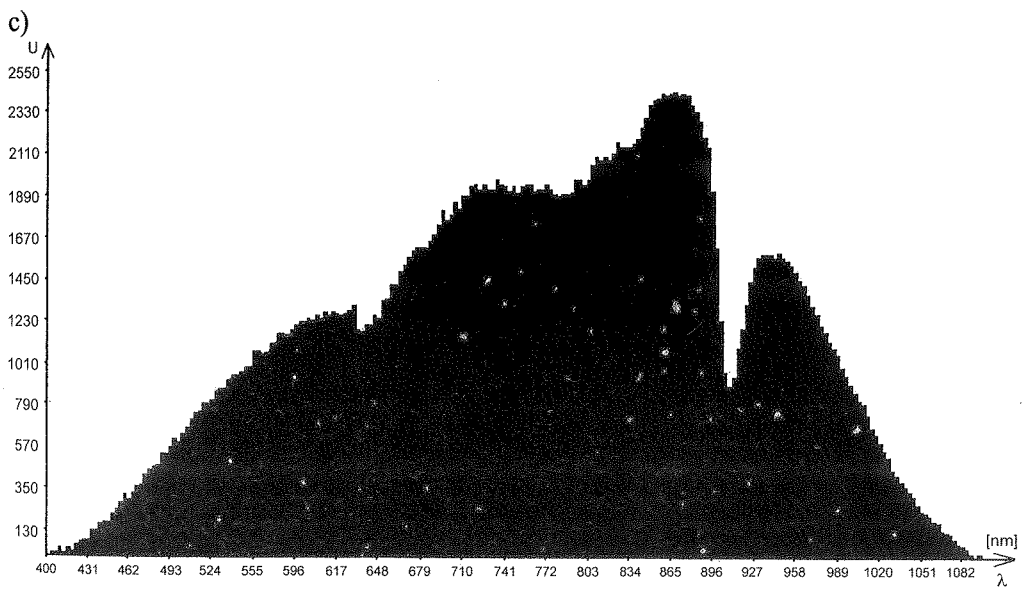






Rys. 10. Pr





Rys. 10. Przykładowe wyniki pomiarów: a) laser wielomodowy, b) dioda LED, c) światło białe po propagacji w światłowodzie polimerowym

6. PODSUMOWANIE

W pracy opisano homodynowy układ pomiaru mocy słabych sygnałów optycznych, który jest przeznaczony do współpracy z monochromatorem. Zamieszczono również krótki opis współpracującego z nim oprogramowania. Najważniejszym problemem projektowym była optymalizacja szumowa fotoodbiornika. Z przeprowadzonej analizy błędów pomiarowych wynika, że oprócz błędów związanych z monochromatorem najistotniejszą rolę odgrywają błędy związane z niedokładnością określenia zależności czułości fotodiody od długości fali.

BIBLIOGRAFIA

1. *CVI Laser Corporation, CM110 1/8 meter monochromator.* Instrukcja przyrządu
2. C.D. Motchenbacher, F.C. Fitchen: *Projektowanie elementów i układów elektronicznych niskoszumowych.* Warszawa, WNT, 1977
3. D.J.H. Maclean: *Optical Line Systems.* Chichester, John Wiley & Sons, 1996
4. Texas Instruments, Linear Circuit Data Book, katalog
5. Maxim, karty katalogowe układów liniowych
6. T. Li [Ed]: *Topics in Lightwave Transmission Systems.* Boston, Academic Press, 1991
7. W.C. Lindsey: *Synchronization systems in communication and control.* Englewood Cliffs, Prentice Hall, 1972.

J. SIUZDAK, R. NOWAK, T. CZARNECKI

Homodyne system for weak optical signal detection

Summary

In the paper, there is described a homodyne system for weak optical signal power detection meant for monochromators. The software used is also shortly described. Main design problem was noise optimisation of the photoreceiver. The measurement errors were also analysed. It follows that monochromator errors and photodetector sensitivity errors play the main role.

Key words: homodyne detection, monochromator, optical spectrum

KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJI, 2000, 46, z. 4, ss. 511–552

Stabilność częstotliwości rezonansowej rezonatorów z ceramiki piezoelektrycznej a ich struktura domenowa

ZYGMUNT SUROWIAK, DARIUSZ BOCHENEK

Katedra Materiałoznawstwa,
Wydział Techniki, Uniwersytet Śląski

SVETLANA V. GAVRILYACHENKO, M.F. KUPRIYANOV

Instytut Fizyki, Uniwersytet Rostowski

Otrzymano 2000.04.03
Autoryzowano 2000.09.03

Piezoceramiczne materiały na bazie PZT znajdują zastosowanie jako pasmowe filtry elektryczne. Materiały takie obok odpowiednich wartości parametrów elektrofizycznych muszą wykazywać wysoką temperaturową stabilność częstotliwości rezonansowej ($\Delta f_r/f_r < 0{,}2\%$). Stabilność f_r piezoceramiki na bazie dwuskładnikowego roztworu $Pb(Zr_{1-x}Ti_x)O_3$ jest niewielka ($\Delta f_r/f_r > 0{,}5\%$). Dlatego też prowadzone są intensywne badania mające na celu zmniejszenie względnej zmiany częstotliwości rezonansowej w roboczym zakresie temperatur (213–358 K). Na podstawie danych literaturowych i badań własnych przedstawiono podstawowe metody zwiększania temperaturowej stabilności f_r, a w tym: długotrwałe wyprażanie, hartowanie, napromieniowanie, kompensowanie, dobór warunków technologicznych, dobór warunków polaryzowania, termocyklowanie, modyfikowanie składu chemicznego. Wykazano, że najbardziej efektywną metodą w tym zakresie jest dobór składu chemicznego i stosowanie wieloskładnikowych roztworów stałych typu $PbTiO_3 - PbZrO_3 - \sum_{i=1}^{n} Pb(B'_{1-a}B''_a)O_3$.

Słowa kluczowe: rezonator piezoelektryczny, ceramika typu PZT, częstotliwość rezonansowa, struktura domenowa.

1. WPROWADZENIE

Filtr elektryczny jest to obwód bierny służący do przekazywania elektrycznej energii sygnałów, przepuszczający prądy o częstotliwościach zawierających się w określonym zakresie i nie przepuszczający prądów o częstotliwościach leżących poza tym zakresem.

W pewnym pasmie częstotliwości (p a s m o p r z e n o s z e n i a) drgania elektryczne przekazywane są z wejścia na wyjście praktycznie bez tłumienia, poza tym pasmem zaś są silnie tłumione.

Filtry elektryczne dzielą się na:

- dolnoprzepustowe (pasmo $0 < \omega \leq \omega_o$);
- górnoprzepustowe (pasmo $\omega > \omega_o$);
- pasmowe (pasmo $\omega_1 \leq \omega \leq \omega_2$);
- zaporowe (pasmo zaporowe $\omega_1 \leq \omega \leq \omega_2$).

Budowa filtrów elektrycznych i zasada ich działania może być różnorodna. Może to być liniowy czwórnik bierny (złożony z oporności czynnej R, kondensatorów C i cewek L), filtr magnetostrykcyjny lub filtr piezoelektryczny. Znajdują one szerokie zastosowanie w aparaturze elektronicznej (radiowej, TV) w łączności przewodowej, telefonii wielokanałowej, telemetrii i innych gałęziach teletechniki, piezoelektroniki itd.

Filtr piezoelektryczny jest to filtr LC mostkowy, różnicowy lub drabinkowy, którego określone grupy elementów LC w postaci obwodów rezonansowych, szeregowych, połączonych równolegle z pojemnością C są zrealizowane za pomocą rezonatorów piezoelektrycznych, tj. odpowiednio wyciętych i zamocowanych płytek piezoelektrycznych (krystalicznych, piezoceramicznych).

Pierwotnie piezoelektryczne filtry budowane były wyłącznie na bazie kwarcu i kryształów ferroelektrycznych (sól Seignette'a, TGS). Kwarc charakteryzuje się dużą dobrocią mechaniczną (Q_m) oraz szczególnie wysoką temperaturową stabilnością częstotliwości rezonansowej (f_r) i dokładnością dostrojenia wyższą od 10^{-4}. Wymagań takich nie jest w stanie spełnić żaden ze znanych materiałów piezoceramicznych i stąd też kwarc w pewnego rodzaju piezoelektrycznych filtrach jest materiałem niemożliwym do zastąpienia.

Materiały piezoceramiczne, charakteryzujące się znacznie wyższą od kwarcu wartością współczynników sprzężenia elektromechanicznego (k_p, k_t, $k_{ij} > 0{,}5$), znalazły jednak szerokie zastosowanie w filtrach pasmowych, w przypadku których dokładność dostrojenia 10^{-3} jest wystarczająca. Szerokość pasma przenoszenia filtrów zbudowanych na bazie piezoceramiki, bez dołączania indukcyjności, jest proporcjonalna do kwadratu współczynnika sprzężenia elektromechanicznego. Tak więc jeśli nie uwzględniać cewki indukcyjnej, filtr kwarcowy ze współczynnikiem $k_{ij} \approx 0{,}1$ może przepuszczać pasmo o szerokości nie większej od 1% średniej częstotliwości, natomiast dysk z piezoceramiki PZT z $k_p = 0{,}5$ przepuszcza pasmo o szerokości 10% od średniej częstotliwości.

Pasmowe filtry z szerokością pasma przepuszczania kilku procentów od średniej częstotliwości znajdują zastosowanie w wielokanałowej telefonii (średnia częstotliwość $\overline{\omega} = 455$ kHz), telemetrii ($\overline{\omega} = 100$ kHz), przy analizie sygnałów, w wzmacniaczach wielostopniowych średniej częstotliwości itd. Możliwość zastosowań piezoceramicznych elementów do budowy filtrów pasmowych jest uwarunkowana temperaturową i czasową stabilnością ich częstotliwości rezonansowej (f_r). Względna zmiana ich częstotliwości rezonansowej ($\Delta f_r / f_r$) nie może przekraczać (0,1–0,2)%. Ostrość pasma przepuszczania dla filtru piezoelektrycznego jest uwarunkowana dobrocią mechaniczną (Q_m) piezoceramiki. Powinna być ona równa odwrotności współczynnika stabilności częstotliwości

rezonansowej, co oznacza, że np. dla $\Delta f_r / f_r = 0{,}001$, $Q_m = 1000$. Obecnie na skalę przemysłową produkowane są filtry piezoceramiczne o różnorodnej częstotliwości, szerokości pasma przenoszenia, ostrości pasma przenoszenia i stopniu temperaturowej stabilności f_r.

Najczęściej średnie częstotliwości filtrów piezoceramicznych mieszczą się w przedziale od 5 kHz do 10 MHz, a szerokość pasma od 0,5 do 10%. Filtry takie zawierają od jednego trójpasmowego elementu do 20 elementów w filtrach wieloczłonowych o optymalnych charakterystykach. W przypadku górnej granicy w/w przedziału częstotliwości, zamiast rezonatorów o drganiach radialnych zastosowano małe rezonatory o drganiach grubościowych, otrzymane poprzez naniesienie elektrod metalicznych o małej powierzchni na płytkę piezoceramiczną o dużej powierzchni. Wystąpić wtedy może czysty rezonans dla drgań grubościowych (zazwyczaj grubościowe rezonanse próbek piezoceramicznych nie mogą być wykorzystywane ze względu na dużą liczbę pobocznych rezonansów, wywołanych w wyniku sprzężenia z niskoczęstotliwościowymi modami drgań). Zarówno w piezoceramicznych, jak i kwarcowych filtrach na drganiach grubościowych wykorzystywuje się zjawisko pułapkowania.

Obszar zastosowań piezoceramiki w filtrach pasmowych poszerza możliwość nadawania polarnej osi różnych kierunków w różnych częściach tego samego piezoelementu, w wyniku odpowiedniego polaryzowania go zewnętrznym polem elektrycznym. Płytka piezoceramiczna może być na przykład z jednego końca polaryzowana w kierunku grubości, a w pozostałej części równolegle do powierzchni. Dla częstości bliskich podłużnego rezonansu taka płytka piezoceramiczna z odpowiednimi elektrodami będzie działać jak rezonansowy tranformator.

Zależność podstawowych parametrów piezoceramiki od temperatury jest zazwyczaj mniej lub bardziej złożona. Dla oceny termicznej stabilności są stosowane: współczynnik temperaturowy częstotliwości rezonansowej TK_{f_r} lub szybkość zmian (f_r) wraz z temperaturą (T) w formie współczynnika α_{f_r}.

Współczynnik TK_{f_r} jest definiowany jako procentowa zmiana f_r odpowiadająca zmianie temperatury o 1 deg:

$$TK_{f_r} = \frac{\Delta f_r}{f_r \Delta T}, \tag{1}$$

gdzie:

$\Delta T = T_2 - T_1$ jest różnicą temperatur, w których mierzono $f_r(T_1)$ i $f_r(T_2)$;

$\Delta f_r = f_r(T_2) - f_r(T_1)$ jest bezwzględną zmianą częstotliwości rezonansowej wywołaną zmianą temperatury; w praktyce najczęściej jedna z temperatur T_1 lub T_2 jest temperaturą pokojową ($T = 293$ K);

$\frac{\Delta f_r}{f_r}$ — względna zmiana częstotliwości rezonansowej wywołana zmianą temperatury o ΔT.

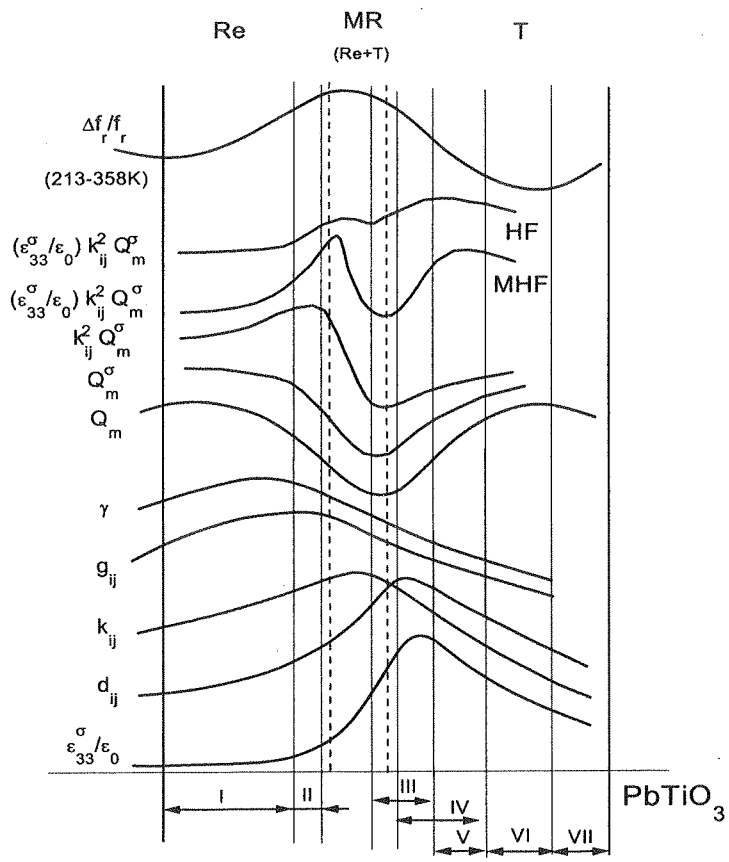
Współczynnik α_{f_r} jest zdefiniowany następująco:

$$\alpha_{f_r} = \frac{f_r(323\ \text{K}) - f_r(293\ \text{K})}{f_r(293\ \text{K})} \cdot 100\%, \tag{2}$$

gdzie:

f_r(323 K) — częstotliwość rezonansowa zmierzona w T_2 = 323 K = 50°C;

f_r(293 K) — częstotliwość rezonansowa zmierzona w T_1 = 293 K = 20°C.



Rys. 1. Zależność podstawowych parametrów piezoceramiki

$PbTiO-PbZrO_{3-}\sum_{i=1}^{n} Pb(B'_{1-a}B'')O_3$ od koncentracji $PbTiO_3$ [1];

Re — faza romboedryczna, T — faza tetragonalna, MR — obszar morfotropowy, HF — twardy ferroelektryk, MHF — średnio twardy ferroelektryk;

I — przetworniki wysokoczęstotliwościowe (małe $\varepsilon^{\sigma}_{33}/\varepsilon_o$); II — przetworniki o wysokiej czułości (duże g_{ij}); nIII — przetworniki niskoczęstotliwościowe (duże $\varepsilon^{\sigma}_{33}/\varepsilon_o$); IV — przetworniki niepodatne na oddziaływanie elektryczne i mechaniczne (ferroelektrycznie twarde, np. piezotransformatory); V — przetworniki niepodatne na oddziaływanie elektryczne i mechaniczne (ferroelektrycznie średnio twarde — np. generatory ultradźwięków); VI — przetworniki o wysokiej temperaturowej stabilności rezonansowej (małe $\Delta f_r/f_r$); VII — przetworniki wysokotemperaturowe (wysokie T_c)

Największa stabilność f_r (najmniejsza wartość $\Delta f_r/f_r$) wykazuje piezoceramika PZT o składach z obszaru tetragonalnego (T) i romboedrycznego (Re), położonych na

diagramie fazowym w znacznej odległości od dwufazowego obszaru morfotropowego MR (ang. *morphotropic region*). W obszarach tych optymalne wartości przyjmują również inne istotne dla działania filtrów parametry elektrofizyczne piezoceramiki (rys. 1). Takie to najczęściej składy PZT poddawane są różnym zabiegom, podwyższającym ich temperaturową stabilność częstotliwości rezonansowej.

Celem niniejszej pracy jest przedstawienie na podstawie danych literaturowych i badań własnych podstawowych metod stabilizowania struktury domenowej piezoceramiki typu PZT. Stabilność struktury domenowej warunkuje temperaturową stabilność częstotliwości rezonansowej piezoceramicznych rezonatorów.

2. WYNIKI BADAŃ I ICH OMÓWIENIE

2.1. KLASYFIKACJA METOD ZWIĘKSZANIA TEMPERATUROWEJ STABILNOŚCI f_r

Temperaturowa stabilność f_r uwarunkowana jest przede wszystkim charakterem struktury domenowej. Wskazuje na to silny związek między wartościami $\Delta f_r/f_r$ i Q_m^{-1} niezależnie od koncentracji $PbTiO_3$ w PZT (rys. 2). Pozwala to sądzić, że temperaturowa stabilność częstotliwości rezonansowej zależy w znacznym stopniu od tych samych czynników co i tarcie wewnętrzne, a więc od [2, 3]:

- budowy struktury domenowej,
- ruchliwości ścian domenowych.

Eksperymentalnym potwierdzeniem istnienia zależności temperaturowej stabilności częstotliwości rezonansowej od struktury domenowej, jest również jej związek z parametrami strukturalnymi δ i η.

Parametr δ charakteryzuje stopień deformacji regularnej komórki elementarnej, zachodzącej podczas przemiany fazy paraelektrycznej (*Pm3m*) w fazę tetragonalną (*P4mm*) lub w fazę romboedryczną (*R3m*) podczas chłodzenia PZT, o określonej koncentracji $PbTiO_3$, poprzez punkt Curie (T_c). W przypadku fazy tetragonalnej (T) jednorodny parametr deformacji wylicza się z wzoru [4]:

$$\delta = \frac{2}{3}\left(\frac{c_T}{a_T} - 1\right), \tag{3}$$

gdzie: a_T, c_T — parametry tetragonalnej komórki elementarnej PZT.

Dla fazy romboedrycznej (Re) jednorodny parametr deformacji przyjmuje wartość następującą [4]:

$$\delta \approx \cos \alpha, \tag{4}$$

gdzie: α — kąt (a_{Re}, c_{Re}),
$a_{Re} = c_{Re}$.

Jednorodność parametru δ pozwala na bezpośrednie porównywanie wartości deformacji komórek elementarnych o różnej symetrii (w naszym przypadku o symetrii układu T i Re).

Drugim parametrem (η), charakteryzującym strukturę domenową ferroelektryków, jest stopień różnych od 180° reorientacji domen, szczątkowych lub pod wpływem oddziaływań zewnętrznych (np. pod wpływem pola elektrycznego w procesie polaryzowania lub przepolaryzowania). Parametr η wyznaczano metodą rentgenowską zgodnie z metodyką opisaną w pracach [5, 6].

Dla fazy tetragonalnej posłużono się wzorem [1]:

$$\eta_T = \left(\frac{I_{002}}{I_{002}+I_{200}}\right)_p \cdot 100\% - \left(\frac{I_{002}}{I_{002}+I_{200}}\right)_0 \cdot 100\%, \tag{5}$$

natomiast dla fazy romboedrycznej wzorem [1]:

$$\eta_{Re} = \left(\frac{I_{222}}{I_{222}+I_{22\bar{2}}}\right)_p \cdot 100\% - \left(\frac{I_{222}}{I_{222}+I_{22\bar{2}}}\right)_0 \cdot 100\%, \tag{6}$$

gdzie:

I_{hkl} — integralne natężenia refleksów charakterystycznych dla fazy tetragonalnej (002/200) i dla fazy romboedrycznej (222/22$\bar{2}$), w stadium wyjściowym (indeks „0") i po przebudowie struktury domenowej (indeks „p").

Parametry δ i η są powszechnie przyjętymi charakterystykami struktury domenowej (rys. 2):

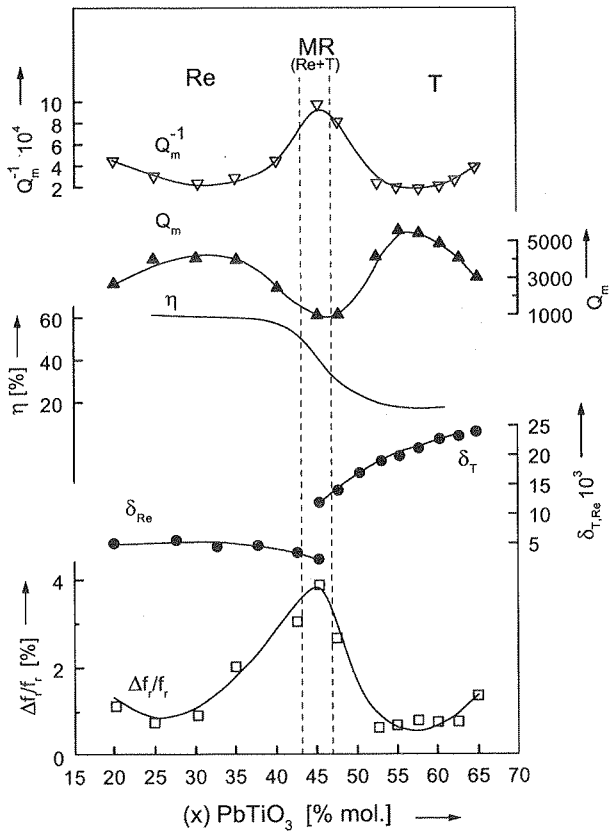
- jednorodny parametr spontanicznej deformacji (δ) można wiązać z ruchliwością ścian domenowych;
- stopień różnych od 180° reorientacji domen, szczątkowych lub podczas polaryzowania względnie przepolaryzowania (η), można wiązać z liczbą ścian domenowych,

Zmiany położenia granic domenowych pod wpływem temperatury, prowadzące do zmiany częstotliwości rezonansowej (f_r), są tym mniejsze, im mniejsze są ruchliwość i liczba ścian domenowych. Tymi dwoma czynnikami można także objaśnić obserwowaną zależność $\Delta f_r/f_r$ od koncentracji $PbTiO_3$ w badanym roztworze stałym (rys. 2).

Zależność dobroci mechanicznej (Q_m) od koncentracji $pbTiO_3$ również jest uwarunkowana charakterem struktury domenowej (ruchliwością ścian domenowych). Dobroć mechaniczna Q_m wzrasta bowiem (rys. 2):

- przy zmniejszeniu ruchliwości ścian domenowych, czyli przy wzroście δ;
- przy zmniejszeniu liczby ścian domenowych, czyli zwiększeniu η.

Opierając się na przytoczonych współzależnościach pomiędzy strukturalnymi parametrami δ i η — z jednej strony, oraz ruchliwością i liczbą ścian domenowych — z drugiej strony, można przyjąć, że wszystkie czynniki stabilizujące strukturę domenową piezoceramiki, a tym samym utrudniające zmianę tej struktury, zwiększają temperaturową i czasową stabilność f_r, Q_m i k_{ij} (k_p, k_t). Główne metody podwyższania temperaturowej stabilności częstotliwości rezonansowej przetworników piezoceramicznych na bazie PZT przedstawiono w tablicy 1.



Rys. 2. Zależność parametrów czteroskładnikowej piezoceramiki $PbTiO_3$–$PbZrO_3$–$PbNb_{2/3}Mn_{1/3}O_3$–$PbW_{1/2}Mg_{1/2}O_3$ od koncentracji (x) $PbTiO_3$[1]: Re – faza romboedryczna; T – faza tetragonalna; MR – dwufazowy obszar morfotropowy (Re+T); Q_m – dobroć mechaniczna, Q_m^{-1} – tarcie wewnętrzne; $\delta_{T,Re}$ – deformacja tetragonalna (T) lub romboedryczna (Re) regularnej komórki elementarnej; $\Delta f_r/f_r$ – względna zmiana częstotliwości rezonansowej f_r w przedziale temperatury 213–358 K; η – liczba różnych od 180° reorientacji domen, szczątkowych lub powstających w procesie polaryzowania.

W dalszej części pracy przedstawione zostaną podstawy wymienionych metod oraz rezultaty w zakresie zwiększania i temperaturowej stabilności f_r piezoceramicznych próbek na bazie PZT, uzyskane przez autorów niniejszej pracy i cytowanych publikacji.

Tablica 1

Metody zwiększania temperaturowej stabilności częstotliwości rezonansowej piezoceramicznych przetworników na bazie PZT

Lp.	Metoda	Przykładowa literatura
I	Dobór koncentracji $PbTiO_3$	[7–10]
II	Modyfikowanie podstawowego składu chemicznego	[11–14]

cd. tablicy 1

III	Dobór warunków syntezy i spiekania	[15–17]
IV	Kontrola mikrostruktury	[18–20]
V	Zwiększenie ferroelektrycznej twardości	[21–25]
VI	Dobór warunków polaryzowania	[16–26]
VII	Długotrwałe wyprażanie	[27–29]
VIII	Hartowanie	[29–30]
IX	Napromieniowanie	[31–33]
X	Kompensowanie	[34]
XI	Termocyklowanie	[35–40]

2.2. DOBÓR KONCENTRACJI $PbTiO_3$ (metoda I)

Rezultaty badań dotyczących wpływu koncentracji $PbTiO_3$ na stabilność parametrów elektrofizycznych przetworników piezoceramicznych na bazie PZT, a szczególnie na temperaturową stabilność f_r, przedstawione zostaną na przykładzie czteroskładnikowego roztworu stałego o następującym składzie chemicznym [7, 8]:

$$PbTiO_3 - PbZrO_3 - PbNb_{2/3}Mn_{1/3}O_3 - PbW_{1/2}Mn_{1/2}O_3. \tag{7}$$

Badania te obejmowały przede wszystkim ustalenie związków między koncentracją $PbTiO_3$ w wymienionym roztworze stałym, a:

- względną zmianą częstotliwości rezonansowej ($\Delta f_r/f_r$) w zakresie temperatury 213–358 K;
- tarciem wewnętrznym (Q_m^{-1});
- dobrocią mechaniczną (Q_m);
- stopniem różnych od 180° reorientacji domen, szczątkowych lub podczas procesu polaryzowania względnie przepolaryzowania w zewnętrznym polu elektrycznym (η);
- jednorodnym parametrem spontanicznej deformacji (δ).

Roztwór stały o składzie określonym równaniem (7) przy koncentracji $PbTiO_3$ w granicach od ~43 do ~47% mol. należy do dwufazowego obszaru morfotropowego (MR), w którym stabilność f_r jest najmniejsza ($\Delta f_r/f_r$ jest największe).

W miarę oddalania się na diagramie fazowym od obszaru morfotropowego (MR) w obszar składów o strukturze tetragonalnej (T) ruchliwość ścian domenowych zmniejsza się w wyniku wzrostu parametru δ (rys. 2). Taka ustabilizowana struktura domenowa decyduje o twardości ferroelektrycznej piezoceramiki, przejawiającej się między innymi we wzroście Q_m (zmniejszeniu Q_m^{-1}) i zmniejszeniu $\Delta f_r/f_r$ a tym samym we wzroście stabilności f_r (rys. 2). Obserwowany w fazie tetragonalnej wzrost

liczby ścian domenowych, na co wskazuje zmniejszenie parametru η, wywiera znacznie słabszy wpływ na stabilność f_r i Q_m piezoceramiki na bazie PZT niż wzrost parametru δ.

Przejściu na diagramie fazowym do składów o koncentracji $PbTiO_3$ niższej od koncentracji odpowiadającej obszarowi morfotropowemu (MR) towarzyszy przemiana struktury PZT w fazę romboedryczną (Re). Po tej przemianie obserwuje się wzrost stabilności częstotliwości rezonansowej f_r (zmniejszenie $\Delta f_r/f_r$) i wzrost dobroci mechanicznej Q_m (zmniejszenie tarcia wewnętrznego Q_m^{-1}). Zmiany te są mniejsze niż w przypadku przejścia od MR do fazy tetragonalnej, a uwarunkowane są silnym zmniejszeniem liczby ścian domenowych (wzrostem parametru η).

W zakresie składów PZT z dwufazowego (Re + T) obszaru morfotropowego $\Delta f_r/f_r$ osiąga maksimum (stabilność f_r jest najmniejsza), natomiast Q_m — minimum. W obszarze MR parametry δ i η mają wartości pośrednie w odniesieniu do faz T i Re. Stan taki charakteryzuje się dużą liczbą ścian domenowych (zmniejszenie η) i dużą ruchliwością ścian domenowych (zmniejszenie δ). Pewien wkład w temperaturową stabilność częstotliwości rezonansowej w dwufazowym obszarze morfotropowym może wnieść również ruch granic międzyfazowych (Re/T), wywołany zmianą temperatury piezoceramiki.

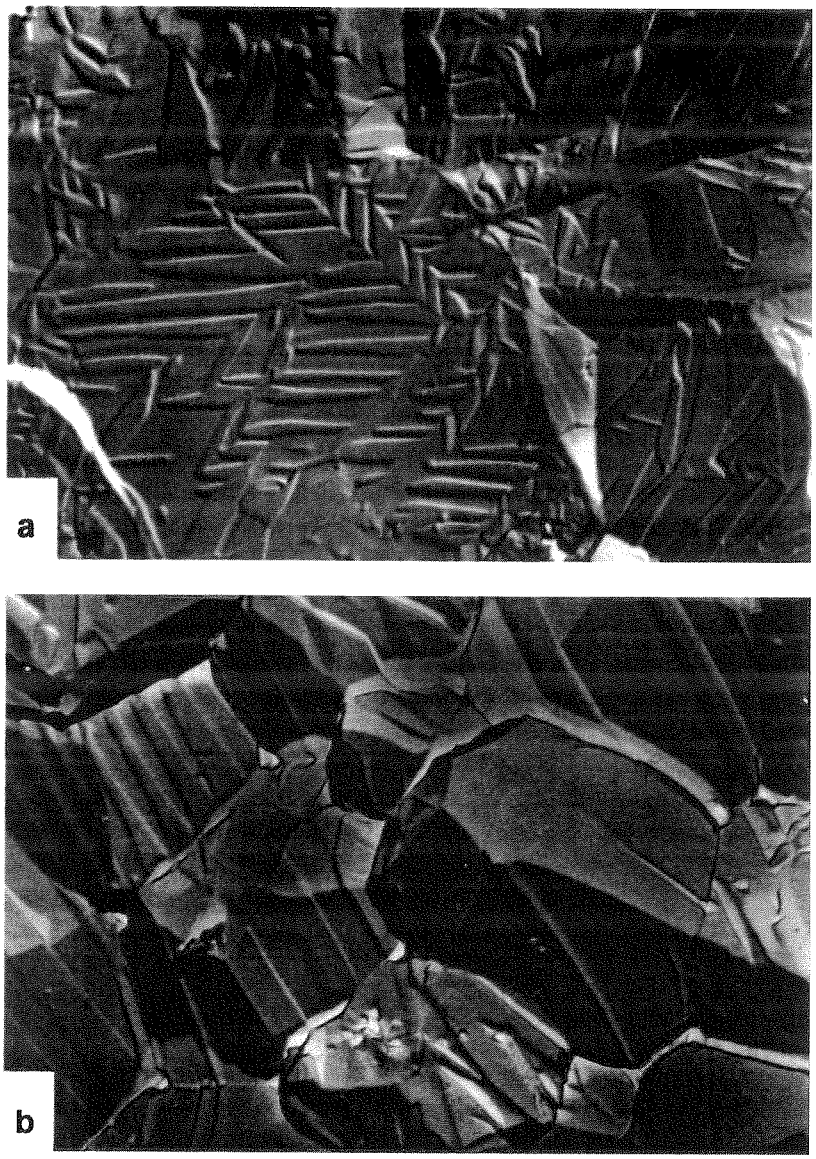
Przedstawione rezultaty badań świadczą pośrednio o ścisłym związku właściwości piezoceramiki na bazie PZT z jej strukturą domenową. Potwierdzają to również bezpośrednie badania struktury domenowej próbek w fazie romboedrycznej (rys. 3), tetragonalnej (rys. 4) i w stanie dwufazowym (Re + T), typowym dla obszaru morfotropowego (rys. 5), przed (a) i po (b) polaryzowaniu w zewnętrznym polu elektrycznym. O wpływie koncentracji (x) $PbTiO_3$ na strukturę domenową polaryzowanej ceramiki typu PZT świadczą mikrofotografie powierzchni (rys. 6) w fazie tetragonalnej (A, B, C) i romboedrycznej (D, E, F), ułożone według malejącego x (A, B...F).

Domenową strukturę piezoceramiki ujawniano metodą chemicznego trawienia powierzchni i przełomów próbek w 5–30% wodnym roztworze HNO_3 z dodatkiem około 10 kropli kwasu fluorowodorowego (HF), przez okres 10–20 s. Ujawnioną strukturę domenową obserwowano za pomocą transmisyjnego mikroskopu elektronowego (TESLA BS–613) wykorzystując metodę replik platynowo–węglowych (Pt/C).

Tetragonalnej fazie odpowiada złożona struktura domenowa z dużą liczbą ścian domenowych. Struktura ta upraszcza się i staje się bardziej uporządkowana w miarę zmniejszania się koncentracji $PbTiO_3$ i zbliżania się do obszaru morfotropowego (rys. 6 A–C). Po przejściu do fazy romboedrycznej obraz struktury domenowej ulega silnej zmianie; zmniejsza się liczba ścian domenowych, a niektóre krystality są bliskie stanu monodomenowego (rys. 6 D–F).

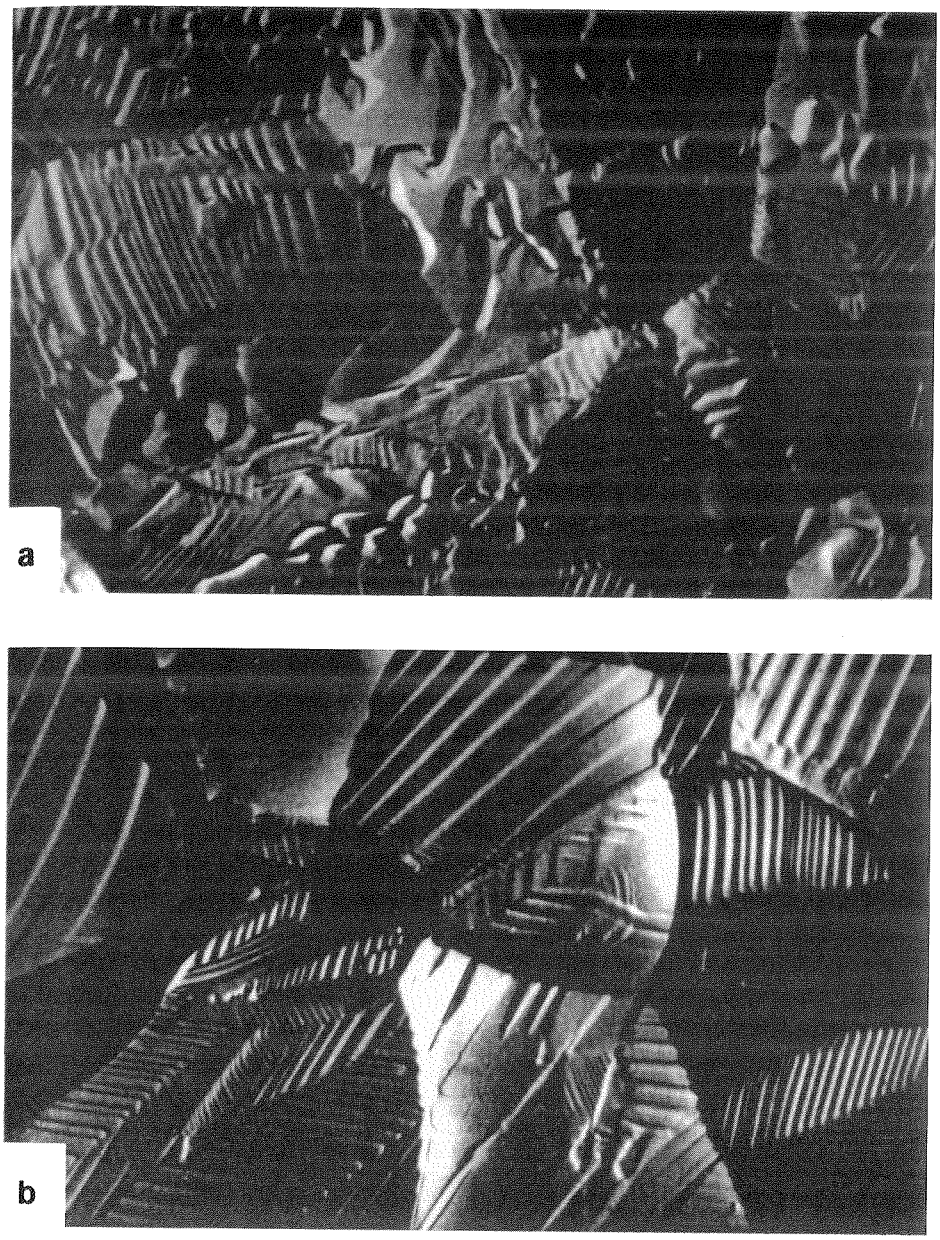
Struktura domenowa próbek przed spolaryzowaniem (rys. 3, 4, 5–a) różni się znacznie od struktury domenowej próbek po spolaryzowaniu (rys. 3, 4, 5–b). Najsilniejszy wpływ na strukturę domenową wywiera proces polaryzowania w przypadku próbek o składach z obszaru morfotropowego (rys. 5), a najsłabszy w przypadku próbek o strukturze tetragonalnej (rys. 4).

W fazie romboedrycznej obserwuje się dobrą korelację między wzrostem parametru η, zmniejszeniem liczby ścian domenowych i wzrostem temperaturowej stabilności



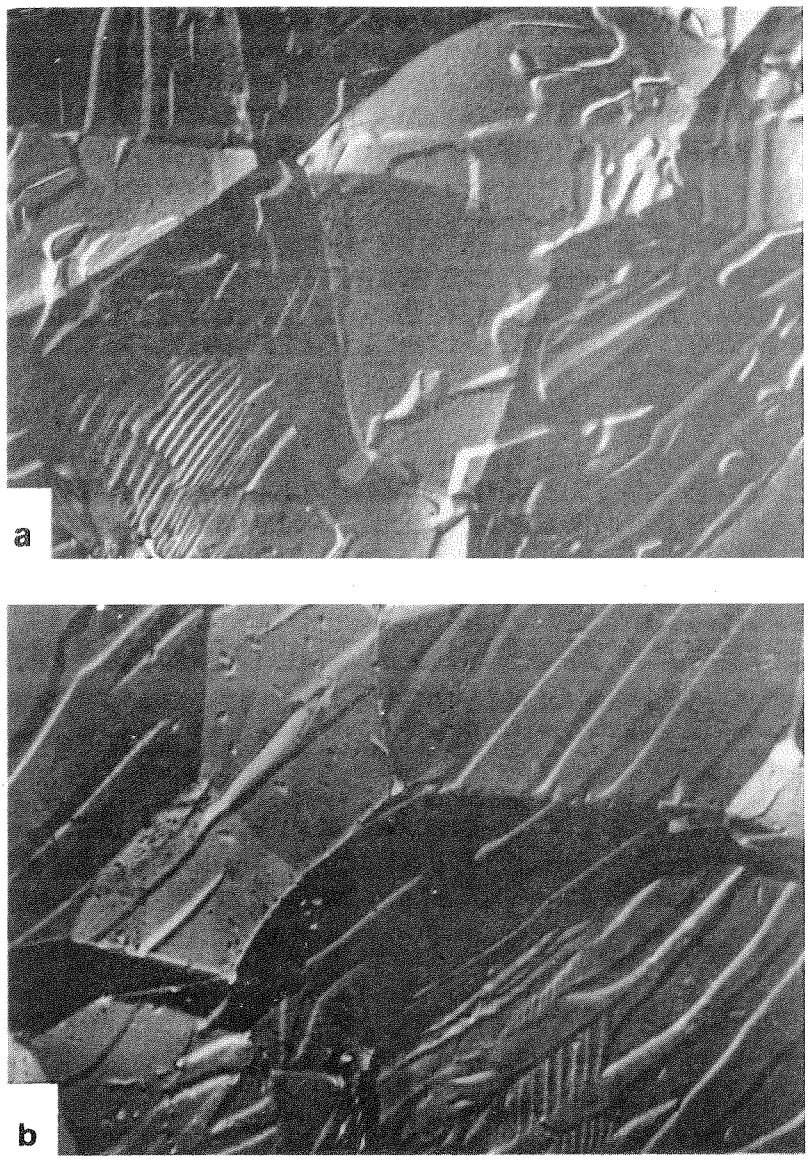
Rys. 3. Domenowo-bliźniacza struktura ceramiki typu PZT w fazie romboedrycznej (Re), ujawniona trawieniem chemicznym powierzchni przełomu:
a) próbka niespolaryzowana; b) próbka spolaryzowana; Powiększenie 10000×(1μm = 10 mm); repliki C/Pt.

częstotliwości rezonansowej. Jednak temperaturowa stabilność f_r składów PZT położonych w obszarze minimum $\Delta f_r/f_r$ w fazie romboedrycznej jest mniejsza niż w fazie tetragonalnej, podczas gdy mechaniczna dobroć w obszarze maksimum Q_m jest najczęściej nieco wyższa w fazie romboedrycznej, niż w fazie tetragonalnej.



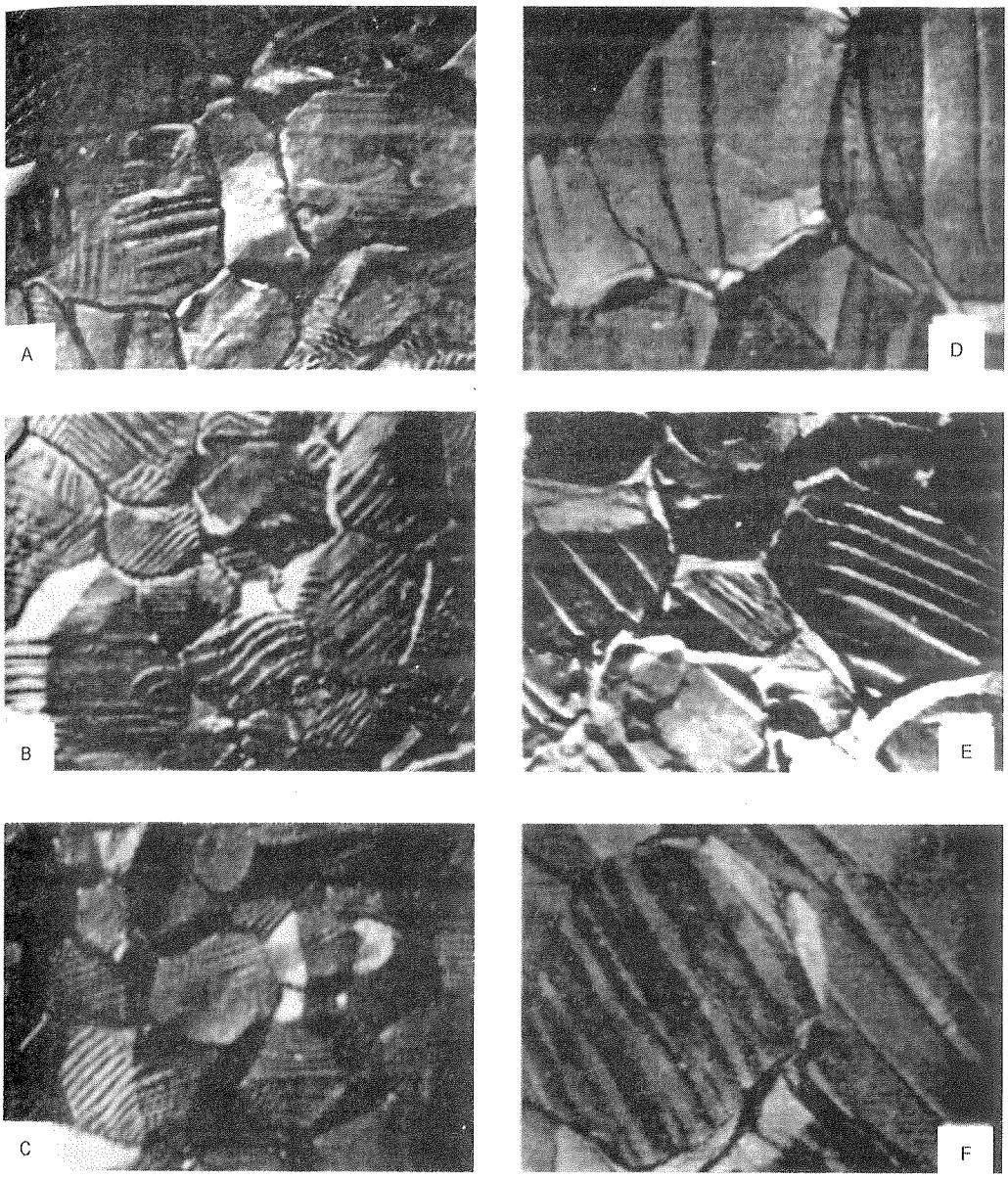
Rys. 4. Domenowo-bliźniacza struktura ceramiki typu PZT w fazie tetragonalnej (T), ujawniona trawieniem chemicznym powierzchni przełomu: a) próbka niespolaryzowana; b) próbka spolaryzowana; Powiększenie 10000×(1μm = 10 mm); repliki C/Pt.

W niektórych pracach przyjmowano, że brak stabilności częstotliwości rezonansowej należałoby wiązać również ze zmianami wartości stałych sprężystych w funkcji temperatury i koncentracji $PbTiO_3$ w piezoceramice na bazie PZT [9]. Rezultaty uzyskane



Rys. 5. Domenowo-bliźniacza struktura ceramiki typu PZT o składzie z obszaru morfotropowego MR (faza Re z małą domieszką fazy T), ujawniona trawieniem chemicznym powierzchni przełomu: a) próbka niespolaryzowana; b) próbka spolaryzowana;
Powiększenie 10000×(1μm = 10 mm); repliki C/Pt.

w niniejszej pracy są jednak bliższe wynikom uzyskanym przez V.A. Isupova [10], wg którego stałe sprężyste w ramach tego samego wieloskładnikowego układu na bazie PZT zmieniają się nieznacznie. Dla piezoceramiki zmiany $\Delta f_r/f_r$, przy przejściu z obszaru



Rys. 6. Domenowo-bliźniacza struktura spolaryzowanej ceramiki typu PZT w fazie tetragonalenj (A, B, C) i w fazie romboedrycznej (D, E, F) dla próbek o coraz niższej koncentraji $PbTiO_3$ (A, B...F) [1]. Powiększenie 5000×(1μm = 5 mm).

morfotropowego do fazy tetragonalnej, są o rząd wartości większe od zmian odwrotności podatności sprężystej $1/S_{11}^E$, czyli współczynnika sztywności $c_{11}^E = \frac{1}{S_{11}^E}$. Oznacza to,

że temperaturowa stabilność f_r drgań radialnych próbek piezoceramicznych jest głównie uwarunkowana procesami domenowo-orientacyjnymi.

Względna zmiana częstotliwości rezonansowej, w miarę oddalania się składów od obszaru morfotropowego w kierunku niższych i wyższych koncentracji $PbTiO_3$, przechodzi najpierw przez minimum a następnie zaczyna nieco wzrastać zarówno w fazie Re, jak i T (rys. 2). Takie zachowanie się $\Delta f_r/f_r$ wiązać należałoby ze zmniejszeniem rozpuszczalności złożonych tlenków na bazie PZT przy większych koncentracjach $PbZrO_3$ lub $PbTiO_3$.

2.3. MODYFIKOWANIE PODSTAWOWEGO SKŁADU CHEMICZNEGO (metoda II)

Od wielu lat prowadzone są badania dotyczące wpływu modyfikowania podstawowego składu PZT na stabilność f_r. Podstawowe metody modyfikacji składu chemicznego PZT przedstawiono w tablicy 2. W większości cytowanych tam prac brak jest fizyko-chemicznej interpretacji obserwowanego wpływu domieszek na stabilność f_r i wartość podstawowych parametrów elektrofizycznych badanych materiałów piezoceramicznych na bazie PZT. Pewne próby w tym zakresie podjęli autorzy w niniejszej pracy.

Kompleksowe badania, dotyczące wpływu modyfikowania (domieszkowania) składu wyjściowego na stabilność f_r przetworników elektroakustycznych na bazie PZT, przeprowadzono na przykładzie trójskładnikowego roztworu stałego

$$PbTiO_3 - PbZrO_3 - PbW_{1/2}Cd_{1/2}O_3, \tag{8}$$

o składzie chemicznym położonym na diagramie fazowym w sąsiedztwie romboedrycznej granicy obszaru morfotropowego (MR).

Stwierdzono, że modyfikatory U^{6+}, Mn^{4+}, Co^{2+}, Cr^{3+} wywołują silne obniżenie względnej zmiany częstotliwości rezonansowej $\Delta f_r/f_r$ w przedziale temperatur 213–358 K. Pozytywny efekt, chociaż słabszy, obserwuje się również po domieszkowaniu Al^{3+}. Modyfikowanie tlenkami Bi^{3+} i Ag^+ pogarsza temperaturową stabilność częstotliwości rezonansowej składu podstawowego. Wpływ domieszki Sb^{3+} na stabilność f_r jest zależny od jej koncentracji (n): pogorszenie stabilności zachodzi dla $n \leq 0{,}9$ at.%, natomiast polepszenie — dla $n > 0{,}9\%$ at.

Dla objaśnienia efektu temperaturowej stabilności częstotliwości rezonansowej w modyfikowanej piezoceramice PZT przeprowadzono badania: rentgenowskie, mikroskopowe i dielektryczne.

Badania rentgenowskie wykazały, że domieszki Mn^{4+}, Cr^{3+}, Co^{2+}, Al^{3+} i Ag^+ nie zmieniają symetrii sieci krystalicznej wyjściowej struktury, lecz przesuwają MR w stronę fazy romboedrycznej, zwiększają jednorodny parametr deformacji tetragonalnej δ_T a zmniejszają jednorodny parametr deformacji romboedrycznej δ_{Re}. Podobne efekty obserwowano w przypadku domieszkowania tlenkami Sb^{3+} dla koncentracji $n = 0{,}43 - 0{,}65$ at.%.

Tablica 2

Zwiększanie temperaturowej stabilności częstotliwości rezonansowej różnymi metodami modyfikowania podstawowego składu PZT

Lp.	Metody modyfikowania	Przykładowa literatura
1.	Częściowe izowalencyjne podstawienia w miejsce Pb^{2+} kationów na Ba^{2+}, Sr^{2+} lub Ca^{2+}	[20, 41, 42]
2.	Modyfikowanie: a) tlenkami pierwiastków: Mn^{2+}, Co^{3+}, Fe^{3+}, Ni^{2+}, Bi^{3+}, Sb^{3+}, U^{4+}, Ge^{2+}, F^{7+}, Ca^{2+}, Zn^{2+}, Cr^{3+}, In^{3+};	[21, 24–28, 41, 43–45]
	b tlenkami pierwiastków ziem rzadkich;	[46]
	c) złożonymi tlenkami na bazie manganu (Mn^{2+}, Mn^{3+}).	[1, 46, 47]
3.	Spiekanie w atmosferze tlenu	[48]
4.	Spiekanie w atmosferze Al_2O_3.	[49]
5.	Wprowadzenie do stechiometrycznego roztworu stałego typu PZT naddatku PbO.	[17]
6.	Wprowadzenie do stechiometrycznego roztworu stałego typu PZT o strukturze perowskitu naddatku fazy pirochlorowej i MnO_2.	[50]

Domieszki Cr^{3+}, U^{6+}, Bi^{3+}, Sb^{3+} i Co^{2+} hamują wzrost ziarn piezoceramiki. Szczególnie efektywne w tym zakresie są domieszki Co^{2+} i Sb^{3+} ($\bar{r}$ = 4–6 μm). Domieszki Mn^{4+}, Al^{3+} i Ag^{+} sprzyjają wzrostowi ziarn ($\bar{r}$ = 16–20μm).

Badanie procesu przepolaryzowania pozwoliły zaobserwować wzrost pola ładunku przestrzennego E_{sc} w modyfikowanej ceramice i ocenić rzeczywiste pole koercji E_c. Stwierdzono, że w modyfikowanej piezoceramice o małej temperaturowej stabilności f_r pole $E_{sc} = (4–12) \cdot 10^4$ V/m jest mniejsze od $E_{sc} = (10–55) \cdot 10^4$ V/m dla piezoceramiki o dużej temperaturowej stabilności f_r. Wynika z tego, że wartość E_{sc} jest jednym z wskaźników temperaturowej stabilności f_r. Ładunek przestrzenny wpływa na ruchliwość ścian domenowych, a tym samym na wartość pola koercji E_c. Ze wzrostem E_{sc} należy oczekiwać zmniejszenia ruchliwości ścian domenowych i wzrostu wartości E_c. Jednak ruchliwość ścian domenowych zależy również od wartości spontanicznej deformacji δ. Ze wzrostem δ ruchliwość ścian domenowych zmniejsza się. W konsekwencji charakter zmian E_c, a tym samym temperaturowa stabilność f_r, są uwarunkowane dwoma czynnikami: (1) polem ładunku przestrzennego E_{sc} i (2) wartością spontanicznej deformacji δ.

Celem poznania natury ładunku przestrzennego przeanalizowano:

- krystalochemiczne charakterystyki wprowadzanych domieszek oraz atomów bazowego składu piezoceramiki;
- trwałość stopnia utlenienia atomów w atmosferze powietrza;

- charakter zmian parametrów komórki elementarnej;
- charakter skrośnego oporu właściwego.

Na tej podstawie stwierdzono, że w przypadku wprowadzenia domieszek Mn^{4+}, Cr^{3+}, Co^{2+}, Al^{3+} i Ag^{+} bilans walencyjności zapewniają powstające dodatkowe wakanse tlenowe V_o, natomiast w przypadku wprowadzenia domieszek Bi^{3+} i Sb^{3+} — dodatkowe wakanse ołowiowe V_{Pb}. Źródłem ładunku przestrzennego są dodatkowe wakanse V_o, natomiast wakanse V_{Pb} powinny zmniejszać ładunek przestrzenny, który hamuje ruch ścian domenowych. Obserwowany wzrost E_{sc}, przy wzroście koncentracji domieszek Bi^{3+} i Sb^{3+} oraz związanych z nimi wakansami V_{Pb}, wiązać należy z silnym zmniejszeniem średniego rozmiaru ziarn ($\overline{r}$) piezoceramiki.

Wartości pól ładunku przestrzennego modyfikowanej piezoceramiki, z lepszą i gorszą temperaturową stabilnością f_r, pokrywają się częściowo. Oznacza to, że o temperaturowej stabilności f_r decyduje nie tylko ruchliwość ścian domenowych, lecz również inne właściwości struktury domenowej krystalitów PZT. Celem wyjaśnienia tego efektu przeprowadzono badania struktury domenowej modyfikowanej i niemodyfikowanej piezoceramiki, metodą mikroskopu elektronowego i dyfrakcji rentgenowskiej. Badania te wykazały, że struktura domenowa bazowej (niemodyfikowanej), spolaryzowanej piezoceramiki typu PZT, o składzie z obszaru morfotropowego, charakteryzuje się szerokimi domenami płytkowymi, nielicznymi klinowymi domenami wzrastającymi od strony granic ziarn i stanem monodomenowym niektórych ziarn [11–13].

Domieszki wywołujące wzrost temperaturowej stabilności f_r (Mn^{4+}, Cr^{3+}, Co^{2+}, Al^{3+}, U^{6+}) powodują, że struktura domenowa modyfikowanej piezoceramiki staje się bardziej złożona: (1) zanikają szerokie domeny płytkowe, (2) pojawia się duża liczba drobnych 90° domen, (3) struktura domenowa przypomina strukturę domenową czystego PZT w fazie tetragonalnej. Ponieważ w obszarze morfotropowym piezoceramika jest dwufazowym roztworem stałym faz T i Re, dlatego dużą koncentrację 90° domem uzasadnić należałoby rodzajem wprowadzanych domieszek i związanych z nimi defektami; powstające w wyniku twardego domieszkowania wakanse tlenowe V_o deformują sieć krystaliczną oraz sprzyjają formowaniu i zamocowaniu 90° struktury domenowej [11–13].

2.4. DOBÓR WARUNKÓW SYNTEZY I SPIEKANIA (metoda III)

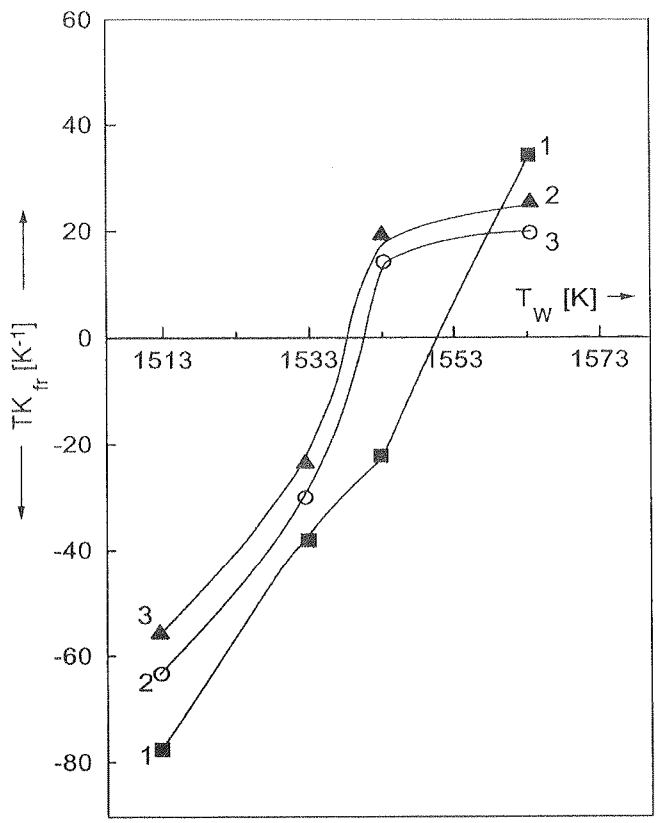
Badania wpływu warunków technologicznych (syntezy i spiekania) na temperaturową stabilność f_r przeprowadzono na próbkach piezoceramicznych o składzie $(Pb_{0,97}Nd_{0,03})(Zr_{0,54}Ti_{0,46})O_3$ [15]. Optymalne wartości TK_{f_r} uzyskano w przypadku otrzymywania piezoceramiki z uprzednio nie zsyntezowanych proszków. Próbki miały dużą porowatość, co uwarunkowane jest rozszerzalnością cieplną w obszarze temperatury formowania się roztworu stałego. Tak otrzymana piezoceramika posiada znacznie niższe wartości k_p i $\varepsilon^o_{33}/\varepsilon_o$, niż piezoceramika otrzymana z proszków uprzednio zsyntezowanych [1].

Minimalna temperatura syntezy (T_s) jest ograniczona temperaturą zakończenia reakcji formowania się roztworu stałego typu PZT (dla ceramiki o podstawowym składzie

T_s = 1073 K). Temperaturowa stabilność częstotliwości rezonansowej w zakresie T = 293–343 K zmniejsza się ze wzrostem temperatury syntezy powyżej 1073 K, ale tylko w przypadku próbek uprzednio zsyntezowanych. W przypadku takich piezoceramicznych przetworników niższa temperaturowa stabilność f_r wiąże się z rozdrobnieniem i deformacją ziarn, powtórnie mielonych próbek po syntezie.

Zmniejszenie szybkości chłodzenia w procesie obróbki cieplnej znacznie poprawia temperaturową stabilność f_r. Powstające przy szybkim ochładzaniu naprężenia cieplne, związane ze słabym przewodnictwem cieplnym piezoceramiki, ujemnie wpływają na temperaturową stabilność częstotliwości rezonansowej próbek piezoceramicznych. Zbyt wolne grzanie, podobnie jak zbyt szybkie chłodzenie w procesie obróbki cieplnej, negatywnie wpływają na wartość TK_{fr}. Natomiast optymalne rezultaty daje szybkie podwyższenie temperatury w piecu (np. 200 K/h) i wolniejsze chłodzenie (np. 100 K/h). Ze wzrostem temperatury wyprażania (T_w) następował wzrost TK_{fr} i zmiana jego znaku od ujemnego do dodatniego (rys. 7).

Stwierdzono, że zmianę temperaturowej stabilności częstotliwości rezonansowej wywołują również odstępstwa od stechiometrii. Przykładowo, mały niedomiar zawarto-



Rys. 7. Wpływ temperatury wyprażania (T_w) na wartość współczynnika temperaturowego częstotliwości rezonansowej (TK_{fr}) dla trzech serii (1, 2, 3) próbek piezoceramicznych na bazie PZT [16]

ści ołowiu w stałym roztworze $(Pb_{0,97})Nd_{0,03})(Zr_{0,53}Ti_{0,47})O_3$ prowadzi do znaczącego polepszenia TK_{fr} bez zmiany wartości k_p [17]. Zastąpienie 3 mol.% ołowiu (Pb^{2+}) jonami neodymu (Nd^{3+}) w stałym roztworze PZT wpływa pozytywnie na parametry piezoelektryczne ferroelektrycznej ceramiki.

2.5. KONTROLA MIKROSTRUKTURY (metoda IV)

Materiał badań stanowiły próbki piezoceramiczne na bazie PZT o składach:

$$\begin{aligned} &\text{I. (x) } PbTiO_3 - PbZrO_3 - PbNb_{2/3}Zn_{1/3}O_3 - PbNb_{2/3}Mn_{1/3}O_3; \\ &\text{II. (x) } PbTiO_3 - PbZrO_3 - PbNb_{2/3}Mn_{1/3}O_3 - PbNb_{1/2}Mg_{1/2}O_3. \end{aligned} \tag{9}$$

Czteroskładnikowy roztwór stały o składzie (I) znany jest jako piezoceramika PKR–13, natomiast o składzie (II) jako piezoceramika PKR–28.

Na przykładzie piezoceramiki PKR–13 badano wpływ średnich rozmiarów ziarn na temperaturową względną zmianę częstotliwości rezonansowej $(\Delta f_r/f_r)$ oraz innych charakterystyk piezoceramicznych, natomiast na przykładzie piezoceramiki PKR–28 badano wpływ porowatości (p) i średnich rozmiarów ziarn $(\overline{r})$ na $\Delta f_r/f_r$.

Strukturę piezoceramiki badano metodami dyfrakcji rentgenowskiej (dyfraktometr rentgenowski DRON 3M, CuK_α), mikrostrukturę — metodą replik platynowo–węglowych i transmicyjnej mikroskopii elektronowej (TESLA, BS–613), przenikalność elektryczną i tangens kąta strat dielektrycznych metodą mostkową ($\nu = 10^3$ Hz), polaryzację orientacyjną i szczątkową oraz pole koercji — metodą Sawyera–Towera (spektrometr ferroelektryczny firmy RADIOPAN), a parametry piezoelektryczne — metodą rezonansu — antyrezonansu.

Ceramiczne próbki o składzie I (PKR–13) otrzymywano metodą prasowania na gorąco, dobierając odpowiednią temperaturę spiekania (T_s), ciśnienie prasowania (p_s), czas spiekania (t_s) oraz charakterystyki $T_s(t)$ i $p_s(t)$ [18]. Otrzymywana piezoceramika miała wysoką gęstość $(\varrho = 7{,}9 - 8{,}0) \cdot 10^3$ kg/m^3, zbliżoną do gęstości rentgenowskiej, i znikomą porowatość $p \leq 2\%$. Rozmiar ziarn piezoceramiki zmieniano dobierając odpowiednio temperaturę spiekania T_s ($T_1 = 1023$ K i $T_2 = 1073$) oraz p_s i t_s.

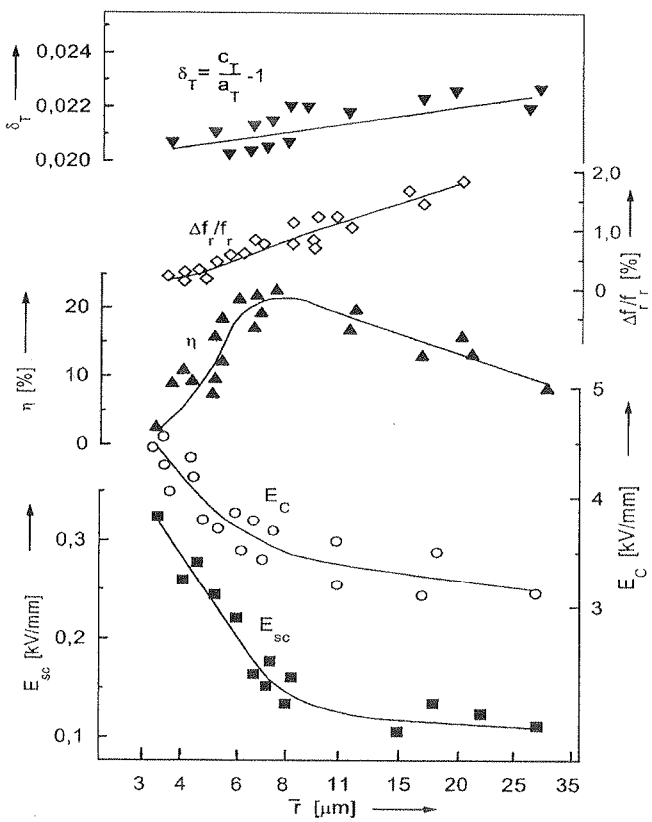
Podstawowe parametry drobnoziarnistej piezoceramiki PKR–13 są następujące: $\varepsilon^\sigma_{33}/\varepsilon_0 = 800$; $k_p = 0{,}4$; $k_{31} = 0{,}23$; $d_{31} = 65 \cdot 10^{-12}$ C/N; $g_{33} = 18{,}5 \cdot 10^{-3}$ Vm/N; $\operatorname{tg}\delta(E^o_\sim = 5$ kV/m) $= 0{,}007$; $Q_m = 2800$; $T_c = 603$ K (330°C); w zakresie 213–358 K $\Delta f_r/f_r = 0{,}2 - 0{,}25\%$, w zakresie 263–323 K $\Delta f_r/f_r \leq 0{,}14\%$; szybkość starzenia a [%/dekadę] dla $\varepsilon^\sigma_{33} - a_{\varepsilon^\sigma_{33}} < 0{,}1$, dla $k_p - a_{k_p} < 0{,}2$ dla $f_r - a_{f_r} = +0{,}02$. Z danych tych wynika, że parametry elektrofizyczne piezoceramiki PKR–13 są porównywalne z parametrami piezomateriałów szeroko stosowanych do konstrukcji pasmowych filtrów elektrycznych, jak np. piezoceramika PZT–6A. Piezoceramika PKR–13, obok małej wartości $\Delta f_r/f_r$ w przedziale 263–323 K, charakteryzuje się wysoką stabilnością czasową f_r i stosunkowo dużymi wartościami dobroci mechanicznej Q_m, od której zależy ostrość amplitudowo–częstotliwościowych charakterystyk w pasmie przepuszczania

i współczynnika sprzężenia elektromechanicznego k_p, od którego zależy szerokość pasma przepuszczania sygnałów elektrycznych.

Piezoceramika PKR–13 charakteryzuje się dużą koncentracją $PbTiO_3$ (47,5 mol.%) i dlatego też na diagramie fazowym jej skład znajduje się w fazie tetragonalnej, oddalonej znacznie od obszaru morfotropowego (MR). Piezomateriały na bazie PZT o takiej koncentracji $PbTiO_3$ charakteryzują się wysokim Q_m i k_p oraz minimalnym $\Delta f_r/f_r$ w roboczym przedziale temperatury (tablica 3).

Stwierdzono, że ze wzrostem temperatury w zakresie 213–358 K względna zmiana częstotliwości rezonansowej $\Delta f_r/f_r$ zwiększa się ze wzrostem rozmiarów ziarn ($\bar{r}$ = 3,5 – 28μm), co oznacza, że ze wzrostem $\bar{r}$ temperaturowa stabilność f_r pogarsza się (rys. 8).

Dla objaśnienia tego efektu przeprowadzono równomiernie pomiary (rys. 8) pola koercji $E_c(\bar{r})$, deformacji tetragonalnej $\delta_T = \left(\frac{c_T}{a_T} - 1\right) = f(\bar{r})$, $\eta(\bar{r})$ i pola ładunku prze-



Rys. 8. Zależność parametrów elektrofizycznych (E_c, E_{sc}, $\Delta f_r/f_r$) i strukturalnych (η, $\delta_T = \frac{c_T}{a_T} - 1$) od średniego rozmiaru ziaren $\bar{r}$ piezoceramiki PKR-13 o składzie: 0,475 $PbTiO_3 - PbNb_{2/3}Zn_{1/3}O_3 - PbNb_{2/3}Mn_{1/3}O_3$.

Tablica 3

Wpływ koncentracji $PbTiO_3$ na podstawowe parametry piezoceramiki o składzie (I): (x) $PbTiO_3-PbZrO_3-PbNb_{2/3}Zn_{1/3}O_3-PbNb_{2/3}Mn_{1/3}O_3$

Lp.	$PbTiO_3$ [mol.%]	$\varepsilon^{\sigma}_{33}/\varepsilon_0$	k_p	tg $\delta \cdot 10^2$ (E = 5 kV/m)	Q_m	T_c [K]	$\Delta f_r/f_r$ (213–358 K)	Faza
1	20,0	420	0,314	0,29	5100	493	–	Re
2	40,0	600	0,495	0,70	2070	573	–	Re
3	40,6	1000	0,562	0,60	1500	589	–	MR(Re+T)
4	41,2	1310	0,577	0,20	1110	597	–	MR(Re+T)
5	46,5	970	0,411	0,96	2000	593	0,99	T
6	47,1	920	0,408	0,80	2400	596	0,65	T
7	47,5 PKR–13	800	0,400	0,70	2800	603	0,20	T
8	48,0	850	0,388	0,80	2500	605	0,25	T
9	48,4	800	0,386	0,76	3000	608	0,36	T

strzennego $E_{sc}(\overline{r})$. Tę ostatnią wielkość obliczono na podstawie charakterystyki procesu przepolaryzowania w oparciu o następujące równania [19]:

$$E_c = \frac{E_{ca}+E_{cr}}{2}, \tag{10}$$

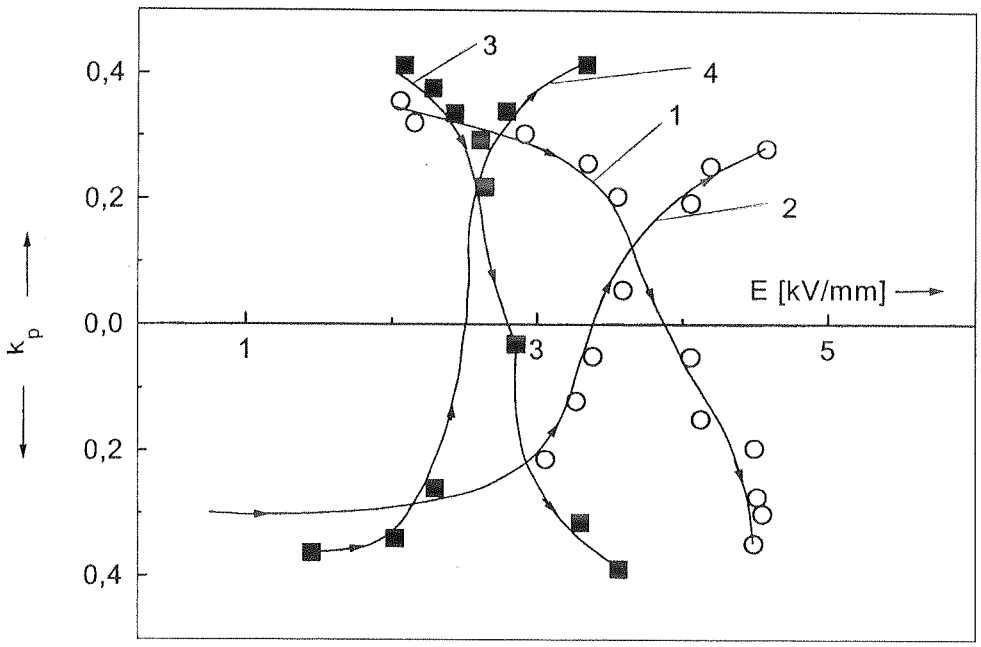
$$E_{sc} = \frac{E_{cr}-E_{ca}}{2}, \tag{11}$$

gdzie:

E_{ca} – pole orientujące polaryzację w kierunku antyrównoległym do wyjściowej polaryzacji;

E_{cr} – pole orientujące polaryzację w kierunku zgodnym z kierunkiem wyjściowej polaryzacji.

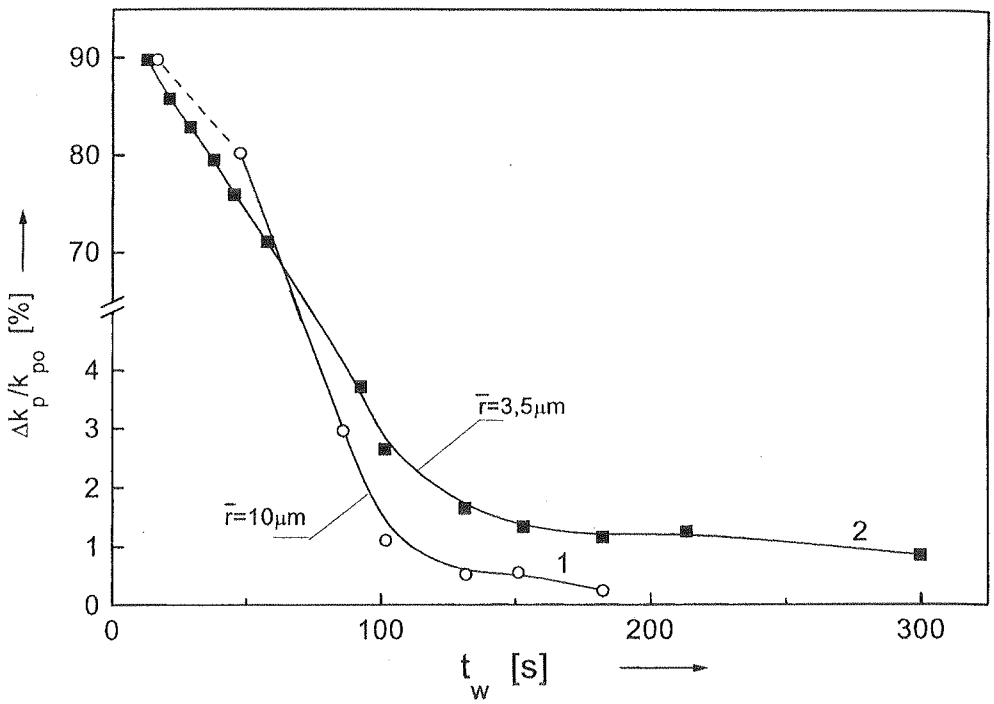
Na rys. 9 przedstawiono krzywe odwracania polaryzacji w polaryzowanych próbkach ceramiki PKR–13 z różnymi rozmiarami ziarn krzywe ($1{,}2-\overline{r}=3{,}5$ μm; $3{,}4-\overline{r}=10$ μm). Obserwuje się tu połączenie prostego i odwrotnego przebiegu charakterystyk przepolaryzowania, uwarunkowane efektem pola ładunku przestrzennego. Efekt ten jest wyraźniejszy w próbkach o mniejszych rozmiarach ziarn. Pole E_{sc} wytworzone w procesie starzenia jest dość stabilne i nie zanika po podgrzaniu piezoceramiki do $T>T_c$. Potwierdzają to badania względnej zmiany współczynnika sprzężenia elektromechanicznego ($\Delta k_p/k_{p_o}$, gdzie k_{p_o} — współczynnik sprzężenia elektromechanicznego wyjściowej piezoceramiki) od czasu (t_w) przetrzymywania w temperaturze T_w =



Rys. 9. Zmiana kierunku polaryzacji w polaryzowanej piezoceramice PKR-13 o składzie: 0,475 $PbTiO_3$–$PbZrO_3$–$PbNb_{2/3}Zn_{1/3}O_3$–$PbNb_{2/3}Mn_{1/3}O_3$: 1, 2 – piezoceramika drobnoziarnista ($\bar{r}$ = 3,5 μm); 3, 4 – piezoceramika gruboziarnista ($\bar{r}$ = 10 μm).

= T_c + 30 K. Współczynnik sprzężenia elektromechanicznego (k_p) charakteryzuje stopień spolaryzowania próbek piezoelektrycznych, dlatego też ze wzrostem czasu przetrzymywania (t_w) w temperaturze wyższej od temperatury Curie ($T_w > T_c$) wartość $\Delta k_p/k_{p_o}$ stopniowo maleje (rys. 10). W przypadku t_w = 300 s wartość $\Delta k_p/k_{p_o}$ dla drobnoziarnistej piezoceramiki zmniejsza się do 1% (rys. 10, krzywa 2). Spadek wartości $\Delta k_p/k_{p_o}$ jest silniejszy dla piezoceramiki gruboziarnistej (rys. 10, krzywa 1).

Z przeprowadzonych badań piezoceramiki PKR–13 wynika, że ze wzrostem $\bar{r}$ zmniejsza się pole ładunku przestrzennego E_{sc} (rys. 9). Początkowo przy wzroście $\bar{r}$ od 3,5 do 8 μm otrzymuje się silne zmniejszanie E_{sc} od 0,35 do 0,1 kV/mm. Po dalszym wzroście $\bar{r}$ do 28 μm zmniejszenie się E_{sc} jest łagodne. Badania mikrostruktury wykazały, że gruboziarnista piezoceramika jest fizycznie niejednorodna, obok dużych ziarn występują również ziarna mniejsze. Jest to z pewnością przyczyną słabych zmian $E_{sc}(\bar{r})$ w zakresie r = 8–28 μm. Charakter zmian $E_{sc}(\bar{r})$ warunkuje podobne zmiany pola koercji (E_c) w miarę wzrostu średnich rozmiarów ziarn ($\bar{r}$). Na krzywej $E_c(\bar{r})$, podobnie jak na krzywej $E_{sc}(\bar{r})$, wyróżnić można wyraźne dwa przedziały: do $\bar{r}$ = 7 μm — silny spadek wartości E_c, natomiast dla $\bar{r}$ > 7 μm łagodne zmniejszanie się E_c w funkcji $\bar{r}$. Spadek $E_{sc}(\bar{r})$ jest znacznie silniejszy od spadku $E_c(\bar{r})$ i to w obydwu zakresach $\bar{r}$. Słabsze zmiany $E_c(\bar{r})$ są uwarunkowane wzrostem spontanicznej deformacji $\left(\delta_T = \frac{c_T}{a_T} - 1\right)$ wraz ze wzrostem średniego rozmiaru ziarn ($\bar{r}$).



Rys. 10. Zależność względnej zmiany współczynnika sprzężenia elektromechanicznego ($\Delta k_p/k_{po}$) spolaryzowanych próbek piezoceramicznych PKR-13 od czasu ich przetrzymywania (t_w) w temperaturze $T_w = 633$ K ($T_w > T_c = 603$ K); skład piezoceramiki: 0,475 $PbTiO_3 - PbZrO_3 - PbNb_{2/3}Mn_{1/3}O_3 - PbNb_{2/3}Zn_{1/3}O_3$

Wynika stąd, że charakter zmian $E_c(\bar{r})$ jest uwarunkowany dwoma czynnikami:
- zmianą pola ładunku przestrzennego $E_{sc}(\bar{r})$;
- zmianą spontanicznej deformacji tetragonalnej ($\delta_T(\bar{r})$).

Dominującym czynnikiem wpływającym na $E_c(\bar{r})$ jest $E_{sc}(\bar{r})$. Zmniejszenie E_c ze wzrostem $\bar{r}$ świadczy o wzroście ruchliwości ścian domenowych, co prowadzi do pogorszenia temperaturowej stabilności częstotliwości rezonansowej f_r. Taki mechanizm zmian temperaturowej stabilności częstotliwości rezonansowej jest słuszny tylko dla piezoceramiki ze średnim rozmiarem ziarn $\bar{r} < 8$ μm. Zwiększanie ruchliwości ścian domenowych jest przyczyną wzrostu parametru $\eta(\bar{r})$ w przedziale $\bar{r} = 3{,}5 - 8$ μm (rys. 8). Wzrost $\eta(\bar{r})$ pociąga za sobą zmniejszenie liczby różnych od 180° ścian domenowych. Gdyby czynnikiem decydującym o stabilności f_r w tym zakresie $\bar{r}$ był wzrost liczby ścian domenowych, to powinno się obserwować zmniejszanie $\Delta f_r/f_r = f(\bar{r})$. Eksperymentalnie stwierdzono wzrost $\Delta f_r/f_r = f(\bar{r})$ wskazuje, że czynnikiem decydującym o stabilności f_r w przypadku $\bar{r} = 3{,}5 - 8$ μm jest wzrost ruchliwości ścian domenowych (wzrost η).

Przy dalszym wzroście średnich rozmiarów ziarn ($\bar{r} = 8$–28 μm), pomimo zmniejszania się E_{sc} i E_c, η silnie maleje ze wzrostem $\bar{r}$ (rys. 8). Powinno to poprawić temperaturową stabilność częstotliwości rezonansowej próbek PKR–13. Spadek wartości $\eta(\bar{r})$ dla $\bar{r} > 8$ μm wiąże się jednak ze wzrostem spontanicznej deformacji tetragonalnej

$\left(\delta_T = \frac{c_T}{a_T} - 1\right)$, która dla $\bar{r} = 24$ μm przyjmuje wysoką wartość $\delta_T = 0{,}23$. W rezultacie wzrasta liczba ścian domenowych, co prowadzi do pogorszenia stabilności f_r w zakresie $\bar{r} = 8$–28 μm (w przypadku piezoceramiki gruboziarnistej wpływ E_{sc} na wartość $\Delta f_r/f_r$ jest bardzo niewielki).

Wpływ porowatości (p) i gęstości (ϱ) na temperaturową stabilność częstotliwości rezonansowej (f_r) badano na przykładzie piezoceramiki PKR–28 o ogólnym wzorze (II):

$$(x)\ PbTiO_3 - PbZrO_3 - PbNb_{2/3}Mn_{1/3}O_3 - PbW_{1/2}Mg_{1/2}O_3. \qquad (12)$$

Podstawowe parametry elektrofizyczne tego roztworu w funkcji koncentracji $PbTiO_3$ zestawiono w tablicy 4.

Tablica 4

Wpływ koncentracji $PbTiO_3$ na podstawowe parametry piezoceramiki o składzie: (x) $PbTiO_3 - PbZrO_3 - PbNb_{2/3}Mn_{1/3}O_3 - PbW_{1/2}Mg_{1/2}O_3$

Lp.	$PbTiO_3$ [mol.%]	$\varepsilon^{\sigma}_{33}/\varepsilon_0$	k_p	tg $\delta \cdot 10^2$ (E = 5 kV/m)	Q_m	T_c [K]	$\Delta f_r/f_r$ (213–358 K)	Faza
1	35,0	430	0,400	0,44	1480	517	–	Re
2	37,5	460	0,504	0,47	1690	538	–	Re
3	42,0	1970	0,662	0,40	900	561	–	MR(Re+T)
4	43,0	2550	0,647	0,20	770	503	–	MR(Re+T)
5	51,5	710	0,346	0,36	2600	603	0,41	T
6	52,0	750	0,340	0,50	2500	598	0,38	T
7	52,5 PKR–28	750	0,331	0,30	3500	598	0,21	T
8	53,0	660	0,327	0,39	3400	598	0,42	T
9	53,5	580	0,300	0,36	2200	603	0,44	T

Piezoceramikę PKR–28 otrzymywano zarówno metodą prasowania na gorąco (HP) w optymalnych warunkach (1) oraz metodą klasyczną (CCM) w warunkach 2–9. W zależności od warunków technologicznych miała ona różne właściwości fizyczne (tablica 5).

Największą gęstość (najmniejszą porowatość) i optymalny rozmiar ziarna otrzymano metodą prasowania na gorąco (tablica 5, warunki (1)). Metoda HP (ang. *hot pressing*) zapewnia dobrą temperaturową stabilność f_r, co jest uwarunkowane małym średnim rozmiarem ziarn ($\bar{r}$ = 4,5 μm) i wysokim polem ładunku przestrzennego (E_{sc} = 0,25 kV/mm).

Tablica 5

Zależność względnej zmiany częstotliwości rezonansowej ($\Delta f_r/f_r$) od średnich rozmiarów ziarn ($\bar{r}$) i gęstości (ϱ) piezoceramiki PKR–28 otrzymywanej w różnych warunkach (T_s)

Warunki spiekania	Metoda spiekania	T_s [K]	$\Delta f_r/f_r$ (248–358 K)	$\bar{r}$ [μm]	$\varrho \cdot 10^3$ [kg/m^3]	E_{sc} [kV/mm]
1	HP	1423	0,15	4,5	8,0	0,25
2	CCM	1333	0,39	2,0	7,6	–
3	CCM	1353	0,20	3,6	7,8	–
4	CCM	1373	0,14	5,7	7,7	–
5	CCM	1393	0,15	5,0	7,8	0,25
6	CCM	1413	0,17	5,0	7,7	0,25
7	CCM	1433	0,16	8,7	7,7	0,24
8	CCM	1453	0,16	9,2	7,6	0,24
9	CCM	1473	0,17	16,4	7,5	0,23

Spiekanie metodą klasyczną CCM (ang. *ceramic classic method*), a więc bez przykładania do próbki ciśnienia zewnętrznego ($p_s = p_{atm.}$) prowadzono w szerokim zakresie temperatur spiekania (T_s = 1333–1473 K), co pozwoliło otrzymać próbki piezoceramiczne o różnej gęstości i porowatości.

Piezoceramika spiekana w warunkach (2), a więc w zbyt niskiej temperaturze, ma ziarna drobne i nieukształtowane oraz charakteryzuje się dużą koncentracją fazy niekrystalicznej. Piezoceramika taka ma słabe właściwości piezoelektryczne i wykazuje małą stabilność f_r.

Próbki piezoceramiczne spiekane w warunkach (3–6) są drobnoziarniste, charakteryzują się zbliżoną gęstością i wykazują wysoką temperaturową stabilność f_r, uwarunkowaną wartościami E_{sc} zbliżonymi do E_{sc} dla próbek otrzymanych w warunkach HT (1). W warunkach (7–9) otrzymuje się piezoceramikę gruboziarnistą, o wzrastającym średnim rozmiarze ziarn ($\bar{r}$) i malejącej gęstości (wzrastającej porowatości) w miarę wzrostu temperatury spiekania. Pole ładunku przestrzennego E_{sc} zachowuje wysokie wartości, co warunkuje dobrą stabilność piezoceramiki, pomimo iż jest ona gruboziarnista. Wynika stąd, że E_{sc} wzrasta nie tylko w miarę zmniejszania się $\bar{r}$ lecz również przy wzroście porowatości piezoceramiki. Oznacza to, że ze wzrostem gęstości piezoceramiki pole ładunku przestrzennego zmniejsza się. W piezoceramice nieporowatej (o dużej gęstości) pola ładunków przestrzennych, otaczających ziarna krystaliczne, częściowo się kompensują. Natomiast w gruboziarnistej piezoceramice porowatej kompensacja ta jest słabsza i stąd E_{sc} przyjmuje duże wartości [20].

2.6. ZWIĘKSZENIE TWARDOŚCI FERROELEKTRYCZNEJ (metoda V)

Do badań wybrano trzy roztwory stałe I, II i III, z których: I — charakteryzuje się podwyższoną twardością ferroelektryczną, II — średnią twardością ferroelektryczną i III

— małą twardością ferroelektryczną [21, 22]:

$$\begin{aligned} &\text{I. (x) } PbTiO_3 - PbZrO_3 - PbNb_{2/3}Zn_{1/3}O_3 - PnNb_{2/3}Mn_{1/3}O_3;\\ &\text{II. (x) } PbTiO_3 - PbZrO_3 - PbNb_{2/3}Mn_{1/3}O_3 - PbTa_{3/4}Li_{1/4}O_3;\\ &\text{III. (x) } PbTiO_3 - PbZrO_3 - PbNb_{2/3}Co_{1/3}O_3 - PbNb_{1/2}Sb_{1/2}O_3. \end{aligned} \tag{13}$$

Próbki piezoceramiczne (13) otrzymywano metodą prasowania na gorąco. Miały one dużą gęstość (ϱ = 7,9–8,0 g/cm^3), małą porowatość (p = 0,2%) i w przybliżeniu jednakowe rozmiary ziarn ($\bar{r}$ = 3,5–4 μm). Składy wszystkich trzech roztworów stałych (I, II, III) należą na diagramie fazowym do obszaru tetragonalnego z wysoką stabilnością częstotliwości rezonansowej. Skład I to znana już piezoceramika PKR–13 (patrz tablica 3, pkt. 7), natomiast podstawowe parametry próbek o składzie II i III przedstawiono w tablicach 6 i 7.

Piezoceramiczne roztwory stałe o składach ułożonych w kolejności III, II, I (rys. 11) wykazują wzrost pola koercji (E_c), parametru spontanicznej deformacji tetragonalnej (δ_T) i dobroci mechanicznej (Q_m) oraz spadek wartości przenikalności elektrycznej ($\varepsilon^{\sigma}_{33}/\varepsilon_0$) i współczynnika sprzężenia elektromechanicznego (k_p). Oznacza to, że kolejność składów III, II, I odzwierciedla wzrost twardości ferroelektrycznej badanych próbek piezoceramicznych. W miarę wzrostu twardości ferroelektrycznej zwiększa się temperaturowa

Tablica 6

Wpływ koncentracji $PbTiO_3$ na podstawowe parametry piezoceramiki o składzie (II):
(x) $PbTiO_3 - PbZrO_3 - PbNb_{2/3}Mn_{1/3}O_3 - PbTa_{3/4}Li_{1/4}O_3$

Lp.	$PbTiO_3$ [mol.%]	$\varepsilon^{\sigma}_{33}/\varepsilon_0$	k_p	$\operatorname{tg}\delta \cdot 10^2$ (E = 5 kV/m)	Q_m	T_c [K]	$\Delta f_r/f_r$ (213–358 K)	Faza
1	45,0	1190	0,596	0,27	1460	568	0,44	T
2	46,5	980	0,457	0,67	2900	608	0,37	T
3	47,5	1000	0,447	0,60	3150	603	0,28	T
4	48,0 (skład II)	950	0,437	0,64	3400	603	0,23	T
5	48,5	840	0,393	0,55	2000	608	0,31	T

Tablica 7

Wpływ koncentracji $PbTiO_3$ na podstawowe parametry piezoceramiki o składzie (III):
(x) $PbTiO_3 - PbZrO_3 - PbNb_{2/3}Co_{1/3}O_3 - PbNb_{1/2}Sb_{1/2}O_3$

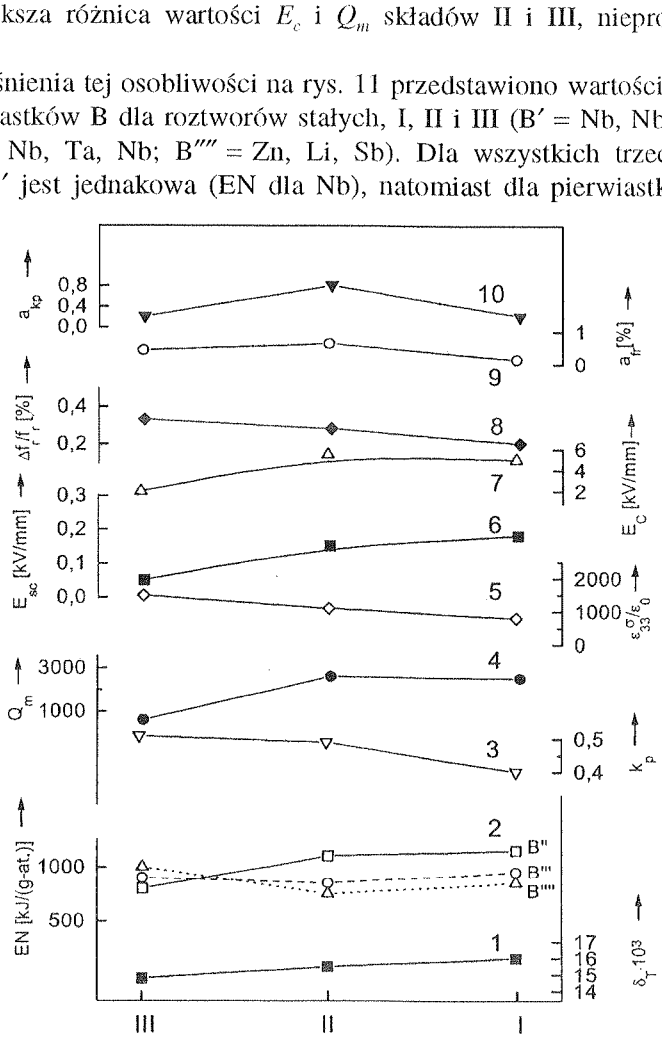
Lp.	$PbTiO_3$ [mol.%]	$\varepsilon^{\sigma}_{33}/\varepsilon_0$	k_p	$\operatorname{tg}\delta \cdot 10^2$ (E = 5 kV/m)	Q_m	T_c [K]	$\Delta f_r/f_r$ (213–358 K) [%]	Faza
1	46,5	1290	0,488	0,98	230	593	0,76	T
2	47,5 (skład III)	1070	0,401	0,99	365	592	0,27	T
3	48,5	1010	0,415	1,01	293	593	1,30	T

stabilność częstotliwości rezonansowej f_r (maleje $\Delta f_r/f_r$) w przedziale temperatury 213–358 K i pole ładunku przestrzennego (E_{sc}). Starzenie f_r i k_p, wyrażone przez a_{fr} i a_{kp} w %/dekadę, jest najsłabsze w przypadku piezoceramiki ferroelektrycznej najtwardszej, o składzie I, a najsilniejsze w przypadku składu II; współczynniki starzenia składu III mają wartość pośrednią.

Przy porównaniu właściwości materiałów o składach I, II i III na uwagę zasługuje:

- mała różnica wartości δ_T składów II i III;
- znacznie większa różnica wartości E_c i Q_m składów II i III, nieproporcjonalna do różnicy δ_T.

Celem objaśnienia tej osobliwości na rys. 11 przedstawiono wartości elektroujemności (EN) pierwiastków B dla roztworów stałych, I, II i III (B′ = Nb, Nb, Nb; B″ = Mn, Mn, Co; B‴ = Nb, Ta, Nb; B⁗ = Zn, Li, Sb). Dla wszystkich trzech składów EN pierwiastków B′ jest jednakowa (EN dla Nb), natomiast dla pierwiastków o zmiennej



Rys. 11. Wpływ ferroelektrycznej twardości na parametry piezoceramicznych roztworów stałych o składach:
I – $PbTiO_3PbZrO_3-PbNb_{2/3}Mn_{1/3}O_3-PbNb_{2/3}Zn_{1/3}O_3$,
II – $PbTiO_3-PbZrO_3-PbNb_{2/3}Mn_{1/3}O_3_PbTa_{3/4}Li_{1/4}O_3$,
III – $PbTiO_3_PbZrO_3-PbNb_{2/3}Co_{1/3}O_3-\bar{P}bNb_{1/2}Sb_{1/2}O_3$;
1 – δ_T; 2 – EN; 3 – k_p; 4 – Q_m; 5 – $\varepsilon^{\sigma}_{33}/\varepsilon_o$; 6 – E_{sc}; 7 – E_c; 8 – $\Delta f_r/f_r$ w zakresie (213-358 K); 9 – a_{fr}; 10 – a_{kp}; ferroelektryczna twardość roztworów stałych wzrasta zgodnie z kolejnością: III, II, I.

wartości, czyli dla $Mn^{II}/Mn^{III}/Mn^{IV}$, Co^{III}/Co^{II} i Sb^{V}/Sb^{III}, przyjęto średnie wartości EN, uwzględniając ich procentowy udział w tworzeniu roztworów.

Przy przejściu od składu I do II EN pierwiastków B″ nie zmienia się, natomiast EN B‴ i B⁗ zmniejszają się. Oznacza to, że musi się zmniejszyć stopień kowalentności wiązań B‴–O i B⁗–O, a w konsekwencji również jednorodny parametr deformacji (δ) składu II, w stosunku do składu I.

Przy przejściu od składu II do III, EN pierwiastków B″ zmniejsza się, a pierwiastków B‴ i B⁗ — wzrasta. Przeciwny charakter zmian EN pierwiastków B″ i B‴ (B⁗) prowadzi do tego, że różnice pomiędzy δ składów II i III są niewielkie δ (III) jest nieco mniejsze od δ (II)).

Brak zgodności między stopniem zmian δ oraz E_c oraz Q_m, przy przejściu od składu II do III, świadczy o tym, że jednorodny parametr deformacji (δ) nie jest w tym przypadku jedynym czynnikiem, warunkującym stan ferroelektrycznej twardości materiałów piezoceramicznych.

Uwzględniając rezultaty pracy [18] przeprowadzono badania procesu przepolaryzowania próbek o składach I, II, III stosując metodykę opracowaną przez I.A. Silantieva et. al. [20]. Na przykładzie tych badań stwierdzono, że dodatkowym czynnikiem, decydującym o ferroelektrycznej twardości i związanych z nią wartościami E_c i Q_m, jest pole ładunku przestrzennego E_{sc}, powstające podczas procesu starzenia piezoceramiki. Charakter zmian E_c i Q_m przy przejściu od składu I do II, jest uwarunkowany zmianami jednorodnego parametru deformacji (δ) i pola ładunku przestrzennego E_{sc}, natomiast przy przejściu od składu II do III — przede wszystkim zmianami E_{sc}.

W przypadku słabych pól $E_{sc} = 0{,}05$ kV/mm następuje znaczący wzrost ruchliwości ścian domenowych, a w konsekwencji znaczne zmniejszenie twardości ferroelektrycznej. Wartość $E_{sc} = 0{,}1$ kV/mm, wyznaczona dla składu II, jest prawdopodobnie najmniejszą graniczną wartością, przy której ściany domenowe mogą być jeszcze dostatecznie silnie zamocowane.

Z wcześniejszych badań wynika, że w roztworach stałych na bazie PZT zachodzi silne oddziaływanie ścian domenowych z wakansami tlenowymi V_o, które *„zakotwiczając"* domeny, zmniejszają ich ruchliwości w zewnętrznym polu elektrycznym [22–25]. W przypadku rozpatrywanych w niniejszej pracy czteroskładnikowych roztworów stałych, wakanse V_o mogą być źródłem pola ładunku przestrzennego. Obecności Li w składzie II towarzyszy zmniejszanie liczby V_o, gdyż Li może częściowo zajmować pozycje międzywęzłowe i w wyniku reakcji [23]:

$$Li_2O \rightarrow 2Li + V''_{Pb} + PbO \tag{14}$$

tworzyć dodatkowe wakanse ołowiowe (V''_{Pb}. W efekcie liczba V_o zmniejsza się. Wyjaśnia to przyczynę zmniejszenia się E_{sc} przy przejściu od składu I do II. Dalsze zmniejszanie E_{sc} przy przejściu od składu II do III, jest uwarunkowane prawdopodobnie obecnością w stałym roztworze III pierwiastków o zmiennej wartości (Sb^{V}/Sb^{III} i Co^{III}/Co^{II}), które przechodząc w stan o większym stopniu utlenienia ($Sb^{II} \rightarrow Sb^{V}$ i $Co^{II} \rightarrow Co^{III}$), wywołują zmniejszenie liczby wakansów V_o.

Tak więc w przypadku różnych materiałów piezoceramicznych na bazie PZT, o zbliżonych rozmiarach ziarn i porowatości, przyczyną różnych wartości E_{sc} może być różny stopień zdefektowania krystalitów, w tym przede wszystkim różna koncentracja wakansów V_o.

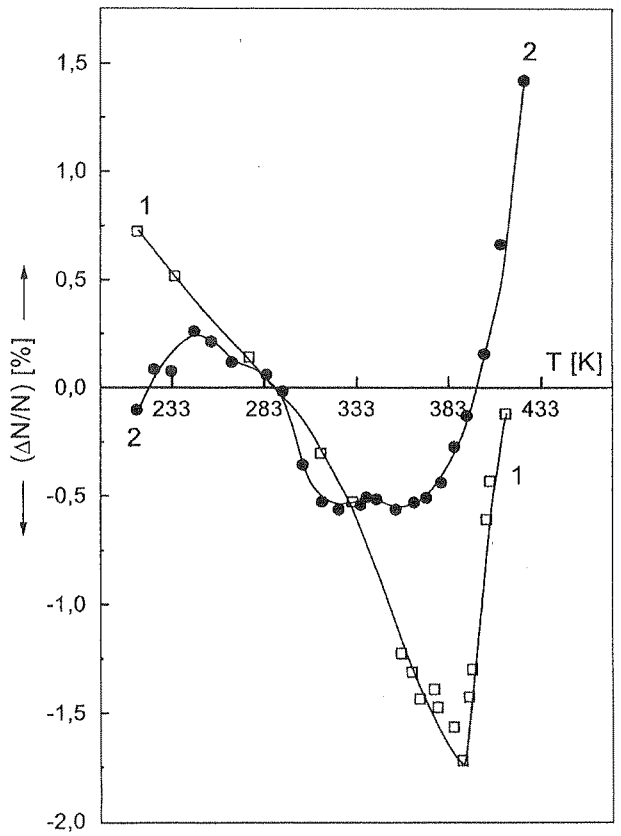
Względna temperaturowa zmiana częstotliwości rezonansowej, przy przejściu od składu III do II i do I, monotonicznie maleje czyli temperaturowa stabilność f_r wzrasta. Uwarunkowane to jest zmniejszeniem ruchliwości ścian domenowych, co przy przejściu od składu III do II związane jest przede wszystkim ze wzrostem ładunku przestrzennego, a przy przejściu od składu II do I — zarówno wzrostem E_{sc} jak jednorodnego parametru deformacji δ.

Przeprowadzono również badania czasowej stabilności f_r, η i k_p, której miarą były szybkości starzenia a_{fr} i a_{kp}, mierzone w procentach na dekady [%/dekada]. Najmniejszymi wartościami a_{fr} i a_{kp} (czyli największą czasową stabilnością f_r) charakteryzuje się skład I, co można związać z maksymalną wartością E_c (uwarunkowana maksymalnymi wartościami δ i E_{sc}), utrudniającą odwrotne reorientacje domen (rys. 11). Przy przejściu od składu I do II szybkości starzenia k_p i f_r wzrastają, co jest uwarunkowane wzrostem ruchliwości ścian domenowych (zmniejszenie E_c), uatwiającej odwrotne reorientacje domen. Czasowa stabilność składu III jest znacznie lepsza niż składu II, nie lepsza niż składu I. Poprawa czasowej stabilności k_p *i* f_r następuje w wyniku dalszego wzrostu ruchliwości ścian domenowych (jeszcze większy spadek E_c), prowadzącego do szybkiej relaksacji naprężeń natychmiast po wyłączeniu pola polaryzującego. Później zachodzi już stosunkowo słabe starzenie.

2.7. DOBÓR WARUNKÓW POLARYZOWANIA (metoda VI)

Na rys. 12 przedstawiono zależność względnych zmian stałej częstotliwościowej ($\Delta N/N$) w funkcji temperatury dla próbek piezoelektrycznych PZT z jednej partii, polaryzowanych w silnych (krzywa 1) i słabych (krzywa 2) zewnętrznych polach elektrycznych [16]. Widać, że warunki polaryzowania silnie wpływają na temperaturową stabilność częstotliwości rezonansowej. W przypadku niskonapięciowego polaryzowania (2) minimum $\Delta N/N$ (T) przesuwa się w stronę temperatury niższej, a w wyniku zmiany przebiegu stałej częstotliwości współczynnik TK_{fr} zmniejsza się prawie dwa razy w przedziale 213–358 K. Różny przebieg $\Delta N/N$ (T) — krzywa 1 i 2 na rys. 12 — jest uwarunkowany przede wszystkim różnym mechanizmem formowania się struktury domenowej podczas polaryzowania w słabych (2) i silnych (1) polach elektrycznych. Przy niskonapięciowej metodzie polaryzowania ($T_{pol} > T_c$) struktura domenowa powstaje w wyniku zarodkowania nowej fazy w kierunku przyłożonego pola polaryzującego podczas chłodzenia poprzez punkt Curie. Natomiast w przypadku wysokonapięciowej metody polaryzowania ($T_{pol} < T_c$) nowa struktura domenowa formuje się w wyniku obrotu domen pod działaniem pola elektrycznego.

Jedną z przyczyn, dla których niskonapięciowa metoda polaryzowania jest bardziej efektywna w procesie stabilizowania struktury domenowej (a tym samym f_r) niż metoda



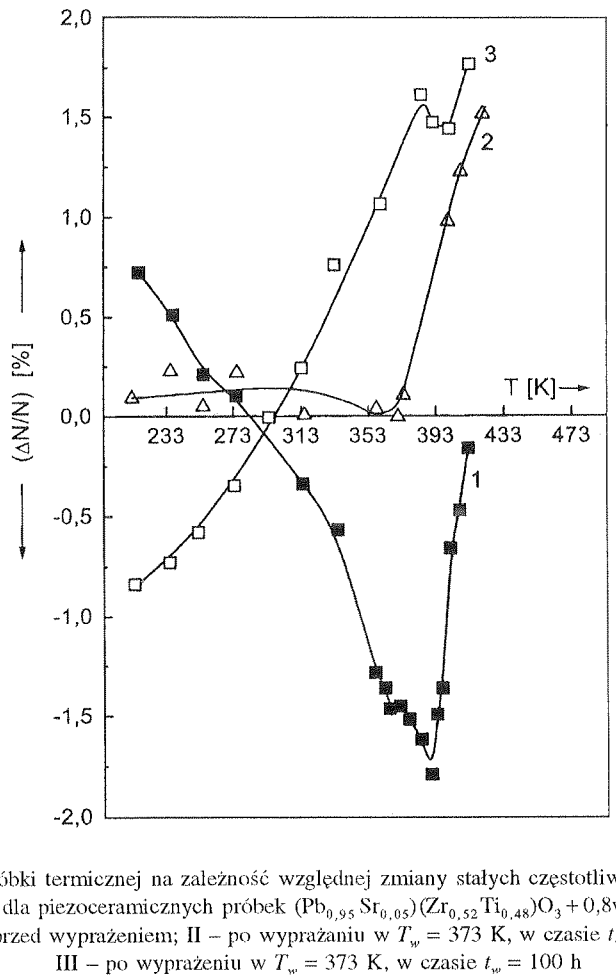
Rys. 12. Zależność względnej zmiany stałej częstotliwościowej ($\Delta N/N$) od temperatury (T) dla jednej serii próbek piezoceramiki typu PZT, polaryzowanych różnymi metodami [16]:
1 – metoda niskotemperaturowa (silne pola polaryzujące E_{pol});
2 – metoda wysokotemperaturowa (słabe pola polaryzujące E_{pol})

wysokonapięciowa, jest zmiana uporządkowania defektów podczas podgrzania próbek do wysokiej temperatury przed włączeniem słabego stałego pola elektrycznego. Równocześnie zachodzi naciekanie ładunków objętościowych, które podobnie jak defekty stabilizują strukturę domenową, zamocowując polaryzację spontaniczną powstającą w punkcie Curie [26].

2.8. DŁUGOTRWAŁE WYPRAŻANIE (metoda VII)

Na rys. 13 przedstawiono zależność względnej zmiany stałych częstotliwościowych ($\Delta N/N$) od temperatury (T) dla próbek piezoceramiki typu PZT o składzie

$(Pb_{0,95}Sr_{0,05})(Ti_{0,48}Zr_{0,52})O_3 + 0{,}8\%$ wag. Cr_2O_3 przed obróbką termiczną (krzywa 1), po wyprażaniu w $T_w = 373$ K przez $t_w = 2$ h (krzywa 2) i po wyprażaniu w $T_w = 373$ K przez $t_w = 100$ h (krzywa 3).



Rys. 13. Wpływ obróbki termicznej na zależność względnej zmiany stałych częstotliwościowych ($\Delta N/N$) od temperatury (T) dla piezoceramicznych próbek $(Pb_{0,95}Sr_{0,05})(Zr_{0,52}Ti_{0,48})O_3 + 0{,}8$wag.% Cr_2O_3 [16]: I – przed wyprażeniem; II – po wyprażaniu w $T_w = 373$ K, w czasie $t_w = 3$h; III – po wyprażeniu w $T_w = 373$ K, w czasie $t_w = 100$ h

W pierwszym przypadku na krzywej $\Delta N/N$ (T) poniżej $T = 373$ K obserwuje się zmniejszenie N, natomiast w zakresie $T > 373$ K — wzrost N ze wzrostem T. Oznacza to, że dla $T > 373$ K temperaturowy współczynnik częstotliwości rezonansowej jest dodatni ($TK_{fr} > 0$); stan taki nazywany jest wysokotemperaturowym. Powtórne pomiary $\Delta N/N$ (T) prowdzone na tych samych próbkach wykazały, że kolejne podgrzewanie do $T = 423$ K silnie wpływa na kształt krzywej 1 (rys. 13): minimum $\Delta N/N$ (T) przesuwa się w stronę niskich temperatur, co oznacza, że w wyniku podgrzewania zachodzi stabilizacja stanu wysokotemperaturowego.

Po obróbce termicznej (rys. 13; krzywe 2 i 3) anomalia $\Delta N/N$ (T) w rejonie $T =$ $= 373$ K jest zachowana, lecz dłuższe wyprażanie powoduje zmniejszenie i przesunięcie minimum $\Delta N/N$ (T) w stronę wyższych temperatur. W wyniku wyprażania przez $t_w = 100$ h (krzywa 3) znak TK_{fr} poniżej $T = 373$ K zmienia się na dodatni. Wynika z tego, że można tak dobrać warunki obróbki termicznej, aby w jej wyniku f_r przestało się praktycznie zmieniać wraz z temperaturą w przedziale 213–358 K (−60 ÷ 85°C). W omawianym przypadku najmniejszą wartość $TK_{fr} \leq (10–15) \cdot 10^{-6}\text{K}^{-1}$ osiąga się w wyniku obrótki w $T_w = 373$ K przez $t_w = 2$ h.

Podobne rezultaty otrzymano również w pracach [27, 28]. W pracy [28] badano wpływ obróbki termicznej w różnych temperaturach w granicach 373–473 K. W wyniku obróbki termicznej w $T_w = 473$ K przez $t_w = 1$ h wartość TK_{fr} zmienia się od $(-30 \div 20) \cdot 10^{-6}\text{K}^{-1}$ do $(+50 \div 100) \cdot 10^{-5}\text{K}^{-1}$. Wyprażanie wpływa na wartość TK_{fr} tylko w tym przypadku, gdy prowadzone ono jest po uprzednim spolaryzowaniu próbki. Podczas obróbki cieplnej zachodzi jednak proces termicznego depolaryzowania, w wyniku którego następuje pogorszenie piezoelektrycznych parametrów próbki piezoceramicznej. Jest to szczególnie niekorzystne w przypadku wykorzystania piezoceramiki do budowy filtrów pasmowych.

W pracy [29] badano wpływ długotrwałego wyprażania ($t_w = 25$; 50; 100; 150 h) na stabilność f_r spolaryzowanej piezoceramiki trójskładnikowej typu PZT, stosując jednak nieco niższą temperaturę wyprażania ($T_w = 343$ K). Również w tym przypadku obserwowano zmiany temperaturowej stabilności częstotliwości rezonansowej próbek piezoceramicznych. Stwierdzono, że temperaturowa stabilność w $T < 273$ K (w zakresie temperatur ujemnych w skali Celsjusza) praktycznie nie zmienia się, natomiast w $T > 273$ K (w zakresie temperatur dodatnich w skali Celsjusza) zmienia się bardzo wyraźnie. Najlepsze rezultaty otrzymano w przypadku $t_w = 100–150$ h.

Stabilizowanie f_r próbek piezoceramicznych metodą obróbki termicznej nie można uznać za sposób optymalny, gdyż:

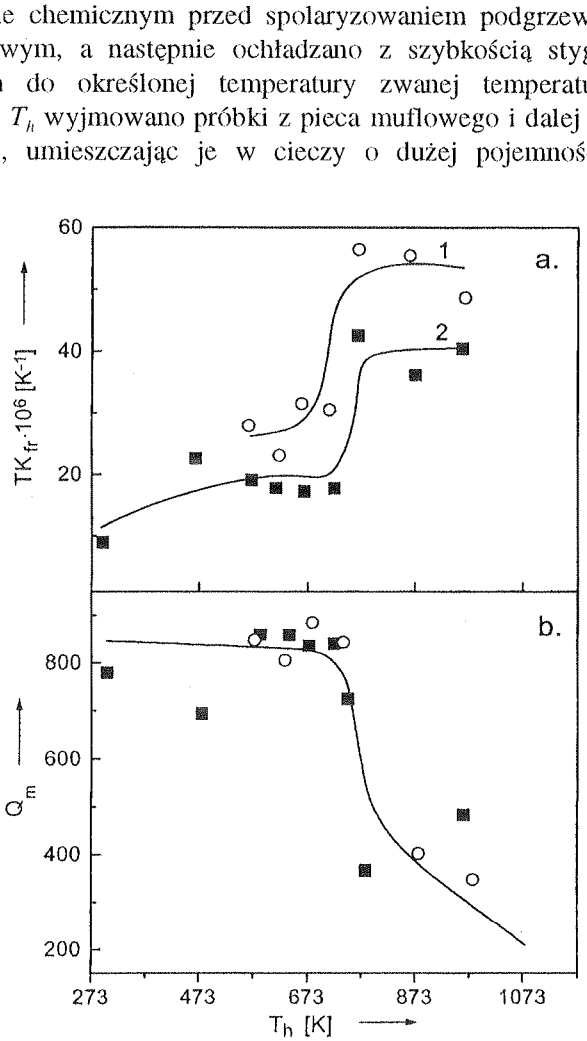
- jest zbyt czasochłonna ($t_w = 100–150$ h), a więc tym samym nieekonomiczna;
- wartość TK_{fr} stabilizuje się tylko na okres 2–3 miesięcy po procesie obróbki.

2.9. HARTOWANIE (metoda VIII)

Innym technologicznym sposobem stabilizowania częstotliwości rezonansowej i wartości piezoelektrycznych parametrów ceramiki typu PZT jest proces hartowania, czyli ochładzanie piezoceramiki z przejściem przez punkt Curie [30]. Struktura domenowa ferroelektrycznej ceramiki jest uwarunkowana dążeniem do minimum energii swobodnej (F). Jednak przy gwałtownym ochładzaniu poprzez punkt Curie (T_c) nie jest spełniony warunek minimum F. W punkcie Curie powstaje duża liczba zarodków nowej (ferroelektrycznej) fazy. Zarodki te zorientowane są w różnych kierunkach, z czym wiąże się wzrost ścian domenowych i sprężystej energii deformacji. Powstająca w ten sposób struktura domenowa stabilizuje się w miarę upływu czasu wskutek migracji ścian

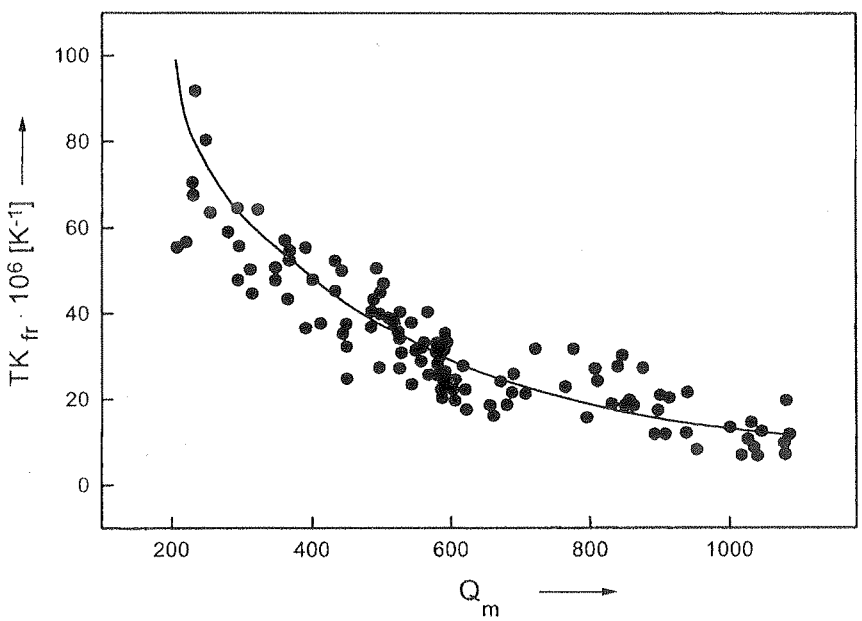
domenowych. Układ przechodzi od stanu metastabilnego w stabilny stan równowagi termodynamicznej, w którym ruchliwość ścian domenowych jest znacznie mniejsza. Czas tego przejścia zależy od jakości piezoceramiki oraz od stopnia tetragonalnej deformacji komórki elementarnej w punkcie Curie ($Pm3m \rightarrow P4mm$).

Wpływ warunków ochładzania poprzez punkt Curie badano na przykładzie piezoceramiki typu PZT o składzie $(Pb_{0,95}Sr_{0,05})(Zr_{0,56}Ti_{0,44})O_3 + 1$ at. % Cr_2O_3 [30]. Próbki o takim składzie chemicznym przed spolaryzowaniem podgrzewano do $T_w = 1073$ K w piecu muflowym, a następnie ochładzano z szybkością stygnięcia muflonu $v_T =$ $= (3–4)$ K/min do określonej temperatury zwanej temperaturą hartowania (T_h). W temperaturze T_h wyjmowano próbki z pieca muflowego i dalej gwałtownie schładzano (hartowano), umieszczając je w cieczy o dużej pojemności cieplnej. Warunki



Rys. 14. Zależność temperaturowego współczynnika częstotliwości rezonansowej TK_{fr} (a) i dobroci mechanicznej Q_m (b) od temperatury hartowania (T_h) piezoceramicznych rezonatorów o składzie: $(Pb_{0,95}Sr_{0,05})(Zr_{0,56}Ti_{0,44})O_3 + 0,1$ wag.% Cr_2O_3 [30]:
a) TK_{fr} 1 – pole polaryzujące $E_{pol} = 30$ kV/cm,
2 – pole polaryzujące $E_{pol} = 40$ kV/cm;
b) $Q_m(T_h)$.

chłodzenia wpływają na wartość TK_{fr} i dobroć mechaniczną (Q_m) piezoceramiki. Ostra zmiana tych parametrów zachodzi dla $T_h > T_c$. W przypadku $T_h > 773$ K zmniejsza się Q_m, natomiast wzrasta TK_{fr} (rys. 14). W zakresie $T_h < T_c$ silnie zmniejszają się parametry piezoelektryczne ceramiki typu PZT. Przykładowo, dla T_h = (673–773 K) zmniejsza się temperaturowa stabilność f_r i pogarsza się dobroć mechaniczna (Q_m). Zmiany te objaśnić można powstaniem dużej liczby defektów i wzrostem naprężeń termicznych. Wartości Q_m i TK_{fr} są bardzo czułe na obecność defektów i naprężeń mechanicznych w piezoceramice i są ściśle ze sobą powiązane: wzrostowi Q_m odpowiada zmniejszenie TK_{fr} (rys. 15). Optymalną temperaturą hartowania, dla otrzymania piezoceramiki z największą wartością dobroci mechanicznej i najlepszą temperaturową stabilnością f_r, okazała się T_h = 603–623 K (praktycznie pokrywająca się z T_c dla wcześniej podanego składu piezoceramiki typu PZT).



Rys. 15. Współzależność temperaturowego współczynnika częstotliwości rezonansowej TK_{fr} i dobroci mechanicznej Q_m hartowanych rezonatorów piezoceramicznych na bazie PZT [29]

Hartowanie nie znalazło szerokiego zastosowania jako metoda zwiększająca temperaturową stabilność f_r ze względu na wpływ składu chemicznego i jakości piezoceramiki na:

- długość czasu osiągania przez strukturę piezoceramiki stanu równowagi termodynamicznej;
- nietrwałość zachowania w piezoceramice stanu wysokiej stabilności temperaturowej f_r.

2.10. NAPROMIENIOWANIE (metoda IX)

W wyniku napromieniowania piezoceramiki promieniowaniem γ zmienia się kształt pętli histerezy i wzrasta pole koercji [31]. Na tej podstawie przyjęto, że promieniowanie γ może być efektywnym czynnikiem stabilizującym parametry piezoceramiki. W pracy [32] stosowano następujące dawki absorbowane promieniowania: 10; 20; 50 i 100 Mrad. Jednak w krótkim czasie od momentu napromieniowania próbek, pomimo tak silnej dawki γ, nie nastąpiła temperaturowa stabilizacja f_r.

Prowadzono również badania dotyczące wpływu innych rodzajów promieniowania na właściwości piezoceramiki [33]. Polepszenie temperaturowej i czasowej stabilności charakterystyk piezoceramicznych przetworników osiągano kosztem pogorszenia wartości ich parametrów piezoelektrycznych. Próbki napromieniowywano:

- promieniowaniem γ (Co^{60}) z mocą dawek ekspozycyjnych 100 i 900 rs^{-1} (rentgenów na sekundę), aż do osiągnięcia integralnej dawki zaabsorbowanej od 0,1 do 1000 Mrad;
- szybkimi neutronami o energii $E = 0{,}7$ MeV, aż do strumienia $\varphi = 3{,}7 \cdot 10^{17}$ N $\cdot$ cm^{-2};
- szybkimi neutronami z reaktora jądrowego o energii $E > 0{,}1$ MeV, aż do strumienia $\varphi = 10^{18}$ N $\cdot$ cm^{-2}.

Wszystkie rodzaje stosowanego napromieniowania wywołują jakościowo jednakowy wpływ na elektrofizyczne właściwości piezoceramiki. Radiacyjna odporność badanych próbek piezoceramicznych silnie zależy od składu roztworu stałego typu PZT. We wszystkich przypadkach napromieniowanie:

- zwiększa ferroelektryczną twardość piezoceramiki (twardość ta przejawia się w zmniejszeniu ruchliwości ścian domenowych oraz zwiększeniu f_r i v_R (gdzie v_R — prędkość dźwięku);
- prowadzi do częściowego zdepolaryzowania piezoceramiki, o czym świadczy zmniejszenie wartości d_{31}.

Efekty obserwowane po napromieniowaniu próbek piezoceramicznych są analogiczne do efektów towarzyszących procesowi starzenia, tzn. czym większa ruchliwość ścian domenowych (im szybciej zachodzi starzenie), tym niższa radiacyjna odporność ceramicznych piezorezonatorów.

2.11. KOMPENSACJA (metoda X)

Dla osiągnięcia wysokiej stabilności f_r w pracy [34] zastosowano metodę kompensacji, polegającą na wykorzystaniu w złożonych przetwornikach piezoelektrycznych rezonatorów metalicznych o różnych wartościach i znakach TK_{fr}. Zmieniając stosunek mas elementów piezoelektrycznych i metalicznego rezonatora zbudowano przetworniki z elementami piezoceramicznymi i piezomonokrystalicznymi, których TK_{fr} znajdował się w zakresie $\pm 10 \cdot 10^{-6}$ K^{-1}. Stwierdzono przy tym wpływ szybkości zmiany temperatury na wartość TK_{fr}. W ujemnym zakresie temperatury ($T < 273$ K) o wynikach pomiaru $\Delta f_r / f_r$ decydują szybkość ochładzania i czas wytrzymywania w temperaturze pomiaru. W przypadku szybkiego ochładzania wyniki pomiarów nie są powtarzalne. Prawdopodobną przyczyną tego efektu jest udar cieplny.

2.12. TERMOCYKLOWANIE (metoda XI)

Temperaturowa stabilność częstotliwości rezonansowej jest uwarunkowana stanem struktury domenowej oraz jej zmianą pod wpływem oddziaływań zewnętrznych. Wśród czynników wpływających na ruchliwość ścian domenowych wymienić należy defekty sieci krystalicznej [35–37]. Wskazuje na to zmniejszenie tarcia wewnętrznego Q_m^{-1} [35] oraz wzrost pola koercji E_c [36] wraz ze wzrostem liczby wakansów w węzłach anionów O^{2-} perowskitowej komórki ABO_3. Defekty te są zaporami dla ruchu ścian domenowych (zamocowują strukturę domenową). Najsilniejszy wpływ na ruchliwość ścian domenowych mają jednak liniowe defekty sieci krystalicznej [37]. Wokół dyslokacji powstają silne pola elektryczne w wyniku naciekania własnych naładowanych defektów punktowych (*atmosfera Debye'a-Hückela*) oraz atomów domieszkowych (*atmosfera Coltrella*). Czym wyższa gęstość dyslokacji i silniejsze elektryczne pola lokalne tym efektywniej struktura dyslokacyjna ogranicza ruchliwość ścian domenowych. Istnieje więc ścisły związek pomiędzy koncentracją liniowych defektów struktury (dyslokacji) a temperaturową i czasową stabilnością częstotliwości rezonansowej piezoceramicznych przetworników elektroakustycznych [38].

Koncentracja dyslokacji w ferroelektrycznej ceramicie może zwiększyć się w wyniku podgrzania próbek do temperatury wyższej od temperatury Curie ($T_w > T_c$), a następnie chłodzenia poprzez punkt Curie do temperatury pokojowej (T_r). W procesie grzania i chłodzenia zachodzą dwie kolejne przemiany fazowe: ferroelektryk ($T < T_c$) $\leftrightarrow$ paraelektryk ($T > T_c$). W czasie przemian fazowych krystality piezoceramiki doznają dużych naprężeń mechanicznych, porównywalnych z naprężeniami pozwalającymi przekroczyć granicę sprężystości materiału. Naprężenia te sprzyjają zmianie rozkładu pierwotnych defektów i zarodkowaniu nowych defektów, a szczególnie dyslokacji [39]. W tej sytuacji ważne stało się zbadanie, jak wielokrotne grzanie i chłodzenie próbki poprzez punkt Curie, czyli cykliczne przemiany fazowe, wpływają na temperaturową stabilność f_r i wartości podstawowych parametrów dielektrycznych i piezoelektrycznych piezoceramiki. Poddawanie ferroelektryków takim wielokrotnym, cyklicznym przemianom fazowym w literaturze nosi nazwę termicznego cyklowania lub t e r m o c y k l o w a n i a [40].

Materiał badań stanowiły dwie grupy materiałów piezoceramicznych typu PZT otrzymanych metodą prasowania na gorąco:

I. Materiały piezoceramiczne o wysokiej stabilności temperaturowej f_r produkowane na skalę półprzemysłową w Zakładach PIEZOPRIBOR w Rostowie nad Donem: PKR-15, PKR-28, PKR-21;

II. Materiały piezoceramiczne na bazie czteroskładnikowych roztworów stałych o niskiej temperaturowej stabilności f_r wyprodukowane w warunkach laboratoryjnych:
   a) $PbTiO_3 - PbZrO_3 - PbNb_{2/3}Co_{1/3}O_3 - PbNb_{2/3}Ni_{1/3}O_3$;
   b) $PbTiO_3 - PbZrO_3 - PbW_{1/2}Co_{1/2}O_3 - PbW_{1/2}Zn_{1/2}O_3$.

Ze względu na to, że wpływ procesu termocyklowania na temperaturową stabilność f_r i parametry piezoelektryczne wszystkich badanych materiałów był jakościowo podobny, w niniejszej pracy zostaną przykładowo przedstawione wyłącznie rezultaty badań piezoceramiki PKR-28.

Próbki PKR-28 posortowane zostały na dwie grupy:

1) próbki o wysokiej temperaturowej stabilności częstotliwości rezonansowej ($\Delta f_r/f_r = 0{,}20-0{,}23\%$);
2) próbki o niskiej temperaturowej stabilności częstotliwości rezonansowej $\Delta f_r/f_r = 0{,}36-0{,}40\%$).

Próbki grupy 2, czyli te o wysokiej wartości $\Delta f_r/f_r$, otrzymano w warunkach odbiegających od optymalnych reżimów procesu technologicznego (najczęściej były to odstępstwa od optymalnych warunków spiekania). Średni rozmiar ziarn w takich próbkach nie przekraczał $\overline{r} = 2-2{,}5$ μm, podczas gdy dla 1 grupy $\overline{r} = 3{,}5-5$ μm.

Próbki poddawano cyklicznej obróbce termicznej ($N_{Tc} = 1$–40 cykli), polegającej na podgrzewaniu próbek do $T_w = 673$ K ($T_c = 573$ K), a następnie chłodzeniu do $T_r = 293$ K (rys. 16 i 17). Po termocyklowaniu dokonywano pomiarów stabilności częstotliwości rezonansowej ($\Delta f_r/f_r$), tangensa kąta strat dielektrycznych (tg δ), dobroci mechanicznej (Q_m), tarcia wewnętrznego (Q_m^{-1}), szerokości połówkowej refleksów rentgenowskich *00l* i *h00* ($\Delta_{00l,\,h00}$) i kąta dezorientacji składowych 90° bliźniaków strukturalnych, czyli 90° domen (α_{bl}) w funkcji liczby cykli (N_{Tc}) grzania i chłodzenia poprzez punkt Curie. Badania $\Delta f_r/f_r$ prowadzone były w zakresie temperatury od 213 do 358 K, natomiast pomiary pozostałych parametrów elektrofizycznych i rentgenowskich prowadzono w temperaturze pokojowej. Szerokość połówkowa refleksów rentgenowskich ($\Delta_{00l,\,h00}$) jest miarą stopnia doskonałości struktury krystalitów (bliźniaków strukturalnych) [39], natomiast zmiany kąta bliźniakowania (α_{bl}) w funkcji N_{Tc} wiązać należy z liniowym naruszeniem translacyjnej symetrii. Dla (1) grupy próbek PKR-28, a więc tych o wysokiej temperaturowej stabilności f_r, zależności jakości struktury krystalicznej ($\Delta_{00l,\,h00}$) i kąta bliźniakowania (α_{bl}) od liczby przemian fazowych (N_{Tc}) silnie z sobą korelują. Na podstawie porównawczych badań $\Delta_{00l,\,h00}$ (N_{Tc}) dla kryształów $PbTiO_3$ i piezoceramiki PKR-28, stopień jakości struktury krystalitów należy wiązać z koncentracją liniowych defektów struktury (dyslokacji).

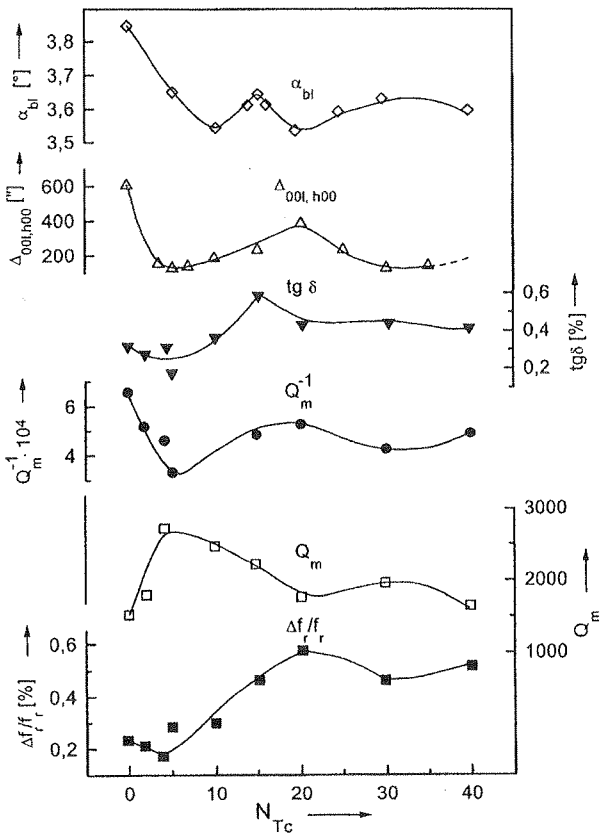
Wartość $\Delta f_r/f_r$ (N_{Tc}) dla próbek grupy 1 zmienia się następująco:

- dla $N_{Tc} = 1$–5, $\Delta f_r/f_r$ zmniejsza się do wartości 0,18%;
- dla $N_{Tc} = 6$–20, $\Delta f_r/f_r$ znacząco wzrasta do wartości 0,60%;
- dla $N_{Tc} = 20$–40, $\Delta f_r/f_r$ płynnie maleje do wartości 0,4–0,5%.

Analogiczny charakter mają zmiany (Q_m^{-1} (N_{Tc}), tg $\delta(N_{Tc})$, $\Delta_{00l}(N_{Tc})$ i $\alpha_{bl}(N_{Tc})$.

Na podstawie tych danych eksperymentalnych można wnioskować, że przy $N_{Tc} = 1$–5 w krystalitach następuje zburzenie istniejącego pierwotnie rozkładu defektów. Procesowi temu towarzyszy przemieszczanie się defektów do granic bliźniakowania, a w konsekwencji zmniejszenie koncentracji defektów we wnętrzu krystalitów. Zamocowanie granic domen strukturalnych przez defekty znacznie zmniejsza ich ruchliwość, co prowadzi do zmniejszenia $\Delta f_r/f_r$, czyli zwiększenia temperaturowej stabilności f_r. Potwierdzeniem tego mechanizmu jest równoczesne zmniejszenie Q_m^{-1} i tg δ w funkcji N_{Tc}, jako wielkości charakteryzujących rozpraszanie energii mechanicznej i elektrycznej.

Następne cykle przemian fazowych ($N_{Tc} = 6$–20) zwiększają koncentrację defektów w objętości ziarn piezoceramiki, co prowadzi do procesów uwarunkowanych wzajemnym oddziaływaniem defektów z objętości, z defektami na granicach (wzrost Q_m^{-1} i tg δ).



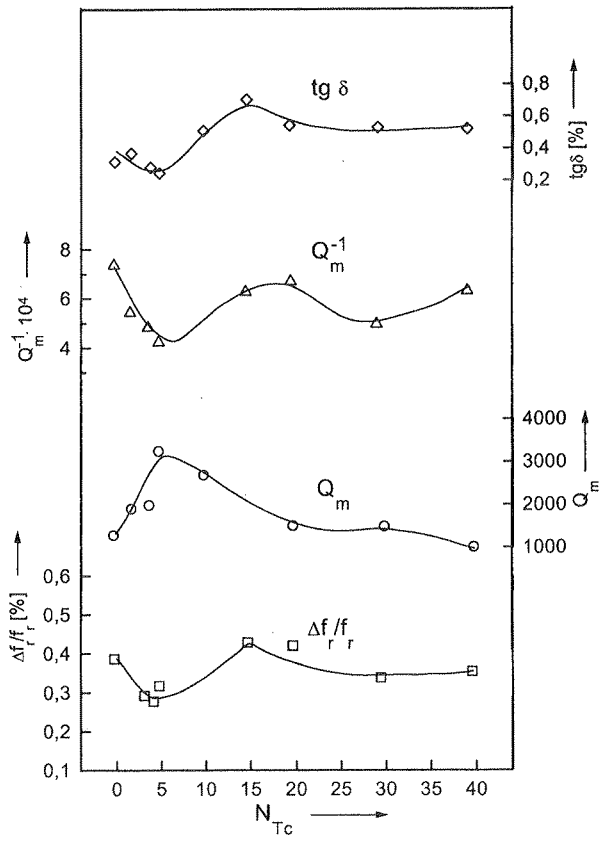
Rys. 16. Zależność elektrofizycznych i strukturalnych parametrów próbek wieloskładnikowej piezoceramiki typu PZT (PKR-28) o wysokiej temperaturowej stabilności częstotliwości rezonansowej ($\Delta f_r/f_r < 0{,}25\%$, w zakresie od 213 do 258 K) od liczby cykli przemian fazowych (N_{Tc}); α_{bl} – kąt dezorientacji składowych 90° bliźniaków strukturalnych; $\delta_{00l,h00}$ – szerokość połówkowa refleksów rentgenowskich *00l*, *h00*; tgδ – tangens kąta strat dielektrycznych; Q_m^{-1} – tarcie wewnętrzne; $\Delta f_r/f_r$ – względna zmiana częstotliwości rezonansowej

Wzrost $\Delta f_r/f_r$ dla $N_{Tc} = 6$–20 (zmniejszenie stabilności f_r) można wiązać z ogólną niestabilnością nieuporządkowanego układu defektów, co wpływa na pogorszenie się charakterystyk strukturalnych i parametrów elektrofizycznych piezoceramiki.

W przypadku $N_{Tc} > 15$–20 prawdopodobnie w objętości ziarn formuje się stabilny układ defektów, który wpływa stabilizująco na właściwości piezoceramiki.

W wyniku dalszego zwiększania cykli przemian fazowych ($N_{Tc} > 40$) zachodzi dalsze nasycenie defektami wnętrza ziarn, co prowadzi do pojawienia się skupisk defektów (tzw. *„ultradyslokacji"*), a w konsekwencji do zniszczenia próbek. Efekt taki nazywany jest zmęczeniem ferroelektryka, wywołanego cyklem kolejnych przemian fazowych.

Badano również wpływ termocyklowania na stabilność f_r oraz wartość charakterystyk strukturalnych i parametrów elektrofizycznych 2 grupy próbek PKR-28, charakteryzujących się mniejszą wyjściową stabilnością częstotliwości rezonansowej. Zależno-



Rys. 17. Zależność elektrofizycznych parametrów próbek wieloskładnikowej piezoceramiki typu PZT (PKR-28) o małej stabilności częstotliwości rezonansowej ($\Delta f_r/f_r \approx 0{,}4\%$ w zakresie od 213 do 258 K) od liczby cykli przemian fazowych (N_{Tc}): tgδ – tangens kąta strat dielektrycznych; Q_m^{-1} – tarcie wewnętrzne; Q_m – dobroć mechaniczna; $\Delta f_r/f_r$ – względna zmiana częstotliwości rezonansowej

ści $\Delta f_r/f_r$ (N_{Tc}), Q_m (N_{Tc}), Q_m^{-1} (N_{Tc}) i tg δ (N_{Tc}) dla 2 grupy próbek są jakościowo podobne do odpowiednich zależności dla 1 grupy próbek PKR-28. Różnice sprowadzają się do:

- szybszego zmniejszania się $\Delta f_r/f_r$ (z 0,39% do 0,27%) dla $N_{Tc} = 4$–5;
- większego wzrostu Q_m (2–3 razy) dla $N_{Tc} = 4$–5.

Osobliwości te związane są z mniejszą liczbą ścian domenowych w drobnioziarnistej piezoceramice PKR-28 (próbki 2 grupy). Przy mniejszej liczbie ścian domenowych napływające na ich granice ładunki elektryczne mogą je silnie zamocować już po $N_{Tc} = 4$–5.

Z przeprowadzonych badań dotyczących wpływu termocyklowania na temperaturową stabilność częstotliwości rezonansowej piezoceramiki typu PZT wynika, że:

- zmiana stopnia zdefektowania i rozkładu defektów struktury w krystalitach podczas przemian fazowych prowadzi do zmian temperaturowej stabilności f_r;
- termocyklowanie piezoceramiki polepsza stabilność f_r przy $N_{Tc} = 5$–6;
- w przypadku piezoceramiki o zbyt niskiej wyjściowej stabilności f_r termocyklowanie

(N_{Tc} = 5–6) poprawia wprawdzie stabilność, ale nie na tyle, aby można ją wykorzystać do budowy pasmowych filtrów elektrycznych.

3. PODSUMOWANIE

Przedstawione metody (III-XI) zwiększania stabilności częstotliwości rezonansowej przetworników piezoceramicznych mają szereg niedostatków:

- dużą czasochłonność (np. przy wyprażaniu przez t_w = 150 h próbki wykazują dobrą stabilność f_r tylko przez 2–3 miesiące po obróbce cieplnej);
- dużą energochłonność (wysokie temperatury hartowania, duże dawki promieniowania itd.);
- obniżenie piezoelektrycznych parametrów w procesie stabilizowania f_r;
- częsta zawodność tych metod w technologii wieloseryjnej.

Z tych względów znacznie efektywne jest poszukiwanie nowych materiałów piezoceramicznych (metody I–II), które dzięki swojej strukturze i składowi chemicznemu charakteryzują się zbiorem parametrów optymalnych ze względu na konkretne zastosowanie w piezoelektronice.

Największy wpływ na stabilność f_r piezoceramiki na bazie PZT ma dobór odpowiedniego stosunku koncentracji Pb/Zr. Przeprowadzone w tym zakresie badania wykazały, że:

- największą temperaturową stabilność f_r wykazują roztwory stałe w fazie tetragonalnej (T), o składach położonych na diagramie fazowym w znacznej odległości od obszaru morfotropowego o dużej koncentracji $PbTiO_3$; wysoka stabilność f_r jest uwarunkowana w tym przypadku dużą wartością parametru spontanicznej deformacji komórki elementarnej (δ_T) i zmniejszeniem ruchliwości ścian domenowych;
- nieco mniejszą, ale również dobrą stabilnością f_r charakteryzują się składy w fazie romboedrycznej, położonej na diagramie fazowym w pewnej odległości od obszaru morfotropowego w stronę małych koncentracji $PbTiO_3$; wysoka stabilność f_r jest w tym przypadku uwarunkowana wzrostem stopnia różnych od 180° reorientacji domen w polu elektrycznym (η) i zmniejszeniem liczby ścian domenowych;
- najmniejszą temperaturową stabilność f_r wykazują dwufazowe (T+Re) składy z obszaru morfotropowego (MR); w obszarze tym δ i η mają wartości pośrednie w stosunku do faz T i Re, co prowadzi do wzrostu ruchliwości i liczby ścian domenowych, a tym samym do wzrostu $\Delta f_r/f_r$; zmniejszeniu stabilności f_r sprzyjają również granice międzyfazowe Re/T przemieszczające się pod wpływem pola pomiarowego i temperatury.

Przez modyfikowanie składu podstawowego $PbTiO_3$–$PbZrO_3$ o minimalnym $\Delta f_r/f_r$ tzw. twardymi domieszkami można dodatkowo zwiększyć stabilność f_r piezoceramiki. Stwierdzono, że:

- domieszki zmniejszające ruchliwość i liczbę ścian domenowych w ziarnach krystalicznych zwiększają temperaturową stabilność częstotliwości rezonansowej f_r piezoceramiki typu PZT;
- domieszki hamujące wzrost ziarn zwiększają temperaturową stabilność f_r piezoceramiki;

- zmniejszenie średnich rozmiarów ziarn i wzrostowi porowatości towarzyszy zwiększenie elektrycznego ładunku przestrzennego, zamocowującego domeny i zmniejszającego ruchliwość ścian domenowych, a w konsekwencji zwiększającego temperaturową stabilność f_r piezoceramiki.

Praktycznym osiągnięciem niniejszej pracy jest otrzymanie na bazie PZT przetworników piezoceramicznych o wysokiej stabilności częstotliwości rezonansowej ($\Delta f_r/f_r < 0{,}2\%$) i wysokich wartościach parametrów dielektrycznych i piezoelektrycznych.

Podziękowania:
Praca była wykonana częściowo w ramach grantu KBN Nr 7 T08D 005 17.

BIBLIOGRAFIA

1. E.G. Fesenko, A.Ya. Danciger, O.N. Razumovskaya: *Novye piezokeramischeskie materialy*. Rostov-na-Donu, RGU 1983.
2. J. Bochenek et. al.: *Structural Acoustics and Vibrations for Technology*. 9, 71 (2000).
3. J. Bochenek et. al.: *Structural Acoustics and Vibrations for Technology*. 9, 77 (2000).
4. E.G. Fesenko: *Semeistvo perovskita i segnetoelektrichestvo*. Moskva, Atomizdat, 1972.
5. *Polaryzacja piezokeramiki*. Red. E.G. Fesenko. Rostov-na-Donu, RGU, 1968.
6. *Titanat bariya*. Red. V.S. Filipiev. Rostov-na-Donu, RGU, 1971.
7. A.Ya. Danciger et. al.: *Tez Dokl. VI Mezhotraslevoi Konf., Donieck*. AN SSSR, 1978, s. 7.
8. E.G. Fesenko et al.: Izv. AN SSSR, ser. Neorg. mater. 16, 12, 2247 (1980).
9. B. Jaffe, W.R. Cook, H. Jaffe: *Piezoelectric ceramics*. London, Academic Press, 1971.
10. V.A. Isupow: FTT. 12, 5, 1380 (1970).
11. *Piezoelektricheskoe priborostroenie. T.I. Fizika segnetoelektricheskoi keramiki*. Red. A.V. Gorish. Moskva, Radiotekhnika, 1999.
12. A.E. Panich, M.F. Kupriyanov: *Fizika i tekhnologiya segnetokeramiki*. Rostov-na-Donu, RGU, 1989.
13. M.F. Kupriyanov, GM. Konstantinov, A.E. Panich: *Segnetoelektricheskie morfotropnye perekhody*, Rostov-na-Donu, RGU, 1992.
14. Y. Xu: *Ferroelectric materials and their applications*. New York, North-Holland, 1991.
15. N.B. Feldman: *Voprosy radioelektroniki*. ser. III, 7, 89 (1962).
16. V.P. Mokolova, N.B. Feldman, T.N. Lezginceva: *Elektronnaya tekhnika*. ser. IV, 1, 29 (1969).
17. N.B. Feldman: *Voprosy radioelektroniki*. ser. III, 7, 83 (1962).
18. *Materialy s piezoelektricheskimi svoistvami*. Patent Japoński Nr 44-17344, opublikowany w MKI SOCh. v. 35/49.
19. *Sostav piezoelektricheskoi keramiki*. Patent Japoński, Nr 44-26508, opublikowany w MKI COCh. v. 35/46.
20. I.A. Silantieva, I.S. Ibraimov, V.A. Golovin: *Piezoelektricheski keramicheski material*. Patent SSSR, Nr 963977, opublikowany w MKI SOCh. v. 35/46.
21. *Piezoelektricheski material*. Patent Japoński, Nr 47-9185, opublikowany w MKI SOCh. v. 35/46.
H. Ouchi, M. Nishida: *Piezoelectric ceramics*. Patent USA, Nr 3400076, opublikowany w MKI SOCh. v. 35/42.
23. *Element iz piezoelektricheskoi keramiki*. Patent Japoński, Nr 45-7467, opublikowany w MKI SOCh. v. 35/46.
24. *Compositions ceramiques piezoelectriques*. Patent Francuski, Nr 1551976, opublikowany w MKI SOCh. v. 35/00.

25. *Piezoelektricheski keramicheski material.* Patent RFN, Nr 1816171, opublikowany w MKI SOCh. v. 35/49.
26. Z. Surowiak: *Specjalny Zeszyt Naukowy Polit. Śląskiej.* ser. Akustyka, Akustooptyka, Optoelektronika, Gliwice, 1998, s. 75.
27. E.G. Smazhevskaya, N.B. Feldman: *Voprosy radioelektroniki.* ser. III, 5, 89 (1965).
28. N.B. Feldman, L.F. Mikhailova, T.N. Lezginceva: *Mater. Semin. „Metody polucheniya i issledovaniya piezo- i segnetomaterialov i iskhodnogo syriya iz nikh".* Moskwa, MDNTP, 1966, s. 116.
29. P.O. Gribovski, I.G. Davidiyan, A.P. Kaushanski: *Mater. Semin. „Metody polucheniya i issledovaniya piezo- i segnetomaterialov i iskhodnogo syriya iz nikh".* Moskwa, MDNTP, 1966, s. 143.
30. N.B. Feldman, T.N. Lezginceva: *Voprosy radioelektroniki.* ser. III, 5, 105 (1965).
31. V.A. Yurin, N.S. Zhelidev: *Tez. Dokl. II Vsesoyuz. Konf. po Fizike Dielektrikov.* Rostov-na-Donu, RGU, 1958, s. 12.
32. M.A. Akselrod, V.M. Elecki, Z.A. Frankova: *Voprosy radioelektroniki.* ser. III, 5, 96 (1965).
33. V.N. Zaicev, M.V Morozov, E.G. Savenkowa: *Mater. Semin. „Novye piezo-i segnetoterialy i ikh primenenie".* Moskwa, MDNTP, 1975, s. 128.
34. I.V. Ushakov, V.S. Kharitanov: *Tez. Dokl. „III Vsesoyuz. Konf. po Fizike Segnetoelektrikov".* T. 1. Moskwa, MGU, 1984, s. 73.
35. Wu Long, Wei Ching-Chuang, Wu Tien-Shon, Lin Hsi-Chuan, J. Phys. C.: *Solid State Phys.* 16, 14, (1983).
36. Wu Long, Wei Ching-Chuang, Wu Tien-Shon, Teng Chain-Chean: J. Phys. C.: *Solid State Phys.* 16, 14, (1983).
37. V.V. Prisedski: *Segnetomiagkost' i segnetozhestkost' piezokeramiki CTS.* W: Segneto- i piezomaterialy v uskorenii nauchno — tekhnicheskovo progressa. Moskva, MGU, 197, s. 61.
38. S.V. Gavrilyachenko, E.S. Cikhocki, E.S. Gagarina: *Tezisy Dokladov XII Vsiezoyuz. Konf. po Fizike Segnetoelektrikov.* R. 1. Rostov-na-Donu, RGU, 1989, s. 48.
39. V.P. Dudkevich, E.S. Cikhocki: *Tezisy Dokladov XII Vsiezoyuz. Konf. po Fizike Segnetoelektrikov.* T. 1. Rostov-na-Donu, RGU, 1989, s. 10.
40. Z. Surowiak, L. Kozielski, E.G. Fesenko: *Ceramics (Poland),* 54, 365 (1997).
41. P. Roy, B.S. Deshipande: *Indian J. Pure and Appl. Phys.* 22, 12, 708 (1984).
42. G.E. Savenkova, O.S. Didkovskaya, V.V. Klimov: *Mater. Semin. „Poluchenie i primenenie segneto- i priezomaterialov v narodnom khozayistve".* Moskwa, MDNTP, 1981, s. 78.
43. K. Najezchleb: *Ferroelectrics,* 26, 1-4 , 779 (1980).
44. H.H. Schnessler: *JEEE Trans. Sonics and Ultrosonics.* 21, 4, 257 (1974).
45. M. Takahaski: *Jap.J.Appl.Phys.* 10, 5, 643 (1971).
46. O.S. Didkovskaya, V.A. Zvonik, N.S. Morozova: *Izv AN SSSR, ser. Neorg. mater.* 13, 12, 2202 (1977).
47. V.V. Klimov et al.: *Ferroelectrics.* 7, 1-4, 383 (1974).
48. O. Toshio, W. Kikno: *J.Ceram.Soc.Jap.* 92, 9, 510 (1984).
49. M. Nicita, H. Duchi: *Piezoelektrishe Keramik.* Patent RFN, Nr 1646739, opublikowano w MKI SOCh. v. 17/00.
50. H. Tsutomu, H. Makoto: *Method of preparing a piezoelectric ceramics.* Patent USA Nr 4087366, opublikowano w MKI SOCh. v. 35/46.

Z. SUROWIAK, D. BOCHENEK, S.V. GAVRILYACHENKO, M.F. KUPRIYANOV

STABILITY OF RESONANCE FREQUENCY OF THE PIEZOCERAMIC RESONATORS AND THEIR DOMAIN STRUCTURE

Summary

Piezoceramic materials on the PZT basis are used as band electric filters. Such materials, apart from suitable values of electrophysical parameters must show high temperature stability of resonance frequency ($\Delta f_r/f_r<0{,}2\%$). Stability f_r of the piezoceramics based on two-component solution $Pb(Zr_{1-x}\ Ti_x)O_3$ is little ($\Delta f_r/f_r>0{,}5\%$). That is why intense examinations are held to decrease relative change of the resonance frequency in the working range of temperatures (213-358 K). Basing on literature data and our examinations there were shown basic methods of the temperature stability (f_r) increase, among these: long-term annealing, hardening, irradiation, compensation, selection of technology conditions, selection of polarization, conditions, thermocycling, chemical composition modification. It was shown that the most effective method of those is chemical composition selection and use of the multicomponent solid solution of $PbTiO_3-PbZrO_3-\sum_{i=1}^{n} Pb(B'_{1-a}B''_a)O_3$.

Key words: piezoceramic resonator, ceramics of PZT-type, resonance frequency, domain structure.

KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJI, 2000, 46, z. 4, ss. 553–587

Implementacja algorytmu konwolucji 2D w układach specjalizowanych VLSI oraz w procesorach ogólnego przeznaczenia i sygnałowych

KAZIMIERZ WIATR, ERNEST JAMRO

Katedra Elektroniki, Akademia Górniczo-Hutnicza w Krakowie

Otrzymano 2000.04.17
Autoryzowano 2000.07.27

W opracowaniu przedstawiono specjalizowane układy VLSI do realizacji obliczeń konwolucji 2D na przykładzie układów firm: Harris, Plessey i InMos. Stosowanie tych układów zapewnia wysokie parametry czasowe realizowanych obliczeń, ale wymaga stosowania dedykowanego otoczenia sprzętowego, związanego z dostarczeniem odpowiednio zorganizowanych danych wejściowych i odebraniem przetworzonych danych. Alternatywą dla tych układów jest realizacja algorytmu konwolucji 2D w procesorach ogólnego przeznaczenia i w procesorach sygnałowych. Dynamiczny rozwój tych procesorów w ostatnim okresie oferuje nowe możliwości w zakresie osiągania wysokich parametrów czasowych związanych z implementacją algorytmu dwuwymiarowej konwolucji na przykładzie obrazu wizyjnego. W szczególności omówiono architekturę superskalarną procesora Pentium z uwzględnieniem różnych jego wersji (P75, P166, P300, MMX) oraz procesory sygnałowe z rodziny DSP56000 firmy Motorola i TMS320C80 firmy Texas Instruments. Przedstawiono zagadnienia wpływu poszczególnych elementów architektury tych procesorów na realizację zadań obliczeniowych jakie występują w trakcie wykonywania programu realizującego algorytm konwolucji. Omówiono także wpływ metod konstruowania algorytmu i sposobu jego realizacji przez program użytkowy (na poziomie kodu maszynowego) na czas realizacji obliczeń. Poszczególne implementacje omówiono ze szczególnym uwzględnieniem tych aspektów, które wpływają na czas realizacji operacji konwolucji na danych obrazowych i innych podobnie zorganizowanych. Ważnym elementem przedstawionego opracowania są załączone wyniki prac eksperymentalnych.

Słowa kluczowe: systemy czasu rzeczywistego, przetwarzanie obrazów, architektury systemów procesorowych, procesory specjalizowane

1. WSTĘP

Operacja dwuwymiarowej konwolucji jest operacją często stosowaną przy filtracji i przetwarzaniu obrazu [1, 2]. Sam proces obliczeniowy można przedstawić za pomocą wzoru (1).

$$b_{y,x} = \frac{1}{K} \sum_{i=0}^{N} \sum_{j=0}^{M} w_{i,j} \cdot a_{y+i-(N-1)/2,\, x+j-(M-1)/2} \tag{1}$$

Pomimo swojej prostoty obliczeniowej operacja konwolucji stanowi poważny problem obliczeniowy ze względu na znaczną liczbę wykonywanych operacji, szczególnie dla dużych bloków danych, a z takimi mamy do czynienia w przypadku obrazów wizyjnych. Dla pojedynczego piksela konwolucja wymaga wykonania $L_M = N \cdot M$ mnożeń oraz $L_A = N \cdot M - 1$ dodawań. Przykładowo; jeżeli dysponujemy obrazem o rozdzielczości $N_X = 512$, $N_Y = 512$ pikseli oraz $N_O = 25$ obrazów na sekundę to wymagana liczba obliczeń dla $N = M = 3$ wynosi: $L_M = N_O \cdot N_X \cdot N_Y \cdot N \cdot M = 58\,982\,400$ mnożeń na sekundę oraz $L_A = N_O \cdot N_X \cdot N_Y \cdot (N \cdot M - 1) = 52\,428\,800$ sumowań na sekundę. Dlatego celem niniejszego opracowania jest analiza przydatności poszczególnych omawianych architektur, pod kątem która z nich będzie zdolna wykonać tak olbrzymią liczbę operacji w czasie rzeczywistym.

Innym ważnym aspektem jest dobór rozmiaru okna konwolucji. Zagadnienie to zostało szczegółowo opisane w [1, 2] i ze względu na tok niniejszych rozważań nie zostanie tutaj przytoczone. Kolejnym bardzo ważnym zagadnieniem są wartości współczynników konwolucji. Należy zwrócić uwagę, że nie tylko rozmiar okna konwolucji N, M (patrz równanie 1), ale także wartości współczynników $w_{i,j}$, K mają duży wpływ na poziom skomplikowania obliczeń. W ogólnym przypadku można przyjąć, że $K = 1$, a współczynniki $w_{i,j}$ są zmiennoprzecinkowe. Jednakże ze względu na uproszczenie architektury układu realizującego algorytm konwolucji (nazywanego często w skrócie układem konwolwera) współczynniki $w_{i,j}$ są liczbami całkowitymi, a współczynnik K często (szczególnie dla filtrów dolnoprzepustowych) przyjmuje wartość wyrażoną wzorem (2).

$$K = \sum_{i=1}^{N} \sum_{j=1}^{M} w_{i,j} \tag{2}$$

Wartość współczynnika K ma szczególnie duże znaczenie dla poziomu komplikacji obliczeń, ponieważ operacja dzielenia jest skomplikowana, a w niektórych architekturach jest nawet bezpośrednio niedostępna. Z tego względu współczynnik K często wybiera się jako wartość potęgi liczby 2, dzięki czemu operacja mnożenia jest

a) K=1/16

1	2	1
2	4	2
1	2	1

b) K=1

-1	-2	-1
0	0	0
1	2	1

c) K=1

1	1	1
1	-8	1
1	1	1

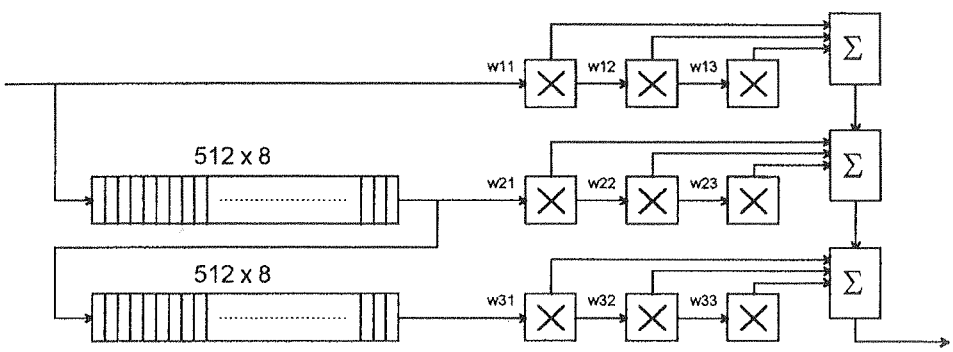
Rys. 1. Przykłady filtrów, które nie wymagają operacji mnożenia:
a) filtr dolnoprzepustowy, b) gradient Sobela, c) Laplasjan

równoważna przesunięciu bitowemu. Niestety założenie to oraz spełnienie równania (2) ogranicza dobór wartości współczynników $w_{i,j}$.

Rozwiązaniem optymalnym pod względem wymaganej architektury jest założenie, że nie tylko współczynnik K, ale również współczynniki $w_{i,j}$ są potęgą liczby 2. W tym wypadku operację mnożenia można zastąpić operacjami dodawania i przesuwania bitowego. Przykłady takich filtrów są przedstawione na rys. 1.

2. SPECJALIZOWANE PROCESORY KONWOLUCJI W TECHNOLOGII VLSI

Efektywną z punktu widzenia parametrów czasowych metodą implementacji operacji konwolucji w czasie rzeczywistym są układy specjalizowane zrealizowane w technologii VLSI. Do realizacji operacji konwolucji przez procesor specjalizowany musi on być wyposażony w linie opóźniające (rys. 2), zapewniające ciągły przepływ potoku danych przy jednoczesnym dostępie do pełnego otoczenia analizowanego punktu (opóźnienie równe czasowi trwania dwóch linii dla okna 3×3). Dokładność operacji kontekstowych zależy od wielkości analizowanego otoczenia punktu (np. 3×3, 5×5, 7×7). Wiąże się to z wyposażeniem procesora w odpowiednią liczbę linii opóźniających (odpowiednio: 2, 4, 6) i układów arytmetycznych (odpowiednio: 9, 25, 49 układów mnożących i 3, 5, 7 wielowejściowych układów sumujących).



Rys. 2. Struktura procesora do sprzętowej realizacji funkcji kontekstowych

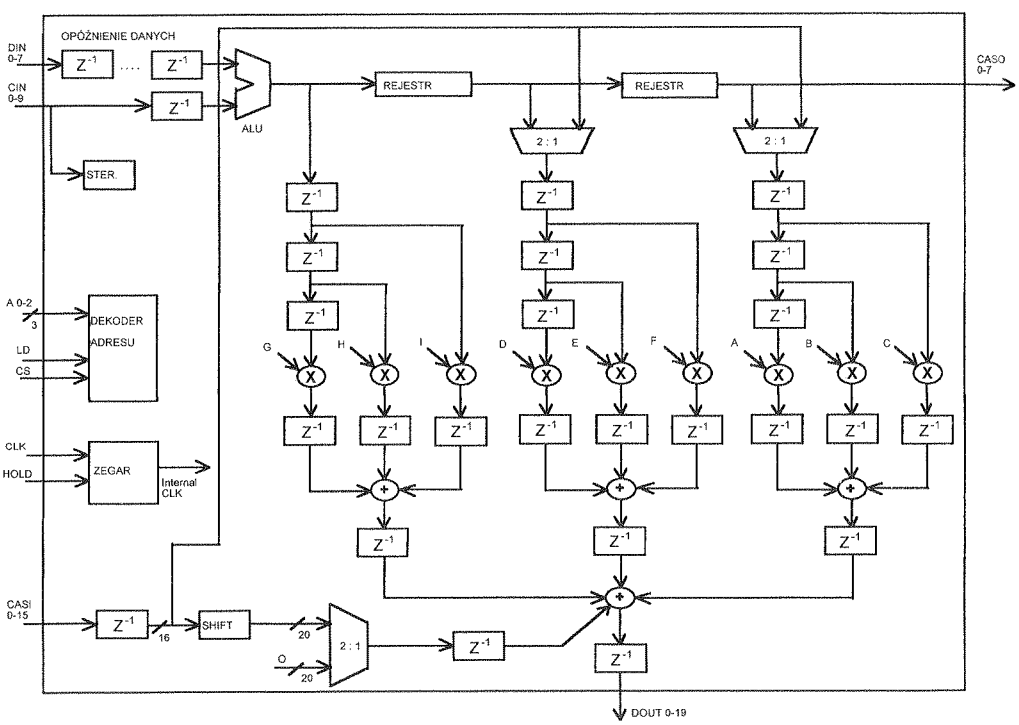
Rys. 2 przedstawia strukturę procesora specjalizowanego do realizacji funkcji konwolucji. Linia opóźniająca zbudowana jest jako rejestr przesuwny o pojemności 512 słów (ilość punktów w linii) 8-bitowych (256 poziomów szarości). Operacja mnożenia danych video z kamery i z linii opóźniających przez współczynniki macierzy danej funkcji odbywa się w miarę pojawiania się danych wejściowych, przy czym mnożenie odbywa się jednocześnie w trzech wierszach i w trzech kolumnach. Wynik po zsumowaniu może być normalizowany.

W ostatnich latach, z uwagi na szybki rozwój dziedziny cyfrowej obróbki sygnałów, rośnie także liczba komercyjnie dostępnych rozwiązań scalonych procesorów kon-

wolucji. W chwili obecnej istnieje już co najmniej kilka typów specjalizowanych procesorów, przeznaczonych wyłącznie do wykonywania tej operacji. Poniżej krótko scharakteryzowane zostaną trzy wybrane procesory konwolucji, pochodzące od różnych producentów.

2.1. PROCESOR HSP-48908

Procesor konwolucji HSP-48908 firmy Harris stanowi stosunkowo ubogie rozwiązanie, jeżeli chodzi o rozmiar aktywnego okna obróbki obrazu [3]. Przy zastosowaniu pojedynczego układu można zaimplementować okno o rozmiarze 3×3 piksele (rys. 3). Istnieje możliwość kaskadowego łączenia układów, ale już dla okna o rozmiarze 5×5 pikseli należałoby zastosować aż 4 takie jednostki. W przypadku tego procesora odpowiedni sposób kaskadowania umożliwia również obróbkę obrazów o większej niż 1024 piksele długości linii (wynikającej z maksymalnej długości wewnętrznych cyfrowych linii opóźniających), co jednak rzadko ma zastosowanie w praktyce. Natomiast szczególnie ciekawą cechą tego procesora jest możliwość poddania strumienia danych wejściowych jednej z 19-tu różnych, programowalnych operacji arytmetyczno-logicznych oraz dodatkowo normalizacji danych (przesunięciu danych maksymalnie do 3 bitów w lewo lub prawo). Drugim argumentem operacji arytmetyczno-logicznych jest stała dla danego ciągu próbek wartość zaprogramowana wcześniej w rejestrze ALU przy

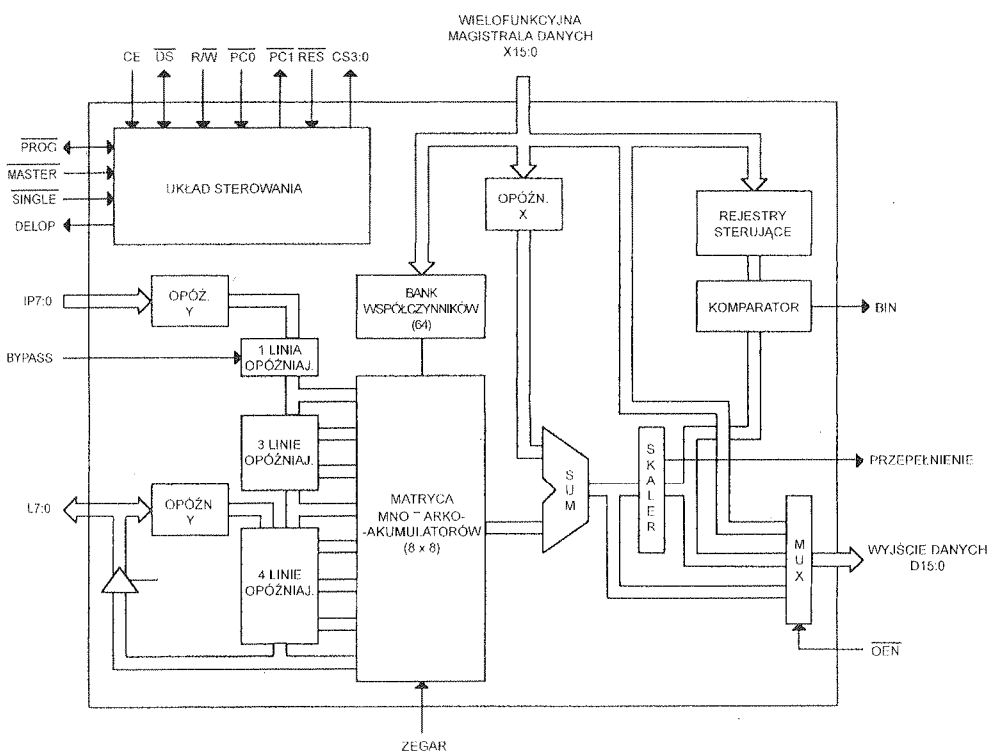


Rys. 3. Struktura blokowa procesora HSP48908

pomocy zewnętrznego procesora nadrzędnego. Układ operuje na danych wejściowych 8-bitowych i 20-bitowej precyzji danej wyjściowej oraz pracuje z częstotliwością zegara 32 MHz.

2.2. PROCESOR PDSP-16488

Całkowicie odmienne rozwiązanie architektury zastosowano w procesorze PDSP-16488 firmy Plessey [4]. Jest to jednostka bardzo dobrze nadająca się do implementacji okien o dużych rozmiarach (aż do 8×8 w pojedynczym układzie) i do obróbki danych wejściowych o dużej dokładności (do 16 bitów). Architektura tego układu jest zaprojektowana w sposób bardzo elastyczny, umożliwiający dostosowanie w szerokich granicach do aktualnych potrzeb użytkownika. Jednostka ta posiada jednak szereg ograniczeń. Ze względu na fakt, iż największą część powierzchni struktury tego typu układów zajmują cyfrowe linie opóźniające, nie jest możliwe równoczesne wykorzystanie pełnej długości linii opóźniających (1024 piksele) i pełnego rozmiaru okna (8×8). Z kolei szczególny sposób wykonania bloku mnożarek, polegający na wielokrotnym wykorzystaniu każdej mnożarki podczas jednego okresu zegara danych, przy zwiększeniu dokładności do 16 bitów implikuje zmniejszenie długości wewnętrznych linii opóźniających i rozmiaru okna. Powoduje to także zmniejszenie maksymalnej prędkości przepływu strumienia

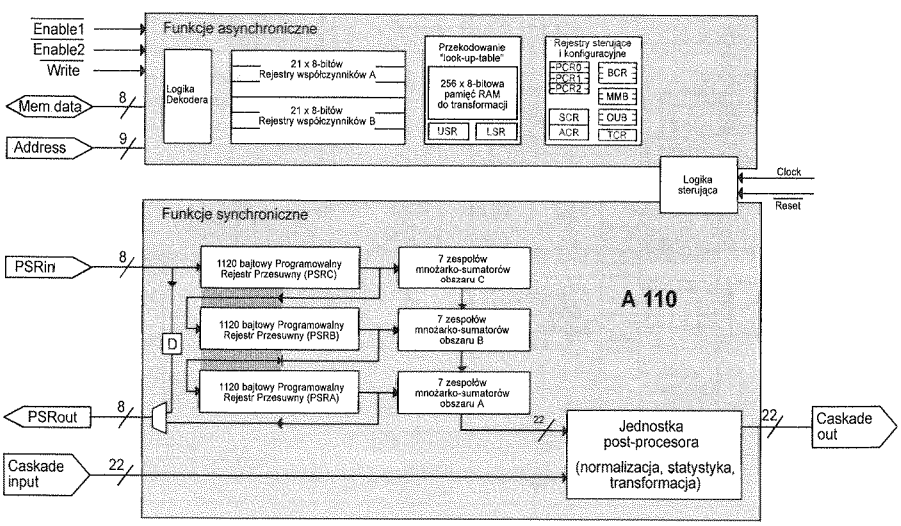


Rys. 4. Architektura wewnętrzna procesora PDSP16488

danych przez układ. Podobnie jak układ HSP48908, tak i ten posiada możliwość kaskadowego łączenia kilku jednostek. Układ ten posiada dodatkowo normalizator danych (układ przesunięć) oraz prosty, dwustanowy komparator danych wyjściowych. Częstotliwość pracy układu PDSP16488 jest zależna od precyzji danych wejściowych i wielkości przetwarzanego okna. Dla danych wejściowych 8-bitowych częstotliwość pracy wynosi: 40 MHz dla okna 4×4, 20 MHz dla okna 8×4 i 10 MHz dla okna 8×8. Dla danych wejściowych 16-bitowych układ pracuje z częstotliwością 20 MHz dla okna 4×4 i 10 MHz dla okna 8×4. Architektura wewnętrzna procesora PDSP-16488 jest przedstawiona na rys. 4.

2.3. PROCESOR IMS-A110

Procesor konwolucji IMS-A110 firmy SGS Thomson -InMos stanowi jedno z najwcześniejszych rozwiązań problemu implementacji konwolucji w strukturze hardware'owej [5]. Architektura wewnętrzna jest przejrzysta, a bardzo cennym jest umieszczenie w strumieniu danych wyjściowych bogato wyposażonego tzw. post-procesora, którego szerokie możliwości ułatwiają w znacznym stopniu zastosowanie tej jednostki w wielu różnorodnych aplikacjach. Zasadniczy człon procesora składa się z trzech linii opóźniających o długości 1120 bajtów oraz trzech siedmiostopniowych mnożarko-sumatorów. Jest on wyposażony (tak jak poprzednio omówione układy) w możliwość kaskadowego łączenia z innymi tego typu jednostkami w celu zwiększenia rozmiaru aktywnego okna lub zwiększenia dokładności danych i/lub współczynników konwolucji.



Rys. 5. Architektura wewnętrzna procesora IMS-A110

Na szczególną uwagę zasługuje wspomniany wcześniej układ postprocesora. Składa się on z czterech bloków: bloku wejściowego (rejestr przesuwny, sumator kaskadowy), bloku monitora statystycznego (detektor wartości minimalnej i maksymalnej oraz komparator z rejestrami przepełnienia/niedoboru), bloku formowania danych (układ transformacji i układ normalizacji danych) oraz bloku wyjściowego (sumatora i multipleksera wyjściowego).

Dodatkowo procesor IMS-A110 wyposażony jest w ośmiobitowy interface mikroprocesora, służący do konfiguracji układu przez zapisanie odpowiednich danych sterujących do rejestrów wewnętrznych. Przestrzeń adresowa procesora obejmuje 512 lokacji adresowych.

Procesor ten posiada układ mnożarko-sumatora o rozmiarach okna całkowicie wystarczających do obliczeń konwolucji 3×3. Maksymalna częstotliwość zegara danych video wynosi 20 MHz, co mieści się w zakresie potrzeb. Na wyborze tego procesora do prac eksperymentalnych zaważył fakt najbardziej rozbudowanej jednostki postprocessing'u, ułatwiającej współpracę z innymi elementami systemu.

W układzie IMS-A110 wewnętrzny mnożarko-sumator zapewnia maksymalne okno konwolucji o rozmiarze 3 linie$\times$7 kolumn. Macierz współczynników konwolucji uporządkowana jest w tablicy 1.

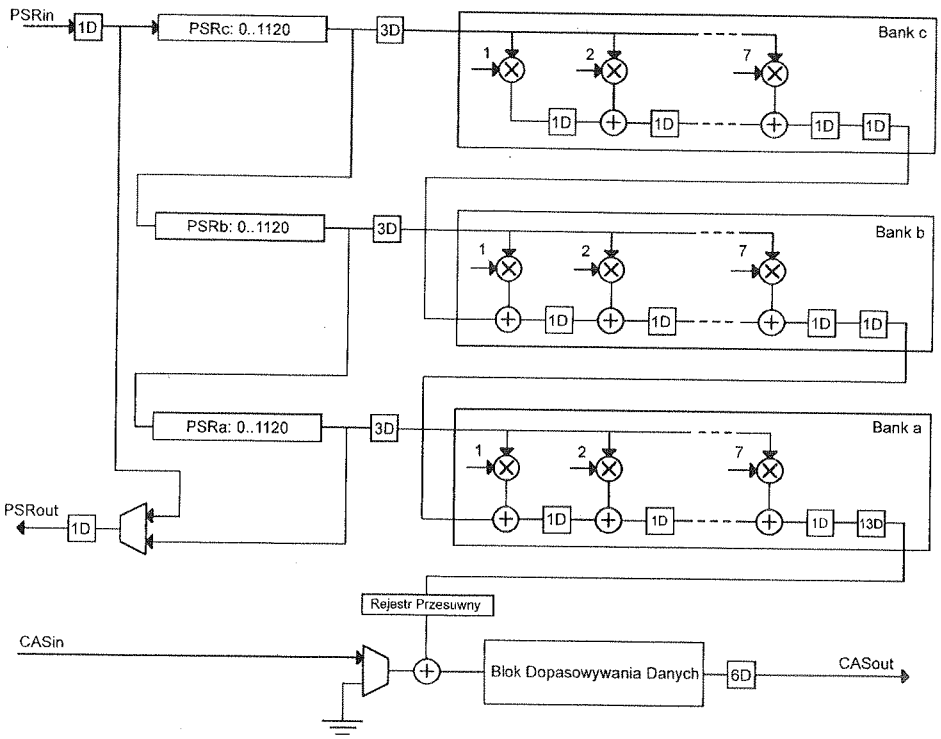
Tablica 1

Macierz współczynników konwolucji

A1	A1	...	A7
B1	B2	...	B7
C1	C2	...	C7

W przypadku zastosowania mniejszego okna konwolucji (np.: 3×3) można wykorzystać dowolne 3 kolejne kolumny. Od miejsca umieszczenia tego okna w oknie maksymalnym zależeć będzie wyłącznie sumaryczne opóźnienie układu (pojawienie się na wyjściu pierwszej wartości, do obliczenia której użyty został chociaż 1 punkt bieżącego obrazu). Jeżeli okno konwolucji zajmie skrajne kolumny o numerach 1 ... 3, to całkowite opóźnienie układu wyniesie 45 taktów zegara (dla $f = 20$ MHz opóźnienie to wyniesie 2,25 μs). W przypadku zajęcia kolumn skrajnych o numerach 5 ... 7, opóźnienie to wyniesie tylko 41 taktów.

Warto zauważyć, że długość cyfrowych programowalnych linii opóźniających w tych przypadkach nie zmienia się, ponieważ zmniejszenie opóźnienia w poprzednim mnożarko-sumatorze jest skompensowane przez dłuższy czas przejścia informacji w następnym. Długość linii opóźniających PSRb i PSRc (rys. 6) powinna wynosić $H_{actv}+7$, gdzie H_{actv} jest długością linii w obrabianym obrazie. Długość linii PSRa nie jest krytyczna i może być dobrana tak, aby dostosować opóźnienie układu do warunków pracy. W zaprojektowanym przez autorów module procesora konwolucji linia ta została użyta do zwiększenia opóźnienia układu do 512 taktów zegara (pełna linia informacji obrazowej) w celu uproszczenia logiki synchronizacji modułu. Zaprogramowana długość tej linii powinna wynosić $H_{actv}-D$, gdzie D jest całkowitym opóźnieniem wnoszonym przez układ.

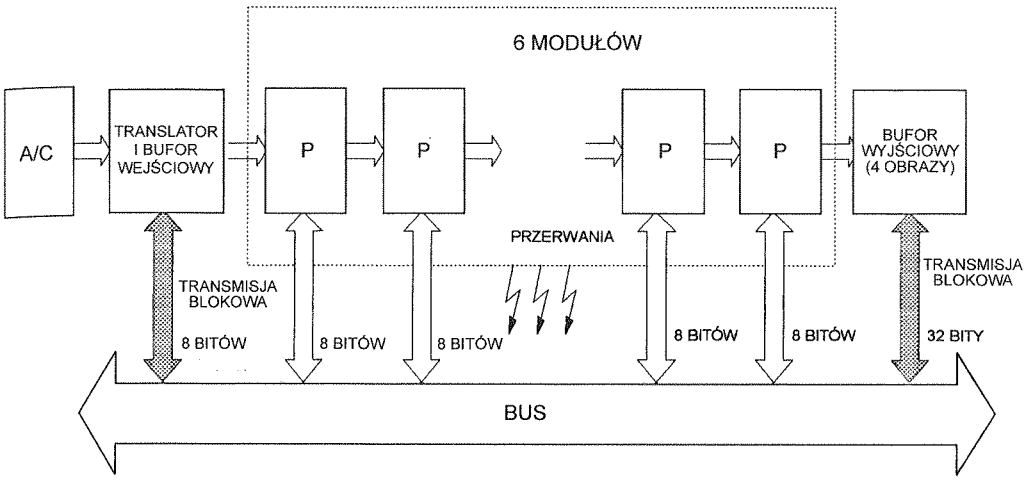


Rys. 6. Blokowy model opóźnień w procesorze IMS-A110

2.4. PODSUMOWANIE

Zastosowanie specjalizowanego procesora konwolucji wiąże się z budową dedykowanego otoczenia zapewniającego prawidłowe dostarczenie danych do przetwarzania i odczytanie danych wynikowych. Schemat blokowy zbudowanego przez autorów modułu realizującego wymogi procesorów specjalizowanych P i dostosowującego ją do otoczenia przedstawiono na rys. 7. Moduł ten umożliwia pobieranie danych do przetwarzania z magistrali systemu lub bezpośrednio z wejścia wizyjnego systemu. Przetwornik A/C przetwarza sygnał analogowy z kamery na postać cyfrową, translator dostosowuje ten sygnał (zorganizowany w międzyliniowe półobrazy) do postaci kolejnoliniowej, a bufor wyjściowy dostosowuje zorganizowanie przetworzonych danych do wymogów standardowej magistrali systemowej. Specjalizowane procesory VLSI posiadają dodatkowo połączenia z magistralą systemu (linie danych, adresów i przerwań) [6]. Opracowany sposób przepływu przetwarzanych danych przez procesory zapewnia ich szeregową i synchroniczną pracę i dlatego taki sposób połączenia procesorów nazwano architekturą potokową.

Moduł zawierający architekturę potokową skupia w sobie aż cztery niejednorodne struktury organizacji danych wizyjnych:



Rys. 7. Schemat blokowy dedykowanego otoczenia procesorów specjalizowanych

1. na wejście przetwornika A/C podawany jest sygnał analogowy z kamery,
2. na wyjściu przetwornika A/C dane 8-bitowe zorganizowane są jako kolejne półobrazy (tzn. międzyliniowo — co jest niezgodne z wymaganiami architektury potokowej i samego konwalwera),
3. wewnątrz modułu dane 8-bitowe przekazywane są potokowo (tzn. kolejnoliniowo),
4. na wyjściu dane 32-bitowe (jako 4 słowa 8-bitowe) przekazywane są na magistralę systemu mikroprocesorowego.

Operacje konwolucji są bardzo przydatne na poziomie wstępnego przetwarzania danych wizyjnych. Dobierając odpowiednie wartości współczynników możemy realizować filtracje dolnoprzepustowe (eliminacja zakłóceń, rozmywanie brzegów obiektów itp.), filtracje górnoprzepustowe (wyostrzanie krawędzi, uwypuklanie izolowanych punktów itp.), detekcja krawędzi i narożników. Ostatni rodzaj filtracji ma charakter kierunkowy. Oznacza to, że skuteczność wykrycia krawędzi pionowej lub poziomej zależy silnie od zestawu zastosowanych współczynników. Aby skutecznie wykrywać zarówno krawędzie pionowe i poziome warto zbudować moduł procesora składający się z dwóch układów konwolwerów przetwarzających te same dane wizyjne: jeden z nich wykrywa krawędzie poziome, drugi zaś pionowe. Na wyjściu sumuje się otrzymane wyniki. W ten sposób z wysoką skutecznością wykrywane są granice obiektów [6].

Każdy z uwidocznionych na rys. 7 procesorów P musi być wyposażony w układ sterowania. Składa się on z dwóch zasadniczych bloków: pierwszy blok (synchronizacji) zapewnia prawidłową synchronizację układu procesora w architekturze potokowej, uwzględniając opóźnienie wyjściowej informacji obrazowej o trzy linie obrazowe (dla konwalwera), drugi blok (interfejsu magistrali systemowej) zapewnia procesorowi nadrzędnemu dostęp do wewnętrznych rejestrów konfiguracji procesora poprzez magistralę systemową.

W zbudowanym dla celów badawczych systemie sterowanie zrealizowano w technologii FPGA wykorzystując układ XC4003-5PG120 firmy Xilinx. Jako specjalizowany procesor operacji kontekstowych zdecydowano zastosować konwolwer IMS-A110 firmy SGS Thomson-InMos [5]. Czas opóźnienia wnoszonego przez moduł procesora konwolucji w takiej dedykowanej architekturze związany jest z zapewnieniem równoczesnego dostępu do pełnego otoczenia przetwarzanego punktu. Dostęp ten zapewniają bufory FIFO znajdujące się zazwyczaj we wnętrzu specjalizowanych układów VLSI. W zbudowanym przez autorów systemie czas ten jest równy czasowi trwania trzech linii obrazowych i wynosi $3 \times 512 \times 1/15$ MHz = 102 μs.

W przedstawionym w tym rozdziale opisie wybranych specjalizowanych procesorów konwolucji zrealizowanych w technologii VLSI dokonano przeglądu wybranych konwolwerów (ze względu na szczupłość miejsca i cel niniejszego opracowania). Nie wyczerpują one pełnej listy tego typu procesorów, a przykładem jednego z kolejnych konwolwerów jest układ TMC2255 firmy Raytheon [7] o dopuszczalnych rozmiarach okna 3×3 lub 5×5 (liczby 8-bitowe) oraz maksymalnej częstotliwości zegara 30 MHz. W ostatnim czasie rozwój specjalizowanych układów procesorów konwolucji uległ pewnemu ograniczeniu co jest wynikiem bardzo wysokich parametrów czasowych procesorów ogólnego przeznaczenia i procesorów sygnałowych. Nowe możliwości w tym zakresie daje także implementacja tych procesorów w układach programowalnych FPGA, jednak omówienie tego problemu wykracza daleko poza ramy niniejszego artykułu.

3. PROCESORY OGÓLNEGO PRZEZNACZENIA

Rozważania dla przypadku realizacji obliczeń konwolucji przez procesor ogólnego przeznaczenia zostaną przeprowadzone na przykładzie obecnie najpopularniejszego procesora typu Pentium. Procesor ten posiada dość skomplikowaną budowę i niezoptymalizowane kodowanie listy rozkazów, co wynika z kompatybilności tego procesora z procesorem 8086 oraz z cech architektury systemu ogólnego przeznaczenia. Pomimo tych cech procesor ten jest często stosowany i dlatego zostanie omówiony poniżej.

Dla zobrazowania problemów realizacji obliczeń konwolucji przez ten mikroprocesor poniżej przedstawiono realizację tego algorytmu w kodzie języka C (listing 1).

Listing 1. Prosty przykład programu w języku C do obliczania konwolucji

```
const int M = 3, N = 3; // rozmiar okna konwolucji: M- w poziomie; N- w pionie
const int Nx = 512, Ny = 512; // rozdzielczość obrazu: Nx- w poziomie, Ny- w pionie
const int K = 16; // współczynnik przeskalowania (patrz wzór 2)
BYTE a[Ny+1+N][Nx]; // obraz źródłowy powiększony o dodatkowe linie w celu kompensacji brzegów
      (właściwy obraz źródłowy znajduje się od linii 1+N/2 do Ny+N/2, pozostałe linie są wypełniane
      w taki sposób aby zminimalizować błąd brzegów)
BYTE b[Ny][Nx]; // obraz wynikowy
int w[N][M] = {1,2,1, 2,4,2, 1,2,1}; // przykładowy filtr jak w tab. 1a.
      void
      convol2()
```

```
{BYTE *pa &a[0][Nx-M/2]; // wskaźnik do danych obrazu źródłowego (wskazuje lewy górny róg źródła)
BYTE *pb = &b[0][0]; // wskaźnik do danych obrazu wynikowego
for(int y = 0; y<Ny; y++) // dla każdego punktu w pionie
{       for(int x = 0; x<Nx; x++, pb++, pa++) // dla każdego punktu w poziomie
        {       register BYTE *pw  = w[0]; // pw -wskaźnik współczynników konwolucji
                register BYTE *pa1 = pa; // wskaźnik aktualnej rozważanej danej źródłowej
                register int sum = K/2; // wstępna wartość redukująca błąd zaokrąglanie przeskalowania
        for(int i = 0; i<N; i++) // konwolucja w pionie
                {for(int j = 0; j<M; j++) // konwolucja w poziomie
                    sum+ =*pw++**pa1++; // sam proces konwolucji
                pa1+Nx-M; // przejście do pierwszej danej następnej linii
                }
                sum/ = K; // przeskalowane (dzielenie jest zastępowane przez kompilator na przesunięcie bitowe)
                *pb = (BYTE) sum; // konwersja z typu int (4 bajty) na zmienną 1 bajtową.
}       }
}
```

Przedstawiony powyżej program można optymalizować pod kątem zwiększenia szybkości realizowanych obliczeń poprzez następujące zabiegi:

- rozpisanie pętli *i*, *j* (ang. *unrolling*),
- zaimplementowanie funkcji konwolucji bezpośrednio w asemblerze,
- optymalizacja kodu ze względu na superskalarną architekturę procesora Pentium,
- wykorzystanie instrukcji SIMD procesora MMX,
- zastosowanie współczynników konwolucji nie wymagających mnożenia.

3.1. INSTRUKCJA PĘTLI

W większości przypadków realizowanych obliczeń z góry znany jest rozmiar okna konwolucji, dlatego w celu przyspieszenia obliczeń zrezygnowano z pętli *i*, *j* (listing 1), dziewięciokrotnie rozpisując operację MAC (mnożenia i dodawania). Sama instrukcja pętli zawiera w przybliżeniu tylko dwie instrukcje asemblerowe: {*dec* (dekrementacji) i *jnz* (skoku warunkowego)}. Jednakże brak instrukcji skoku warunkowego przynosi znaczące przyspieszenie czasu wykonywania konwolucji ze względu na architekturę procesora, co zostanie w skrócie omówione poniżej.

W celu zwiększenia częstotliwości pracy procesora stosuje się coraz bardziej zaawansowany proces technologiczny wytwarzania układu scalonego procesora poprzez zmniejszenie długości kanału tranzystora CMOS, zwiększenie liczby warstw metalizacji, itd. Oprócz tego, w celu zwiększenia szybkości zegara, stosuje się bardziej zaawansowaną architekturę procesora, w szczególności wykorzystuje się architekturę potokową (ang. *pipeline*), przez co każda instrukcja rozłożona jest na szereg etapów. Procesor równocześnie wykonuje szereg instrukcji, które są na różnym etapie wykonywania. Dzięki temu średni czas wykonywania instrukcji jest w przybliżeniu równy czasowi

wykonywania instrukcji na pojedynczym etapie, a nie sumie czasów wykonywania instrukcji na wszystkich etapach. Dokładniejszy opis architektury potokowej znajduje się w [8]. W procesorze Pentium stosuje się 5 etapów potokowości [9], które są przedstawione w tablicy 2.

Tablica 2

Etapy potokowości dla procesora Pentium 75 MHz

1	2	3	4	5
Pobranie Instrukcji	Dekodowanie (etap 1)	Dekodowanie (etap 2)	Wykonanie instrukcji	Zapis Rejestrów

W wyniku zastosowania architektury potokowej zyskuje się znacznie na szybkości pracy mikroprocesora, ale z drugiej strony, każda instrukcja skoku powoduje, że instrukcje znajdujące się na etapie 1, 2, 3 (tabl. 2) mogą być nieaktualne. Dzieje się tak dlatego, że instrukcja skoku jest wykonywana dopiero w końcowym etapie, czyli instrukcje na etapach początkowych są pobrane bez uwzględnienia skoku. Dlatego w przypadku procesorów 80486 każda instrukcja skoku powoduje zatrzymanie wykonywania instrukcji znajdujących się za instrukcją skoku do czasu kiedy zostanie całkowicie wykonana instrukcja skoku. Powoduje to dodatkowe postoje procesora. Aby temu zapobiec w mikroprocesorze Pentium zastosowano zaawansowaną technikę anulowania zmian, która polega na tym, że procesor niezależnie od wyniku instrukcji skoku może kontynuować wykonywanie dalszych instrukcji. W konsekwencji jeżeli skok został wykonany (założeniem było, że skok nie będzie wykonany) procesor musi anulować instrukcje wykonane na początkowych etapach potokowości oraz wczytać od początku uaktualnione instrukcje po dokonaniu skoku. W konsekwencji powoduje to, że przez 3 lub 4 cykle zegarowe (w zależności od bloku 'U' lub 'V' dla P75) procesor praktycznie nie wykonuje żadnej instrukcji.

Ponadto istnieje pewna prawidłowość w wykonywaniu instrukcji skoku. Śledząc wybraną w programie instrukcję skoku można zauważyć, że dany skok jest przez procesor z reguły (z pewnym dość dużym prawdopodobieństwem) wykonywany albo też nie wykonywany. W konsekwencji, aby ograniczyć postój procesora spowodowany przez instrukcje, która prawie zawsze wykonuje skok, stosuje się logikę predykcji skoku (ang. *branch prediction*). Logika ta działa równolegle z układem dekodowania instrukcji dzięki czemu już na początkowym etapie dekodowania instrukcji skoku warunkowego, kiedy nie jest jeszcze znany wynik warunku, procesor wykonuje (lub nie wykonuje) skok na podstawie zapisu w tabeli BTB (ang. *Branch Target Buffer*) [9, 10]. Tabela BTB jest pamięcią cache typu *look aside*, która zawiera osobny zapis dla różnych instrukcji skoku. Tabela BTB zawiera zapis o historii (prawdopodobieństwie) skoków w postaci czterech stanów: mocny skok (ang. *strongly taken*), słaby skok (ang. *weakly taken*), słaba kontynuacja (ang. *weak not taken*), mocna kontynuacja (ang. *strong not taken*). Stany te są modyfikowane po wykonaniu instrukcji skoku według następującej logiki: jeżeli skok zostanie wykonany, to zapis o historii skoku jest zmieniany o jeden stan

w stronę mocnego skoku (np. dla stanu słaba kontynuacja zostanie zmieniony na słaby skok) i na odwrót jeżeli instrukcja skoku nie zostanie wykonana to zapis jest zmieniany w kierunku mocnej kontynuacji. Predykcja skoku mimo tak prostej logiki w ogólnym przypadku poprawnie przewiduje wykonanie skoku (dla mikroprocesora Pentium z prawdopodobieństwem około 90%).

Analizując kod programu realizującego algorytm konwolucji (listing 1), można przyjąć, że procesor będzie domniemywał, że każda pętla będzie wykonywana w nieskończoność, czyli każde wyjście z pętli będzie okupione postojem procesora. Aby przyspieszyć wykonywanie obliczeń należy zrezygnować z pętli na tyle na ile jest to możliwe. W przytoczonym przypadku (listing 1) korzystne jest zastąpienie pętli konwolucji w pionie i w poziomie za pomocą $N \cdot M$ instrukcji {sum+=*pw++**pa1++;}.

Warto także zwrócić uwagę, że ilość cykli bezczynnych na skutek błędnej predykcji ulega wydłużeniu w nowszych procesorach na wskutek zwiększenia liczby etapów potokowości np. dla PentiumPro liczba etapów wynosi 12 [11]. Jest to uwidocznione w tabl. 3, gdzie dla P75 stosunek czasu wykonywania konwolucji dla opcji a i b (z pętlą i po rozpisaniu pętli) wynosi: 1.42, natomiast dla P300 Celeron aż 1.77.

3.2. ARCHITEKTURA SUPERSKALARNA MIKROPROCESORA PENTIUM

W celu przyspieszenia pracy mikroprocesora Pentium zastosowano w nim zrównoleglenie architektury. Mikroprocesor Pentium składa się z trzech bloków wykonawczych:

- jednostki stałoprzecinkowej 'U',
- jednostki stałoprzecinkowej 'V',
- jednostki zmiennoprzecinkowej (FPU — ang. *Floating Point Unit* oraz procesora MMX).

Dodatkowo w mikroprocesorze PentiumPro wprowadzono dodatkowe jednostki wykonawcze: drugą jednostkę zmiennoprzecinkową oraz dwie dodatkowe jednostki do kalkulacji adresu (ang. *address generation units*) [11].

W przypadku realizacji algorytmu konwolucji są wykorzystywane tylko jednostki stałoprzecinkowe. Wykorzystując bloki 'U' oraz 'V' mikroprocesor Pentium może wykonywać dwie niezależne instrukcje jednocześnie. Jednakże istnieją pewne ograniczenia w równoległej pracy tych dwóch jednostek wynikające z braku całkowicie symetrycznej architektury. Przykładem tej asymetrii jest fakt, że jednostka 'V' jest pozbawiona układu przesunięcia bitowego (ang. *barrel shifter*) oraz mikroprocesor Pentium nie może równocześnie wykonywać dwóch operacji mnożenia [9, 10]. Ponadto istnieją pary instrukcji, które nie mogą być wykonywane równocześnie, co wynika z ogólnych ograniczeń architektury równoległej, np. gdy argumentem instrukcji w bloku 'V' jest wynik instrukcji w bloku 'U', która jest właśnie wykonywana (ang. *register contention*). Wszystkie opisane powyżej problemy powodują, że zastosowanie dwóch równoległych jednostek stałoprzecinkowych nie powoduje realnego podwojenia szybkości pracy całego mikroprocesora Pentium. Dla przykładu stosunek czasu obliczeń dla

procesorów 80486/P75 dla algorytmu bez mnożenia (tabl. 3f) wynosi 1.9, a dla algorytmu z mnożeniem (tabl. 3d) tylko 1.19. Wysoki wskaźnik 1.9 świadczy o dużym zrównolegleniu wykonywanych operacji oraz o ogólnym przyspieszeniu pracy P75 w stosunku do 80486. Z drugiej strony niski wskaźnik 1.19 potwierdza, że operacja mnożenia nie może być wykonywana równolegle w P75.

Najnowsze wersje mikroprocesora Pentium są bardziej udoskonalone poprzez zastosowanie technologii dynamicznego wykonywania programu (ang. *Dynamic Execution*) [11], która polega między innymi na możliwości przestawiania kolejności wykonywania instrukcji w celu ograniczenia liczby postojów procesora. Oprócz tego liczba cykli zegarowych na wykonywanie operacji mnożenia została skrócona i możliwe jest wykonywanie dwóch operacji mnożenia równolegle, w wyniku czego stosunek czasów P75/P300 w tabl. 3d (z mnożeniem) wynosi: 16, a stosunek czasów dla tabl. 3f (bez mnożenia) wynosi: 4.5.

W związku z powyższymi cechami najnowszych mikroprocesorów, w celu zwiększenia szybkości wykonywanych instrukcji należy tak układać kod programu, aby poszczególne instrukcje mogły być wykonywane równocześnie, czyli należy zadbać o to, aby wybrane instrukcje nie następowały bezpośrednio po sobie (procedury ważne dla P75). Są to następujące instrukcje:

- instrukcje mnożenia *imul* (brak zrównoleglenia dla instrukcji mnożenia),
- instrukcje, których argumentem jest wynik poprzedniej instrukcji (ang. *register contention*),
- instrukcje odwołujące się równocześnie do tego samego banku pamięci cache (bank = 32 bajty).

Ponadto należy uwzględnić czas wykonywania instrukcji, np. instrukcja mnożenia jest wykonywana dłużej niż instrukcja dodawania, dlatego w trakcie wykonywania instrukcji mnożenia może być wykonanych parę równoległych instrukcji pod warunkiem, że nie odnoszą się one do wyniku mnożenia. Czasy wykonywania instrukcji zmieniają się w zależności od modelu mikroprocesora i są rzadko podawane w katalogach. Ponieważ przy realizacji operacji konwolucji instrukcja mnożenia jest kluczowa i wynik mnożenia jest potrzebny do następnych operacji mnożenia dlatego uwzględnienie tego aspektu w tym przypadku jest praktycznie niemożliwe.

W wyniku dokonanych ulepszeń mikroprocesora Pentium operacje MAC są przeplatane (np. w momencie kiedy wykonywane jest mnożenie $w_i \cdot a_i$ pobierana jest dana a_{i+1}) co przedstawia listing 2.

Listing 2. Przykład fragmentu kodu asemblerowego dla konwolucji 3×3:
a) kodu pisanego dla architektury skalarnej,
b) kodu pisanego dla architektury superskalarnej.

a)

```
// piksel 3 (prawy górny róg)
xor edx,edx // zeruj edx
mov dl, byte ptr [ecx+2] // załaduj daną a (piksel 3)
imul edx,dword ptr [edi+8h] // pomnóż piksel3 *w[0][2]
add eax, edx // dodawaj wynik mnożenia edx do eax
```

```
// piksel 4 (lewy środkowy piksel)
xor edx,edx // zeruj edx
mov dl, byte ptr [ecx+200h] // załaduj piksel 4
imul edx, dword ptr [edi+0Ch] // edx = pel4 *w[1][0]
add eax, edx // dodaj wynik mnożenia edx do eax
```

b)

```
xor edx, edx //pel 3(zeruj)
imul ebx, dword ptr [edi+/4]//pel2 ebx = pel2*w[0][1]
mov dl, byte ptr [ecx+2] // pel 3 dl = pel3
add eax, ebx // koniec pel2 eax+= ebx
imul edx, dword ptr [edi+8] // pel 3 edx = pel3*w[0][2]
xor ebx, ebx // pel4 zeruj
add eax, edx // koniec pel 3 eax+= edx
mov bl, byte ptr [ecx+200h] // pel4 bl = pel4
```

Warto zwrócić uwagę na fakt, że w wyniku dokonanych ulepszeń przyspieszono czas wykonywania programu tylko o około 10%. Przyczyną tego są następujące okoliczności:

- kod niezoptymalizowany do pracy równoległej jest również wykonywany równolegle (ale nie optymalnie),
- najbardziej czasochłonną operacją jest operacja mnożenia (*imul* zabiera 10 cykli zegarowych [12]), która nie może być wykonywana równolegle. Niestety co czwartą instrukcją jest instrukcja mnożenia, dlatego nie jest możliwa pełna optymalizacja programu.

W przypadku mikroprocesora Pentium Celeron nie uzyskano przyspieszenia pracy, co jest spowodowane więcej niż dwudrożną architekturą superskalarną oraz blokiem do dynamicznego wykonywania instrukcji (ang. *Dynamic Execution*). Blok ten zapewnia ciągłą pracę procesora poprzez optymalizację kolejności wykonywania instrukcji.

Warto w tym miejscu zwrócić uwagę na duże ograniczenia architektury superskalarnej. Architektura superskalarna ulega szybkiemu nasyceniu — wzrost liczby równoległych jednostek obliczeniowych nie powoduje odpowiedniego wzrostu szybkości. W praktyce nie powinno się stosować więcej niż 2–5 jednostek równoległych [13]. Ponadto architektura superskalarna wymaga rozbudowanego bloku dekodowania instrukcji. Alternatywnym rozwiązaniem jest architektura VLIW (ang. *Very Long Instuction Word*) [13], dla której dekompozycja programu na poszczególne jednostki równoległe odbywa się już na poziomie kompilacji. W konsekwencji, pojedyncza instrukcja posiada odpowiednio rozlokowane rozkazy dla wszystkich jednostek procesora, wynikiem czego jest duża szerokość (rzędu 64–1024 bitów) pojedynczej instrukcji. W architekturze VLIW blok dekodowania instrukcji ulega znacznemu uproszczeniu, dzieje się to jednak kosztem nadmiernego kodu programu — np. często jakaś jednostka procesora nie wykonuje żadnej operacji, a przez to przydzielony jej fragment instrukcji jest niewykorzystywany.

Ciekawe rozwiązanie zostało wprowadzone w procesorze Crusoe firmy Transmeta [14], który jest całkowicie kompatybilny z procesorami x86. Jądro tego procesora jest jednak procesorem VLIW, całkowicie niezależnym od kodu procesorów rodziny x86. Translacja kodu (ang. *code morphing*) z x86 na wewnętrzny kod VLIW odbywa się częściowo software'owo oraz częściowo hardware'owo tylko jeden raz — raz przetłumaczony kod jest zapamiętywany w pamięci translacji (ang. *translation cache*). Warto zwrócić uwagę, że translacja ta jest uzależniona od stopnia wykorzystania programu: translacja jest bardziej efektywna, a przez to bardziej czasochłonna dla części programu wykonywanego wiele razy oraz mało efektywna dla fragmentu programu, który jest wykorzystywany tylko raz (translacja jest udoskonalana za każdym wykonaniem danej części programu, aż do osiągnięcia określonej optymalności). Ponadto predykcja skoku może być również częściowo uwzględniana na etapie translacji poprzez zmiany docelowego programu na podstawie historii skoków. W wyniku zastosowanej architektury znacząco uproszczony został dekoder instrukcji, a budowa procesora uniezależniona została od kompatybilności z procesorem 8086. W rezultacie zmniejszony został również o 60–70% pobór mocy [14].

3.3. CACHE

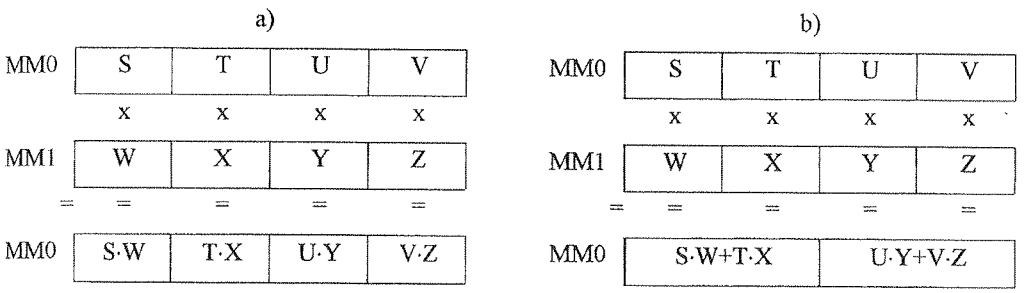
Analizując kod języka maszynowego opisującego operację konwolucji (listing 2) można zauważyć, że co druga instrukcja tego programu odwołuje się do pamięci. W konsekwencji procesor bardzo często komunikuje się z dużo wolniejszą pamięcią, co powoduje znaczne opóźnienia w wykonywaniu instrukcji programu. Aby temu zapobiec stosuje się pamięć typu *cache*, która w znaczący sposób przyspiesza wykonywanie instrukcji komunikujących się z pamięcią. W przypadku konwolucji można zauważyć, że kod programu, współczynniki konwolucji oraz wynik konwolucji transferowane są z (do) pamięci typu *cache*. Niestety w pamięci *cache* nie znajduje się cały obraz wizyjny przeznaczony do przetwarzania (tablica *a* w listing 1). Jest to spowodowane po pierwsze zbyt dużą liczbą tych danych (pamięć *cache* L1 danych ma rozmiar od 8 kB dla P75 do 32kB dla K6), a po drugie aby dana znajdowała się w pamięci *cache* musi być przynajmniej jeden raz wczytana. Jednakże analizując rozmiar linii (512 bajtów) można stwierdzić, że w pamięci *cache* znajdują się poprzednio odczytywane dane tablicy *a* (dla konwolucji 3×3 wymagana jest pamięć *cache* danych $>2\times512$ bajtów czyli 2 linie). Dzięki temu procesor odczytuje bezpośrednio z pamięci (zamiast z szybkiej pamięci typu *cache*) tylko jeden piksel (prawy dolny róg konwolucji). Zakładając, że przy odczycie z pamięci procesor odczytuje bank 32 bajtowy (wielkość banku *cache* wynosi 32 bajty), można przyjąć, że tylko co 32 operacja konwolucji będzie musiała odwoływać się do wolniejszej pamięci zewnętrznej. W rezultacie opóźnienie pracy procesora spowodowane komunikacją z zewnętrzną pamięcią jest nieznaczne.

Ilustruje to tablica 3 dla opcji *f*, *g*. Opcja *g* jest zmodyfikowaną opcją *f* — nie następuje inkrementacja adresu wejściowego i wyjściowego, czyli konwolucja operuje

cały czas na 9-ciu tych samych pikselach wejściowych i jednym pikselu wyjściowym, które znajdują się w pamięci *cache*; w konsekwencji czego algorytm konwolucji nie komunikuje się z pamięcią zewnętrzną. Różnica pomiędzy opcją *f* a opcją *g* wynosi tylko około 4%.

3.4. TECHNOLOGIA MMX

W celu przyspieszenia operacji multimedialnych, które z reguły używają dużej liczby nieskomplikowanych operacji np. 8 bitowego dodawania, zastosowano architekturę SIMD (ang. *Single Instruction-stream Multiple Data-stream*), która umożliwia wykonanie wielu (do 8) takich samych operacji równolegle. Operacje te są wykonywane na jednym rejestrze 64 bitowym, przy czym rejestr ten jest podzielony na 2, 4 lub 8 części w zależności od długości pojedynczego słowa [11]. Dla przykładu dla słowa 16 bitowego rozkaz PMULLW wykonuje równolegle 4 niezależne operacje mnożenia (rys. 8) z zachowaniem tylko 16 najmłodszych bajtów wyniku.



Rys. 8. Przykłady operacji MMX: a) operacja czterokrotnego równoległego mnożenia (z odrzuceniem dwóch najbardziej znaczących bajtów): PMULLW MM0, MM1; b) operacja MAC- pomnóż i akumuluj: PMADDWD MM0, MM1

Wzrost mocy obliczeniowej procesora MMX jest uzyskiwany dodatkowo dzięki zastosowaniu architektury superskalarnej; procesor może wykonywać dwie instrukcje MMX jednocześnie, jeżeli tylko instrukcje te używają różnych zasobów jednostki MMX. Dlatego jednostka MMX potrafi wykonywać nawet 16 operacji ośmiobitowego dodawania równolegle.

Analizując kod programu konwolucji przeznaczonego dla procesora MMX (listing 3) liczba instrukcji asemblerowych w pętli została zredukowana z 42 (opcja *c* tabl. 3) do 21 (opcja *m* tabl. 3 — obliczenia dla pojedynczego piksela). Dla opcji *m*, która bazuje na operacji MAC (instrukcja PMADDWD — patrz rys. 8), z 21 instrukcji tylko 11 instrukcji jest instrukcjami MMX. Jest to spowodowane tym, że rezultat operacji MAC jest uzyskiwany niezależnie w górnej i dolnej połowie rejestru MMX (patrz rys. 8), dlatego wymagane jest dodatkowe przetworzenie w jednostce stałoprzecinkowej. Jest to spowodowane tym, że dane wejściowe nie są właściwie rozlokowane z punktu widzenia instrukcji MMX — instruk-

cja PMADDWD operuje na 4 danych wejściowych, a dostarczane są tylko trzy ważne dane wejściowe w jednej linii, dlatego czwarty nieważny piksel jest mnożony przez wartość zero. To z kolei powoduje, że procesor MMX nie jest w pełni wykorzystywany.

Listing 3 przedstawia fragmenty programu w języku asemblera na procesor MMX dla przykładowej konwolucji (np. z rys. 1a), przy czym przedstawiono dwa warianty: obliczenia prowadzone są dla pojedynczego piksela i obliczenia prowadzone są dla czterech pikseli równocześnie.

Listing 3. Fragmenty programu w języku asemblera na procesor MMX dla konwolucji z rys 1a-opcje: m) obliczenia dla pojedynczego piksela; n) obliczenia dla czterech pikseli równocześnie

m)

```
beg: // etykieta startu pętli
movd mm0, dword ptr [ecx] // załaduj wiersz 0
movd mm1, dword ptr [ecx+200h] // załaduj
    wiersz 1
movd mm2, dword ptr [ecx+400h] // załaduj
    wiersz 2

punpcklbw mm0, mm5 // konwertuj byte->word
punpcklbw mm1, mm5 // mm5 — zawiera same
    zera
punpcklbw mm2, mm5

pmaddwd mm0, mm6 // operacja MAC; mm6 = 0,
    1,2,1
pmaddwd mm1, mm7 // mm7 = 0,2,4,2
pmaddwd mm2, mm6

    paddd mm0, mm1 // dodanie poszczególnych
    wyników MAC
paddd mm0, mm2 // wynik w mm0
// operacje stałoprzecinkowe
movq [edi], mm0 // wyślij rejestr MMX do pamię-
    ci
movd eax, mm0 // pobierz dolną połówkę rejestru
    MMX
add eax, [edi+4] // dodaj dolną i górną połowę
    rejestrów MMX
add eax, 8 // dodanie w celu zmniejszenia błędu
    zaokrąglenia
inc esi // inkrementuj adres piksela wynikowego
sar eax, 4 // dzielenie przez 16- przeskalowanie
inc ecx // inkrementacja adresu piksela źródło-
    wego
mov byte ptr [esi], al. // załadowanie wyniku do
pamięci
dec ebp // dekrementacja licznika pętli
jnz beg // sprawdzenie wyjścia z pętli
```

n)

```
beg: // etykieta startu pętli
movd mm0, dword ptr [ecx] // pobranie piksela (0,0)
punpcklbw mm0, mm7 // konwertuj byte->word;
    mm7 = 0

movd mm1, dword ptr [ecx+1] // pobranie piksela
    (0,1)
pmullw mm0, mm6 // mnożenie dla piksela (0,0),
    mm6 = 1,1,1,1
punpcklbw mm1, mm7 // konwertuj byte->word, pik-
    sela (0,1)

movd mm2, dword ptr [ecx+2] // pobranie piksela
    (0,2)
pmullw mm1, mm5 // mnożenie dla piksela (0,1);
    mm5 = 2,2,2,2
punpcklbw mm2, mm7 // konwertuj byte->word dla
    piksela (0,2)
paddw mm0, mm1 // dodanie wyników dla pikseli
    (0,0) i (0,1)
// podobne operacje dla pozostałych pikseli
...
pmullw mm2, mm6 // mnożenie dla piksela (2,2),
    mm6 = 1,1,1,1
paddw mm0, mm2 // końcowy wynik

paddw mm0, mm3 // dodanie w celu zmniejszenia
    błędu // zaokrąglenia mm3 = 8,8,8,8
psrlw mm0, 4 // podziel przez 16
packuswb mm0, mm7 // konwersja wyniku word-
    >byte
add ecx, 4 // zwiększenie adresu dla następnych 4 pik-
    seli
movd dword ptr [esi], mm0 // zachowanie wyniku
add esi,4 // zwiększenie adresu wyniku
dec ebp // dekrementacja licznika pętli
jnz beg // sprawdzenie wyjścia z pętli
```

W celu ograniczenia nadmiarowych operacji, obliczenia wykonywane są równocześnie dla czterech pikseli obrazu wynikowego (opcja *n* tabl. 3 — listing 3n). Dla tej opcji wszystkie cztery piksele wejściowe są mnożone przez ten sam współczynnik i odpowiednio dodawane. W rezultacie liczba instrukcji asemblerowych w pętli przypadających na jeden piksel została zmniejszona do 43/4 = 10.7, przy czym tylko 4/4 = 1 to instrukcje przeznaczone dla procesora stałoprzecinkowego. Ponadto dla opcji *n* wykorzystywana może być instrukcja MMX konwersji 16 bitów do 8 bitów z nasyceniem (ang. *saturation*), dzięki której np. dla danej 16 bitowej większej od 255, wynikiem jest wartość 255, a nie jak to z reguły ma miejsce: wartość znajdująca się na 8 mniej znaczących bitach. Warto zauważyć, że osiągnięty przez autorów czas obliczeń operacji konwolucji jest mniejszy od danych podawanych w [15], gdzie wynosi on 38.7 ms dla operacji stałoprzecinkowej i 12.6 ms dla opcji MMX procesora Pentium II 266MHz.

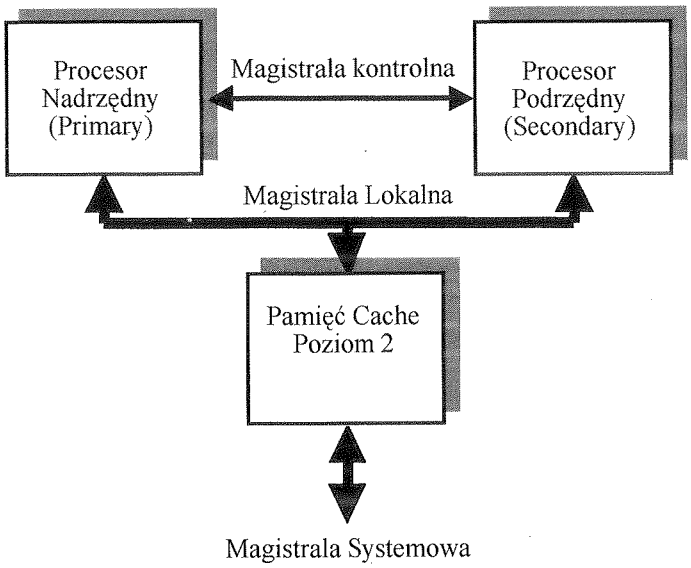
Podsumowując, czas wykonywania konwolucji przy użyciu koprocesora MMX wynosi około 1/4 czasu bez użycia koprocesora MMX (opcje *d*, *n* tabl. 3) dla mikroprocesora Pentium 166 MHz oraz około 40% dla mikroprocesora Celeron. W przypadku mikroprocesora Celeron osiągnięty zysk z zastosowania operacji MMX jest mniejszy, co jest spowodowane skróceniem czasu wykonywania operacji mnożenia — stosunek czasu wykonywania konwolucji z mnożeniem (opcja *d*) do konwolucji bez mnożenia (opcja *f*) wynosi: 4.97 dla procesora P166 oraz 1.45 dla procesora P300). Oprócz tego mikroprocesor Celeron posiada lepszą architekturę superskalarną, która znacząco wzmacnia stałoprzecinkową moc obliczeniową tego procesora.

3.5. ŚRODOWISKO WIELOPROCESOROWE

Mikroprocesor Pentium posiada możliwość pracy w środowisku dwuprocesorowym i wówczas dwa mikroprocesory Pentium mogą wykonywać dwa niezależne zadania. Architektura taka została przedstawiona na rys. 9. Architektura ta posiada wspólną magistralę do komunikacji z pamięcią zewnętrzną [9], co może znacząco spowolnić pracę procesorów. Warto zwrócić uwagę, że architektura taka nie wymaga dodatkowego układu kontrolującego współpracę mikroprocesorów.

Realizując algorytm konwolucji na dwóch mikroprocesorach Pentium, należałoby rozpisać program realizujący operację konwolucji w sposób zupełnie inny niż w przypadku architektury superskalarnej. Dla architektury superskalarnej wiele jednostek arytmetyczno — logicznych wykorzystuje te same zasoby procesora (rejestry, pamięć *cache* itd.). W przypadku architektury wieloprocesorowej zasoby te są niezależne. Dlatego należy całkowicie rozdzielić realizację procesu konwolucji na dwa niezależne procesy.

Warto zwrócić uwagę na właściwości procesu konwolucji, której dane wyjściowe są od siebie niezależne, czyli wynik operacji piksela poprzedniego nie wpływa na wynik operacji aktualnie obliczanego piksela. Ta cecha umożliwia łatwe zrównoleglenie obliczania operacji konwolucji, gdzie dwa procesory mogą obliczać dane dla niezależ-



Rys. 9. Schemat połączeń dla współpracy dwóch mikroprocesorów Pentium

nych pikseli, np. procesor nadrzędny oblicza piksele dla linii od 0 do 255 a procesor podrzędny piksele dla linii od 256 do 511. W tym wypadku, pomijając wpływ konfliktów w komunikacji pomiędzy procesorami a pamięcią *cache* L2, uzyskuje się podwojenie mocy obliczeniowej układu.

Wadą układu wieloprocesorowego oprócz utrudnionej komunikacji z otoczeniem, jest jego uzależnienie od systemu operacyjnego, który nie zawsze potrafi obsługiwać więcej niż jeden procesor, lub też nieefektywnie przydziela (alokuje) zadania na poszczególne procesory. Z tego też powodu rzadko stosuje się systemy wieloprocesorowe z wykorzystaniem mikroprocesorów Pentium.

W tablicy 3 zestawiono liczby instrukcji asemblerowych w pojedynczej pętli i czasy ich realizacji dla opisanych przypadków. Oznaczenie poszczególnych opcji w tablicy jest następujące:

a) opcja podstawowa — język C (listing 1),
b) po rozpisaniu pętli *i*, *j* — język C,
c) jak opcja *b* ale program napisany bezpośrednio w asemblerze,
d) jak opcja *c* ale po zrównolegleniu algorytmu dla architektury superskalarnej,
e) dla stałych współczynników będących potęgą liczby 2 (bez operacji mnożenia, filtr rys. 1a),
f) jak opcja *e* ale po zrównolegleniu algorytmu dla architektury superskalarnej,
g) jak opcja *f* ale obliczenia tylko dla jednego piksela (brak odczytów z/do pamięci zewnętrznej),
m) zastosowaniu koprocesora MMX dla pojedynczego piksela (napisane w asemblerze),
n) zastosowanie koprocesora MMX — obliczenia równolegle dla 4 pikseli (napisane w asemblerze).

Tablica 3

Liczba instrukcji asemblerowych w pojedynczej pętli oraz czasy wykonywania konwolucji 3x3 i obrazie 512x512 na procesorze 486DX4 — 100 MHz, Pentium 75MHz, Pentium 166MHz MMX oraz Pentium 300MHz Celeron

Opcja	a	b	c	d	e	f	g	m	n
Liczba instrukcji asembl. w pętli	—	—	42	42	37	35	35	21	43/4
Czas [ms] 486DX4–100	670	465	460	435	147	134	131	—	—
Czas [ms] P75	587	414	403	366	95	70.3	67.6	—	—
Czas [ms] P166	247	175	169	159	45	32	30	60	26
Czas [ms] P300	48.6	25.8	22.6	22.8	16.6	15.6	14.0	22.1	9.1

4. SPECJALIZOWANE PROCESORY DSP

Analizując kod procesora ogólnego przeznaczenia można łatwo zauważyć, że procesor wykonuje operację konwolucji nieefektywnie zarówno pod kątem czasu obliczeń jak i przyjętej architektury. Dlatego dla operacji konwolucji bardzo często stosuje się układy procesorów sygnałowych DSP. Istnieje wiele modeli procesorów DSP i dlatego w celu porównania celowości użycia procesorów DSP do realizacji obliczeń algorytmu konwolucji zostaną omówione dwa z nich. Pierwszym jest procesor sygnałowy z rodziny DSP56000 firmy Motorola, który jest szczególnie atrakcyjny dla operacji konwolucji ze względu na prostotę przyjętej architektury. Drugim jest procesor sygnałowy TMS320C80 firmy Texas Instruments, który pod względem mocy obliczeniowej i struktury wewnętrznej szczególnie dobrze nadaje się do przetwarzania obrazu.

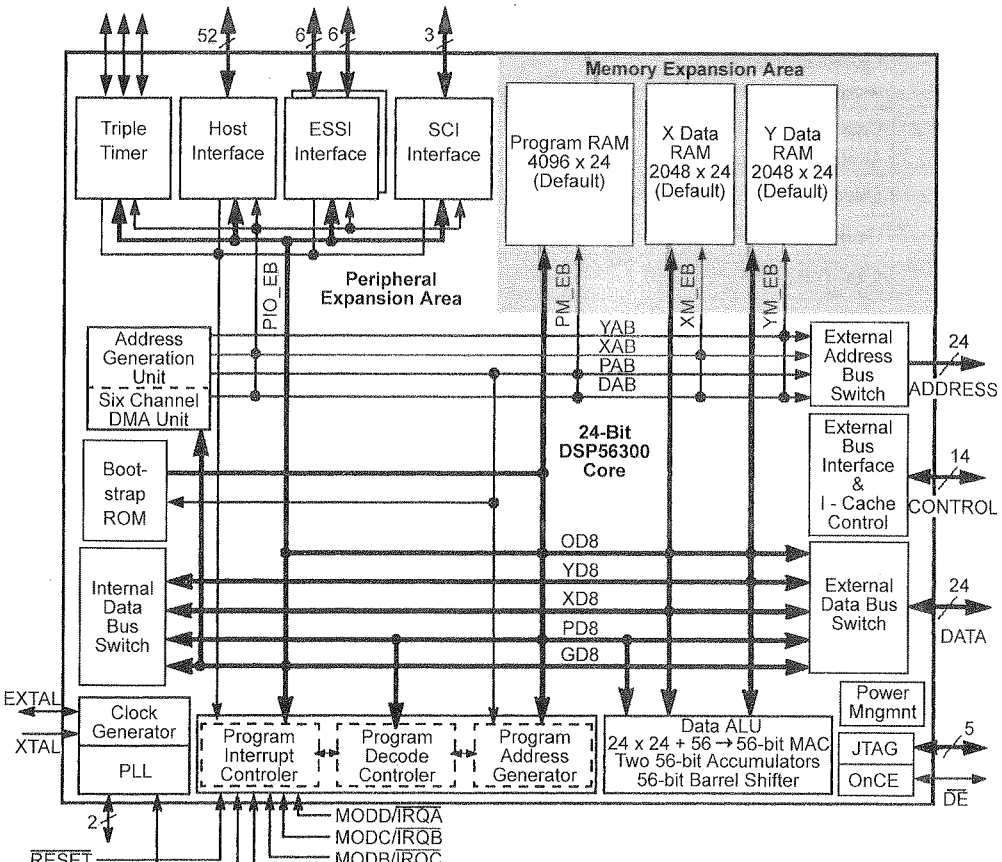
4.1. PROCESORY Z RODZINY DSP56000

Schemat blokowy przykładowego rdzenia procesorów DSP z serii 56300 [16] jest przestawiony na rys. 10. Procesor ten posiada architekturę 24 bitową, co zapewnia zakres dynamiki 144 dB, który jest szczególnie ważny dla obróbki dźwięku. Przyspieszenie operacji konwolucji uzyskuje się dzięki zastosowaniu pięciu niezależnych magistral danych i czterech magistral adresowych oraz trzech niezależnych wewnętrznych bloków pamięci:

- blok pamięci programu oraz połączone z nią magistrale: PDB — magistrala danych programu, PAB- magistrala adresowa,
- blok pamięci danych Y: YDB — magistrala danych, YAB- magistrala adresowa,
- blok pamięci danych X: XDB — magistrala danych, XAB- magistrala adresowa.

Dzięki zastosowaniu trzech niezależnych magistral oraz przyporządkowanych im trzech bloków pamięci: jednego bloku programu oraz dwóch bloków pamięci danych

możliwe jest pobranie w jednym cyklu zegara: instrukcji, danej X, danej Y, czyli wszystkich operacji dostępu do pamięci potrzebnych do wykonania instrukcji MAC($Z = Z + X \cdot Y$).



Rys. 10. Rdzeń procesorów sygnałowych serii DSP56300

Procesor ten posiada trzy niezależne jednostki wykonawcze, które również zostały zaprojektowane z myślą o procesie konwolucji:

- Jednostka Arytmetyczno Logiczna ALU (ang. *Arithmetic Logic Unit*) — jednostka ta umożliwia wykonanie operacji MAC($Z = Z + X \cdot Y$) w jednym takcie zegara oraz posiada cztery 24-bitowe rejestry wejściowe X0, X1, Y0, Y1 oraz dwa 56-bitowe rejestry wyjściowe A, B. Dzięki zastosowaniu 56-bitowych rejestrów wyjściowych uzyskuje się duży poziom bezpieczeństwa przed przepełnieniem (wynik mnożenia dwóch liczb 24-bitowych jest 48-bitowy, co daje 8-bitowy zakres bezpieczeństwa). Oprócz operacji MAC, ALU umożliwia wykonanie operacji: samego mnożenia (bez

dodawania), dodawania, odejmowania, przesunięć bitowych, operacji logicznych na bitach oraz operacji skalowania i zaokrąglania.

- Jednostka Generacji Adresu AGU (ang. *Address Generator Unit*) — jednostka ta składa się z dwóch równolegle pracujących jednostek kalkulacji adresu i jest bezpośrednio podłączona do dwóch z wybranych magistral adresowych (w procesie konwolucji najczęściej do XAB, YAB). Każda z dwóch jednostek zawiera 3*4 16-bitowe (nowsze wersje 24 bitowe) rejestry: (M0-3, N0-3, R0-3), (M4-7, N4-7, R4-7). Dzięki zastosowaniu dwóch jednostek generacji adresu, obliczanie adresu pobrania danych X, Y odbywa się bez udziału jednostki ALU, tak więc jednostka ALU może wykonywać tylko operacje arytmetyczne na danych.
- Jednostka Kontroli Programu PCU (ang. *Program Control Unit*) — jednostka ta jest odpowiedzialna za: obliczanie adresu programu, dekodowanie instrukcji, sprzętową kontrolę pętli oraz obsługę przerwań. Na szczególną uwagę zasługuje sprzętowa obsługa pętli, zawierająca rejestr adresu końca pętli (ang. *Loop Address Register*), który wskazuje adres ostatniej instrukcji w pętli oraz licznika pętli (ang. *Loop Counter*), który jest dekrementowany po wykonaniu ostatniej instrukcji pętli oraz porównywany z wartością 1 (jeżeli jest równy 1 następuje wyjście z pętli). Dzięki takiej budowie procesora instrukcja pętli nie wpływa znacząco na szybkość wykonywania programu.

Algorytm konwolucji dwuwymiarowej na procesorze sygnałowym serii DSP56300 można zaimplementować następująco: współczynniki konwolucji należy umieścić w pamięci Y, a współczynniki danych wejściowych w pamięci X. Procesor DSP56301 posiada pamięć X o wielkości 2048 słów (w następnych procesorach pamięć ta uległa zwiększeniu), która jest niewystarczająca do załadowania całego obrazu 512×512 (256k słów), jednakże dla potrzeb omawianego algorytmu wystarczy, aby umieść w niej 4 linie obrazu. Reasumując dla konwolucji 3×3 wymagane są tylko dwa transfery do/z pamięci zewnętrznej na jeden piksel obrazu wynikowego. Pierwszy w celu pobrania prawego dolnego piksela, drugi w celu zapamiętania wyniku. Instrukcje transferu nie muszą jednak bezpośrednio angażować procesor główny ponieważ układy procesorów sygnałowych serii DSP56300 posiadają rozbudowane układy bezpośredniego dostępu do pamięci DMA.

Typowa instrukcja konwolucji na procesorze sygnałowym serii DSP56300 wygląda następująco:

Listing 4. Fragment kodu konwolucji na procesor sygnałowy serii DSP56300

```
a) MAC X0, Y0, A  X: (R0)+, X0  Y: (R4)+, Y0  (lub)
b) MAC X0, Y0, A  X: (R0)+N0, X0  Y: (R4)+, Y0
```

Taka typowa instrukcja konwolucji (listing 4) wykonuje trzy niezależne operacje:

- operację MAC($A = A + X0 \cdot Y0$) w jednostce ALU,
- operację załadowania rejestru wejściowego X0 daną z pamięci X wskazaną przez rejestr R0 oraz dla przykładu a) postinkrementację rejestru R0 (rejestr R0 będzie

wskazywał następny piksel w linii) lub dla przykładu b) zwiększenie wartości rejestru R0 tak, aby wskazywał on na pierwszy piksel w następnej linii (rejestr N0 zawiera stałą wartość równą np. 510 = 512 − 3 + 1 dla konwolucji 3×3 i długości linii 512),

- operację załadowania rejestru wejściowego Y0 daną z pamięci Y wskazaną przez rejestr R4 oraz postinkrementację rejestru R4.

Warto podkreślić, że dzięki trybom adresowym modulo istnieje łatwa kontrola granicy adresowania poprzez operację zawijania. Przykładowo przy braku adresacji modulo, w celu ograniczenia pobierania danych tylko z wbudowanej pamięci X (np. 0-2047) po każdej operacji sumy {(R0)+lub+(R0)+N0} należałoby każdorazowo dokonać porównania z liczbą graniczną 2047 i ewentualnie wykonać korektę co ilustruje listing 5. Adresowanie modulo umożliwia łatwe stosowanie bufora FIFO, który jest konieczny do implementacji operacji konwolucji korzystającej tylko z wbudowanej pamięci procesora.

Listing 5. Przykład kodu w języku C zastępującego operację adresowania Modulo

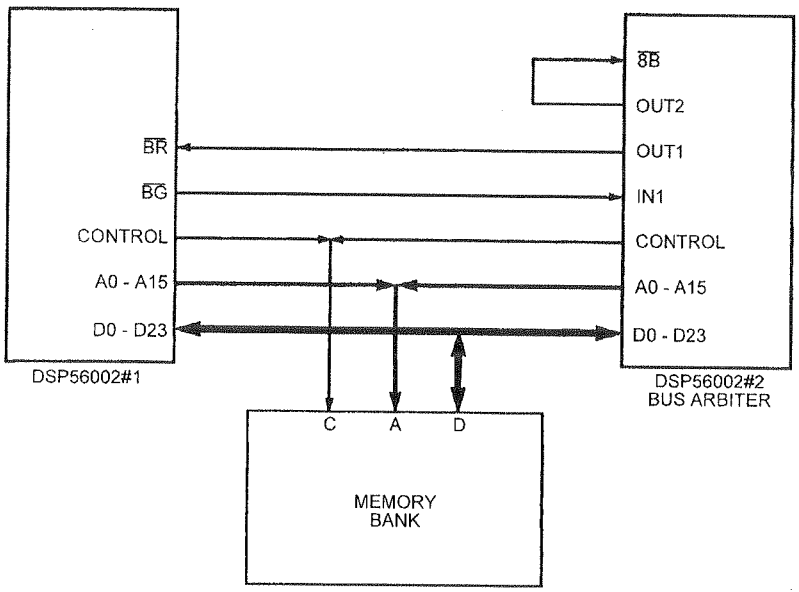
```
N0++;
if(N0>2047) N0-=2047;
```

Procesor sygnałowy DSP56301 posiada również duże możliwości komunikacji z otoczeniem poprzez zastosowanie sześciu kanałów DMA oraz następujących portów :

- Port A — równoległy port (24 linie danych i 24 linie adresowe) — port ten umożliwia proste podłączenie do pamięci SRAM, DRAM. Oprócz tego port A posiada arbitraż magistrali, dzięki czemu można łatwo budować systemy wieloprocesorowe.
- Port B — magistrala PCI.
- Porty szeregowe C, D, E.

Operacja konwolucji, realizowana na procesorze sygnałowym serii DSP56300, wymaga około 10 cykli zegarowych dla konwolucji 3×3 (przy braku postoju związanego z transmisją danych wejściowych i wyjściowych). Przy częstotliwości zegara równej 100 MHz daje to czas konwolucji 3×3 dla obrazu 512×512 równy: 26.3 ms. Czas ten jest porównywalny do czasów osiągniętych dla procesora ogólnego przeznaczenia i pozwala na przetwarzanie obrazu w czasie rzeczywistym ($t<40$ ms). Warto jednak zwrócić uwagę, że zwiększenie rozdzielczości obrazu, ilości przetwarzanych obrazów na sekundę lub też zwiększenie rozmiaru okna konwolucji może spowodować, że powyższy procesor będzie miał za małą moc obliczeniową. Możliwe jest wtedy zastosowanie układu wieloprocesorowego, którego przykład jest przedstawiony na rys. 11.

Przedstawiona na rys. 11 architektura posiada wadę utrudnionej komunikacji z pamięcią. Problem ten może być rozwiązany poprzez zastosowanie pamięci wieloportowej lub też poprzez zastosowanie matrycy przełączanych magistral (ang. *crossbars*). Jednakże istnieją inne układy wieloprocesorowe jednoukładowe bardziej przystosowane do pracy wieloprocesorowej (np. układ TMS320C80).



Rys. 11. Architektura złożona z dwóch procesorów DSP56002 wykorzystujących port A i wbudowany arbitraż magistrali

Procesor sygnałowy rodziny DSP56000 jest procesorem zoptymalizowanym pod kątem realizacji operacji konwolucji, jednakże nie jest zoptymalizowany pod kątem przetwarzania obrazu. Na przykład architektura procesora jest 24-bitowa (istnieją również procesory 16-bitowe: seria 56100 lub 56800), a większość operacji na obrazie to operacje 8-bitowe lub nawet jednobitowe.

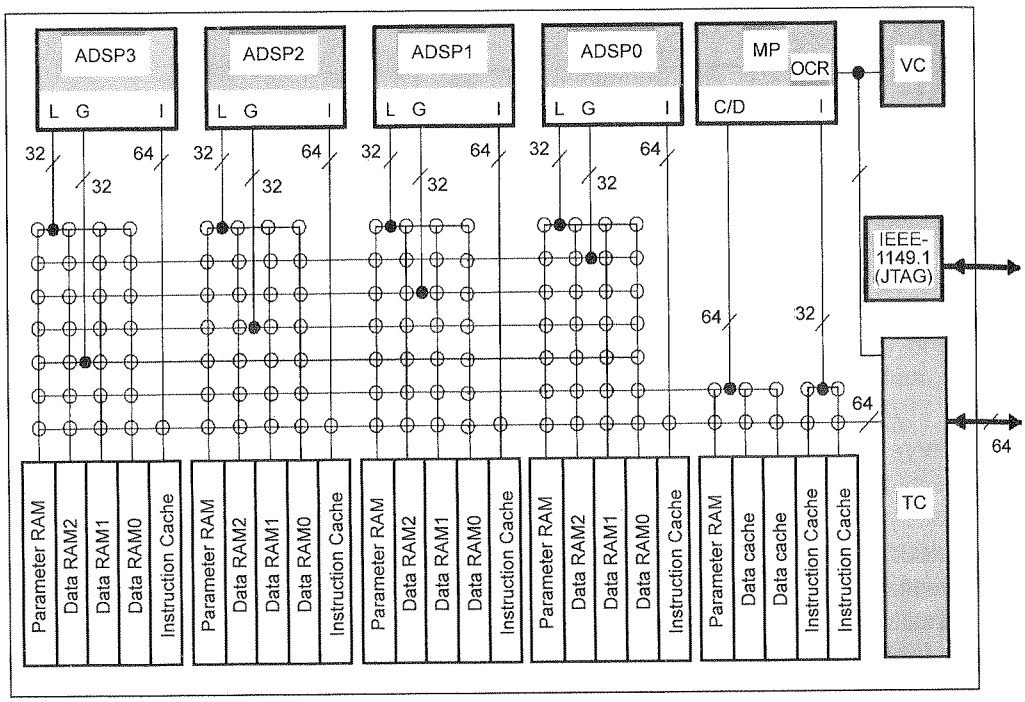
4.2. PROCESOR TMS 320C80

Procesor sygnałowy TMS320C80 [17] jest układem wieloprocesorowym zawierającym:

- Procesor nadrzędny zmiennoprzecinkowy MP (ang. *Master Processor*).
- Cztery podrzędne procesory równoległe stałoprzecinkowe PP (ang. *Parallel Processor*) ADSP0-3.
- Jednostkę kontroli transferu TC (ang. *Transfer Controller*) przeznaczoną do kontroli komunikacji z otoczeniem oraz nadzorującą pamięć *cache*. Dzięki TC wszelki transfer pomiędzy pamięcią zewnętrzną, a pamięciami wewnętrznymi odbywa się bez udziału procesorów. Maksymalna szybkość transferu wynosi 400 MB/s.
- Jednostkę kontroli video VC (ang. *Video Controller*) będącą układem interfejsu z dwoma systemami wizyjnymi, zawierającą dwa timery, układ synchronizacji, wygaszania, logikę dla obrazów z lub bez przeplotu itd.
- Wbudowaną pamięć RAM 50kB oraz układ przełączników (ang. *crossbar*), dzięki którym została ograniczona komunikacja z pamięcią zewnętrzną. Wbudowana pa-

mięć *cache* programu, osobna dla każdego procesora, umożliwia szybkie wykonywanie powtarzających się instrukcji.

Schemat blokowy układu TMS320C80 jest przedstawiony na rys. 12.



Rys. 12. Schemat blokowy układu TMS320C80 ze szczególnym uwzględnieniem wbudowanej pamięci i układów przełączników (ang. *crossbar*)

Procesor równoległy PP

Schemat wewnętrzny procesora PP przedstawiony jest na rys. 13. Składa się on z bloku danych DA (ang. *Data Unit*), jednostki generacji adresów lokalnych i globalnych: (ang. *Local* i *Global Address Unit*) oraz jednostki kontroli programu PFCU (ang. *Program Flow Control Unit*). Jednostka komunikuje się z otoczeniem za pomocą trzech niezależnych portów programu, danych globalnych i danych lokalnych.

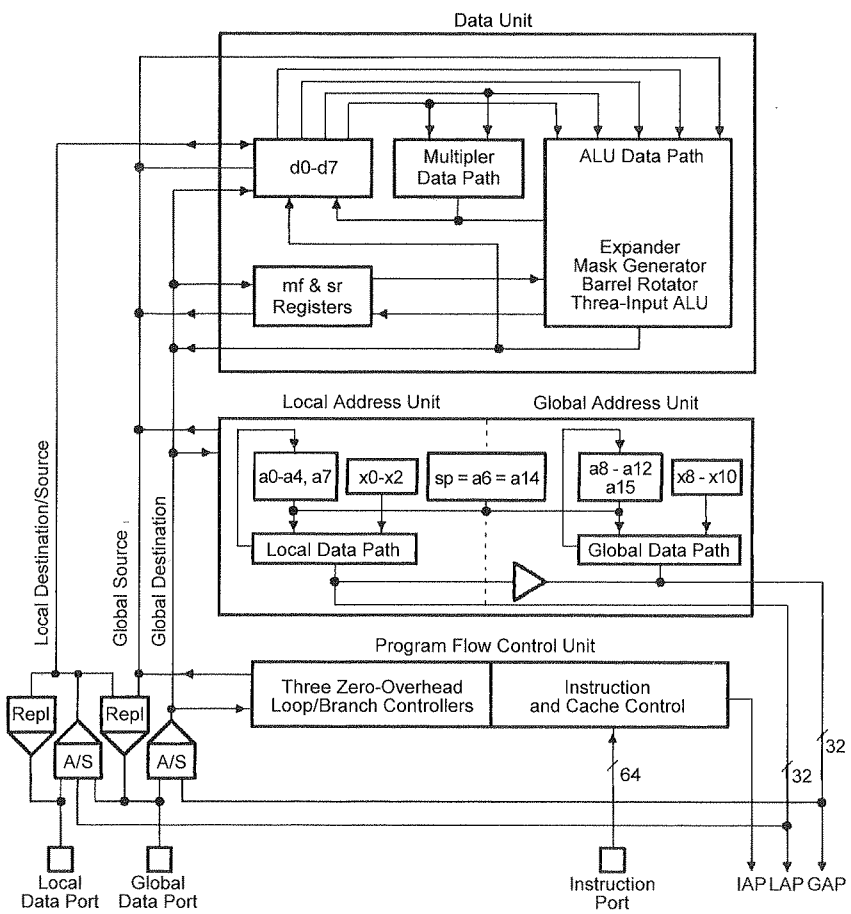
Blok danych DA jest podzielony na dwie ścieżki, które działają niezależnie od siebie:

- Blok mnożący.
- Blok ALU.

Głównym zadaniem bloku mnożącego jest wykonywanie operacji mnożenia, która może być wykonana w pojedynczym cyklu zegara. Oprócz tego blok mnożący może wykonać niezależną operację skalowania i zaokrąglania. Normalnie układ mnożący wykonuje pojedynczą operację mnożenia $16 \times 16 = 32$ bity, jednakże posiada on również możliwość równoczesnego wykonywania dwóch rozdzielonych operacji mnożenia

(ang. *split multiply*) na mniejszych słowach (tj. $8 \times 8 = 16$), czyli może wykonywać operacje SIMD, wyglądające podobnie jak operacje wykonywane w koprocesorze MMX. Warto zwrócić uwagę, że argumentami mnożenia może być jedna zmienna ze znakiem (ang. *signed*), a druga zmienna bez znaku (ang. *unsigned*). Dzięki temu możliwe jest wykonanie operacji konwolucji na obrazie, który z reguły jest 8 bitową zmienną bez znaku, natomiast współczynniki konwolucji są zmiennymi ze znakiem.

Blok ALU jest zoptymalizowany pod kątem operacji bitowych, przyjmuje trzy 32--bitowe dane wejściowe i wyprowadza jedną 32-bitową dana wyjściową. Dla przykładu, ALU potrafi wykonać operacje $(A \& \sim C)|(B \& C)$ (A i nie C lub B i C) w pojedynczym cyklu zegara. Ponadto umożliwia wykonywanie operacji arytmetycznych: dodawania i odejmowania na dwóch argumentach. Operacje te mogą być również wykonywane na zmiennych rozdzielonych (ang. *split operation*) o szerokości słowa 32, 16 albo 8 bitów, czyli blok ALU może wykonać w jednym cyklu zegara do czterech operacji dodawania lub odejmowania na słowie 8-bitowym. Warto podkreślić, że równolegle z wyżej



Rys. 13. Schemat blokowy procesora równoległego PP

opisanymi operacjami blok ALU umożliwia rotacje, maskowanie bitowe i podobne operacje.

Każdy procesor równoległy PP posiada dwa niezależne bloki adresowe: lokalny i globalny, dzięki czemu w procesie konwolucji zarówno adresy danych i współczynników konwolucji mogą być obliczane niezależnie od bloku danych.

Architektura procesora PP ma cechy architektury VLIW (ang. *Very Long Instruction Word*), gdzie każda instrukcja ma długość 64 bitów. W konsekwencji jedna instrukcja zawiera rozkazy dla dwóch bloków danych (bloku mnożącego i ALU) oraz dla dwóch jednostek kalkulacji adresu (globalnej i lokalnej). Ponadto możliwe są przesłania i instrukcje warunkowe, jednakże w tym przypadku nie można korzystać z dwóch przesłań jednocześnie.

Blok kontroli programu PFCU, w porównaniu z wcześniej prezentowanymi procesorami (Pentium i DSP56300), posiada widoczną od strony użytkownika strukturę potokową, z trzema etapami potokowości:

1. pobrania instrukcji (ang. *instruction fetch*),
2. kalkulacji w bloku adresowym (ang. *address unit computation*),
3. kalkulacji w bloku danych i komunikacji z pamięcią (ang. *execute data unit operation and memory transfers*).

Widoczność procesu potokowości przerzuca na programistę konieczność zmiany składni pętli i instrukcji skoku. Dla przykładu, listing 6 przedstawia typową pętlę software'ową. Dla tego przykładu, kiedy instrukcja skoku zostanie wykonana (na trzecim etapie potokowości), pobrane i częściowo wykonane zostaną już instrukcje: Instrukcja 2 (w momencie wykonania instrukcji skoku znajduje się ona na drugim etapie potokowości) oraz Instrukcja 3 (na pierwszym etapie). Procesor PP nie posiada możliwości anulowania zmian wprowadzonych przez częściowo wykonane instrukcje: Instrukcja 2–3; dlatego instrukcje te zostaną wykonane do końca, pomimo że instrukcja skoku wystąpiła przed tymi instrukcjami.

Listing 6. Przykład pętli software'owej

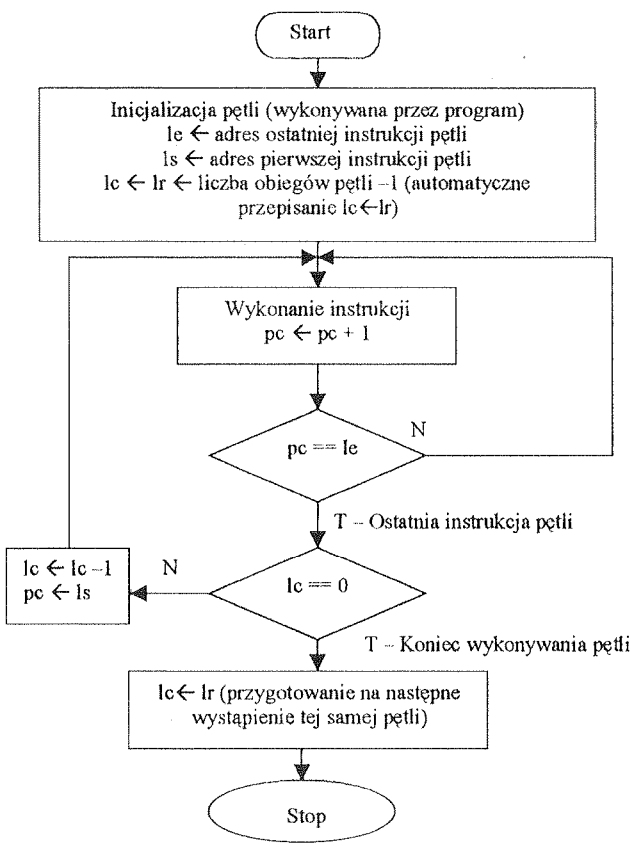
```
Skok 1: Instrukcja 1; Pierwsza instrukcja pętli
br = Skok 1;    Instrukcja skoku — skok na początek pętli
Instrukcja 2;   Druga instrukcja pętli
Instrukcja 3;   Trzecia i ostatnia instrukcja pętli
Instrukcja 4;   Instrukcja poza zakresem pętli
```

Zatem dla instrukcji pętli pisanej na procesor z anulowaniem zmian instrukcja skoku znajduje się na końcu pętli, czyli po Instrukcji 3. Natomiast dla procesora PP mikroprocesora TMS320C80 instrukcja skoku musi znajdować się o dwie instrukcje wcześniej jak to przedstawia listing 6. Zaletą tego typu rozwiązania jest brak opóźnień spowodowanych postojem procesora na wykonanie instrukcji skoku, a w porównaniu z procesorem Pentium również brak skomplikowanej logiki predykcji skoku. Wadą tego rozwiązania jest nietypowa kolejność instrukcji, która może powodować błędy podczas pisania programu. Ponadto ściśle określona liczba stopni potokowości utrudnia wszelkie

poważne zmiany architektury, ponieważ konsekwencją zmiany liczby stopni potokowości jest konieczność zmian programu. Następna wada ujawnia się w przypadku gdy wynik skoku warunkowego jest znany dopiero przed samą instrukcją skoku; w tym przypadku instrukcja skoku musi być ostatnią instrukcją, a dwie następujące po niej instrukcje, instrukcjami *nop* (ang. *no operation* — nic nie rób). Zatem większy ciężar programowania procesora PP jest przerzucony na barki programisty lub też kompilatora.

Procesor PP układu TMS320C80 posiada również sprzętowe wykonywanie pętli poprzez zastosowanie rejestrów:

- le0-2 — rejestry adresu końca pętli (ang. *loop end*),
- ls0-2 — rejestry adresu startu pętli (ang. *loop start*),
- lc0-2 — liczniki pętli (ang. *loop counter*) — automatycznie zmniejszane o jeden po dojściu do ostatniej instrukcji pętli,
- lr0-2 — rejestry przeładowujące licznik pętli po zakończeniu pętli (ang. *loop reload*),
- lctl — rejestr kontrolny (ang. *loop control*).



Rys. 14. Diagram demonstrujący zachowanie się pętli hardware'owej

Diagram ilustrujący zachowanie się pętli sprzętowej przedstawia rys. 14. Wszystkie operacje pętli (z wyjątkiem operacji inicjalizujących) są wykonywane równolegle i niezależnie od innych funkcji procesora, dlatego nie wymagają dodatkowego zaangażowania procesora poza procesem inicjalizacji, zatem rozpisanie pętli nie przyniesie żadnego przyspieszenia pracy procesora. Należy zauważyć, że równoległa praca bloku obsługi pętli nie jest włączona w proces potokowej pracy procesora, dzięki czemu nie występuje opóźnienie skoku o dwie instrukcje jak to ma miejsce w przypadku skoku software'owego. Z tego powodu blok obsługi pętli może być również wykorzystywany do szybkiej obsługi skoku bezwarunkowego lub warunkowego, kiedy to wynik warunku jest zapisywany do rejestru *lc*. Jeżeli wartość rejestru *lc* jest różna od zera dla adresu instrukcji równego wartości rejestru *le*, nastąpi skok pod adres zapisany w rejestrze *ls*.

Procesor PP posiada trzy zestawy rejestrów obsługi pętli przez co możliwe staje się zagnieżdżenie trzech pętli, a dzięki funkcji rejestru przeładowującego *lr* nie jest konieczna powtórna inicjalizacja rejestru *lc* po zakończeniu wykonywania pętli (patrz rys. 14).

Porównując procesory sygnałowe serii DSP56300 i TMS320C80 PP należy stwierdzić, że procesor serii DSP56300 posiada tylko dwa rejestry obsługi pętli odpowiadające rejestrom *lc*, *le*. W wyniku tego procesor serii DSP56300 może w sposób sprzętowy obsługiwać tylko pojedynczą pętlę, a rejestr *lc* po zakończeniu pętli musi zostać powtórnie przeładowany. Ponadto rolę rejestru startu *ls* pełni stos, do którego procesor musi się odwołać po dojściu do ostatniej instrukcji pętli.

Operacja konwolucji w procesorze równoległym PP

Pojedynczy procesor PP w jednym takcie zegara może wykonać dwie operacje MAC. W tym celu procesor wykorzystuje blok mnożący do wykonania dwóch rozdzielonych operacji (ang. *split*) mnożenia na danych wejściowych 8-bitowych oraz blok ALU do wykonania dwóch rozdzielonych operacji dodawania na słowach 16-bitowych. Przy konwolucji o rozmiarze okna 3×3 występuje jednak problem rozłożenia danych przy operacjach rozdzielonych: w jednej linii tylko trzy dane wejściowe znajdują się koło siebie, a procesor pobiera dane parami. Dlatego podobnie jak dla koprocesora MMX, obliczenia należy wykonywać dla dwóch pikseli równolegle. Fragment kodu konwolucji [18] dla procesora PP przedstawia listing 7.

Listing 7. Pojedyncza instrukcja procesora PP wykonywana w jednym cyklu zegara

```
mult = m pixel *filter||          ; Pomnóż piksel przez współczynnik konwolucji
result+ = mult||                  ; Dodaj poprzedni wynik mnożenia
filter = b *Ga_filt++||           ; Załaduj następny współczynnik konwolucji (dana 8 bitowa ze znakiem)
pixel = ub *(La_pt++= offset1)    ; Załaduj następny piksel wejściowy i przejdź do następnej linii (8 bitów
                                    bez znaku)
```

Przeprowadzono obliczenia dotyczące liczby cykli zegarowych i odpowiadającego im czasu potrzebnych do wykonania operacji konwolucji dla okna o rozmiarach 3×3 i obrazu 512×512 pikseli. Wyniki obliczeń przedstawia tablica 4. Przedstawione wyniki obliczeń porównano w tablicy z czasem zmierzonym doświadczalnie, który uległ dalszemu skróceniu [15].

Tablica 4

Realizacja konwolucji w układzie TMS320C80 50MHz

Opis	Operacje nierozdzielone	Operacje rozdzielone
Liczba cykli zegarowych na jeden piksel [18]	11	8.5
Czas realizacji operacji konwolucji (okno 3×3, obraz 512×512) dla podanej liczby cykli zegarowych (idealizowany dla czterech PP)	14.4 ms	11.2 ms
Wynik doświadczalny [15]	10.3 ms	

Środowisko pracy procesora równoległego

W systemie wieloprocesorowym komunikacja pomiędzy procesorami stanowi poważne ograniczenie mocy obliczeniowej systemu. W przypadku układu TMS320C80 przyjęto architekturę matryc przełączających *crossbar*, która charakteryzuje się dobrym czasem dostępu do pamięci. Wadą tego rozwiązania jest występowanie konfliktów w momencie, kiedy więcej niż jeden procesor komunikuje się z tym samym bankiem pamięci. Aby zminimalizować liczbę konfliktów dla jednego procesora przydzielono trzy banki pamięci danych, jeden bank pamięci parametrycznej oraz jeden bank pamięci *cache* programu, każdy o wielkości 2kB (rys. 12). Pamięć parametryczna zawiera między innymi: wektory przerwań, dane do pakietowego przesyłania danych oraz pamięć stosu, która może być dzielona z danymi użytkownika.

Każdy procesor równoległy PP posiada trzy porty do komunikacji z pamięcią:

- 64-bitowy port programu (*instruction port I*),
- 32-bitowy port danych globalny (*global port G*),
- 32-bitowy port danych lokalnych (*local port L*).

Dzięki układowi kluczy przełączających *crossbars*, port globalny może komunikować się prawie z każdym bankiem pamięci wewnętrznej układu TMS320C80 (rys. 12). Port lokalny posiada silne ograniczenia — może się komunikować tylko z czterema bankami pamięci, które zostały przydzielone danemu procesorowi PP. Dlatego w przypadku instrukcji odwołującej się do dwóch obszarów pamięci, które wykraczają poza zakres pamięci lokalnej, procesor PP używa dwukrotnie portu globalnego, przez co dana instrukcja jest wykonywana o jeden cykl zegara dłużej.

Bardzo ważną rolę w układzie TMS320C80 spełnia blok Kontroli Transferu TC (ang. *Transfer Controller*), który zarządza komunikacją układu TMS320C80 z pamięcią zewnętrzną. Magistrala pamięci zewnętrznej ma szerokość do 64 bitów i maksymalną częstotliwość 50 MHz (maksymalny transfer 400 MB/s). TC może dostarczać dane bezpośrednio do danego procesora poprzez procedurę DEA (ang. *Direct External Access*), jeżeli procesor odwołuje się do danej, która nie znajduje się w pamięci

wbudowanej. Jednakże procedura ta nie jest polecana ze względu na długi czas dostępu do danych. Dlatego preferowany jest transfer blokowy PT (ang. *Packet Transfer*), kiedy to TC bez udziału procesora PP transferuje blok danych pomiędzy pamięcią zewnętrzną, a dowolną pamięcią lokalną.

Do celów synchronizacji procesów wykonywanych na czterech procesorach PP i kontroler transferu TC może służyć Procesor Nadrzędny MP (ang. *Master Procesor*). W tym celu zastosowano specjalny pseudo-rejestr *cmnd*. Poprzez zapis do rejestru *cmnd* procesor MP może wydać następujące komendy dla ustalonego procesora (procesorów) PP:

- restartu (ang. *reset*),
- zatrzymania (ang. *halt*),
- wyjścia ze stanu zatrzymania (ang. *unhalt*),
- przerwania zadania (ang. *task interrupt*),
- przerwania w celu odebrania komunikatu (ang. *massage interrupt*),
- wyzerowania pamięci *cache* programu (ang. *instruction-cache reset*).

Do rejestru *cmnd* może również zapisywać dowolny procesor PP, ale zakres wydawanych komend jest ograniczony ze względu na bezpieczeństwo wykonywanych procesów, np. procesor PP może tylko wydać komendę przerwania do innego procesora w celu odebrania komunikatu.

Porównując wyniki teoretyczne dla procesora równoległego PP oraz dane otrzymane doświadczalnie dla całego układu TMS320C80 (tabl. 4) można zauważyć, że szybkość realizacji obliczeń algorytmu konwolucji jest ograniczona głównie przez moc obliczeniową procesorów PP, a nie przez ograniczenia wynikające z komunikacji i przesyłania danych pomiędzy procesorami.

4. PODSUMOWANIE

W opracowaniu zostały przedstawione zagadnienia związane z implementacją algorytmu konwolucji 2D w specjalizowanych układach VLSI, procesorach ogólnego przeznaczenia oraz w rozbudowanych procesorach sygnałowych. Jako przykładowe rozwiązania specjalizowanych układów VLSI do realizacji obliczeń konwolucji 2D przedstawiono układy firm: Harris, Plessey i SGS Thomson -InMos. Warto zauważyć, że procesor specjalizowany wykonany jako układ VLSI nie wymaga pobierania i dekodowania rozkazu. Program jest zapamiętany w strukturze układu i jest wykonywany bez przestojów. Jest to ważna cecha układów VLSI, zważywszy jak wielki problem dla układów procesorowych stanowi pogodzenie szybkiego wykonywania rozkazów, struktury potokowej procesora oraz wykonywanie instrukcji skoków. W konsekwencji oprócz skomplikowanego dekodera rozkazów, procesory ogólnego przeznaczenia posiadają specjalizowane układy predykcji skoku, logikę anulowania zmian na skutek błędnej predykcji, logikę dynamicznego wykonywania programu, układy sprzętowej implementacji pętli, osobną magistralę dla programu czy też szybkiej pamięci programu typu *cache*. Ponadto zrównoleglenie algorytmu przebiega dużo prościej w układzie spec-

jalizowanym VLSI — bez dodatkowych konfliktów magistrali i dodatkowych procesów synchronizujących pracę poszczególnych procesorów równoległych.

Stosowanie specjalizowanych układów VLSI zapewnia wysokie parametry czasowe realizowanych obliczeń, ale wymaga stosowania dedykowanego otoczenia sprzętowego, związanego z dostarczeniem odpowiednio zorganizowanych danych wejściowych i odebraniem przetworzonych danych. Stosowanie procesorów dedykowanych VLSI powoduje także, że jakakolwiek zmiana w projekcie może pociągnąć za sobą bezużyteczność danej platformy sprzętowej. Przykładem jest np. rodzaj rozszerzenia na brzegach obrazu: zmiana ze spiralnego (używanego przez układ dedykowany) na rozszerzenie cylindryczne lub lustrzane.

Alternatywą dla układów specjalizowanych VLSI jest realizacja algorytmu konwolucji 2D w procesorach ogólnego przeznaczenia i w procesorach sygnałowych. Dynamiczny rozwój tych procesorów w ostatnim okresie oferuje nowe możliwości w zakresie osiągania wysokich parametrów czasowych związanych z implementacją algorytmu dwuwymiarowej konwolucji na przykładzie obrazu wizyjnego. Procesor Pentium jest przedstawicielem procesorów ogólnego przeznaczenia. Posiada on rozbudowaną architekturę poprzez zastosowanie struktury superskalarnej, dynamicznego wykonania programu, predykcji skoku czy też stosowanie wbudowanej pamięci *cache*. Warto zauważyć, że architektura procesora ogólnego przeznaczenia ulega w pewnym stopniu nasyceniu — dodatkowe rozbudowanie architektury procesora nie powoduje proporcjonalnego wzrostu szybkości. Aby przyspieszyć operacje multimedialne, które wykonują takie same operacje na wielu tego samego rodzaju danych, wprowadzono dodatkowy koprocesor MMX. Procesor ten znacząco przyspiesza operacje konwolucji aczkolwiek nie stanowi przełomu w uzyskiwanym czasie obliczeń — jego znaczenie jest coraz mniejsze w porównaniu z udoskonalanymi jednostkami stałoprzecinkowymi nowszych procesorów. Podsumowując, procesor ogólnego przeznaczenia w obliczeniach konwolucji nie wykorzystuje swoich możliwości optymalnie — iloczyn czasu realizacji operacji konwolucji i skali skomplikowania układu jest wysoki. Dlatego usprawiedliwieniem stosowania procesora ogólnego przeznaczenia do operacji konwolucji jest jego powszechność, a przez to niska cena.

Dobrą alternatywą dla procesora ogólnego przeznaczenia są procesory sygnałowe DSP, których architektura przeznaczona jest do tego typu zastosowań. W opracowaniu omówiono architekturę dwu procesorów sygnałowych: z rodziny DSP56300 firmy Motorola oraz TMS320C80 firmy Texas Instruments. Procesor z rodziny DSP56300 może wykonywać operacje konwolucji o rozmiarze okna 3×3 i rozdzielczości obrazu 512×512 w czasie rzeczywistym. Jednakże zwiększenie rozdzielczości lub rozmiaru okna konwolucji może spowodować, że konieczne stanie się równoległe łączenie dwóch lub więcej procesorów DSP, a to pociąga za sobą dodatkowe skomplikowanie układu. Lepszym rozwiązaniem tego zagadnienia jest zastosowanie procesora DSP TMS320C80, który posiada cztery procesory równoległe PP oraz dodatkowy procesor nadrzędny MP. Porównując budowę procesorów PP z budową procesora ogólnego przeznaczenia, można zauważyć, że procesor PP w przeciwieństwie do procesora Pentium ma widoczną od strony programującego architekturę potokową oraz zamiast skomplikowanej struktury predykcji

skoku posiada sprzętowe wspomaganie pętli (skoku). Ponadto w strukturze superskalarnej procesor Pentium jest odpowiedzialny za rozczłonkowanie programu na poszczególne procesory równoległe, natomiast w przypadku procesora TMS320C80 odpowiedzialność ta spada na barki programisty. Cecha ta upraszcza architekturę procesora TMS320C80 i służy jego lepszemu wykorzystaniu, jednakże komplikuje pisanie programu — szczególnie dla algorytmów bardziej skomplikowanych niż proces konwolucji. Warto zauważyć, że aż do tej pory trwają prace nad kompilatorem języka C na procesor TMS320C80, który potrafiłby optymalnie wykorzystać procesor główny i jego cztery procesory równoległe. Należy także zauważyć, że stosowanie procesorów sygnałowych daje znacznie większą przewidywalność czasów wykonania poszczególnych operacji w porównaniu z najnowszymi procesorami ogólnego przeznaczenia, w których efekt działania pamięci *cache* uniemożliwia precyzyjne wyliczenia tych czasów.

Należy w tym miejscu zaznaczyć, że opisane procesory ogólnego przeznaczenia i procesory sygnałowe nie wyczerpują pełnej listy tych procesorów. Warto wspomnieć o procesorach PowerPC, głównie generacji G4 z technologią AltiVec (posiadającą pewne podobieństwa do mechanizmów MMX) oraz o procesorach sygnałowych rodziny TigerSHARC firmy Analog Devices (stosującym technikę SIMD zbliżoną także do mechanizmów MMX) i procesorach TMS320C6xxx firmy Texas Instruments. Pełne omówienie całej listy procesorów stałoby się sporym opracowaniem monograficznym, podczas kiedy autorzy artykułu mieli zamiar przedstawić jedynie zakres problemów jakie stają przed użytkownikami procesorów ogólnego przeznaczenia i procesorów sygnałowych, zamierzających stosować te procesory do szybkich obliczeń w systemach wizyjnych czasu rzeczywistego.

Podsumowując należy stwierdzić, że współczesne procesory ogólnego przeznaczenia, a szczególnie procesory sygnałowe, dobrze radzą sobie z realizacją obliczania konwolucji 2D w czasie rzeczywistym dla obrazu o rozmiarach 512×512 pikseli i oknie konwolucji 3×3. Niestety w tym czasie procesor taki nie może równolegle wykonywać żadnych innych czynności, a zwiększenie okna konwolucji lub rozmiarów obrazu (co ostatnio ma często miejsce nawet w rozwiązaniach komercyjnych np. obrazy HDTV) powodują, że obliczenia te nie są realizowane w czasie rzeczywistym. Wad tych nie posiadają układy specjalizowane VLSI, które z drugiej strony wymagają budowy dedykowanego otoczenia sprzętowego. Nowe możliwości w tym zakresie są związane ze stosowaniem do tego typu obliczeń układów programowalnych, szczególnie układów typu FPGA. Jednak szczegółowe omówienie problemów związanych ze stosowaniem kładów programowalnych FPGA wykracza poza ramy niniejszego opracowania.

BIBLIOGRAFIA

1. R. Gonzalez, P. Wintz: *Digital Image Processing*. Addision–Wesley, 1987
2. K.R. Castleman: *Digital Image Processing*. Prentice Hall, 1996
3. HARRIS: *Digital Signal Processing — Databook*. Harris Semiconductor Inc., 1994
4. PLESSEY: *Digital Signal Processing — IC Handbook*. GEC Plessey Semiconductor, 1990
5. SGS THOMSON: *Image Processing — Databook*. SGS Thomson Microelectronics, 1990

6. K. Wiatr: *Architektura specjalizowanych procesorów potokowych do obróbki wstępnej danych wizyjnych w czasie rzeczywistym.* Kwartalnik Elektroniki i Telekomunikacji, Warszawa, PAN, 1997, z. 1, s. 121–148
7. RAYTHEON: *Data Book.* Raytheon Semiconductor, 1994
8. V.K. Madisetti: *VLSI Digital Signal Processors.* Butterworth Heinemann, 1995
9. D. Anderson, T. Shanley: *Pentium Processor System Architecture.* Addision-Wesley, 1995
10. INTEL: *Intel Architecture Software Developer's. Manual Volume 3: System Programming Guide. (Order Number 243192).* Intel Corporation, 1997
11. INTEL: *Intel Architecture Software Developer's. Manual Volume 1: Basic Architecture. (Order Number 243190).* Intel Corporation, 1997
12. B.B. Brey: *The Intel Microprocessors.* Prentice-Hall, 1997
13. K. Hwang: *Advanced Computer Architecture: Parallelism, Scalability, Programmability.* New York, McGraw-Hill, 1993
14. A. Klaiber: *The Technology behind Crusoe Processors.* Transmeta Co. Jun, 2000 http://www.transmeta.com/crusoe/download/pdf/crusoetechwp.pdf
15. MATROX: *Matrox MIL-32 Library.* Matrox Electronic Systems Ltd, http://www.matrox.com /imgweb/
16. MOTOROLA: *DSP Technical Documentation.* Motorola Inc., http://www.mot.com
17. TEXAS INSTRUMENTS: *TMS320C80 (MVP), User's Guide.* Texas Instruments Inc., http://www.ti.com/
18. TEXAS INSTRUMENTS: *Implementation of an Image Processing Library for TMS320C8x (MVP).* Texas Instruments Inc., 1997

K. WIATR, E. JAMRO

IMPLEMENTATION OF 2D CONVOLUTION ALGORITHMS IN DEDICATED VLSI DEVICES AND GENERAL-PURPOSE PROCESSORS AND DSPs

Summary

This paper discusses usage of different architectures and algorithms for a 2D convolution. At first dedicated VLSI devices are presented, giving the example of Harris, Plessey and InMos chips. Employing these devices results in the low-time calculation, however requires dedicated hardware and appropriate input / output data interface. An alternative solution is the use of general-purpose processors or Digital Signal Processors (DSP). In the recent years, a rapid development of these processors is observed and consequently these processors are often capable (fast enough) of carrying out 2D convolution, e.g. 2D real-time image convolution. In particular, different models of superscalar Pentium architectures (486, P75, P166, P300, MMX) have been discussed. Further, DSPs: Motorola DSP56000 and Texas Instruments TMS320C80 have been approached. The influence of the processor architectures on the 2D convolution calculation process has been extensively investigated. Furthermore, methods of constructing 2D convolution (in particular image convolution) algorithms to reduce calculation time have been investigated and as the result, examples of programs (C and assembler language level) and corresponding calculation times have been given.

Key words: real-time systems, image processing, multiprocessor systems architecture, specialised processors

w uk
sprzę
miesz
istotn
odpo
takich
komó
sprzę
mode
przed
larny
wzro
ryzac
w pr
przyl
łożo

Słow
ne, f

Istnie
liwościac
go, jak i
to: odbic

KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJI, 2000, 46, z. 4, ss. 589–613

Metody modelowania sprzężenia podłożowego w układach scalonych

MARIUSZ SOLECKI, ZBIGNIEW FELENDZER

Zakład Układów Elektronicznych, Politechnika Gdańska

Otrzymano 2000.02.01
Autoryzowano 2000.10.27

Temat artykułu stanowią zagadnienia związane z modelowaniem sprzężenia podłożowego w układach scalonych. W pierwszej części pracy uzasadniono konieczność modelowania sprzężenia podłożowego w procesie projektowania układów. Zaprezentowano kilka przykładów mieszanych analogowo-cyfrowych układów scalonych, w których zakłócenia podłożowe mają istotny wpływ na jakość pracy układu. Zwrócono również uwagę na potrzebę włączenia odpowiednich modeli sprzężenia przez podłoże w proces symulacji dla układów scalonych, takich jak pętla synchronizacji fazowej czy też układy przeznaczone do zastosowań w telefonii komórkowej. W dalszej części pracy omówiono dwie alternatywne metody modelowania sprzężenia podłożowego: metodę opartą na elektrostatycznej funkcji Greena oraz metodę modelowania podłoża z użyciem konstrukcji geometrycznej zwanej mozaiką Voronoi. Z kolei przedstawiono wybrane problemy modelowania sprzężenia podłożowego w krzemowych bipolarnych układach scalonych pracujących w zakresie bardzo wielkich częstotliwości. Wraz ze wzrostem częstotliwości pracy układów i systemów elektronicznych oraz z postępującą miniaturyzacją technologii, problem zakłóceń sprzężenia podłożowego staje się coraz bardziej istotny w procesie projektowania układów scalonych. Stąd też, zasadniczym celem artykułu jest przybliżenie czytelnikom wybranych zagadnień związanych z modelowaniem sprzężenia podłożowego w układach scalonych.

Słowa kluczowe: metody modelowania, sprzężenie podłożowe, układy scalone CMOS i bipolarne, funkcja Greena, mozaika Voronoi.

1. WPROWADZENIE

Istnieje szereg pasożytniczych zjawisk, które pojawiają się przy wysokich częstotliwościach pracy systemów elektronicznych, zarówno w przypadku montażu drukowanego, jak i MCM (*multichip moduls*) oraz na poziomie pojedynczego układu scalonego. Są to: odbicia, przesłuchy, emisja promieniowana, a także szum jednoczesnego przełą-

czania. Zjawiska te można uznać za swego rodzaju globalne problemy w projektowaniu układów i systemów elektronicznych. Muszą być one szczegółowo rozpoznane i przeanalizowane, aby przed wykonaniem projektu można było wnieść do układu odpowiednie modyfikacje mające na celu efektywną redukcję wymienionych powyżej zjawisk pasożytniczych a w konsekwencji wysoką jakość pracy układu.

W ostatnich latach wiele uwagi poświęcono analizie szumu jednoczesnego przełączania. Szum ten, wywołany dużą liczą przełączających się jednocześnie cyfrowych komponentów układu, propaguje się nie tylko w obwodach masy i zasilania dając w efekcie tzw. *ground bounce* oraz *Vdd-bounce*, ale również poprzez podłoże układów scalonych. Wówczas określamy go jako zakłócenia sprzężenia podłożowego. Sprzężenie podłożowe ma szczególnie istotne znaczenie w mieszanych: analogowo-cyfrowych układach scalonych, w których prąd wstrzykiwany do wspólnego podłoża układu z dużej liczby szybko przełączających się elementów cyfrowych może zakłócić poprawne działanie wrażliwych podzespołów analogowych układu [np. 7,10,11,17,25,26,29-32]. Dzieje się tak, ponieważ wstrzykiwany z przełączającej się części cyfrowej do podłoża krzemowego prąd oddziałuje na lokalny potencjał podłoża elementów analogowych powodując zaburzenia napięcia progowego V_t każdego z tranzystorów i w konsekwencji wrażliwe elementy analogowe nie działają zgodnie z projektem. Wielu autorów wskazuje na konieczność odpowiedniego modelowania zakłóceń spowodowanych przez sprzężenie podłożowe [np. 4-8,10-20,22,24,27-31] oraz proponuje różne techniki minimalizacji tych zakłóceń na etapie projektowania układu [np. 4,7,9,12,17,25].

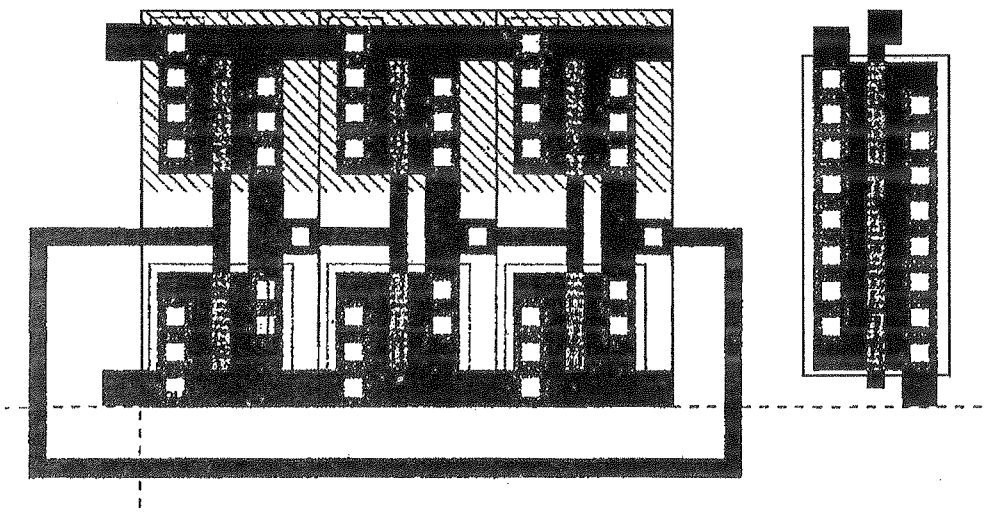
2. UZASADNIENIE POTRZEBY MODELOWANIA SPRZĘŻENIA PODŁOŻOWEGO

Aby zilustrować wpływ sprzężenia podłożowego na jakość pracy układu i uzasadnić potrzebę ekstrakcji dokładnych modeli sprzężenia podłożowego w procesie projektowania, przedstawiamy najpierw dwa przykłady prostych analogowo-cyfrowych układów.

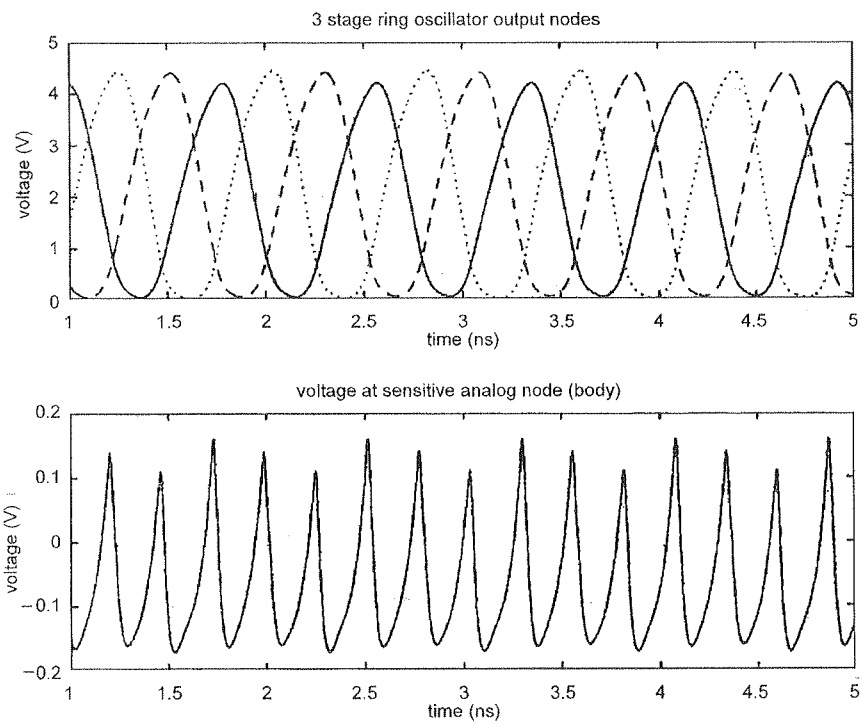
2.1. OSCYLATOR PIERŚCIENIOWY I POJEDYNCZY TRANZYSTOR „ANALOGOWY"

W pierwszym z rozważanych przypadków [10,11,31] cyfrową część układu stanowi trójstopniowy oscylator pierścieniowy a część analogową – umieszczony w pobliżu pojedynczy tranzystor reprezentujący wrażliwy analogowy wzmacniacz. Topografię analizowanego układu przedstawiono na rys. 1.

Układ został użyty do analizy wpływu szybko przełączających się końcówek części cyfrowej (tzn. wejść/wyjść poszczególnych inwerterów w oscylatorze) na tranzystor „analogowy", który został spolaryzowany tak, aby przy nieobecności szumu podłoża dostarczać prąd o stałej wartości. Zacisk podłożowy tranzystora (body terminal) został poddany monitorowaniu. Przebiegi napięcia na poszczególnych stopniach oscylatora oraz przebieg napięcia V_{body} dla tranzystora przedstawiono na rys. 2. W symulacji układu uwzględniono model sprzężenia przez podłoże [10,11,31].



Rys. 1. Topografia 3-stopniowego oscylatora pierścieniowego z umieszczonym w pobliżu tranzystorem „analogowym" [31]

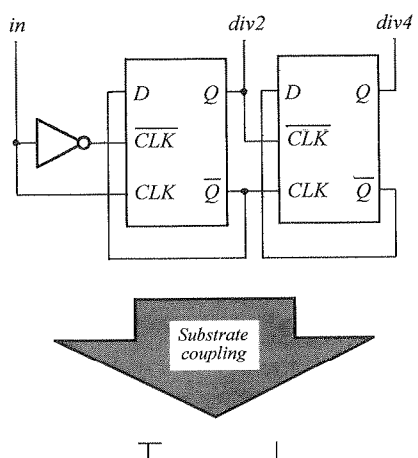


Rys. 2. Wpływ sprzężenia podłożowego na „analogowy" tranzystor znajdujący się blisko 3-stopniowego oscylatora pierścieniowego [31]

Przebiegi napięciowe dla oscylatora są takie, jakich można oczekiwać dla szybkiego oscylatora. Gdy w procesie ekstrakcji i symulacji (SPICE) nie jest uwzględnione sprzężenie podłożowe, napięcie na zacisku podłożowym „analogowego" tranzystora ma wartość stałą (0V), jeżeli jednak do listy połączeń opisującej układ wstawiony jest wyekstrahowany model sprzężenia przez podłoże, wyniki pokazują że napięcie na zacisku podłożowym oscyluje ze znaczną amplitudą (około 150 mV). Zauważmy, że piki w przebiegu V_{body} odpowiadają momentom czasu, w których końcówki wyjściowe oscylatora zmieniają swój stan najgwałtowniej. Te oscylacje przebiegu V_{body} spowodują zmiany w napięciu progowym tranzystora V_t, co może oznaczać znaczne obniżenie jakości jego pracy a w przypadku bardziej złożonego układu analogowego zmianę wartości niektórych jego parametrów. Oczywiście, stosując w procesie projektowania odpowiednie metody zabezpieczeń możemy uniknąć lub przynajmniej zminimalizować ten problem. Na tym etapie rozważań istotne jest to, że symulacje wykonane bez wyekstrahowanego modelu sprzężenia przez podłoże dadzą nieprawidłowy wynik, a mogą skłonić projektanta, by uwierzył, że jego projekt jest prawidłowy. Dla weryfikacji poprawności układu konieczne jest zatem, by do każdej symulacji włączyć modele sprzężenia przez podłoże.

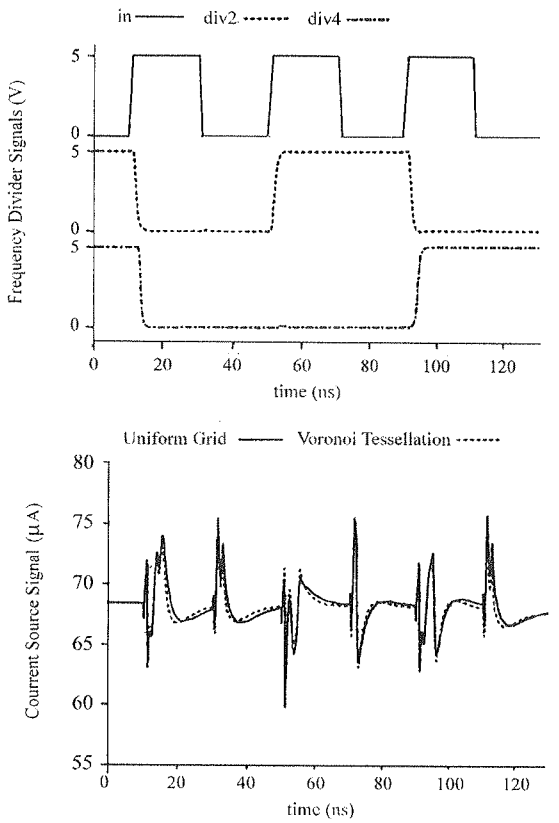
2.2. DZIELNIK CZĘSTOTLIWOŚCI I ANALOGOWE ŹRÓDŁO PRĄDOWE

Rys. 3. Przedstawia schemat dzielnika częstotliwości CMOS i analogowego źródła prądowego, które reprezentuje bliskie komórki w układzie scalonym z mieszanymi sygnałami.



Rys. 3. Cyfrowy dzielnik częstotliwości i sąsiednie analogowe źródło prądowe [10,11]

Układ był zaprojektowany w 2-mikrometrowej technologii MOSIS CMOS [10,11]. Źródło prądowe zgodnie z projektem pobiera prąd 68 μA. Fizyczna separacja między dzielnikiem i źródłem wynosi 150 μm. W symulacji układu z użyciem programu SPICE został uwzględniony model podłoża wyekstrahowany na podstawie specyfikacji layoutu oraz danych technologicznych (N, Z_j) przy pomocy specjalnego symulatora MEDICI [3], który w modelowaniu sprzężenia podłożowego bazuje na konstrukcji zwanej mozaiką Voronoi (patrz rozdział 4). Dla źródła przedstawiono również przebieg uzyskany z modelowania opartego na metodzie dokonującej dyskretyzacji podłoża z regularną prostopadłościenną siatką, opisanej w rozdziale 3. Wyniki symulacji przedstawiono na rys. 4.

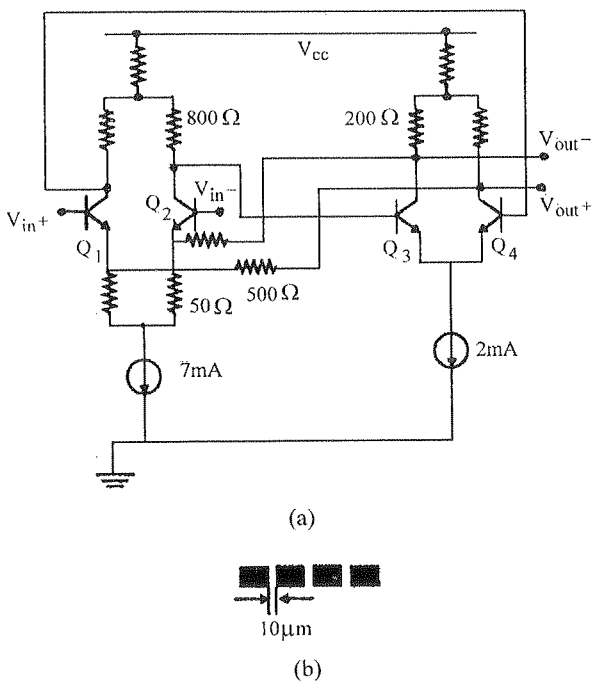


Rys. 4. Przebiegi napięciowe dla dzielnika częstotliwości z rys. 3 oraz przebieg prądu I_{src} dla sąsiedniego źródła analogowego [10,11]

Układ dzielnika powoduje przepływ prądu przez podłoże do źródła głównie podczas stanów przejściowych. W tych chwilach czasu szum jest szczególnie duży i dlatego źródło prądowe, którego wydajność przy braku sprzężenia podłożowego jest stała, w przypadku uwzględnienia sprzężenia podłożowego odchyla się od tej wartości o ponad 20% [10,11].

2.3. REDUKCJA SZEROKOŚCI PASMA W DWUSTOPNIOWYM WZMACNIACZU ZE SPRZĘŻENIEM ZWROTNYM

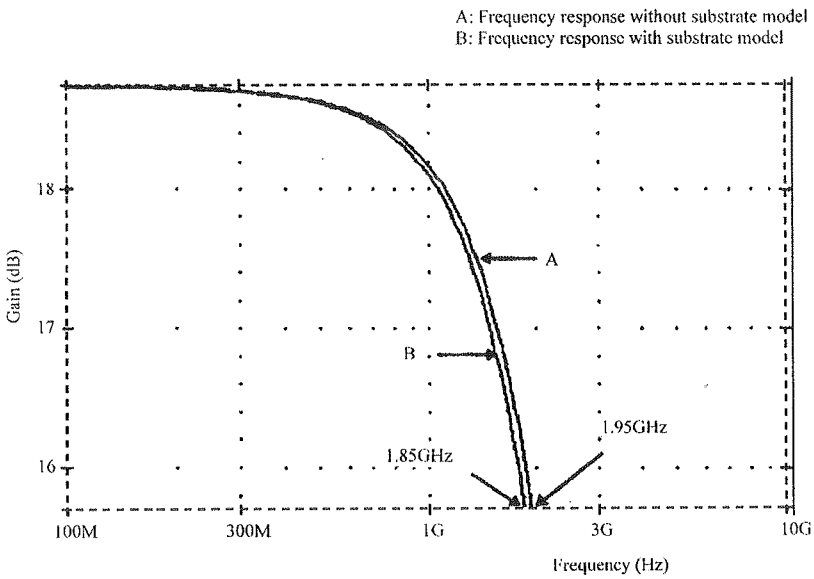
Kolejny przykład, którego celem jest uzasadnienie konieczności modelowania sprzężenia podłożowego, pokazuje wpływ skończonej rezystancji podłoża na szerokość pasma dwustopniowego wzmacniacza. Schemat układu pary różnicowej z napięciowym sprzężeniem zwrotnym jest pokazany na rys. 5a, a rys. 5b przedstawia topografię złącz kolektor-podłoże tranzystorów $Q_1 \div Q_4$.



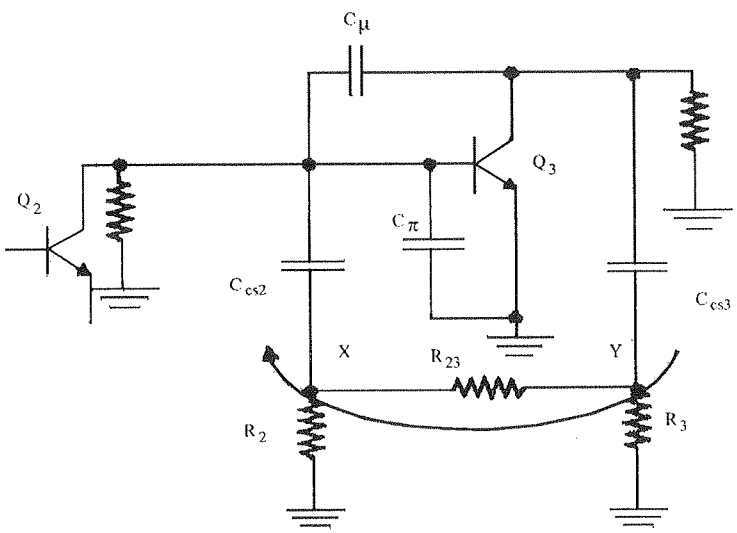
Rys. 5. Schemat układu i layout kontaktów kolektor-podłoże wzmacniacza różnicowego z napięciowym sprzężeniem zwrotnym [12]

Cztery tranzystory ($Q_1 \div Q_4$) są w projekcie topografii ułożone w rzędzie, a odległość pomiędzy sąsiednimi tranzystorami wynosi 10 μm. Charakterystyka częstotliwościowa wzmacniacza z uwzględnionym modelem sprzężenia przez podłoże i bez uwzględnienia tego modelu jest pokazana na rys. 6.

Częstotliwość trzydecybelowa wzmacniacza bez modelu sprzężenia przez podłoże wynosi 1,95 GHz. Włączenie modelu w proces symulacji układu obniża częstotliwość 3-decybelową o 100 MHz, do 1,85 GHz [12]. W przypadku pojedynczego wzmacniacza ta 5% różnica nie ma może dużego znaczenia, ale w przypadku projektów filtrów aktywnych z wykorzystaniem wielu takich wzmacniaczy odchyłka o takiej wartości może mieć już zauważalny wpływ na charakterystyki częstotliwości całego filtru. Spadek ten można



Rys. 6. Charakterystyka częstotliwościowa wzmacniacza z- (B) i bez- (A) modelu sprzężenia przez podłoże [12]



Rys. 7. Zmiennoprądowy schemat zastępczy połówki układu wzmacniacza z włączeniem modelu sprzężenia przez podłoże [12]

wyjaśnić analizując zmiennoprądowy schemat zastępczy połówki układu przedstawiony na rys. 7.

Bez modelu sprzężenia przez podłoże węzły X i Y na rys. 7 są na potencjale masy. Jeżeli model sprzężenia podłożowego zostanie włączony, to swój wkład do dominujące-

go bieguna otwartej pętli w układzie udział wnosi nie tylko pojemnościowe obciążenie kolektorów tranzystorów, na które składają się pojemności: baza-emiter (C_π) tranzystora Q_3, kolektor-podłoże (C_{CS2}) tranzystora Q_2 oraz baza-kolektor (C_μ) tranzystora Q_3 zwielokrotniona ze względu na efekt Millera przez wzmocnienie tranzystora Q_3, ale również wynikające z włączenia modelu sprzężenia przez podłoże rezystancje R_2, R_3 i R_{23}. Ich uwzględnienie prowadzi do obniżenia częstotliwości bieguna dominującego otwartej pętli, co z kolei powoduje zmniejszenie szerokości pasma wzmacniacza widoczne na rys. 5. Redukcja szerokości pasma może być zminimalizowana przez odpowiednią zmianę wartości poszczególnych elementów, co można osiągnąć przez odpowiedni wybór materiałów podłoża (np. przez zastosowanie technologii o innych wartościach rezystywności materiału podłoża) [12].

2.4. INNE PRZYKŁADY UZASADNIAJĄCE POTRZEBĘ MODELOWANIA SPRZĘŻENIA PODŁOŻOWEGO

Istnieją publikacje, w których autorzy opisują specjalne struktury eksperymentalne zaprojektowane w celu zweryfikowania wprowadzonych metod modelowania sprzężenia podłożowego, zarówno dla technologii CMOS [np. 4,7,10-12,16,31,32], bipolarnej [18,19,28], jak i Bi-CMOS [22]. Wielu autorów prezentuje również wyniki symulacji oraz wyniki pomiarów projektów analogowo-cyfrowych układów scalonych, dla których na etapie ekstrakcji i symulacji układu uwzględniono modele sprzężenia przez podłoże. Na początku lat 90, analizowano wpływ sprzężenia podłożowego na jakość pracy przetworników sigma-delta [1], czy też innych układów cyfrowych przetwarzania sygnałów [np. 2]. Szczególne znaczenie ma wpływ zakłóceń wywołanych sprzężeniem podłożowym na pracę pętli synchronizacji fazowej PLL, co przeanalizowano szczegółowo w pracy [25], gdzie w analizie pracy pętli zastosowano model nieidealnego podłoża wygenerowany z użyciem programu MEDICI [3] aby uwidocznić efekty sprzężenia podłożowego. PLL są dobrymi obiektami do badań sprzężenia podłożowego, ponieważ zawierają zarówno składniki analogowe jak i cyfrowe pracujące na stosunkowo wysokich częstotliwościach. Miarą wydajności układu PLL jest tzw. *jitter* – fluktuacja odpowiadająca odchyleniu okresu sygnału wejściowego od przebiegu idealnego. Częstotliwość systemów zegarowych jest ograniczona przez *jitter* generatora zegara, który decyduje o wielkości błędu okresu drgań. Przez symulacje oraz testy wykonane na zrealizowanym układzie [25] autorzy uznali, że główną przyczyną drgań *jitter* jest właśnie sprzężenie podłożowe. W pracy [25] można znaleźć szczegółowe wyjaśnienie mechanizmu sprzężenia podłożowego oraz rozważania dotyczące metod redukcji zakłóceń podłożowych w prezentowanym układzie.

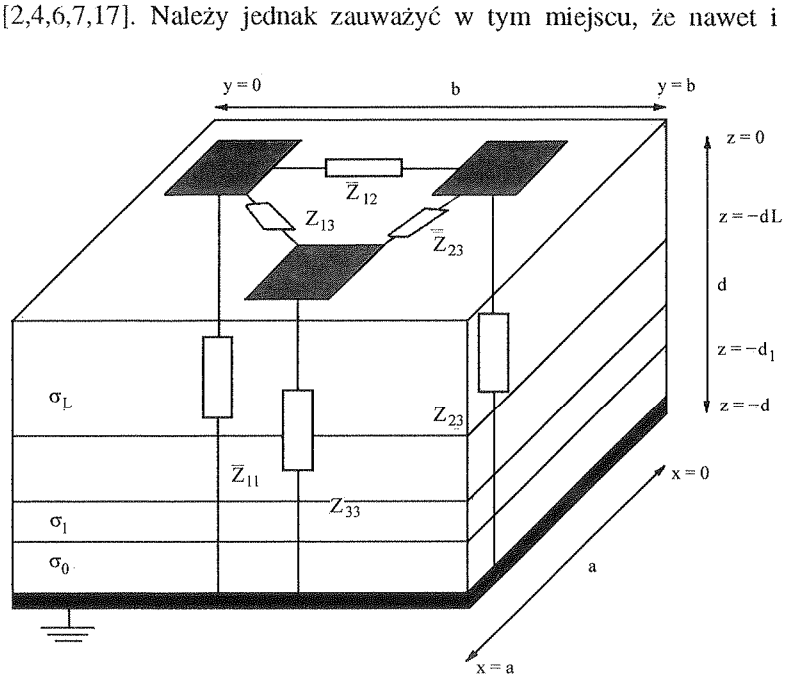
Szczególne znaczenie ma modelowanie zakłóceń wywołanych sprzężeniem podłożowym dla układów stosowanych w telefonii komórkowej [np. 29]. Największe przeszkody, aby scalić składniki telefonu komórkowego występują w kanale odbiornika, ponieważ coraz to więcej kanałów jest pakowanych w ten sam zakres częstotliwości i konieczne staje się stosowanie metod projektowych eliminujących sygnały zakłócające.

W telefonii komórkowej jest to szczególnie trudne z powodu wymaganego dużego zakresu dynamiki [26]. Telefon komórkowy musi czysto odbierać sygnał ze stacji bazowej odległej o mile, podczas gdy z sąsiednich kanałów wciskają się sygnały zakłócające. Jednym ze źródeł tych zakłóceń jest sprzężenie podłożowe. Wspólne podłoże może działać jako stałoprądowa i małosygnałowa ścieżka sprzężenia zwrotnego powodując wahania w stałoprądowych warunkach działania i parametrach małosygnałowych takich jak wzmocnienie i szerokość pasma [16]. W torze odbiornika sprzężenie podłożowe może spowodować przedostanie się sygnału LO (lokalnego oscylatora) do wejścia nisko-szumowego wzmacniacza LNA (tzw. „przeciek LO"), gdzie on zostaje wzmocniony i następnie zmieszany ze sobą („mieszanie własne") dając w wyniku przesunięcia DC. Podobny efekt może pojawić się, jeżeli duży sygnał zakłócający przenika z LNA lub wejścia mieszacza i jest pomnożony przez siebie [23]. Oprócz tego, strata sygnału w podłożu ma duże znaczenie w procesie realizacji cewek o wysokiej dobroci, które projektujemy w strukturze układu scalonego [21]. Projektowanie mieszanych (analogowo-cyfrowych) bezprzewodowych układów scalonych do zastosowań telefonii komórkowej jest bardzo trudnym problemem ze względu na fakt, że sygnały odbierane przez układ są często w zakresie zaledwie kilkudziesięciu mikrowoltów. Przy mikrowoltowych poziomach sygnału układy odbiorników radiowych są już wrażliwe na podstawowy szum przyrządów półprzewodnikowych i jeszcze bardziej wrażliwe na zakłócenia spowodowane sprzężeniem podłożowym. W pracy [29] można znaleźć kilka zaleceń projektowych, które muszą być wzięte pod uwagę przy realizacji radiowych układów scalonych (zawierających zarówno bloki analogowe jak i cyfrowe). Możliwość symulacji komputerowych z uwzględnieniem modeli sprzężenia podłożowego jest dla tej klasy układów w decyzjach projektowych widoczną pomocą. Oczywiście, aby na etapie symulacji projektu móc w stanie przewidzieć jego zachowanie się po ewentualnym wykonaniu, oprócz modelowania sprzężenia podłożowego należy również opracować i zaimplementować modele innych składników układu scalonego, w tym szczególnie model jego obudowy.

W pracy [17] opisano natomiast modelowanie zakłóceń wywołanych sprzężeniem podłożowym w odniesieniu do weryfikacji projektu potrójnego ośmiobitowego potokowego analogowo-cyfrowego konwertera video z dużymi prędkościami przetwarzania sygnału video. Układ ten jest produktem firmy *Texas Instruments* i został nazwany TVP57OO. TVP5700 konwertuje złożone sygnały analogowe luminancji i chrominancji (BY i RY) na 8 cyfrowych sygnałów luminancji oraz 4 multipleksowane cyfrowe sygnały chrominancji. Na podstawie symulacji i pomiarów autorzy projektu uznali, że udział zakłóceń spowodowanych w projekcie sprzężeniem podłożowym był istotny, gdyż w jego pierwszej wersji zmiana napięć progowych V_t tranzystorów zmieniała poziom wyzwalania komparatorów, co pogarszało wymaganą dokładność analogowo-cyfrową. Symulacje projektu z uwzględnieniem modelu sprzężenia podłożowego autorzy uznali za konieczny element procesu projektowego mający w konsekwencji zapewnić poprawną pracę układu.

3. METODA MODELOWANIA PODŁOŻA OPARTA NA ELEKTROSTATYCZNEJ FUNKCJI GREENA

Wiele narzędzi projektowych (np. SCA – *Substrate Coupling Analysis*, wchodzący w skład pakietu *Cadence*) podczas ekstrakcji pasożytniczych wartości elementów podłoża układu scalonego dokonuje dyskretyzacji podłoża za pomocą metody używającej elektrostatycznej funkcji Greena. Dyskretyzacja jest dokonywana na podstawie danych geometrycznych layoutu oraz danych technologicznych specyficznych dla konkretnego procesu. Istnieje wiele publikacji, w których jest zaprezentowana mniej lub bardziej szczegółowo sama metoda [np. 7,12,16,24,27,29,31], jak i też rozmaite algorytmy mające na celu poprawienie jej zbieżności [np. 17,27,31]. W metodzie tej podłoże jest traktowane jako dielektryk i scharakteryzowane przez poszczególne warstwy o różnych wartościach przenikalności elektrycznej ε_k (lub konduktywności σ_k), gdzie k jest numerem warstwy [rys. 8]. Analityczny model podłoża bazuje na uproszczonej formie równania Maxwella odniesionego do poszczególnych warstw, na jakie dyskretyzowane jest podłoże (w kierunku z) oraz na warunkach granicznych dla składowych wektora natężenia pola elektrycznego, jakie muszą być spełnione na granicy tych warstw. W ogólności metoda daje w wyniku impedacyjny (RC) model sprzężenia podłożowego, jakkolwiek w wielu zastosowaniach praktycznych (tzn. w symulatorach działających w oparciu o tę metodę) generowany jest jedynie model rezystancyjny sprzężenia podłożowego, co zdaniem niemal wszystkich projektantów jest wystarczające przy częstotliwości pracy układów poniżej kilku GHz [2,4,6,7,17]. Należy jednak zauważyć w tym miejscu, że nawet i w takim



Rys. 8. Przekrój poprzeczny podłoża pokazujący trójwymiarowy model jako homogeniczny wielowarstwowy system z kontaktami na górze podłoża [31]

przypadku modelowana jest sieć RC, ponieważ niezależnie od wygenerowania modelu rezystancyjnego, w ekstrakcji i symulacji są uwzględniane pewne pojemności, np. pojemności pomiędzy poszczególnymi końcówkami tranzystora a podłożem.

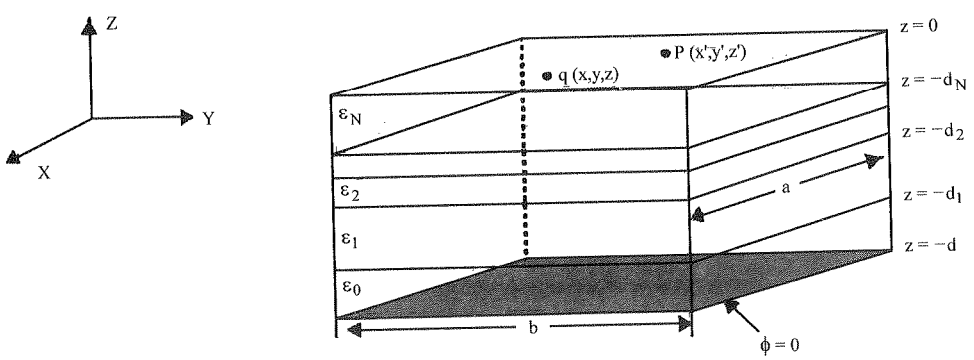
Zakładając aproksymację elektrostatyczną, podłoże układu scalonego może być modelowane jako środowisko skomponowane z poszczególnych homogenicznych warstw scharakteryzowanych przez ich konduktywność (lub przenikalność dielektryczną), jak pokazano na rys. 8.

Na górze tego stosu warstw definiowana jest liczba kontaktów (inaczej: portów), które odpowiadają powierzchni, gdzie zaprojektowany układ oddziałuje z podłożem. Kontakty są zatem zdefiniowane jako złącza element-podłoże. Przykłady kontaktów obejmują możliwe źródła szumu lub receptory, takie jak kontakty z podłoża lub wysp do linii zasilania, powierzchnie drenu, źródła i kanału tranzystorów itp. Rys. 8 pokazuje przekrój poprzeczny podłoża jako wielowarstwowe środowisko z kontaktami zdefiniowanymi na górze powierzchni i przykład impedancyjnego modelu sprzężenia. Kontakty na górze powierzchni są zazwyczaj przyjmowane jako płaskie, tzn. dwuwymiarowe. Spód podłoża jest również dołączony do ustalonego napięcia (zazwyczaj masy) przez pewien duży kontakt lub zostawiony swobodnie. Niekiedy stosowane są podłoża z płaszczyzną *backplane*, która dostarcza lepszej izolacji podłoża, jednakże technologie z warstwą *backplane* należą do stosunkowo drogich [4]. Niekiedy również jako wyróżniona warstwa modelowany jest klej stosowany do połączenia podłoża z warstwą *backplane* [31]. W każdej zdefiniowanej warstwie potencjał Φ spełnia równanie Laplace'a:

$$\nabla(\sigma \cdot \nabla\Phi) = 0 \tag{1}$$

W celu dokonania analitycznego opisu metody na rys. 9 na górnej powierzchni struktury zdefiniowano pewien punkt obserwacyjny q(x,y,z) oraz punkt źródłowy P(x',y',z').

Funkcja Greena dla środowiska podłoża G(x,y,z,x',y',z') jest zdefiniowana jako potencjał w pewnym punkcie obserwacji q(x,y,z) wywołany przez ładunek punktowy umieszczony w punkcie źródłowym P(x,y,z). Tak zdefiniowana funkcja Greena spełnia warunki graniczne i powierzchniowe w podłożu. Zgodnie z teorią Greena, potencjał Φ



Rys. 9. Poprzeczny przekrój podłoża pomocny w sformułowaniu funkcji Greena [16]

w dowolnym punkcie q(x,y,z) spodowowany przez dowolny rozkład ładunku o gęstości ρ(x',y',z') może być obliczony z funkcji Greena przez splot rozkładu ładunku z funkcją Green'a jak pokazano poniżej:

$$\Phi(x,\ y,\ z) = \int_{V'} \varrho(x',\ y',\ z') \cdot G(x,\ y,\ z,\ x',\ y',\ z')\, dv' \tag{2}$$

gdzie V' jest objętością rozkładu ładunku.
Wszystkie punkty źródłowe i obserwacyjne znajdują się przy kontaktach podłoża, które są na górze podłoża (z = z' = 0) i zostały przyjęte jako dwuwymiarowe, stąd też całkę objętościową (2) sprowadzamy do całki powierzchniowej po sumie powierzchni kontaktów podłoża S:

$$\Phi(x,\ y) = \int_{S} \varrho_S(x',\ y') \cdot G(x,\ y,\ x',\ y')\, ds \tag{3}$$

gdzie ρ_S jest gęstością powierzchniową prądu.
Wzór opisujący funkcję Greena dla podłoża jest dany jako [np. 12,13,16,31]:

$$G = \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} f_{mn} \cdot C_{mn} \cdot \cos(\delta \cdot x) \cdot \cos(\delta \cdot x') \cdot \cos(\xi \cdot y) \cdot \cos(\xi \cdot y') \tag{4}$$

gdzie $\delta = (m \cdot \pi)/a$, $\xi = (n \cdot \pi)/b$.
a i b są bocznymi wymiarami podłoża a funkcje f_{mn} są określone jako:

$$f_{mn} = \frac{1}{a \cdot b \cdot \gamma_{mn} \cdot \varepsilon_N} \times \frac{\beta_N \cdot \tanh(\gamma_{mn} \cdot d) + \Gamma_N}{\beta_N + \Gamma_N \cdot \tanh(\gamma_{mn} \cdot d)} \tag{5}$$

oraz $C_{mn} = 4$ dla $(m, n) > 0$, $C_{mn} = 2$ dla $m = 0$ lub $n = 0$ i $C_{mn} = 0$ dla $m = n = 0$.
Współczynniki γ_{mn} są określone jako:

$$\gamma_{mn} = \sqrt{\left(\frac{m \cdot \pi}{a}\right) + \left(\frac{n \cdot \pi}{b}\right)^2} \tag{6}$$

β_N i Γ_N są wyprowadzone z następującego równania macierzowego:

$$\begin{bmatrix} \beta_k \\ \Gamma_k \end{bmatrix} = \begin{bmatrix} (\varepsilon_{k-1}/\varepsilon_k) - \theta_k^2 & (\varepsilon_{k-1}/\varepsilon_k - 1) \cdot \theta_k \\ (1 - \varepsilon_{k-1}/\varepsilon_k) \cdot \theta_k & 1 - (\varepsilon_{k-1}/\varepsilon_k) \cdot \theta_k^2 \end{bmatrix} \cdot \begin{bmatrix} \beta_{k-1} \\ \Gamma_{k-1} \end{bmatrix} \tag{7}$$

gdzie $\theta_k = \tanh(\gamma_{mn} \times (d - d_k))$, $\beta_0 = 1.0$, $\Gamma_0 = 0$ i $k \in [1,N]$.
Przy powierzchni, dla $m = n = 0$, mamy:

$$G = (1/a \cdot b \cdot \varepsilon_N) \times (\Gamma_N/\beta_N) \tag{8}$$

gdzie:

$$\begin{bmatrix}\beta_k \\ \Gamma_k\end{bmatrix} = \begin{bmatrix}(\varepsilon_{k-1}/\varepsilon_k) & 0 \\ (\varepsilon_{k-1}/\varepsilon_k)\cdot d_k & 1\end{bmatrix} \cdot \begin{bmatrix}\beta_{k-1} \\ \Gamma_{k-1}\end{bmatrix} \tag{9}$$

Gdy funkcja Greena jest już określona, zależność (3) może być użyta, aby dla zbioru N kontaktów wyznaczyć kwadratową macierz o wymiarach N×N wiążącą wektor potencjału kontaktu [Φ] z wektorem ładunku [Q]. (Na każdym i-tym kontakcie rozprowadzony jest równomiernie ładunek jednostkowy a średni potencjał $\Phi_{śr}$ na wszystkich pozostałych kontaktach i na i-tym kontakcie jest obliczany ze wzoru (3); ponieważ krok ten jest powtarzany dla wszystkich par kontaktów, otrzymujemy macierz kwadratową o wymiarach N×N). Macierz ta jest nazywana macierzą współczynników potencjału [P] poprzez analogię do macierzy [Z] w opisie prądowo-napięciowym. Odwrócenie macierzy [P] daje macierz, która wiąże wektor ładunku kontaktu z wektorem potencjału kontaktu i jest znana jako macierz współczynników indukcji [c], która jest podobna do macierzy [Y].

Obliczenie macierzy współczynników potencjału [P] jest pierwszą częścią procesu ekstrakcji. Druga część polega na odwróceniu macierzy [P], aby wygenerować macierz [c]. Jak wspomniano wyżej, macierz [P] ustala zależność średnich potencjałów na kontaktach od całkowitego ładunku na nich (zmiany potencjału na kontaktach wywołane zmianą ładunku w pewnym węźle są uśrednione) a macierz [c] jest macierzą zależności ładunków na różnych węzłach od średnich potencjałów zdefiniowanych w tych węzłach. Z opisu ładunkowo-potencjałowego układu możemy przejść na tradycyjny opis napięciowo-prądowy, tzn.: dla zbioru kontaktów szukamy modelu, który wiąże prądy I_C na tych kontaktach z ich rozkładem napięć V_C:

$$[Q] = [c]\cdot[\Phi] = [P]^{-1}\cdot[\Phi]\ [I_C] = [Y_C]\cdot[V_C] = [I_C] = [Z_C]^{-1}\cdot[V_C] \tag{10}$$

gdzie $[Y_C]$ jest macierzą admitancyjną systemu. Model opisany przez zależność (10) jest siecią admitancyjną, w której kontakty są węzłami sieci a pozycja (i, j) w macierzy $[Y_C]$ reprezentuje admitancję między węzłami „i" i „j". Włączenie takiego modelu w standardowy symulator układu, taki jak SPICE lub SPECTRA jest zadaniem bardzo prostym.

Dotychczasowe rozważania zakładały, że kontakt (port) posiada stałą gęstość rozkładu ładunku (stałą gęstość prądu, który przez niego płynie). Założenie takie jest w istocie błędne i prowadzi do rezultatów, które mogą różnić się o 10% od rzeczywistego rozwiązania. Wymagana dokładność czyni koniecznym, aby kontakt dyskretyzować na zbiór paneli, gdzie liczba paneli jest dużo większa od liczby kontaktów. Zbiór równań wiążących prądy i potencjały na wszystkich panelach może być wyrażony jako [31]:

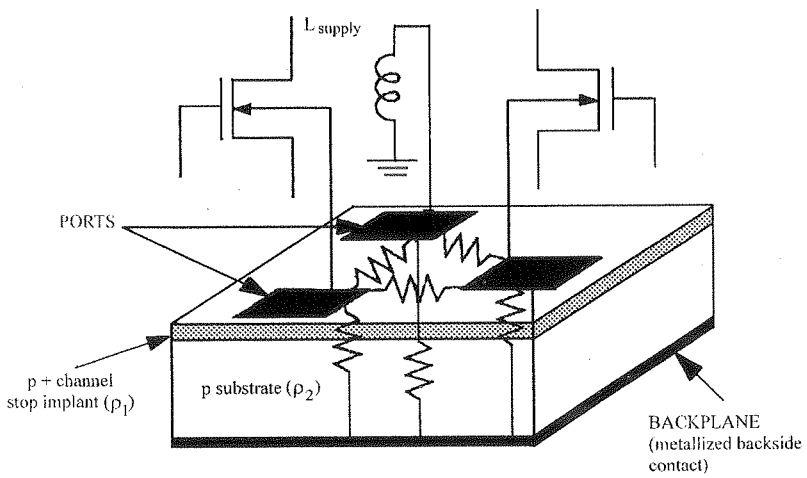
$$\Phi_p = z_p \cdot I_p \tag{11}$$

gdzie $[z_p]$ jest macierzą impedencyjną zbioru paneli, w której każda pozycja jest określona jako:

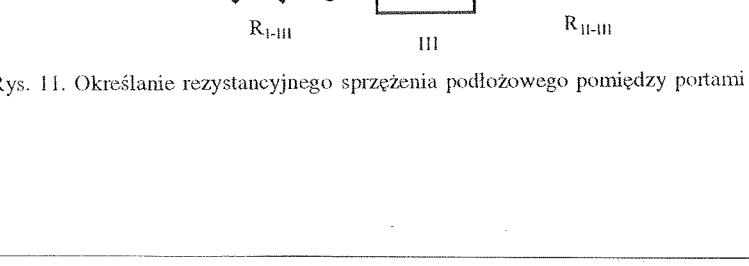
$$Z_{Pij} = \frac{1}{S_i \cdot S_j} \int_{S_i} \int_{S_j} G(x, y, x', y')da' da \tag{12}$$

gdzie S_i S_j są odpowiednio powierzchniami paneli „i" i „j".

Więcej informacji dotyczących rozwiązania oraz algorytmy poprawy zbieżności można znaleźć w pracach [16, 31] a szczegółową analizę matematyczną w rozprawie doktorskiej [33] natomiast na rys. 10 i 11 zilustrowano problem z trzema kontaktami (ograniczając się do jedynie rezystancyjnego modelu sprzężenia przez podłoże – większość narzędzi projektowych, np. SCA wersja 4.4.3 generuje właśnie model rezystancyjny, a nie impedancyjny), z których jeden został podzielony na dwa panele, a zależności (13) i (14) pokazują, jak są obliczane wartości rezystancji w modelu sprzężenia.



Rys. 10. Rezystancyjnie połączone ze sobą porty podłoża [29]



Rys. 11. Określanie rezystancyjnego sprzężenia podłożowego pomiędzy portami [29]

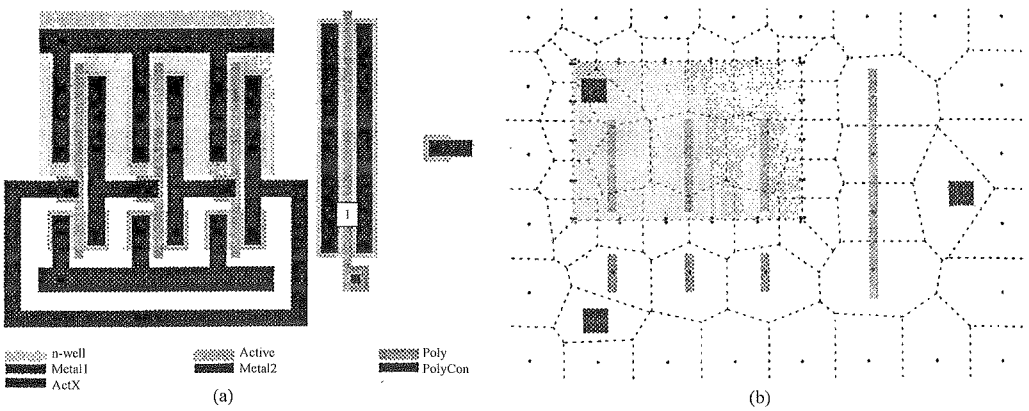
$$\begin{bmatrix} y_{11} & y_{12} & y_{13} & y_{14} \\ y_{12} & y_{22} & y_{23} & y_{24} \\ y_{13} & y_{23} & y_{33} & y_{34} \\ y_{14} & y_{24} & y_{34} & y_{44} \end{bmatrix} = \begin{bmatrix} z_{11} & z_{12} & z_{13} & z_{14} \\ z_{12} & z_{22} & z_{23} & z_{24} \\ z_{13} & z_{23} & z_{33} & z_{34} \\ z_{14} & z_{24} & z_{34} & z_{44} \end{bmatrix} \tag{13}$$

$$\begin{aligned} R_{I-II} &= \frac{1}{y_{13}+y_{23}} \\ R_{II-III} &= \frac{1}{y_{34}} \\ R_{I-III} &= \frac{1}{y_{14}+y_{24}} \\ R_{I-0} &= \frac{1}{y_{11}+y_{22}} \\ R_{II-0} &= \frac{1}{y_{33}} \\ R_{III-0} &= \frac{1}{y_{44}} \end{aligned} \tag{14}$$

4. MODELOWANIE PODŁOŻA Z UŻYCIEM MOZAIKI VORONOI I TRIANGULACJI DELAUNAY'A

Metoda modelowania sprzężenia podłożowego prezentowana w niniejszym rozdziale została zaproponowana przez Ivana Wemple i Andrew Yanga w roku 1995 [10] i stanowi alternatywę dla metod używających funkcji Greena, dla których znamienna jest dyskretyzacja podłoża oparta na prostopadłościennej siatce. W rozdziale tym skoncentrowano się na sposobie wyznaczania rezystancyjnego modelu sprzężenia podłożowego, natomiast w literaturze [10, 11] znaleźć można również opis wyznaczania wartości pojemności w podłożu, co daje w konsekwencji model impedancyjny. Omawiana technika modelowania bazuje na konstrukcji geometrycznej zwanej diagramem Voronoi i jej uzupełnieniem tzw. triangulacją Delaunay'a. Najpierw, na rys. 12, przedstawiony zostanie przykład ilustrujący metodę: topografia 3-stopnia oscylatora pierścieniowego i wrażliwego pojedynczego tranzystora „analogowego'' oraz otrzymana na podstawie tej topografii mozaika Voronoi. Następnie zostanie omówiony sposób generowania mozaiki.

Procedura modelowania sprzężenia podłożowego z użyciem mozaiki Voronoi wytwarza trójwymiarowy obwód węzłów między którymi umieszczone są elementy R i C, których wartości są obliczane na podstawie geometrii layoutu oraz danych technologicznych procesu. Dane layoutu definiują topologię obwodu dla każdej płaszczyzny węzłów (x,y). Wielokrotne płaszczyzny węzłów (x,y) są następnie układane na stosie i łączone, aby rozciągnąć obwód w kierunku z, to jest prostopadle do podłoża. Na górnej płaszczyźnie węzłów obwód jest podłączony elektrycznie z końcówkami elementów aktywnych.



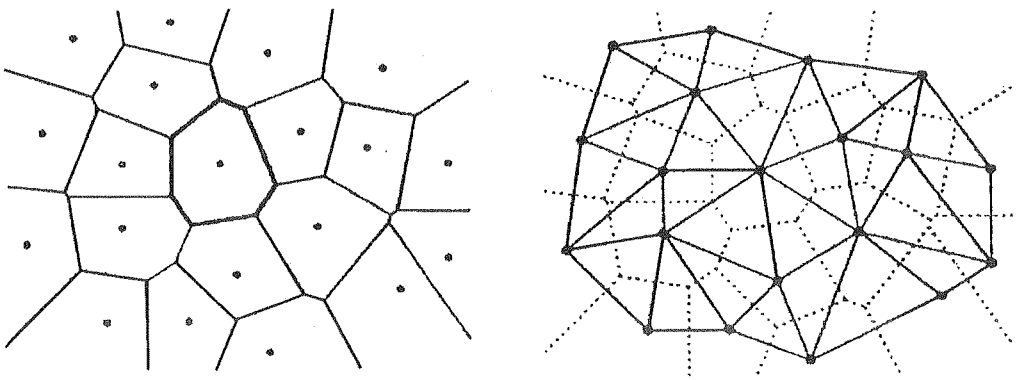
Rys. 12. Topografia prostego układu cyfrowo-analogowego (a) i odpowiadająca jej mozaika Voronoi (b) [10, 11]

Rozpatrzymy najpierw tylko górną płaszczyznę węzłów (x,y) i jej topologiczną zależność od danych layoutu (niżej leżące płaszczyzny węzłów będą kolejno osiągać taką samą topologię). Aby zamodelować sprzężenie przez podłoże należy uwzględnić w modelu geometrię i parametry technologiczne (domieszkowanie) tranzystorów, wysp, dyfuzji oraz kontaktów omowych. Obszary te są w modelu rozpatrywane przez wielokąty wyprowadzone na podstawie informacji o layoucie. Istotnym pojęciem w tej metodzie jest „miejsce" – definiowane jako każdy węzeł układu związany z terminalem podłoża lub kontaktem do podłoża. Do każdego obszaru kanału jest generowane przynajmniej jedno miejsce. Wielokąty o dużych rozmiarach i wielokąty nieregularnego kształtu są rozpatrywane przez kilka miejsc. Blisko granic wysp oraz na krawędziach granicy layoutu są generowane dodatkowe miejsca. Pozycje wszystkich tych miejsc są otrzymane wyłącznie z danych layoutu. Oprócz tego są generowane tzw. „miejsca wypełniające", to znaczy dodatkowe miejsca świadomie dodane w obszarach powierzchni layoutu, gdzie gęstość miejsc jest mała. (Chodzi na przykład o te powierzchnie topografii układu, które nie zawierają elementów aktywnych ale są przeznaczone do trasowania kanałów).

Położenie N miejsc może być użyte do zdefiniowania nie-prostokątnej siatki, która umożliwia sformułowanie obwodu podłoża zawierającego o wiele mniej węzłów niż inne metody modelowania [10,11]. Mozaika Voronoi przyporządkowuje każdemu miejscu wypukły wielokąt obejmujący obszar płaszczyzny, który jest bliższy danemu miejscu niż innemu miejscu.

N wielokątów dzieli zatem płaszczyznę na zbiór N wypukłych obszarów nazywany diagramem Voronoi. Każdy taki obszar nazywa się wielokątem Voronoi. Rys. 13a przedstawia diagram (mozaikę) Voronoi dla przypadkowego zbioru 20 miejsc. Teoretycznie mozaika rozciąga się do nieskończoności, ale wielokąty Voronoi dla wypukłych granicznych miejsc (tych przyległych do granic layoutu) są odcięte na granicy prostokąta obejmującego layout. Mozaika Voronoi posiada dużo unikalnych, interesujących

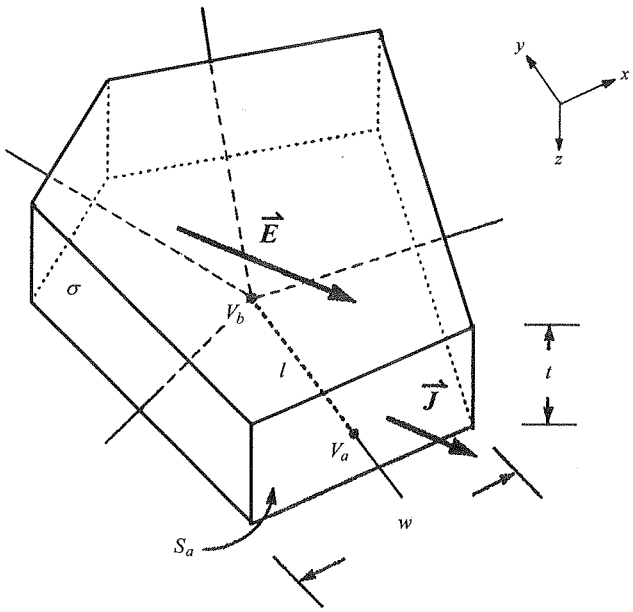
właściwości. W naszym przypadku, najważniejszą z nich jest istnienie konstrukcji będącej uzupełnieniem diagramu Voronoi, nazywanej tiangulacją Delaunay'a. Konstrukcję tę otrzymuje się przez połączenie każdej pary miejsc, które mają wspólną granicę wielokąta Voronoi. Krawędzie wielokątów Voronoi są prostopadłe do trójkątnych krawędzi Delaunay'a łączących sąsiednie miejsca. Ta właściwość okaże się użyteczna podczas formułowania elementów układu w modelu sprzężenia podłożowego, które łączą odpowiednie miejsca. Mozaika Voronoi i triangulacja Delaunay'a są z matematycznego punktu widzenia konstrukcjami dualnymi, ponieważ każda z nich może być wyprowadzona z drugiej i nie są wymagane żadne dodatkowe informacje lub obliczenia.



Rys. 13. Mozaika Voronoi (a) i odpowiadające jej triangulacje Delaunay'a dla przykładu 20 miejsc (b) [10,11]

W omawianym podejściu rozpatywane są pojedyncze wielokąty Voronoi oraz półkrawędzie trójkąta Delaunay'a i na tej podstawie obliczane są wartości elementów modelujących sprzężenie przez podłoże. W pierwszym kroku każdemu wielokątowi jest przypisana grubość płyty t. Wyznaczona w ten sposób objętość reprezentuje obszar krzemu otaczającego miejsce Voronoi. Każde miejsce staje się elektrycznym węzłem makromodelu układu. Dla każdej płyty Voronoi robimy dwa założenia a priori dotyczące jej elektrycznych właściwości. Po pierwsze: chwilowa wartość natężenia pola elektrycznego E ma wewnątrz płyty stałą wartość. Oznacza to, że dla dowolnie dużych wielokątów Voronoi wymagana dokładność modelu może nie być osiągnięta. Z tego powodu w obszarach płaszczyzny, gdzie gęstość miejsc jest mała, dodawane są dodatkowe miejsca wypełniające. Po drugie: konduktywność σ wewnątrz objętości płyty jest przyjęta jako stała. Drugie założenie ogranicza grubość płyty t, która musi być wybrana tak, aby koncentracja domieszki wewnątrz płyty mogła być bezpiecznie modelowana przez jedną średnią wartość (szczegóły dotyczące generowania dodatkowych miejsc wypełniających oraz zagadnienie doboru grubości płyty t są omówione w pracach [10,11].

Na rys. 14 przedstawiono płytę Voronoi powstałą przez przypisanie wielokątowi Voronoi grubości t oraz związane z nią krawędzie triangulacji Delaunay'a.



Rys. 14. Płyta Voronoi otaczająca miejsce mozaiki użyta do określenia wartości rezystancji w modelu sprzężenia podłożowego: segmenty triangulacji Delaunay'a są prostopadłe do krawędzi płyty [10,11]

Naszym celem jest obliczenie wartości rezystora, który modeluje przepływ prądu przez przednią płaszczyznę płyty S_a. Definiujemy potencjał V_b w miejscu Voronoi. V_a jest potencjałem w punkcie, w którym triangulacja Delaunay'a przecina krawędź płyty. Wówczas:

$$V_b - V_a = \int_a^b \vec{E} \cdot \vec{dl} \tag{15}$$

gdzie całka liniowa jest obliczana wzdłuż półkrawędzi Delaunay'a l.

Podobnie, prąd na przedniej ścianie płyty może być obliczony jako:

$$I = \int_{S_a} \vec{J} \cdot \vec{ds} \tag{16}$$

gdzie S_a jest powierzchnią przedniej ściany płyty.

Ponieważ wektor gęstości prądu jest iloczynem konduktywności σ i wektora natężenia pola elektrycznego E, to:

$$R = \frac{V_b - V_a}{I} = \frac{\int_a^b \vec{E} \cdot \vec{dl}}{\sigma \int_{S_a} \vec{E} \cdot \vec{ds}} \tag{17}$$

Krawędź triangulacji jest prostopadła do przedniej ściany płyty i wzór (17) może być uproszczony do postaci:

$$R = \frac{l}{\sigma \cdot t \cdot w} = \frac{l}{\sigma \cdot A} \tag{18}$$

Otrzymujemy w ten sposób wzór na numeryczną wartość dla gałęziowego rezystora pomiędzy miejscem i krawędzią płyty dla sąsiednich miejsc na płaszczyźnie węzłów (x,y). l jest połową długości krawędzi Delaunay'a a A jest powierzchnią przedniej ściany płyty. Wartości rezystora węzeł-węzeł (miejsce-miejsce) dla sąsiednich miejsc są więc określone przez zsumowanie wartości dwóch rezystorów połączonych z każdą krawędzią Delaunay'a. Dla modelowania rezystancji między płaszczyznami, gdy płaszczyzny węzłów są ułożone na stosie, zależność (18) daje się zastosować, przy oznaczeniu przez l odległości pomiędzy płaszczyznami a przez A – powierzchni wielokąta Voronoi.

Analogicznie, tak jak w przypadku rezystorów, triangulacja Delaunay'a może być również użyta do określenia sieci kondensatorów. Szczegółowy opis wyznaczania wartości pojemności modelujących sprzężenie podłożowe można znaleźć w [10,11].

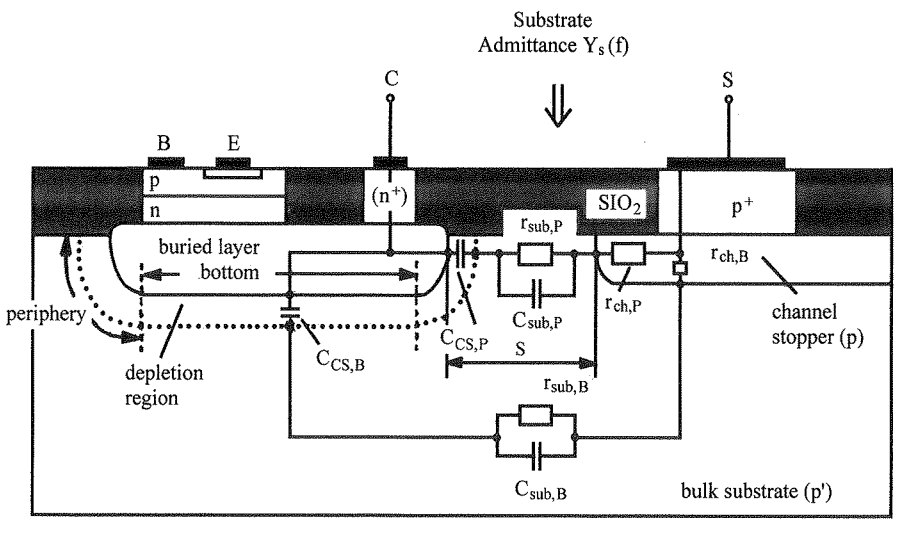
Prezentowana metoda modelowania została zaimplementowana w symulatorze o nazwie MEDICI [3] i zweryfikowana przez symulacje i pomiary zarówno dla prostych [10,11], jak i o dużej skali integracji układach scalonych [np. 25].

5. MODELOWANIE PODŁOŻA W KRZEMOWYCH UKŁADACH BIPOLARNYCH PRACUJĄCYCH W ZAKRESIE B.W.CZ. (do 40 GHz)

Dzisiejsze krzemowe bipolarne układy scalone osiągają prędkości działania, które kilka lat temu wydawały się zarezerwowane jedynie dla specjalnych drogich technologii półprzewodnikowych. Również dla tych układów istnieje konieczność opracowywania i weryfikacji odpowiednich modeli sprzężenia podłożowego [18,19,28]. Modelowanie sprzężenia podłożowego jest zazwyczaj bardziej złożone dla podłoży wysoko-rezystywnych (typu p-), które są preferowane w krzemowych technologiach bipolarnych, chociaż metody modelowania opracowanego dla tego typu podłoży (prezentowane w tym rozdziale) mogą być również zastosowane do podłoży nisko-rezystywnych [19]. W niniejszym rozdziale przedstawiono, jak pasożytnicze sieci podłoża mogą być modelowane do zakresu bardzo wysokich częstotliwości (40 GHz) przez użycie schematów zastępczych złożonych z kilku skupionych elementów. Przy tak wysokich częstotliwościach pracy, w przeciwieństwie do metod opisanych w poprzednich rozdziałach, założenie czysto rezystywnego charakteru podłoża nie jest już spełnione. W takim zakresie częstotliwości sprzężenie podłożowe musi być modelowane jako sieć RC lub RLC [18,19,28]. Zdaniem autorów pracy [28] wpływ indukcyjności połączeń w strukturze półprzewodnikowej na sprzężenie podłożowe jest raczej słaby, jednak w pewnych specyficznych warunkach (np. w przypadku struktur z ekranowaniem), indukcyjności połączeń mają znaczny udział w sprzężeniu i muszą być uwzględnione w modelach

sprzężenia podłożowego. W zakresie b.w.cz. dyskutowana jest również przez projektantów ewentualność uwzględnienia w modelowaniu zakłóceń wywołanych sprzężeniem podłożowym oddziaływań pól magnetycznych [28]. We wszystkich prezentowanych w pracy metodach oddziaływania pola magnetycznego są pominięte, gdyż ich uwzględnienie – zdaniem wielu projektantów – prowadziłoby do bardzo skomplikowanej analizy podłoża spowalniając proces ekstrakcji i symulacji i nie dając zauważalnych różnic w porównaniu z sytuacją, w której przedmiotem rozważań są jedynie oddziaływania pola elektrycznego. Jednak przy częstotliwościach gigabitowych coraz częściej pojawiają się rozbieżności między wynikami symulacji i pomiarów, wskazujące być może na konieczność kompletnej analizy pola elektrycznego i magnetycznego.

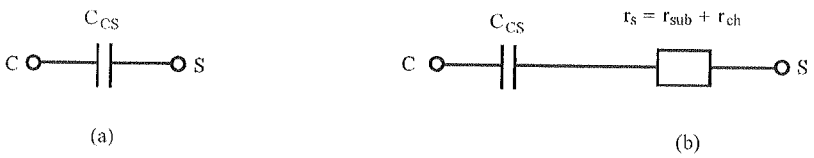
Na rys. 15 przedstawiono schematyczny przekrój poprzeczny struktury krzemowego tranzystora bipolarnego z izolacją LOCOS oraz wysokorezystywnym podłożem oraz omówiony w dalszej części schemat zastępczy dla podłoża.



Rys. 15. Schematyczny przekrój poprzeczny struktury badanego tranzystora; C jest kontaktem kolektora a S kontaktem podłoża [19]

Odległość s pomiędzy warstwą zgrzebaną (podkolektorową) i warstwą *channel stoper*, w zależności od struktury testowej, wynosi albo 3,5 μm albo w przybliżeniu 0,25 μm [28]. Dla modelowania sprzężenia podłożowego wystarczające jest, aby rozpatrzyć: warstwę zagrzebaną (n^+) mającą potencjał kolektora (C), obszar zubożony złącza kolektor-podłoże, podłoże *bulk* (p^-) warstwę *channel stoper* (p) oraz kontakt podłoża (S). Celem jest znalezienie schematu zastępczego dla całego obszaru izolującego tranzystor, który dokładnie modeluje admitancję $Y_s(f)$ pomiędzy kolektorem i kontaktem podłoża do zakresu wysokich częstotliwości. Stąd też dodatkowo oprócz pojemności złącza kolektor-podłoże musi być rozpatrzona kompletna pasożytnicza sieć podłoża. W najprostszym

schemacie zastępczym, pokazanym na rys. 16a brany jest pod uwagę jedynie obszar zubożony wokół warstwy zagrzebanej. Jest on modelowany przez skupiony kondensator złączowy C_{CS}, którego wartość może być łatwo określona na podstawie geometrii warstwy zagrzebanej i na podstawie pojemności dolnej i bocznej (obrzeżowej) złącza o określonej powierzchni i długości. Takie jednak podejście ograniczone jest do stosunkowo niskich częstotliwości, ponieważ pomija ono wpływ rezystancji podłoża r_{sub} oraz rezystancji warstwy *channel stoper* r_{ch}. Aby je rozpatrzyć, dodatkowy rezystor $r_s = r_{sub} + r_{ch}$ jest zazwyczaj włączany szeregowo do C_{CS}, jak pokazano na rys. 16b.

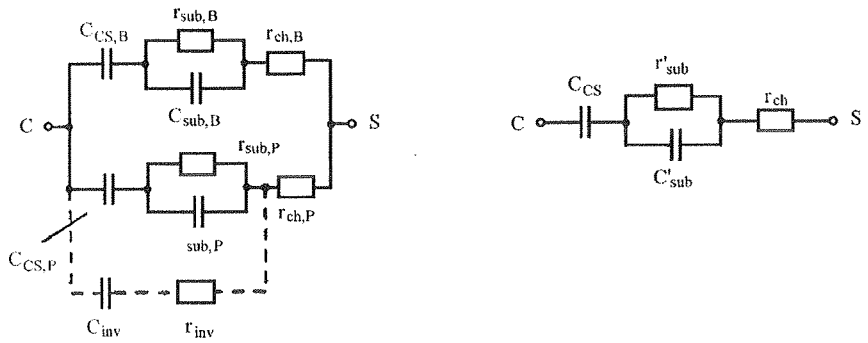


Rys. 16. Powszechnie stosowany schemat zastępczy dla modelowania sprzężenia przez podłoże [19]

Schemat z rys. 16b, zaimplementowany np. w modelu tranzystora w symulatorze SPICE, posiada dwie niedogodności. Po pierwsze: w większości przypadków nie ma łatwego sposobu dla projektanta, aby określić prawidłową wartość r_s, po drugie: pominięte jest dielektryczne zachowanie podłoża *bulk*, co daje znaczny błąd w wyniku przy bardzo dużych częstotliwościach. Z tego powodu kondensator C_{sub}, którego wartość zależy od przenikalności podłoża ε_{si} musi być zbocznikowany rezystorem podłoża r_{sub}. Faktem uwzględnianym w modelowaniu winno być również to, że obrzeże obszaru izolującego tranzystor zachowuje się inaczej niż obszar dolny. Dzieje się tak z powodu bliskości warstwy *channel stoper*, która skraca ścieżkę przepływu prądu obrzeżowego (bocznego). Dlatego nie można założyć, że potencjał przy granicy pomiędzy obszarem zubożonym i podłożem *bulk* jest taki sam zarówno dla dołu jak i dla obrzeża (części bocznej) obszaru izolującego tranzystor. W oparciu o te rozważania obszar izolujący tranzystor jest w procesie modelowania podzielony na część dolną i boczną (obrzeżową), jak zilustrowano na rys. 15. Dół jest definiowany jako obszar, który leży poniżej płaskiej części warstwy zagrzebanej, obrzeże (część boczna) jest pozostałą częścią. Oprócz tego zakładamy, że każda część może być modelowa przez pojedynczą gałąź zawierającą elementy omówione poprzednio: kondensator złączowy C_{SC}, rezystor podłoża r_{sub} zbocznikowany przez kondensator C_{sub} oraz rezystor r_{ch} warstwy *channel stoper*. Indeksy B i P na rys. 15 oznaczają odpowiednio gałąź dolną i boczną. Całkowity schemat zastępczy wstawiony do przekroju poprzecznego z rys. 15 jest pokazany na rys. 17a.

Część dolna może być modelowana przez jedynie trzy skupione elementy, ponieważ z powodu swej małej wartości $r_{ch,B}$ może być często pominięta.

Kolejne rozwinięcie proponowanego schematu zastępczego polega na uwzględnieniu potencjalnej warstwy inwersyjnej przy powierzchni krzemu. Jak pokazano na rys. 17a, jedynie pojedyncza gałąź zawierająca kondensator C_{inv} i szeregowy rezystor r_{inv}



Rys. 17. Dokładny (a) i uproszczony (b) schemat zastępczy do modelowania obszaru izolującego tranzystor; $C_{CS} = C_{CS,B} + C_{CS,P}(+C_{inv})$ [19]

musi bocznikować sieć zbudowaną przez $C_{CS,P}$, $r_{sub,P}$, i $C_{sub,P}$. C_{inv} jest spowodowane przez złącze pn pomiędzy podłożem *bulk* i częścią końcową domieszkowania warstwy *channel stoper* z jednej strony i warstwą inwersyjną z drugiej strony. Rezystancja warstwy inwersyjnej jest modelowana przez r_{inv}. Jeżeli warstwa *channel stoper* prawie dotyka warstwy zagrzebanej (małe „s'' na rys. 15), warstwa inwersyjna zanika a odpowiadająca jej przerywana gałąź na rys. 17a może być zastąpiona przez pojedynczą pojemność złączową. Reprezentuje ona obszar zubożony blisko powierzchni, który jest teraz znacznie zredukowany w szerokości. Dlatego, też pojemność ta często dominuje w całkowitej bocznej pojemności złączowej.

Chociaż schemat zastępczy, który dokładnie odzwierciedla procesy fizyczne, pomaga zrozumieć zależność sprzężenia podłożowego zarówno od geometrii jak i od technologicznych i fizycznych parametrów, dla symulacji układu (np. za pomocą SPICE) preferowane są prostsze schematy zastępcze, aby zredukować czas symulacji. Stosowane podejście jest pokazane na rys. 17b. W tym przypadku, różne gałęzie modelu z rys. 17a są połączone w pojedynczą gałąź. Oczywiście, w takim przypadku fizyczny charakter modelu jest częściowo stracony, jakkolwiek ten prosty schemat jest nadal prawidłowy w zakresie bardzo wysokich częstotliwości, nawet jeżeli istnieje warstwa inwersyjna powierzchni.

Sposoby wyznaczania wartości poszczególnych elementów zaprezentowanych schematów zastępczych są opisane szczegółowo w literaturze [18,19,28], gdzie poprzez symulacje i wyniki pomiarów specjalnych struktur testowych udowodniono ich słuszność do bardzo wysokich częstotliwości.

6. PODSUMOWANIE

W artykule przedstawiono wybrane zagadnienia związane z modelowaniem sprzężenia podłożowego w układach scalonych. Rozważany problem sprzężeń ma istotne znaczenie w mieszanych: analogowo-cyfrowych układach scalonych, zwłaszcza gdy są

one przeznaczone do pracy w zakresie częstotliwości radiowych. Zaprezentowane metody modelowania oraz próba uzasadnienia konieczności modelowania sprzężeń podłożowych przez projektantów układów scalonych stanowią jedynie swego rodzaju wstęp do omawianej tematyki. W kolejnych pracach autorzy mają nadzieję przedstawić liczne własne przykłady analogowo-cyfrowych układów scalonych, dla których przeprowadzono badania pozwalające określić, w sposób ilościowy i jakościowy, zakłócający wpływ przełączających się elementów cyfrowych układu na parametry i charakterystyki wrażliwych na to szkodliwe oddziaływanie bloków analogowych a także zaprezentować różnego rodzaju struktury zabezpieczające, których celem jest redukcja zakłóceń wywołanych w układzie przez sprzężenie podłożowe.

BIBLIOGRAFIA

1. B.P. Del Signore, D.A. Kerth, N.S. Sooch, E.J. Swanson: *A Monolithic 20-b Delta-Sigma A/D Converter.* IEEE J. Solid-State Circuits, vol. SC-25, Dec. 1990, pp. 1311–1317
2. T.J. Schmerbeck, R.A. Richetta, L.D. Smith: *A 27 MHz Mixed A/D Magnetic Recording Channel DSP Using Partial Response Signalling with Maximum Likelihood Detection,* Digest of the IEEE International Solid-State Circuits Conference, Feb. 1991, p. 136–137
3. TMA MEDICI User's Manual, Technology Modeling Associates, Inc. 1992
4. D.K. Su, M.J. Loinaz, S. Masui, B.A. Wooley: *Experimental Results and Modeling Techniques for Substrate Noise in Mixed-Signal Integrates Circuits.* IEEE J. Solid State Circuitts, vol. 28, April 1993, pp. 420–430
5. N.K. Verghese, D.J. Allstot, S. Masui: *Rapid Simulation of Substrate Coupling Effects in Mixed-Mode IC's.* Proc. IEEE Custom Integrates Circuits Conf., May 1993, pp. 18.1.1.–18.1.4
6. B.R. Stanisic, N.K. Verghese, D.J. Allstot, R.A. Rutenbar, L.R. Carley: *Addressing Substrate Coupling in Mixed-Mode ICs: Simulation and Power Distribution Synthesis.* IEEE J. Solid-State Circuits, vol. 29, Mar. 1994, pp. 226–238.
7. N.K. Verghase, T.J. Scmerbeck, D.J. Allstot: *Simulation Techniques and Solutions for Mixed-Signal Coupling in Integrated Circuits.* Norwell, MA Kluwer Academic, 1994
8. R. Gharpurey, R.G. Meyer: *Modeling and Analisis of Substrate Coupling in IC's,* Proc. IEEE Custom Integrated Circuits Conf., May 1995, pp. 125–128
9. B.R. Stanisic, R.A. Rutenbar, L.R. Carley: *Addressing Noise Decoupling in Mixed-Signal IC's: Power Distribution Design and Cell Customization.* IEEE Journal of Solid-State Circuits, vol. 30, March 1995, pp. 321–326.
10. I.L. Wemple, A.T. Yang: *Mixed-Signal Switching Noise Analysis Using Voronoi-Tesselated Substrate Macromodels.* Proc 32nd Design Automation Conference, San Francisco, CA, USA, June 1995, pp. 439–444
11. I.L. Wemple, A.T. Yang: *Integrated Circuits Substrate Coupling Models Based on Voronoi Tesselation.* IEEE Transations on Computer-Aided Design, Dec. 1995, pp. 1459–1469
12. R.G. Gharpurey, R.G. Meyer: *Analysis and Simulation of Substrate Coupling in Integrated Circuits.* Int. J. Circuits Theory and Applications, vol. 23, Aug. 1995, pp. 381–394
13. N.K. Verghese, D.J. Allstot, M. Wolfe: *Fast Parasitic Extraction for Substrate Coupling in Mixed-Signal IC's.* Proc. IEEE Custom Integrated Circuits Conf., May 1995, pp. 7.2.1–7.2.4
14. X. Aragones, F. Moll, M. Roca, A. Rubbio: *Analysis and Modeling of Parasitic Substrate Coupling in CMOS Circuits.* IEEE Proc. Circuits Devices systems, vol. 142, Oct. 1995, pp. 307–312
15. T. Smedes, N.P. van der Meijs, A.J. van Genderen: *Extraction of Circuits Models for Substrate Crosstalk.* Proc. Int. Conference on Computer-Aided Design, Nov. 1995, pp. 199–206
16. R. Gharpurey, R.G. Meyer: *Modeling and Analysis of Substrate Coupling in Integrated Circuits.* IEEE J. Solid-State circuits, vol. 31, No. 3, March 1996, pp. 344–353

17. N.K. Verghese, D.J. Allstot, M.A. Wolfe: *Verification Techniques for Substrate Coupling and their Application to Mixed-Signal IC Design.* IEEE J. Solid-State Circuits, vol. 31, No 3, March 1996, pp. 354–365
18. M. Pfost, H-N. Rain, T. Holzwarth: *Modeling of the Substrate Effect in High-Speed Si-Bipolar IC's.* Proc. IEEE 1995 BCTM, Minneapolis MN, Oct. 1995, pp. 182–185
19. M. Pfost, H-N. Rain, T. Holzwarth: *Modeling of the Substrate Effect in High-Speed Si-Bipolar IC's.* IEEE Solid-State Circuits, vol. 31, Oct. 1996, pp. 1493–1501
20. T. Blalack, J. Lau, F.J.R. Clement, B.A. Woley: *Experimental Results and Modeling of Noise Coupling in a Lightly Doped Substrate.* Proc. IEEE 1996, IEDM, Dec. 1996, pp. 623–626
21. J.R. Long, M.A. Copeland: *The Modeling, Characterization, and Design of Monolitic Inductors for Silicon RF IC's.* IEEE J. Solid-State Circuits, vol. 32, Mar. 1997, pp. 357–369
22. J.M. Casalta, X. Aragones, A. Rubbio: *Substrate Coupling Evaluation in BiCMOS Technology.* IEEE Journal of Solid-State Circuits, vol. 32, No. 4, April 1997, pp. 598–603
23. B. Razavi: *Design Considerations for Direct-Conversion Receivers.* IEEE Trans, Circuits Syst.-II, vol. 44, June 1997, pp. 428–435.
24. E. Chiprout: *Interconnect and Substrate Modeling and Analysis: An Overview.* IEEE Journal of Solid-State Circuits. vol. 33, No. 9, Sept. 1998, pp. 1445–1452
25. R.J. Welch, A.T. Yang: *Substrate Coupling Analysis and Simulation for Ana Industrial Phase-Locked Loop.* IEEE 1998
26. S. Kiaei, D.J. Allstot, K. Hansen, N.K. Verghese: *Noise Considerations for Mixed-Signal RF IC Transceivers.* ACM J. Wireless Networks, Mar./Apr. 1998
27. A.M. Niknejad, R. Gharpurey, R.G. Meyer: *Numerically Stable Green Funtion for Modeling and Analysis of Substrate Coupling in Integrated Circuits.* IEEE Trans. Computer Aided-Design of Integrated Circuits, vol. 17, April 1998, pp. 305–315
28. M. Pfost, H-M. Rein: *Modeling and Measurement of Substrate Coupling in Si-Bipolar IC's up to 40GHz.* IEEE Journal of Solid-State Circuits, vol. 33, No. 4, April 1998, pp. 582–591
29. N.K. Verghese, D.J. Allstot: *Computer-Aided Design Considerations for Mixed-Signal Coupling in RF Integrated Circuits.* IEEE of Solid-State Circuits, vol. 33, No. 3, March 1998, pp. 314–323
30. E. Charbon, P. Miliozzi, L.P. Carloni, A. Ferrari, A. Sangiovanni-Vincentelli: *Modeling Digital Substrate Noise Injection in Mixed-Signal IC's.* IEEE Transations on Computer-Aided Design of Integrated Circuits and Systems, vol. 18, No. 3, March 1999, pp. 301–310
31. J.P. Costa, M. Chou, L.M. Silveira: *Efficient techniques for Accurate Modeling and Simulation of Substrate Coupling in Mixed-Signal IC's.* Transations on Computer-Aided Design of Integrated Circuits and Systems, vol. 18, No. 5, May 1999, pp. 597–607
32. P. Pawłowski, A. Guziński: *Investigations of Substrate Noise in Mixed Analog-Digital Systems.* XXII KKTOiUE, Warszawa-Stare Jabłonki, Oct. 1999, pp. 203–208
33. R. Gharpurey: *Modeling and Analysis of Substrate coupling in Integrated Circuits.* Ph.D. dissertation, University of California, Berkeley 1995, praca dostępna w internecie http:/bwrc.eecs.berkeley.edu/Publications/Theses.htm

M. SOLECKI, Z. FELENDZER

MODELING METHODS OF SUBSTRATE COUPLING IN INTEGRATED CIRCUITS

Summary

The main subject of this paper is problems of a modeling of a substrate coupling in inegrated circuits. First, we give reasons why a modeling of the substrate coupling in design-process is necessary. A few basic mixed-signal examples have been presented to show an injurious influence of the substrate coupling noise on

an analog part of a design. These examples are a ring oscillator with a single transistor and a digital frequency divider with a current source. The substrate coupling noise is important in mixed analog-digital IC's, because the current injected to the chips substrate from a large number of simultaneously switching digital elements can cause a malfunction of sensitive analog elements of the system. Current injected to the silicon substrate influences the local potential of the substrate of analog elements thus modulates the threshold voltage V_t of the transistors and as a consequence of that, the sensitive analog modules do not operate properly. In addition, a few industrial designs like a phase-locked loop, video converter or RF-circuits have been considered to indicate the importance of substrate coupling problem.

In the next part of this paper we focus on two alternative modeling methods of substrate coupling. Thase are: the Green function method and the method based on geometrical construction called Voronoi Tessellation. (The latter was showed the first time by I. Wemple and A. Yang in 1995). These methods perform IC substrate discretization basing on the geometric layout of the circuit and technology parameters specific for the given process. In Green function method the substrate is treated as layers of dielectrics characterized by varying electrical permitivity ε_k, where k is the index of the layer. An analytic model of the substrate is based on a simplified form of Maxwell equations referring to consecutive substrate layers and on boundary conditions and finally on the electric field intensity vector component, which have to be fulfilled at layer boundaries. In the method based on Voronoi Tessellation the substrate of IC is discretisized on irregular volumes with constant values of the conductivity and the electric field intensity. Each of these methods can provide designers with a parasitic model of the substrate.

In the final part, we discuss some problems of a modeling of the substrate coupling in Si-Bipolar ICs up tp very high frequencies (40GHz).

Key words: modeling methods, substrate coupling, CMOS and bipolar integrated circuits, Green function, Voronoi Tessellation.

KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJI, 2000, 46, z. 4, ss. 615–632

Wielkosygnałowy dynamiczny makromodel tranzystora Darlingtona mocy do analizy elektrotermicznych stanów przejściowych

JANUSZ ZARĘBSKI, KRZYSZTOF GÓRECKI

Wyższa Szkoła Morska w Gdyni, Katedra Radioelektroniki Morskiej

Otrzymano 2000.08.27
Autoryzowano 2000.10.12

W pracy zaproponowano wielkosygnałowy dynamiczny elektrotermiczny makromodel tranzystora Darlingtona mocy oraz algorytm wyznaczania wartości parametrów elektrycznych tego makromodelu. Zaproponowany w pracy makromodel sformułowano na bazie uprzednio opracowanego makromodelu stałoprądowego, uzupełnionego o elementy modelujące inercję elektryczną i termiczną. Zgodnie z zaproponowanym w pracy algorytmem estymacji wartości parametrów makromodelu, wartości te są obliczane na podstawie współrzędnych punktów pomiarowych, leżących na odpowiednich odcinkach izometrycznych charakterystyk dynamicznych badanego tranzystora, dla których słuszny jest uproszczony model lokalny, opisany za pomocą jednego lub kilku parametrów makromodelu. Rozważania teoretyczne zilustrowano zweryfikowanymi doświadczalnie wynikami symulacji nieizometrycznych charakterystyk dynamicznych prostego układu elektronicznego zawierającego tranzystor Darlingtona mocy.

Słowa kluczowe: tranzystor Darlingtona mocy, modelowanie elementów elektronicznych, estymacja wartości parametrów

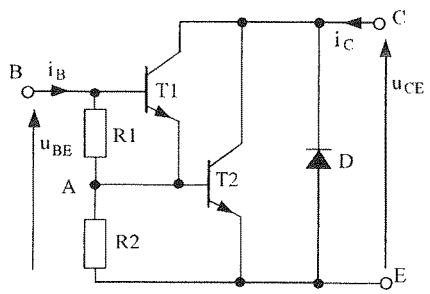
1. WSTĘP

Problem modelowania elementów półprzewodnikowych jest zagadnieniem trudnym i na ogół czasochłonnym, obejmującym m.in. wybór i sposób opisu wybranych zjawisk fizycznych, estymację wartości parametrów, implementację komputerową oraz weryfikację modelu. Podstawowym problemem jest jednak sformułowanie modelu zapewniającego jak najlepszą zgodność wyników pomiarów i symulacji przy możliwie najmniejszym stopniu komplikacji modelu.

Szybko postępująca miniaturyzacja struktur, wzrost gęstości upakowania elementów w strukturze półprzewodnikowej i dążenie do wydzielania dużych mocy powodują, że fundamentalnego znaczenia nabierają efekty termiczne wynikające z samonagrzewania

i wzajemnych oddziaływań termicznych między elementami składowymi układu scalonego. Jednoczesne uwzględnienie w analizach komputerowych zjawisk elektrycznych i termicznych wymaga zastosowania specjalnych modeli elektrotermicznych, których interpretacją graficzną są charakterystyki nieizometryczne. Każdemu punktowi, leżącemu na takiej charakterystyce odpowiada inna wartość temperatury wnętrza, w przeciwieństwie do charakterystyk izometrycznych, na których wszystkim punktom odpowiada jednakowa wartość temperatury. Aby umożliwić praktyczną realizację obliczeń wybranego układu elektronicznego z uwzględnieniem obu wymienionych wyżej zjawisk termicznych, rozważane modele powinny być opracowane w formie akceptowalnej przez uznane programy do analizy układów elektronicznych, np. program SPICE. Analizy komputerowe układów elektronicznych z wykorzystaniem modeli elektrotermicznych nazywane są analizami elektrotermicznymi, natomiast analizy, w których nie uwzględnia się zjawisk termicznych nazywane są analizami izotermicznymi.

Współczesne przyrządy półprzewodnikowe mocy to często struktury złożone z wielu elementów składowych. Jednym z takich przyrządów jest tranzystor Darlingtona mocy, stanowiący połączenie dwóch tranzystorów bipolarnych T1 i T2, dwóch rezystorów R1 i R2 oraz diody antyrównoległej D (rys. 1). Rozważany tranzystor Darling-



Rys. 1. Struktura tranzystora Darlingtona mocy

tona typowo pracuje w warunkach wydzielania dużych mocy elektrycznych, a zatem należy oczekiwać, że efekty termiczne będą miały istotny wpływ na przebieg statycznych i dynamicznych charakterystyk rozważanego przyrządu półprzewodnikowego, przy czym pod pojęciem charakterystyk dynamicznych autorzy rozumieją zarówno charakterystyki małosygnałowe jak i czasowe wielkosygnałowe przebiegi napięć i prądów zaciskowych. A zatem, do opisu elementów składowych oraz całego tranzystora Darlingtona rozumianego jako układ scalony, powinno się stosować odpowiednio modele oraz makromodele elektrotermiczne.

Niniejsza praca stanowi kontynuację prac [1, 2, 3], w których autorzy zaproponowali stałoprądowy makromodel tranzystora Darlingtona oraz procedurę estymacji wartości parametrów tego makromodelu. Celem pracy jest zaprezentowanie opracowanego przez autorów, na bazie wersji stałoprądowej, wielkosygnałowego dynamicznego elektrotermicznego makromodelu tranzystora Darlingtona (DEMTD) oraz algorytmu estymacji war-

tości jego parametrów elektrycznych. Opracowanie algorytmu estymacyjnego nie było zadaniem trywialnym, ponieważ nie wszystkie końcówki elementów składowych są dostępne dla użytkownika tranzystora Darlingtona, a zatem nie można dokonać bezpośredniego pomiaru wartości prądów płynących przez te końcówki. Dostępne są jedynie wartości prądów i napięć na zaciskach całego tranzystora Darlingtona (rys. 1). Tak więc, wyznaczenie wartości parametrów elektrycznych poszczególnych elementów składowych wymaga odpowiedniego doboru warunków polaryzacji tranzystora Darlingtona, umożliwiających określenie rozpływu prądów w strukturze i lokalną estymację wartości parametrów oraz wymaga wyprowadzenia końcówki bazy tranzystora T2 (punkt A na rys. 1).

W rozdziale drugim przedstawiono sposób konstrukcji dynamicznego elektrotermicznego makromodelu tranzystora Darlingtona dla programu SPICE, w rozdziale trzecim zaprezentowano algorytm estymacji wartości parametrów opracowanego makromodelu ESDYN, natomiast w rozdziale czwartym przedstawiono wyniki weryfikacji eksperymentalnej opracowanego makromodelu i algorytmu estymacji jego parametrów.

2. POSTAĆ MAKROMODELU

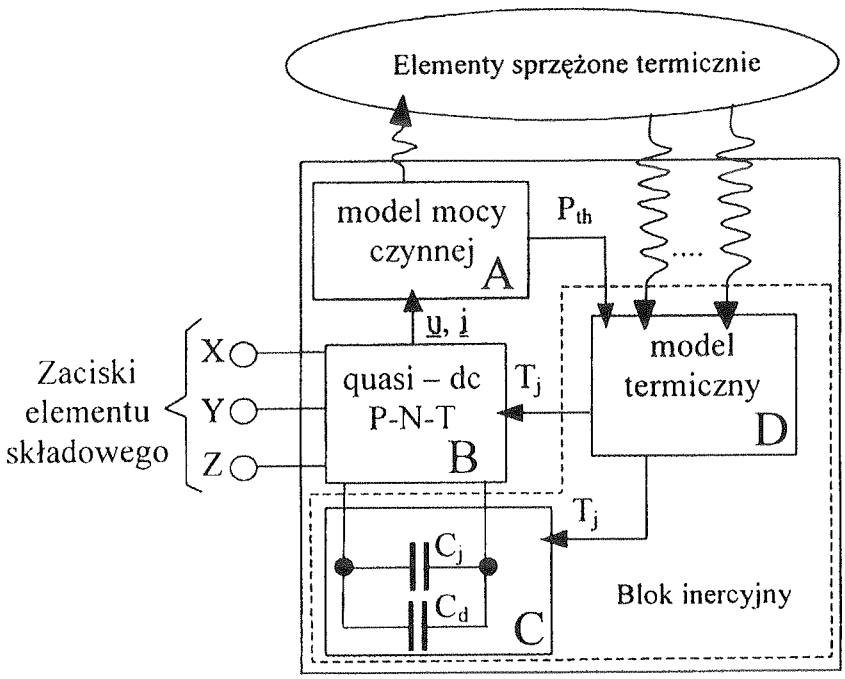
Konstrukcja dynamicznego elektrotermicznego makromodelu tranzystora Darlingtona opiera się na odpowiednim połączeniu elektrotermicznych modeli poszczególnych elementów składowych rozważanej struktury tranzystora Darlingtona zawierających:

- zależności prądowo-napięciowo-temperaturowe (P-N-T), opisujące stałoprądową postać modelu elektrycznego z parametrami zależnymi od temperatury [1] oraz odpowiednie pojemności złączowe C_j i dyfuzyjne C_d, występujące w rozważanej strukturze,
- skupiony model termiczny,
- zależność mocy czynnej od napięć i prądów zaciskowych.

Schemat blokowy tak zdefiniowanego modelu elementu składowego przedstawiono na rys. 2. Część bezinercyjna modelu (bloki A i B) została szczegółowo opisana we wcześniejszych pracach autorów [1, 3, 4] i nie będzie tu przytaczana, natomiast bloki C i D związane z inercją tranzystora Darlingtona są omówione w dalszej części pracy.

Modelowanie i analiza pracy przyrządu półprzewodnikowego w warunkach pobudzenia sygnałem zmiennym (praca impulsowa) wymaga uwzględnienia w jego makromodelu, elementów inercyjnych występujących w strukturze półprzewodnikowej. Modelowanie z uwzględnieniem oddziaływań elektrotermicznych wymaga zatem uwzględnienia dwóch rodzajów inercji: inercji elektrycznej, reprezentowanej przez pojemności złączowe (C_j) i dyfuzyjne (C_d) związane z każdym złączem p-n struktury oraz inercji termicznej, reprezentowanej przez odpowiednie termiczne stałe czasowe τ_{th}.

Znane, literaturowe zależności opisujące pojemności elektryczne uwzględniają wpływ temperatury, w tym przypadku temperatury wnętrza elementu T_j, na te pojemności. W pracy przyjęto klasyczny model pojemności złączowej [5]



Rys. 2. Struktura elektrotermicznego modelu elementu skladowego tranzystora Darlingtona

$$C_j(T_j) = \frac{C_{j0}}{\sqrt[n]{1 - \frac{u}{V_j(T_j)}}} \tag{1}$$

oraz pojemności dyfuzyjnej

$$C_d = \tau \cdot \frac{\partial i}{\partial u} = \tau \cdot \frac{i(u, T_j) + \mathrm{I}_S(T_j)}{h \cdot T_j} \tag{2}$$

gdzie, we wzorze (1), n = 2 dla złącz emiterowych oraz n = 3 dla złącz kolektorowych i dla diody antyrównoległej, C_{j0} jest wartością pojemności złączowej przy zerowym napięciu u na złączu, V_j oznacza napięcie bariery, h jest ilorazem stałej Boltzmanna przez ładunek elektronu, we wzorze (2) i(u, T_j) opisuje charakterystykę statyczną złącza, τ jest parametrem zależnym od czasu przelotu nośników mniejszościowych przez bazę złącza oraz od czasu życia tych nośników, a niezależnym od temperatury, natomiast I_s jest prądem nasycenia złącza, opisanym zależnością

$$I_S(T_j) = I_0 \cdot \left(\frac{T_j}{T_0}\right)^2 \cdot \exp\left(-\frac{U_{go}}{hT_j}\right) \tag{3}$$

gdzie $U_{go} = 1{,}206$ V dla krzemu, T_0 jest temperaturą odniesienia, natomiast I_0 to parametr, którego wartość jest wprost proporcjonalna do powierzhni złącza [6].

Jak wynika z pracy [7], zależność pojemności złączowej C_j od temperatury, związana z zależnością $V_j(T)$ jest słaba. Temperaturowy współczynnik zmian pojemności złączowej przyjmuje wartości od około 0,01%/K przy zaporowej polaryzacji złącza do 1%/K dla złącza spolaryzowanego przewodząco [7]. Dlatego we wzorze (1) pominięto zależność napięcia bariery V_j od temperatury.

W celu zapewnienia ciągłości modelu pojemności złączowej danego wzorem (1) dla dowolnych wartości napięć i temperatur, dla napięć na złączu spełniających warunek $u \geq V_j - U_{gr}$, zastosowano aproksymację liniową tej zależności o postaci [3]

$$C_j = C_{j0} \cdot \sqrt[n]{\frac{V_j}{U_{gr}}} \cdot \frac{1}{n} \cdot \left(n + 1 - \frac{V_j - u}{U_{gr}} \right) \tag{4}$$

Napięcie U_{gr} powinno być małe, aby w jak najszerszym zakresie napięć mogła być wykorzystywana zależność (1). W obliczeniach przyjęto $U_{gr} = 0{,}2$ V.

W celu praktycznego zastosowania przytoczonych modeli pojemności złączowych i dyfuzyjnych, przy implementacji makromodelu tranzystora Darlingtona w programie SPICE, równolegle do wszystkich złączy, włączono sterowane źródła prądowe, których wydajność I_C opisana jest za pomocą opcji ABM (Analog Behaviour Modelling) następującym wzorem

$$I_C = (C_j(u, T_j) + C_d(u, T_j)) \cdot \frac{du}{dt} \tag{5}$$

gdzie wartość pochodnej napięcia na złączu du/dt wyznaczana jest za pomocą układu różniczkującego RC [3, 8], w którym wartość stałej czasowej RC jest rzędu pojedynczych ps.

Z kolei istnienie inercji termicznej w elemencie powoduje konieczność uwzględnienia tego zjawiska w modelu termicznym. Znając postać przejściowej impedancji termicznej można wyznaczyć termiczną odpowiedź elementu na pobudzenie dowolnym przebiegiem mocy p(t) wykorzystując w tym celu, np. całkę splotu o postaci [7]

$$T_j(t) = T_a + \int_0^t Z'(t - x) \cdot p(x) \cdot dx \tag{6}$$

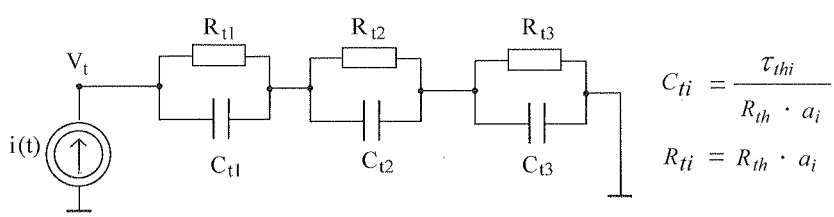
gdzie T_a oznacza temperaturę otoczenia, natomiast Z'(t) jest czasową pochodną przejściowej impedancji termicznej Z(t), której przebieg jest typowo modelowany za pomocą sumy funkcji eksponencjalnych [9]

$$Z(t) = R_{th} \cdot \left[1 - \sum_{i=1}^{L} a_i \cdot \exp\left({}^{-t}/_{\tau_{thi}} \right) \right] \tag{7}$$

gdzie R_{th} oznacza rezystancję termiczną, τ_{thi} – termiczną stałą czasową (i = 1, 2,...,L), $a_1 + a_2 + ... + a_L = 1$, natomiast L zależy od liczby warstw tworzących konstrukcję elementu i przyjmuje zwykle wartość 2 lub 3 [10].

W literaturze można znaleźć wygodne do implementacji komputerowej, zdyskretyzowane modele oparte na zależności (6) nazywane algorytmami bezpamięciowymi [7, 11, 12]. Zasadniczą zaletą tych algorytmów jest fakt, że w celu obliczenia wartości temperatury wnętrza za pomocą tych algorytmów w kroku k-tym, wymagana jest znajomość wartości temperatury w kroku (k-1)-szym oraz wartości mocy niezbędnych do wyznaczenia wartości całki splotu w k-tym kroku czasowym, w przeciwieństwie do algorytmów klasycznych, w których potrzebne są wartości mocy we wszystkich poprzednich chwilach czasu.

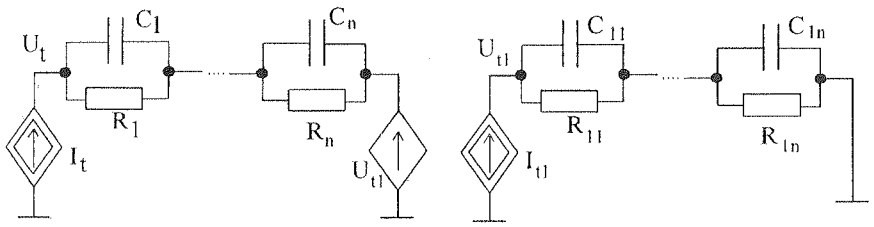
Niestety, dostępne dla użytkownika instrukcje programu SPICE nie pozwalają na realizację tak sformułowanego termicznego modelu elementu w postaci całki splotu. Bezpamięciowe algorytmy obliczania całki splotu, omówione szczegółowo w pracy [7], nie mogą być wykorzystane, gdyż ich realizacja wymaga znajomości wartości mocy w poprzednim kroku czasowym – jak zaznaczono wcześniej – niedostępnych dla użytkownika programu SPICE. Akceptowalnym przez program SPICE odpowiednikiem całki splotu są obwody elektryczne w postaci łańcucha (struktura kanoniczna Fostera – rys. 3) lub drabinki RC (struktura kanoniczna Cauera). Przyjęcie modelu przejściowej impedancji termicznej Z(t) zgodnie z wzorem (7) pozwala na wyznaczenie temperatury wnętrza za pomocą analogu obwodowego RC (rys. 3), w którym iloczyn pojemności C_{ti} oraz rezystancji R_{ti} jest równy termicznej stałej czasowej τ_{thi}, wydajność dołączonego generatora prądowego i(t) równa jest wartości mocy czynnej p(t) wydzielanej w elemencie, natomiast V_t jest wartością nadwyżki temperatury wnętrza T_j ponad temperaturę otoczenia T_a.



Rys. 3. Reprezentacja modelu termicznego w postaci łańcucha

Przyjęcie w rozważanych obwodach trzech ogniw RC jest wystarczające, co potwierdzają też wyniki pomiarów zamieszczone w pracy [3]. Ponieważ w strukturze tranzystora Darlingtona występuje 5 elementów wzajemnie na siebie oddziaływających termicznie, więc w modelu termicznym każdego elementu składowego powinno wystąpić 5 obwodów o strukturze pokazanej na rys. 3. Jeden z obwodów modeluje samonagrzewanie, natomiast cztery pozostałe – przyrost temperatury wnętrza wywołany wydzielaniem mocy czynnej w innych elementach składowych. Jak wynika z pracy [3],

otrzymane z pomiarów wartości wzajemnych przejściowych impedancji termicznych w tranzystorze Darlingtona są zbliżone, dlatego model termiczny każdego elementu składowego można uprościć do postaci pokazanej na rys. 4. Na rysunku tym, źródło prądowe I_t ma wydajność odpowiadającą mocy czynnej wydzielanej w rozważanym elemencie składowym, U_t jest nadwyżką temperatury wnętrza tego elementu ponad temperaturę otoczenia, U_{tI} jest częścią tej nadwyżki wywołaną przez wzajemne sprzężenia termiczne z innymi elementami składowymi, natomiast I_{tI} jest sumą mocy czynnych wydzielanych w elementach sprzężonych termicznie.



Rys. 4. Obwodowy model termiczny elementu składowego przy założeniu jednakowych impedancji termicznych wzajemnych między wszystkimi elementami składowymi

3. ALGORYTM ESTYMACJI WARTOŚCI PARAMETRÓW DEMTD

Przedstawiony w niniejszym rozdziale algorytm estymacyjny ESDYN, przeznaczony do wyznaczania wartości parametrów dynamicznych bloku C (rys. 2), stanowi rozszerzenie algorytmu IDEN [2] wyznaczania wartości parametrów bloku B (rys. 2). Nieporuszony tutaj problem wyznaczania wartości parametrów modelu termicznego (blok D na rys. 2) został rozwiązany przez autorów wcześniej. Zaproponowany w pracach [7, 13] algorytm sekwencyjny umożliwia wyznaczenie wartości rezystancji termicznej R_{th} elementu, termicznych stałych czasowych τ_i oraz współczynników wagowych a_i, na podstawie odpowiednio przekształconego przebiegu przejściowej impedancji termicznej.

Punktem wyjścia przy realizacji algorytmu ESDYN są pomiary odpowiednich dynamicznych charakterystyk zaciskowych – małosygnałowych i wielkosygnałowych, zrealizowane w warunkach izotermicznych przy ustalonej temperaturze otoczenia i określonej polaryzacji tranzystora. Wyznaczanie wartości kolejnych parametrów wymaga właściwego wyboru współrzędnych punktu pomiarowego i stosownych obliczeń z wykorzystaniem sformułowanych przez autorów zależności algebraicznych lub iteracyjnego doboru wartości parametrów, zapewniających uzyskanie zadawalającej zgodności obliczonego i pomierzonego czasowego przebiegu napięć i prądów zaciskowych tranzystora Darlingtona.

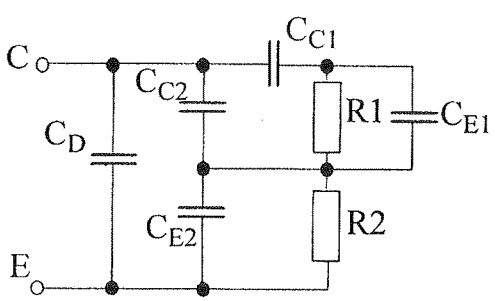
W opisie procedury estymacyjnej oznaczenie literowe elementu składowego, występujące w indeksie dolnym odpowiedniego parametru, określa z jakim elementem

składowym związany jest rozważany parametr, np. $C_{je0(T1)}$ oznacza pojemność złącza baza-emiter tranzystora T1 przy zerowej polaryzacji. Realizacja algorytmu wymaga następującego postępowania pomiarowo-obliczeniowego.

1. Pomiar modułu małosygnałowej impedancji wyjściowej $Z_{wy}(f)$ tranzystora Darlingtona w funkcji częstotliwości przy $i_B = 0$ dla napięcia kolektor-emiter $u_{CE} > 0$. Dla wybranego punktu, leżacego na opadającej części tej charakterystyki o współrzędnych $Z_{wy} = Z_1$, $f = f_1$, należy wyliczyć wartość pojemności wyjściowej

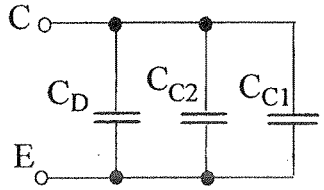
$$C_{wy1} = \frac{1}{2 \cdot \pi \cdot f_1 \cdot |Z_1|} \tag{8}$$

Uproszczony małosygnałowy schemat zastępczy tranzystora Darlingtona dla dodatniego napięcia kolektor-emiter, przy zerowym prądzie bazy przedstawiono na rys. 5. Na rysunku tym, symbole kondensatorów oznaczają pojemności złączowych poszczególnych elementów składowych: C_{E1} oraz C_{C1} to pojemności odpowiednio złączy emiterowego i kolektorowego tranzystora T1, C_{E2} oraz C_{C2} – pojemności złączy tranzystora T2, natomiast C_D – pojemność diody D.



Rys. 5. Uproszczony małosygnałowy schemat zastępczy tranzystora Darlingtona dla dodatniego napięcia kolektor-emiter, przy zerowym prądzie bazy

Jak widać pojemność C_{E1} jest zbocznikowana przez rezystor R1 i moduł jej impedancji dla częstotliwości f_1 jest znacznie mniejszy od rezystancji R1. Z kolei pojemność C_{E2} jest też zbocznikowana przez rezystor R2 i moduł jej impedancji przy częstotliwości f_1 jest znacznie mniejszy od rezystancji R2. Dodatkowo moduł impedancji przy częstotliwości f_1 wnoszony przez pojemność C_{C1} jest dużo większy niż suma rezystancji R1 i R2. Zatem w modelu małosygnałowym, przedstawionym na rys. 5, można pominąć pojemności związane ze złączami emiterowymi tranzystorów składowych oraz rezystory bocznikujące. Otrzyma się wówczas model zredukowany przestawiony na rys. 6. Jak widać, wypadkowa pojemność widziana między zaciskami kolektora i emitera jest równa sumie pojemności złączowych diody oraz złączy kolektorowych tranzystorów składowych.



Rys. 6. Zredukowany małosygnałowy schemat zastępczy tranzystora Darlingtona dla dodatniego napięcia kolektor-emiter, przy zerowym prądzie bazy

Ponieważ wszystkie trzy wymienione złącza są wykonane w jednym procesie technologicznym, więc zasadne jest przyjęcie założeń, że $\frac{C_D}{C_{C1}} = \frac{I_{0(D)}}{I_{0(T1)}} = n_1$ oraz $\frac{C_{C2}}{C_{C1}} = \frac{I_{0(T2)}}{I_{0(T1)}} = n_2$, gdzie wartości parametru I_0 poszczególnych elementów składowych są proporcjonalne do powierzchni odpowiednich złączy. Przy polaryzacji zaporowej wszystkich złączy w tranzystorze Darlingtona, napięcie wsteczne na trzech rozważanych złączach jest praktycznie takie samo.

Zatem wartości pojemności złączonych przy zerowej polaryzacji dla złączy kolektorowych tranzystorów T1 ($C_{jc0(T1)}$) i T2 ($C_{jc0(T2)}$) oraz dla diody D ($C_{j0(D)}$) należy wyznaczyć ze wzorów

$$C_{jco(T1)} = (1 + n_1 + n_2)^{-1} \cdot C_{wy1} \cdot \sqrt[3]{1 - \frac{u_{CE}}{V_{jC(T1)}}} \tag{9}$$

$$C_{jco(T2)} = n_2 \cdot C_{jco(T1)} \tag{10}$$

$$C_{j0(D)} = n_1 \cdot C_{jco(T1)} \tag{11}$$

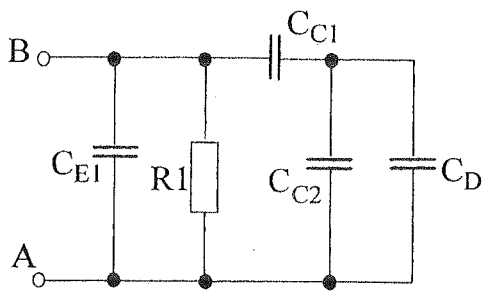
gdzie $V_{jC(T1)}$ jest napięciem bariery złącza kolektorowego tranzystora T1, którego wartość należy wyznaczyć za pomocą algorytmu IDEN [2].

2. Pomiar modułu małosygnałowej impedancji wejściowej $Z_{we1}(f)$ tranzystora Darlingtona przy prądzie kolektora $i_C = 0$ oraz zwarciu bazy tranzystora T2 (punkt A na rys. 1) z zaciskiem emitera, dla dwóch wartości napięcia baza-emiter $u_{BE2} < u_{BE3} < 0$. Dla wybranych dwóch punktów A(u_{BE2}, f_2, Z_2) i B(u_{BE3}, f_3, Z_3), leżących na opadających częściach tych charakterystyk, należy obliczyć wartość pojemności wejściowych tranzystora Darlingtona ze wzorów

$$C_{we1} = \frac{\sqrt{\frac{1}{|Z_2|^2} - \frac{1}{R1^2}}}{2 \cdot \pi \cdot f_2} - \frac{C_{C1} \cdot (C_{C2} + C_D)}{C_{C2} + C_D + C_{C1}} \tag{12}$$

$$C_{we2} = \frac{\sqrt{\frac{1}{|Z_3|^2} - \frac{1}{R1^2}}}{2 \cdot \pi \cdot f_3} - \frac{C_{C1} \cdot (C_{C2} + C_D)}{C_{C2} + C_D + C_{C1}} \tag{13}$$

Powyższe wzory wynikają z małosygnałowego modelu tranzystora Darlingtona, który w tym przypadku można przedstawić w postaci schematu zastępczego, pokazanego na rys. 7.



Rys. 7. Uproszczony małosygnałowy schemat zastępczy tranzystora Darlingtona dla ujemnego napięcia baza-emiter, przy zerowym prądzie kolektora i zwarciu bazy tranzystora T2 (punkt A) z emiterem

Wartości pojemności złączowych złączy kolektorowych tranzystorów składowych C_{C1}, C_{C2}, oraz diody C_D należy wyliczyć na podstawie wzoru (1) przy wykorzystaniu wartości parametrów wyznaczonych w pierwszym punkcie algorytmu, natomiast wartość rezystancji R1 wyznaczono za pomocą algorytmu IDEN [2]. Napięcie na złączu baza-kolektor tranzystora T2 oraz napięcie na diodzie są w przybliżeniu równe zeru, natomiast napięcie na obu złączach tranzystora T1 jest w przybliżeniu równe u_{BE1}.

Wartości parametrów $C_{jeo(T1)}$, $V_{jE(T1)}$ opisujących pojemność złączową złącza emiterowego tranzystora T1 wyliczane są ze wzorów

$$V_{jE(T1)} = \frac{u_{BE3} \cdot \frac{C_{we2}^2}{C_{we1}^2} - u_{BE2}}{\frac{C_{we2}^2}{C_{we1}^2} - 1} \tag{14}$$

$$C_{jeo(T1)} = C_{we1} \cdot \sqrt{1 - \frac{u_{BE2}}{V_{jE(T1)}}} \tag{15}$$

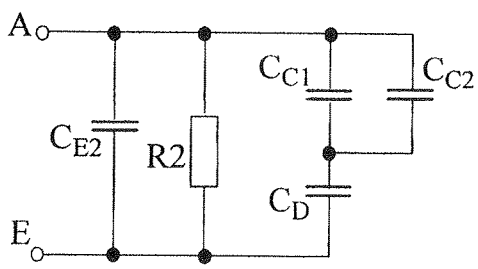
3. Powtórzyć pomiary z punktu 2 przy $i_C = 0$ oraz zwarciu bazy tranzystora T2 (punkt A na rys. 1) z zaciskiem bazy, dla dwóch wartości napięcia baza-emiter $u_{BE4} < u_{BE5} < 0$. Dla analogicznie usytuowanych dwóch punktów pomiarowych A1

(u_{BE4}, f_4, Z_4) i B1 (u_{BE5}, f_5, Z_5), należy obliczyć wartość pojemności wejściowych tranzystora Darlingtona ze wzorów

$$C_{we3} = \frac{\sqrt{\frac{1}{|Z_4|^2} - \frac{1}{R2^2}}}{2 \cdot \pi \cdot f_4} - \frac{C_D \cdot (C_{C2} + C_{C1})}{C_{C2} + C_D + C_{C1}} \tag{16}$$

$$C_{we4} = \frac{\sqrt{\frac{1}{|Z_5|^2} - \frac{1}{R2^2}}}{2 \cdot \pi \cdot f_5} - \frac{C_D \cdot (C_{C2} + C_{C1})}{C_{C2} + C_D + C_{C1}} \tag{17}$$

Powyższe wzory wynikają z małosygnałowego modelu tranzystora Darlingtona, który w tym przypadku można przedstawić w postaci schematu zastępczego (rys. 8). Wartość rezystancji R2 wyznaczona została za pomocą algorytmu IDEN [2]. Napięcie na złączach baza-kolektor tranzystorów składowych jest w przybliżeniu równe u_{BE4}, natomiast napięcie na diodzie jest w przybliżeniu równe zeru.



Rys. 8. Uproszczony małosygnałowy schemat zastępczy tranzystora Darlingtona dla ujemnego napięcia baza-emiter, przy zerowym prądzie kolektora i zwarciu bazy tranzystora T2 (punkt A) z emiterem

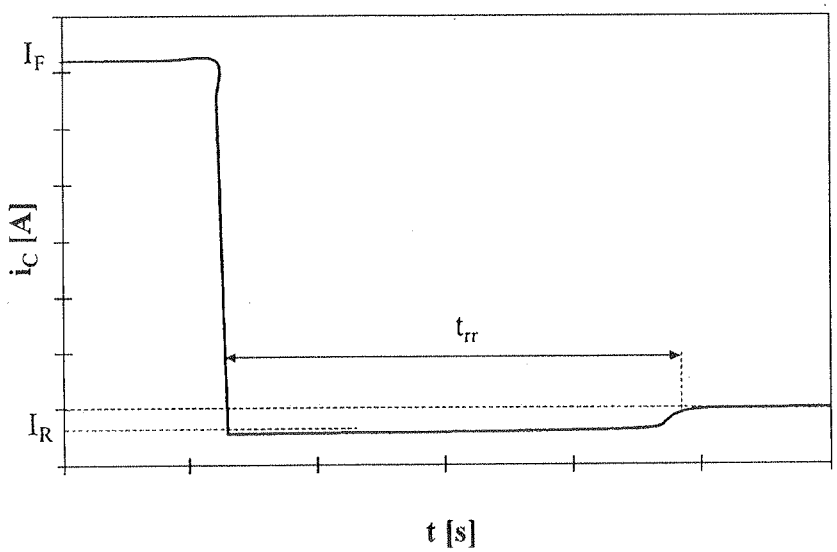
Wartości parametrów $C_{jeo(T2)}$, $V_{jE(T2)}$ opisujących pojemność złączową złącza emiterowego tranzystora T2 wyznaczane są ze wzorów

$$V_{jE(T2)} = \frac{u_{BE5} \cdot \frac{C_{we4}^2}{C_{we3}^2} - u_{BE4}}{\frac{C_{we4}^2}{C_{we3}^2} - 1} \tag{18}$$

$$C_{jeo(T2)} = C_{we3} \cdot \sqrt{1 - \frac{u_{BE4}}{V_{jE(T2)}}} \tag{19}$$

4. Pomiar odpowiedzi czasowej $i_C(t)$ tranzystora Darlingtona na pobudzenie ujemnym impulsem napięcia u_{CE} przy prądzie bazy równym zero (rys. 9). W czasie impulsu następuje załączenie i wyłączenie diody antyrównoległej. Z charakterystyki $i_C(t)$

należy odczytać wartości I_F, I_R oraz czas odzyskiwania zdolności zaworowej diody t_{rr} [14].



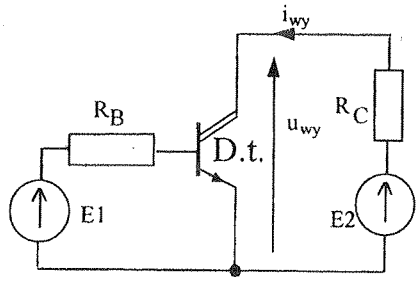
Rys. 9. Odpowiedź tranzystora Darlingtona na pobudzenie ujemnym impulsem napięcia u_{CE}

Ponieważ w czasie pomiaru żadne ze złączy tranzystorów składowych nie jest spolaryzowane przewodząco, można pominąć pojemności dyfuzyjne związane z tymi złączami, a wartość parametru τ obliczana jest ze wzoru [14]

$$\tau = \frac{t_{rr}}{\ln(I_F/I_R + 1)} \tag{20}$$

5. Pomiar odpowiedzi czasowej $i_C(t)$ tranzystora Darlingtona, pracującego w konfiguracji klucza równoległego z obciążeniem rezystancyjnym (rys. 10), na pobudzenie impulsem prostokątnym prądu bazy.

Pojemności dyfuzyjne złączy emiterowych i kolektorowych tranzystorów składowych są opisane odpowiednio za pomocą parametrów charakteryzujących czas życia nośników mniejszościowych i czas przelotu tych nośników przez bazę w kierunku normalnym τ_F oraz inwersyjnym τ_R. W celu wyznaczenia wartości tych parametrów, należy wykonać szereg obliczeń impulsowych charakterystyk tranzystora Darlingtona, pracującego w rozważanym układzie, przy pobudzeniu tego przyrządu półprzewodnikowego prądem bazy w kształcie impulsu prostokątnego, dla modyfikowanych w kolejnych krokach wartości parametrów τ_F oraz τ_R. Warunkiem zakończenia obliczeń iteracyjnych jest uzyskanie zadawalającej zgodności wyników symulacji i pomiarów. Ponieważ oba tranzystory składowe zostały wykonane w jednym procesie technologicznym, parametry τ_F oraz τ_R dla obu tranzystorów mają takie same wartości.



Rys. 10. Analizowany układ klucza równoległego

4. WYNIKI OBLICZEŃ I POMIARÓW

Przedstawiony w pracy dynamiczny elektrotermiczny makromodel tranzystora Darlingtona zaimplementowano w programie PSPICE w postaci podukładu. Celem oceny jego wiarygodności i dokładności, przeprowadzono szereg symulacji prostych układów zawierających rozważany tranzystor, przyjmując w obliczeniach wartości parametrów makromodelu wyznaczone za pomocą procedury ESDYN. Wybrane wyniki obliczeń zostały zweryfikowane doświadczalnie.

Oceny poprawności działania procedury ESDYN dokonano w oparciu o wyniki prostych eksperymentów numerycznych, zrealizowanych dwuetapowo. Najpierw, przyjmując arbitralnie „wzorcowe" wartości parametrów makromodelu, obliczono odpowiednie (zgodnie z wymogami algorytmu) małosygnałowe i wielkosygnałowe dynamiczne charakterystyki zaciskowe. Następnie, traktując tak otrzymane charakterystyki jako wejściowe dane „pomiarowe" wyznaczono wartości parametrów dynamicznych zgodnie z punktami 1÷4 algorytmu ESDYN. Niezbędne w tego typu eksperymencie numerycznym wartości parametrów stałoprądowych zostały wyznaczone wcześniej za pomocą algorytmu IDEN opisanego w pracy [2]. Przykładowe wyniki obliczeń dla wybranych parametrów tranzystora Darlingtona zamieszczono w tablicy 1.

Tablica 1

Wartości wybranych „wzorcowych" i obliczonych algorytmem ESDYN parametrów tranzystora Darlingtona

Parametr	$C_{je0(T1)}$[pF]	$C_{jc0(T1)}$[pF]	$V_{jE(T1)}$[V]	$C_{je0(T2)}$[pF]	$C_{jc0(T2)}$[pF]	$V_{jE(T2)}$[V]	$C_{j0(D)}$[pF]	$\tau_{(D)}$[ns]
Wartość wzorcowa	100	30	0,8	400	120	0,8	24	100
Wartość obliczona	98,83	30,62	0,85	403,2	122,5	0,85	24,5	97,35

Jak widać, uzyskano dobrą zgodność wartości parametrów „wzorcowych" i obliczonych. Różnice nie przekraczają kilku procentów. W tablicy nie zamieszczono dobieranych iteracyjnie (punkt 5 algorytmu ESDYN) wartości parametrów τ_F oraz τ_R.

Algorytm ESDYN wykorzystano także do wyznaczenia wartości parametrów elektrycznych dynamicznego makromodelu wybranych typów tranzystorów Darlingtona

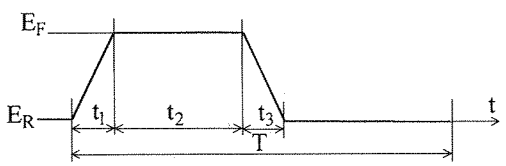
mocy. Przykładowo, w tablicy 2 przedstawiono wyznaczone za pomocą tego algorytmu wartości parametrów dynamicznych tranzystora BU323A.

Tablica 2

Wartości parametrów tranzystora Darlingtona BU323A

parametr	$C_{jeo(T1)}$ [pF]	$C_{jco(T1)}$ [pF]	$V_{jE(T1)}$ [V]	$C_{jeo(T2)}$ [pF]	$C_{jco(T2)}$ [pF]	$V_{jE(T2)}$ [V]
wartość	97	32,3	0,5	485	161,5	0,5
parametr	$C_{jo(D)}$[pF]	$\tau_{(D)}$[ns]	$\tau_{F(T1)}$[ns]	$\tau_{R(T1)}$[ns]	$\tau_{F(T2)}$[ns]	$\tau_{R(T2)}$[ns]
wartość	32,3	166,7	166,7	1000	166,7	1000

Przy wykorzystaniu wartości parametrów otrzymanych za pomocą algorytmu ES-DYN wykonano szereg obliczeń układów elektronicznych zawierających tranzystor Darlingtona mocy. Wyniki tych badań przedstawiono m.in. w pracach [3, 15, 16]. W niniejszej pracy, w charakterze przykładu, zaprezentowano wyniki badań układu klucza równoległego, przedstawionego na rys. 10, który pobudzono napięciem E1, o kształcie pokazanym na rys. 11.

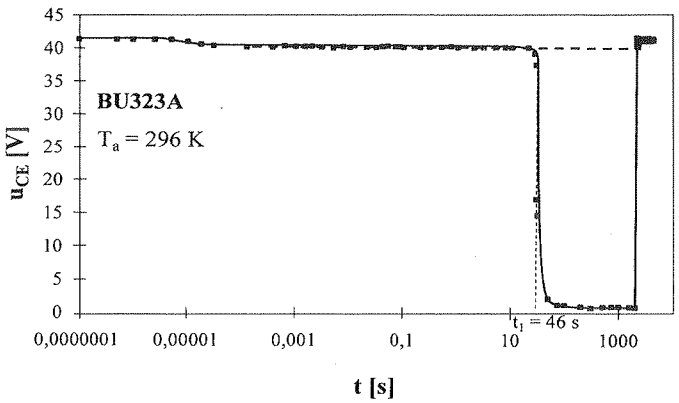


Rys. 11. Kształt impulsu pobudzającego

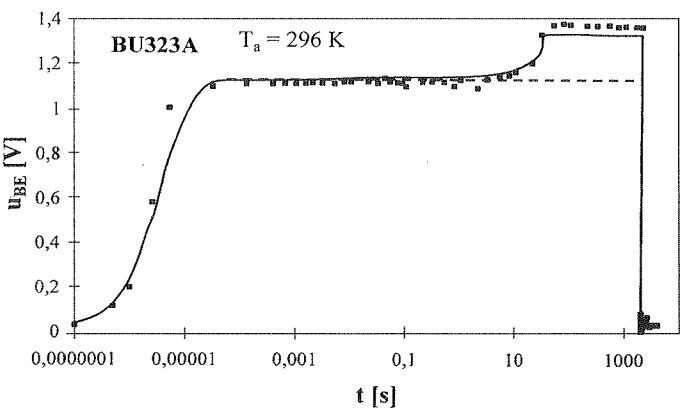
Na rys. 12 przedstawiono wyniki analiz elektrotermicznych (linia ciągła) i pomiarów (kwadraty) oraz, dla porównania, wyniki analiz izometrycznych (linia kreskowa) rozważanego układu z tranzystorem BU323A. Przyjęto następujące wartości parametrów impulsu sterującego: $E_F = 16{,}16$ V, $E_R = 0{,}5$ V, $t_1 = t_3 = 2\mu$s, $t_2 = 2000$ s, $T = 4000$ s oraz pozostałych elementów obwodu: $E2 = 41{,}5$ V, $R_0 = 51{,}3$ Ω, $R_B = 5{,}087$ kΩ. Dodatkowo na rys. 13 przedstawiono obliczone przebiegi temperatury wnętrza diody antyrównoległej D oraz tranzystorów składowych T1 i T2 przy założonym pobudzeniu układu klucza równoległego.

Jak widać na rys. 12a, w przypadku analizy izometrycznej klucz tranzystorowy nie zostaje włączony, to znaczy, napięcie u_{CE} pozostaje dużo większe, równe praktycznie E2, od napięcia nasycenia tranzystora Darlingtona nie przekraczającego wartości kilku woltów [17], natomiast w przypadku analizy elektrotermicznej, na skutek efektu samonagrzewania, po około 46 s od chwili włączenia impulsu, następuje załączenie klucza. Poprawność uzyskanego wyniku analizy elektrotermicznej potwierdzają wyniki pomiarów. Warto zauważyć, że zmiana temperatury wnętrza tranzystora Darlingtona o zaledwie kilkanaście kelwinów (rys. 13) jest wystarczająca do uzyskania istotnej

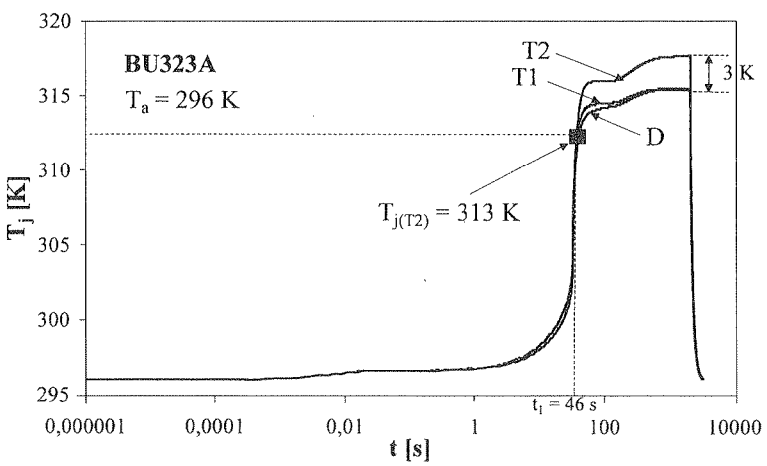
a)



b)



Rys. 12. Odpowiedź klucza na pobudzenie pojedynczym impulsem napięcia E1



Rys. 13. Przebieg temperatury wnętrza tranzystora Darlingtona w rozważanym układzie

różnicy jakościowej w przebiegu charakterystyk rozważanego klucza. W rozważanym przypadku, w strukturze tranzystora Darlingtona panuje praktycznie równomierny rozkład temperatury. Obliczone różnice temperatur wnętrza elementów składowych nie przekraczają 3 K.

5. WNIOSKI KOŃCOWE

W pracy zaproponowano dynamiczny elektrotermiczny makromodel tranzystora Darlingtona mocy oraz prosty i efektywny algorytm wyznaczania wartości jego parametrów elektrycznych. Opracowany makromodel stanowi połączenie elektrotermicznych modeli elementów składowych, zgodnie z topologią przedstawiono na rys. 1, przy czym w opisie poszczególnych elementów składowych zastosowano uproszczenia wynikające ze specyfiki tranzystora Darlingtona. Makromodel opracowano w formie, umożliwiającej wygodną jego implementację w programie SPICE przy wykorzystaniu źródeł sterowanych, których wydajność opisana jest za pomocą opcji ABM (Analog Behaviour Modelling).

Prezentowany w pracy algorytm wyznaczania wartości parametrów jest prosty w realizacji, a niezbędna aparatura pomiarowa dostępna jest w typowym laboratorium. Pewną niedogodnością zaproponowanego algorytmu jest konieczność dostępu do wyprowadzeń wszystkich elementów składowych badanego tranzystora Darlingtona, co sprawia, że w celu realizacji tego algorytmu konieczne jest otwarcie obudowy badanego tranzystora oraz wykonanie dodatkowego wyprowadzenia z wnętrza jego struktury.

Przedstawione w pracy wyniki obliczeń i pomiarów potwierdzają skuteczność i wiarygodność zarówno proponowanego makromodelu tranzystora Darlingtona, jak i algorytmu wyznaczania jego parametrów. Umożliwiają one poprawną analizę elektrotermicznych stanów przejściowych w układach elektronicznych zawierających ten przyrząd półprzewodnikowy. Przedstawiony w pracy przykład obliczeniowy dowodzi, że zmiana temperatury wnętrza tranzystora Darlingtona o zaledwie kilkanaście kelwinów powoduje jakościową zmianę przebiegu charakterystyk dynamicznych układu zawierającego rozważany przyrząd półprzewodnikowy. A zatem, nawet w przypadkach, gdy nadwyżka temperatury wnętrza tranzystora Darlingtona ponad temperaturę otoczenia wynosi zaledwie kilkanaście kelwinów, w celu uzyskania wiarygodnych wyników obliczeń należy zastosować elektrotermiczny makromodel rozważanego przyrządu półprzewodnikowego.

BIBLIOGRAFIA

1. J. Zarębski, K. Górecki: *Modelowanie tranzystora Darlingtona mocy z uwzględnieniem oddziaływań elektrotermicznych.* Kwartalnik Elektroniki i Telekomunikacji, t. 45, z. 3-4, 1999, s. 455.
2. J. Zarębski, K. Górecki: *Algorytm estymacji wartości parametrów elektrotermicznego stałoprądowego makromodelu tranzystora Darlingtona mocy.* Kwartalnik Elektroniki i Telekomunikacji, t. 46, z. 3, 2000, s. 299.
3. K. Górecki: *Elektrotermiczny makromodel tranzystora Darlingtona do analizy układów elektronicznych.* Praca doktorska, Politechnika Gdańska, 1999.

4. J. Zarębski, K. Górecki, W.J. Stepowicz: *Electrothermal Breakdown Modes in the Darlington Power Transistor.* Proc. of the European Conference on Circuit Theory and Design (ECCTD'99), Stresa (Włochy), 1999, Vol. 1, p. 567.
5. W. Marciniak: *Przyrządy półprzewodnikowe i układy scalone.* Warszawa, WNT, 1984.
6. W. Kuźmicz: *Projektowanie analogowych układów scalonych.* Warszawa, WNT, 1981.
7. J. Zarębski: *Modelowanie, symulacja i pomiary przebiegów elektrotermicznych w elementach półprzewodnikowych i układach elektronicznych.* Prace Naukowe Wyższej Szkoły Morskiej w Gdyni, Gdynia, 1996.
8. J. Zarębski, K. Górecki, W.J. Stepowicz: *Simulations of Electrothermal Transients in the Darlington Power Transistor.* Third International Scientific and Technical Conference on Unconventional Electromechanical and Electrical Systems (UEES'97), Alushta (Ukraina), 1997, Vol. 2, p. 721.
9. A. Nowakowski: *Badanie procesów termicznych w przyrządach półprzewodnikowych.* Zesz. Nauk. Polit. Gdanskiej, Elektronika LX, nr 389, 1984.
10. H.E. Hanrahan: *A Laboratory Experiment on the Thermal Behaviour of a Power Transistor.* Int. Journal of Electrical Engineering Education, Vol. 17, 1980, p. 1.
11. W. Janke: *Modele elektrotermiczne elementów półprzewodnikowych.* Zesz. Nauk. Polit. Gdańskiej, nr 374, Elektronika LVII, 1984.
12. J. Zarębski: *The Time Dependence Calculations of the Junction Temperature of Semiconductor Devices.* Int. Journal of Microcomputer Applications, Vol. 13, No. 2, June 1994, p. 62.
13. J. Zarębski, K. Górecki: *Thermal Transients in Bipolar Transistors; Measurements and Simulations.* Int. Conf. on Information, Systems, Methods Applied to Engineering Problems, Malta, 1993, Vol. 2, p. 111.
14. J. Kołodziejski, L. Spiralski, E. Stolarski: *Pomiary przyrządów półprzewodnikowych.* Warszawa, WKiŁ, 1990.
15. J. Zarębski, K. Górecki, S. Bieniecki: *Nieizometryczne charakterystyki przetwornicy obniżającej z tranzystorem Darlingtona mocy.* IV Konferencja Naukowo-Techniczna Zastosowanie Komputerów w Elektrotechnice ZKwE'99, Kiekrz-Poznań, 1999, s. 265.
16. J. Zarębski, K. Górecki, W.J. Stepowicz: *Influence of Thermal Effects on Performance of the Bipolar Darlington Power Transistor Switch.* XVI Symposium Electromagnetic Phenomena in Nonlinear Circuits EPNC'2000, Kraków, p. 147.
17. R. Kudyba: *Informacja techniczna, tranzystory BUP 323, BUP 323A.* Elementy półprzewodnikowe i układy scalone; zastosowania, układy analogowe. Przemysłowy Instytut Elektroniki, Nr 3, Warszawa 1984, s. 22.

J. ZARĘBSKI, K. GÓRECKI

A DYNAMIC POWER DARLINGTON TRANSISTOR MACROMODEL FOR THE ANALYSIS OF ELECTROTHERMAL TRANSIENTS

Summary

In the paper the dynamic electrothermal macromodel of the Darlington power transistor and the algorithm ESDYN for the estimation of all electrical parameters of this macromodel are presented. The proposed algorithm was formulated on the basis of the earlier elaborated d.c. macromodel in which new components, modelling electrical and thermal interia, were included. In the ESDYN algorithm the idea of the local parameter values estimation is used; it means that the values of parameters are calculated on the basis of the coordinates of the selected points lying on the proper parts of the isothermal a.c. characteristics of the investigated device. These part of characteristics can be described by the use of the estimated parameters. The theoretical considerations were illustrated by experimentally verified calculations results of electrothermal transients in the simple network including the power Darlington transistor.

Key words: Darlington power transistor, modeling electronic devices, parameters values estimation

KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJI, 2000, 46, z. 4, ss. 633–650

Badanie właściwości wybranych algorytmów splotowych do analizy termicznej układów elektronicznych

KRZYSZTOF GÓRECKI, JANUSZ ZARĘBSKI

Katedra Radioelektroniki Morskiej, Wyższa Szkoła Morska w Gdyni

Otrzymano 2000.09.04
Autoryzowano 2000.10.12

W pracy zaprezentowano analizę podstawowych właściwości wybranych bezpamięciowych algorytmów splotowych, należących do grupy algorytmów całkowania numerycznego. Zbadano stabilność i zbieżność tych algorytmów, przyjmując jako funkcje testowe funkcję eksponencjalną oraz jednostkowy uskok mocy, a także opracowano zależności opisujące wartość lokalnego błędu obcięcia oraz wyznaczono obszary stabilności bezwzględnej rozważanych algorytmów. Rozważania teoretyczne zilustrowano przykładami obliczeniowymi.

Słowa kluczowe: analiza układów elektronicznych, algorytmy splotowe

1. WPROWADZENIE

Analiza termiczna układu elektronicznego służy do wyznaczania przebiegu czasowego temperatury wnętrza wybranego elementu półprzewodnikowego, wchodzącego w skład układu elektronicznego, na podstawie znajomości czasowego przebiegu mocy czynnej wydzielanej w poszczególnych elementach tego układu. W analizie termicznej wykorzystywany jest model termiczny, który dla pojedynczego elementu elektronicznego niesprzężonego termicznie z innymi elementami w układzie może być wyrażony za pomocą całki splotu o postaci [1]

$$\Delta T = T_j - T_a = \int_0^t Z'(t-\nu)\cdot p(\nu)\,d\nu \tag{1}$$

gdzie ΔT oznacza przyrost temperatury wnętrza elementu T_j ponad temperaturę otoczenia T_a, $p(t)$ jest czasowym przebiegiem mocy wydzielanej w badanym elemencie, natomiast $Z'(t)$ to czasowa pochodna przejściowej impedancji termicznej elementu $Z(t)$,

typowo opisanej za pomocą sumy funkcji eksponencjalnych [2]

$$Z(t) = R_{th} \cdot \left[1 - \sum_{i=1}^{N} a_i \cdot \exp\left(-\frac{t}{\tau_i}\right)\right] \tag{2}$$

gdzie R_{th} jest rezystancją termiczną, τ_i oznacza i-tą termiczną stałą czasową, a_i to współczynniki spełniające zależność $\sum_{i=1}^{N} a_i = 1$, natomiast N jest liczbą termicznych stałych czasowych, przy czym w praktyce wystarczające jest przeważnie przyjęcie $N = 3$ [3].

W celu wyznaczenia nadwyżki temperatury ΔT równanie splotowe (1) może być rozwiązane przy wykorzystaniu klasycznych algorytmów całkowania numerycznego. Jednak w takim przypadku konieczne jest uwzględnienie w n-tym kroku czasowym wartości mocy odpowiadającej $(n-1)$ poprzednim chwilom czasu. Sformułowane w latach osiemdziesiątych, np. w pracy [1], i następnie rozwijane w latach dziewięćdziesiątych, np. w pracach [4, 5, 6], tak zwane bezpamięciowe algorytmy splotowe umożliwiły uproszczenie tych obliczeń dla przypadku, gdy jedna z funkcji splatanych jest funkcją eksponencjalną. Wyznaczenie wartości temperatury w kroku n-tym przy zastosowaniu omawianych algorytmów wymaga znajomości jedynie wartości temperatury w kroku $(n-1)$-szym oraz wartości mocy niezbędnych do wyznaczenia wartości całki splotu w n-tym kroku czasowym.

Dokładność rozważanych algorytmów splotowych była wielokrotnie weryfikowana przez porównanie wyników obliczeń otrzymanych za pomocą tych algorytmów z wynikami otrzymanymi innymi metodami, np. analitycznie, jednak w znanej autorom literaturze nie przeanalizowano formalnie stabilności i zbieżności bezpamięciowych algorytmów splotowych, które to parametry decydują o praktycznej użyteczności algorytmów całkowania numerycznego.

Niniejsza praca jest rozszerzeniem materiału zawartego w pracy [7], w której zaprezentowano wyniki badania stabilności i zbieżności wybranych algorytmów splotowych dla pobudzenia mocą w postaci uskoku jednostkowego. W niniejszej pracy przeanalizowano właściwości trzech wybranych bezpamięciowych algorytmów splotowych opisanych m.in. w pracach [4, 5]: algorytmu punktowego — PMA (Point Memoryless Algorithm), algorytmu polowego — FMA (Field Memoryless Algorithm) oraz algorytmu interpolacyjnego — IMA (Integral Memoryless Algorithm). W rozdziale drugim przedstawiono postać rozważanych bezpamięciowych algorytmów splotowych, w rozdziale trzecim przeprowadzono analizę ich stabilności i zbieżności, natomiast w rozdziale czwartym wyznaczono zależności opisujące wartość lokalnego błędu obcięcia tych algorytmów. Rozważania teoretyczne zilustrowano w rozdziale piątym wynikami obliczeń czasowych przebiegów temperatury wnętrza elementu elektronicznego dla wybranego przebiegu wydzielanej w nim mocy, otrzymanymi za pomocą rozważanych algorytmów bezpamięciowych przy różnych wartościach kroku czasowego oraz, dla porównania, otrzymanych z rozwiązania analitycznego równania (1).

2. BEZPAMIĘCIOWE ALGORYTMY SPLOTOWE

Postać rozważanych w niniejszej pracy bezpamięciowych algorytmów splotowych przedstawiono m.in. w pracach [1, 4, 5, 6]. Wyznaczony z wykorzystaniem tych algorytmów przyrost temperatury wnętrza elementu w n-tym kroku czasowym ΔT^n dany jest zależnością

$$\Delta T^n = \sum_{i=1}^{N} (\xi_i^n \cdot \Delta T^{n-1} + S_i^n) \tag{3}$$

gdzie ΔT^{n-1} oznacza przyrost temperatury wnętrza elementu w $(n-1)$-szym kroku czasowym, S_i^n — wartość całki splotu wyznaczoną w przedziale od t_{n-1} do t_n, odpowiadającą i-tej termicznej stałej czasowej τ_i, natomiast parametr ξ_i^n opisany jest wzorem

$$\xi_i^n = \exp\left(-\frac{h_n}{\tau_i}\right) \tag{4}$$

gdzie h_n oznacza długość n-tego kroku czasowego.

Wartość całki S_i^n można wyznaczyć stosując wybrany algorytm całkowania numerycznego. W szczególności, dla algorytmu trapezów

$$S_i^n = \alpha_{i2}^n \cdot p_{n-1} + \alpha_{i1}^n \cdot p_n \tag{5}$$

Dla algorytmu PMA, współczynniki α_{i1}^n oraz α_{i2}^n dane są wzorami [4, 5]

$$\alpha_{i1}^n = R_{th} \cdot \frac{a_i \cdot h_n}{2 \cdot \tau_i} \tag{6}$$

$$\alpha_{i2}^n = R_{th} \cdot \frac{a_i \cdot h_n}{2 \cdot \tau_i} \cdot \xi_i^n \tag{7}$$

oraz wzorami (8, 9) dla algorytmu FMA

$$\alpha_{i1}^n = R_{th} \cdot a_i \cdot \left(1 - \sqrt{\xi_i^n}\right) \tag{8}$$

$$\alpha_{i2}^n = R_{th} \cdot a_i \cdot \left(1 - \sqrt{\xi_i^n}\right) \cdot \sqrt{\xi_i^n} \tag{9}$$

W przypadku algorytmu IMA, wartość całki S_i^n wyznaczana jest półanalitycznie, to znaczy przebieg mocy $p(t)$ w ostatnich k przedziałach jest aproksymowany wielomianem stopnia k, gdzie k jest liczbą naturalną i analitycznie wyznaczana jest wartość splotu wielomianu aproksymującego moc $p(t)$ i pochodnej przejściowej impedancji termicznej $Z'(t)$ [5]. Wówczas wartość S_i^n wyznaczana jest ze wzoru [5]

$$S_i^n = \int_{t_{n-1}}^{t_n} R_{th} \cdot \frac{a_i}{\tau_i} \cdot \exp\left(-\frac{t_n - x}{\tau_i}\right) \cdot \left(\sum_{j=0}^{k} b_j \cdot x^j\right) dx \tag{10}$$

gdzie b_j oznaczają współczynniki wielomianu aproksymacyjnego, których wartości można wyznaczyć na przykład, przez rozwiązanie układu równań liniowych o postaci

$$\begin{bmatrix} 1 & t_n & \dots & t_n^k \\ 1 & t_{n-1} & \dots & t_{n-i}^k \\ \dots & \dots & \dots & \dots \\ 1 & t_{n-k} & \dots & t_{n-k}^k \end{bmatrix} \cdot \begin{bmatrix} b_0 \\ b_1 \\ \dots \\ b_k \end{bmatrix} = \begin{bmatrix} p_n \\ p_{n-1} \\ \dots \\ p_{n-k} \end{bmatrix} \tag{11}$$

Rozwiązując równania (10) i (11) dla wybranego stopnia k wielomianu aproksymacyjnego, otrzymuje się odpowiednio algorytmy bezpamięciowe IMA rzędu k-tego. W pracy [5] przytoczono jawną postać algorytmów IMA dla $k = 1$ oraz dla $k = 2$. Przykładowo, dla $k = 1$, w algorytmie IMA wartość całki S_i^n wyznaczana jest ze wzoru

$$S_i^n = R_{th} \cdot a_i \cdot b_1 \cdot \left((t_{n-1} - \tau_i) \cdot (1 - \xi_i^n) + h_n\right) + R_{th} \cdot a_i \cdot b_0 \cdot (1 - \xi_i^n) \tag{12}$$

gdzie współczynniki wielomianu aproksymującego wyrażone są wzorami

$$b_1 = \frac{p_n - p_{n-1}}{t_n - t_{n-1}} = \frac{p_n - p_{n-1}}{h_n} \tag{13}$$

$$b_0 = p_{n-1} - b_1 \cdot t_{n-1} \tag{14}$$

Postać algorytmu IMA wyrażona wzorami (12–14) jest stosunkowo skomplikowana. Dodatkowo, wartość całki S_i^n jest jawną funkcją czasu t_{n-1}. Stosując proste przekształcenia algebraiczne, postać wzoru (10) można zapisać w prostszej postaci. Przykładowo, dla algorytmu rzędu zerowego wartość całki S_i^n opisana jest zależnością identyczną jak dla algorytmu FMA skojarzonego z algorytmem prostokątów [5]

$$S_i^n = R_{th} \cdot a_i \cdot p_n \cdot (1 - \xi_i^n) \tag{15}$$

natomiast dla algorytmu IMA rzędu pierwszego, który inni autorzy nazywają algorytmem półanalitycznym [6] wartość całki S_i^n opisana jest zależnością

$$S_i^n = R_{th} \cdot a_i \cdot \left[p_n \cdot \left(1 - \frac{1 - \xi_i^n}{h_n} \cdot \tau_i\right) + p_{n-1} \cdot \left(\frac{1 - \xi_i^n}{h_n} \cdot \tau_i - \xi_i^n\right)\right] \tag{16}$$

Jak widać, stosując algorytmy bezpamięciowe, opisane wzorami (3)–(16), można wyznaczyć wartość przyrostu temperatury ΔT^n w kroku n-tym, znając jedynie wartość tego przyrostu w kroku $(n-1)$-szym, wartości mocy odpowiadających krokowi n-temu p_n i $(n-1)$-szemu p_{n-1} oraz wartości parametrów modelu przejściowej impedancji termicznej $Z(t)$.

3. STABILNOŚĆ I ZBIEŻNOŚĆ BEZPAMIĘCIOWYCH ALGORYTMÓW SPLOTOWYCH

Rozważane w artykule bezpamięciowe algorytmy splotowe należą do grupy wielokrokowych algorytmów całkowania numerycznego. Zgodnie z kryterium podanym w pracy [8] algorytm całkowania jest stabilny, gdy przy długości kroku dążącej do zera, wartość całkowitego błędu algorytmu, po wykonaniu dowolnie dużej liczby kroków czasowych, jest skończona. Z kolei o zbieżności algorytmu mówi się wtedy, gdy otrzymane za jego pomocą rozwiązanie jest jednostajnie zbieżne do rozwiązania dokładnego w rozważanym przedziale czasowym po wykonaniu $n \to \infty$ kroków. Oznacza to, że algorytm jest zbieżny w przedziale czasowym $t \in (0; T)$ wtedy, gdy dla dostatecznie małej wartości kroku obliczeń, równej $h = T/n$, otrzymane za jego pomocą wyniki różnią się dowolnie mało od rozwiązania dokładnego dla czasu t.

Do badania stabilności i zbieżności algorytmów całkowania stosowane jest często spotykane w opisie zjawisk fizycznych testowe zagadnienie początkowe, dla którego znane jest dokładne rozwiązanie analityczne [8]. Zagadnienie to ma typowo postać

$$\dot{x}(t) = -\lambda \cdot x, \qquad x(0) = 1 \tag{17}$$

gdzie λ jest w ogólności liczbą zespoloną.

Jednak, jak wynika z poprzedniego rozdziału, równanie o postaci (17) nie może być rozwiązane przy użyciu bezpamięciowych algorytmów splotowych. W związku z tym, biorąc pod uwagę kryteria podane w pracy [8], zaproponowano, aby do zbadania stabilności i zbieżności algorytmów bezpamięciowych wykorzystać równanie (1) przy założeniu, że moc $p(t)$ ma postać odpowiedniej funkcji testowej. Ponieważ każdą funkcję ciągłą można w wybranym przedziale przybliżyć z dowolną dokładnością za pomocą sumy funkcji stałej i funkcji eksponencjalnych [9], więc wykazanie stabilności i zbieżności algorytmów splotowych dla funkcji stałej oraz funkcji eksponencjalnej implikuje stabilność i zbieżność tych algorytmów dla dowolnej funkcji ciągłej.

W dalszej części rozdziału przedstawiono badania stabilności i zbieżności rozważanych algorytmów dla mocy opisanych funkcją eksponencjalną oraz uskokiem o wartości P

$$p(t) = P \cdot 1(t) \tag{18}$$

Dla uzyskania lepszej przejrzystości dalszych rozważań, założono, że przejściowa impedancja termiczna opisana jest jedną termiczną stałą czasową ($N = 1$). Nie ogranicza to ogólności rozważań, ponieważ, jak wynika z postaci wzorów (1) i (2), zarówno dla przypadku $N = 1$, jak i $N > 1$ wartość przyrostu temperatury wnętrza elementu jest opisana sumą funkcji eksponencjalnych, różniących się tylko wartościami parametrów a_i oraz τ_i. A zatem, ze stabilności i zbieżności rozważanych algorytmów dla przypadku $N = 1$, wynika stabilność i zbieżność tych algorytmów dla $N > 1$. Dla uproszczenia rozważań założono także, że obliczenia realizowane są ze stałym krokiem czasowym h.

Dla rozważanego przebiegu mocy, opisanego wzorem (18), wyznaczony analitycznie przyrost temperatury $\Delta T(t)$ ma postać

$$\Delta T(t) = R_{th} \cdot P \cdot \left[1 - \exp\left(-\frac{t}{\tau}\right)\right] \tag{19}$$

A zatem, zależność analityczną opisującą nadwyżkę temperatury dla n-tego kroku czasowego można przedstawić w postaci

$$\Delta T^n = R_{th} \cdot P \cdot \left[1 - \exp\left(-\frac{n \cdot h}{\tau}\right)\right] \tag{20}$$

Z kolei wykorzystując do wyznaczenia wartości ΔT^n rozważane algorytmy bezpamięciowe i uwzględniając, że dla mocy opisanej wzorem (18) składnik S_i^n we wzorze (3) ma w każdym kroku czasowym taką samą wartość, można wykazać, że otrzymana w n-tym kroku obliczeń wartości ΔT^n opisana jest sumą n-elementowego ciągu geometrycznego, którego pierwszy wyraz ma wartość S_i^n, a iloraz równy jest ξ_i^n. A zatem, dla mocy opisanej wzorem (18), zależności opisujące wartość przyrostu temperatury w kroku n-tym, można przedstawić w następującej postaci:

dla algorytmu PMA

$$\Delta T^n = R_{th} \cdot P \cdot \frac{h}{2 \cdot \tau} \left[2 \cdot \frac{\exp\left(-\frac{(n-1) \cdot h}{\tau}\right) - 1}{\exp\left(-\frac{h}{\tau}\right) - 1} - 1\right] \tag{21}$$

dla algorytmu FMA

$$\Delta T^n = R_{th} \cdot P \cdot \left[1 - \exp\left(-\frac{(n-0{,}5) \cdot h}{\tau}\right)\right] \tag{22}$$

oraz dla algorytmu IMA dowolnego rzędu

$$\Delta T^n = R_{th} \cdot P \cdot \left[1 - \exp\left(-\frac{n \cdot h}{\tau}\right)\right] \tag{23}$$

W celu zbadania stabilności rozważanych algorytmów na podstawie definicji, wyznaczono wartości graniczne wyrażeń (21)–(23) dla liczby kroków $n \to \infty$. Otrzymano następujące zależności:

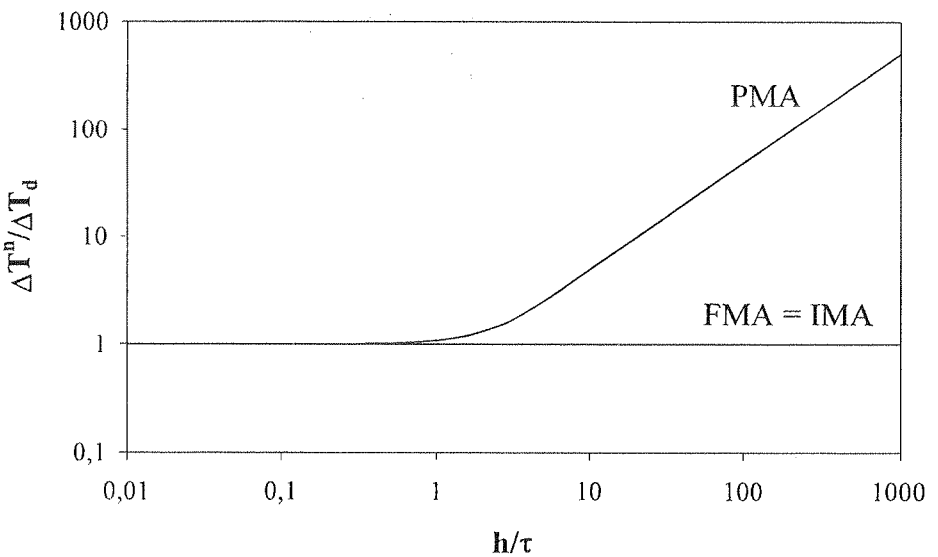
dla algorytmu PMA

$$\lim_{n\to\infty} \Delta T^n = 0{,}5 \cdot R_{th} \cdot P \cdot \frac{h}{\tau} \cdot \left[\frac{1+\exp\left(-\frac{h}{\tau}\right)}{1-\exp\left(-\frac{h}{\tau}\right)} \right] \tag{24}$$

dla algorytmów FMA oraz IMA

$$\lim_{n\to\infty} \Delta T^n = R_{th} \cdot P \tag{25}$$

Jak widać z powyższych zależności, wartości przyrostu temperatury w stanie ustalonym otrzymane za pomocą algorytmów FMA, PMA oraz IMA dla kroku obliczeń $h \to 0$, dążą do wartości równej $R_{th} \cdot P$, różniącej się o skończoną wartość od rozwiązania dokładnego. Tak więc, zgodnie z definicją, algorytmy PMA, FMA oraz IMA są stabilne.



Rys. 1. Zależność unormowanej wartości przyrostu temperatury w stanie ustalonym od ilorazu h/τ

Dodatkowo łatwo zauważyć, że dla algorytmu PMA wartość przyrostu temperatury wyznaczona z zależności (24) jest funkcją ilorazu długości kroku h i wartości termicznej stałej czasowej τ, natomiast, jak widać z wzoru (25), wyniki uzyskiwane w stanie ustalonym za pomocą algorytmów IMA i FMA nie zależą od długości kroku obliczeń i są dokładnie równe rozwiązaniu analitycznemu. Na rys. 1 przedstawiono obliczone za pomocą rozważanych algorytmów przebiegi nadwyżki temperatury w stanie ustalonym ΔT^n, unormowane względem wyznaczonej analitycznie ze wzoru (19) wartości dokład-

nej $\Delta T_d = R_{th} \cdot P$, w funkcji ilorazu długości kroku obliczeń h przez termiczną stałą czasową τ.

Jak widać, wyniki obliczeń za pomocą algorytmów IMA oraz FMA, dla dowolnej wartości ilorazu h/τ, są zgodne z wynikami obliczeń analitycznych, natomiast błąd obliczeń za pomocą algorytmu PMA jest nieakceptowalnie duży dla wartości kroku obliczeń h większego od termicznej stałej czasowej τ. Dla $h/\tau \gg 1$, wartość ilorazu $\Delta T^n / \Delta T_d$ w stanie ustalonym, rośnie proporcjonalnie do h/τ.

Z kolei w celu zbadania zbieżności rozważanych w pracy bezpamięciowych algorytmów splotowych należy, zgodnie z definicją [8], wyznaczyć zależności opisujące czasowy przebieg przyrostu temperatury w n-tym kroku czasowym $\Delta T^n(t)$ i wyznaczyć ich wartość graniczną dla $n \to \infty$. W tym celu w zależnościach (21)–(23) długość kroku h należy zastąpić ilorazem t/n. Wówczas zależności opisujące $\Delta T^n(t)$ mają następującą postać:

dla algorytmu PMA

$$\Delta T^n(t) = R_{th} \cdot P \cdot \frac{t}{2 \cdot \tau \cdot n} \left[2 \cdot \frac{\exp\left(-\frac{(n-1) \cdot t}{n \cdot \tau}\right) - 1}{\exp\left(-\frac{t}{n \cdot \tau}\right) - 1} - 1 \right] \tag{26}$$

dla algorytmu FMA

$$\Delta T^n(t) = R_{th} \cdot P \cdot \left[1 - \exp\left(-\frac{(n-0{,}5) \cdot t}{n \cdot \tau}\right) \right] \tag{27}$$

oraz dla algorytmu IMA

$$\Delta T^n(t) = R_{th} \cdot P \cdot \left[1 - \exp\left(-\frac{t}{\tau}\right) \right] \tag{28}$$

Wyznaczając z powyższych wzorów wartości graniczne przyrostu temperatury $\Delta T^n(t)$ dla $n \to \infty$, otrzymano dla wszystkich rozważanych w pracy algorytmów zależności o postaci identycznej z rozwiązaniem dokładnym, wyrażonym wzorem (19). A zatem, dla mocy o postaci opisanej wzorem (18), rozważane bezpamięciowe algorytmy splotowe są zbieżne dla dowolnego czasu $t \in (0; \infty)$.

Drugą funkcją testującą stabilność i zbieżność rozważanych algorytmów splotowych jest funkcja eksponencjalna o postaci

$$p(t) = \exp(-\lambda \cdot t) \tag{29}$$

gdzie λ jest liczbą zespoloną.

Dla rozważanego przebiegu mocy wyznaczony analitycznie przyrost temperatury $\Delta T(t)$ ma postać

$$\Delta T(t) = \frac{R_{th}}{1-\lambda\cdot\tau}\cdot\left[\exp(-\lambda\cdot t)-\exp\left(-\frac{t}{\tau}\right)\right] \tag{30}$$

Zależność analityczną opisująca przyrost temperatury dla n-tego kroku czasowego można przedstawić w postaci

$$\Delta T^n(t) = \frac{R_{th}}{1-\lambda\cdot\tau}\cdot\left[\exp(-n\cdot\lambda\cdot h)-\exp\left(-n\cdot\frac{h}{\tau}\right)\right] \tag{31}$$

Z kolei, wykorzystując do wyznaczenia wartości przyrostu temperatury ΔT^n rozważane algorytmy bezpamięciowe, łatwo można zauważyć, że wartość tego przyrostu jest sumą częściową ciągu geometrycznego. Stąd dla mocy o postaci danej wzorem (29), zależności opisujące wartość przyrostu temperatury w kroku n-tym, można przedstawić w następującej postaci:

dla algorytmu PMA

$$\Delta T^n = R_{th}\cdot\frac{h}{2\cdot\tau}\left[2\cdot\frac{3\cdot\exp\left(-\frac{(n+1)\cdot h}{\tau}\right)\cdot\exp(\lambda\cdot h)-\exp\left(-\frac{h}{\tau}\right)\cdot\exp(-(n-1)\lambda\cdot h)-\exp(-n\cdot\lambda\cdot h)-\exp\left(-\frac{n\cdot h}{\tau}\right)}{\exp\left(-\frac{h}{\tau}\right)\cdot\exp(\lambda\cdot h)-1}\right] \tag{32}$$

dla algorytmu FMA

$$\Delta T^n = R_{th}\cdot\left[1-\exp\left(-\frac{h}{2\cdot\tau}\right)\right]\cdot\exp\left(-\frac{(n-0{,}5\cdot\mathrm{h})}{\tau}\right)\cdot\left(1+\exp\left(-\frac{h}{2\cdot\tau}\right)\cdot\exp(-\lambda\cdot h)\right)\cdot$$
$$\cdot\frac{\exp\left(\frac{n\cdot h}{\tau}\right)\cdot\exp(-n\cdot\lambda\cdot h)-1}{\exp\left(\frac{h}{\tau}\right)\cdot\exp(-\lambda\cdot h)-1} \tag{33}$$

dla algorytmu IMA rzędu 0

$$\Delta T^n = R_{th}\cdot\left[1-\exp\left(-\frac{h}{\tau}\right)\right]\cdot\frac{\exp\left(\frac{n\cdot h}{\tau}\right)\cdot\exp(-\lambda\cdot h)-\exp(-n\cdot\lambda\cdot h)}{\exp\left(-\frac{h}{\tau}\right)-\exp(-\lambda\cdot h)} \tag{34}$$

oraz dla algorytmu IMA rzędu 1

$$\Delta T^n = R_{th} \cdot X \cdot \exp(-n \cdot \lambda \cdot h) \cdot \frac{1 - \exp\left(-\frac{n \cdot h}{\tau}\right) \cdot \exp(n \cdot \lambda \cdot h)}{1 - \exp\left(-\frac{h}{\tau}\right) \cdot \exp(\lambda \cdot h)} \tag{35}$$

gdzie

$$X = 1 - \frac{1 - \exp\left(-\frac{h}{\tau}\right)}{h} \cdot \tau + \exp(\lambda \cdot h) \left(\frac{1 - \exp\left(-\frac{h}{\tau}\right)}{h} \cdot \tau - \exp\left(-\frac{h}{\tau}\right) \right) \tag{36}$$

W celu zbadania stabilności algorytmów PMA, FMA oraz IMA dla mocy o postaci funkcji testowej danej wzorem (29) na podstawie definicji, wyznaczono wartości graniczne wyrażeń (32)–(35) dla liczby kroków $n \to \infty$. Otrzymano następujące zależności:
dla algorytmu PMA

$$\lim_{n \to \infty} \Delta T^n = R_{th} \cdot \frac{h}{2 \cdot \tau} \cdot \frac{1 + \exp\left(-\frac{h}{\tau}\right) \cdot \exp(\lambda \cdot h)}{1 - \exp\left(-\frac{h}{\tau}\right) \cdot \exp(\lambda \cdot h)} \cdot \lim_{n \to \infty} \exp(-n \cdot \lambda \cdot h) \tag{37}$$

dla algorytmu FMA

$$\lim_{n \to \infty} \Delta T^n = R_{th} \cdot \frac{\exp\left(-\frac{h}{2 \cdot \tau}\right) - \exp\left(-\frac{h}{\tau}\right)}{\exp(-\lambda \cdot h) - \exp\left(-\frac{h}{\tau}\right)} \cdot \left(1 + \exp\left(-\frac{h}{2 \cdot \tau}\right) \cdot \exp(-\lambda \cdot h)\right) \cdot \lim_{n \to \infty} \exp(-n \cdot \lambda \cdot h) \tag{38}$$

dla algorytmu IMA rzędu 0

$$\lim_{n \to \infty} \Delta T^n = R_{th} \cdot \frac{1 - \exp\left(-\frac{h}{\tau}\right)}{\exp(-\lambda \cdot h) - \exp\left(-\frac{h}{\tau}\right)} \cdot \lim_{n \to \infty} \exp(-n \cdot \lambda \cdot h) \tag{39}$$

oraz dla algorytmu IMA rzędu 1

$$\lim_{n \to \infty} \Delta T^n = \frac{R_{th} \cdot X}{1 - \exp(\lambda \cdot h) \cdot \exp\left(-\frac{h}{\tau}\right)} \cdot \lim_{n \to \infty} \exp(-n \cdot \lambda \cdot h) \tag{40}$$

Jak widać, wartości graniczne przyrostu temperatury ΔT^n zależą od wartości parametru λ. W zależności od znaku części rzeczywistej parametru λ otrzymuje się wartości graniczne ΔT^n równe odpowiednio nieskończoności — dla $\mathrm{Re}(\lambda)<0$, rezystancji termicznej R_{th} — dla $\mathrm{Re}(\lambda)=0$ oraz zero — dla $\mathrm{Re}(\lambda)>0$. Wartości te równe są odpowiednim rozwiązaniom analitycznym dla stanu ustalonego. A zatem zgodnie z definicją, algorytmy PMA, FMA oraz IMA rzędu 0 oraz 1 są stabilne.

Niezbędne do zbadania zbieżności rozważanych w pracy bezpamięciowych algorytmów splotowych zależności opisujące czasowy przebieg przyrostu temperatury w n-tym kroku czasowym $\Delta T^n(t)$ otrzymuje się podstawiając w zależnościach (32–35) iloraz t/n zamiast długości kroku h. Wówczas zależności opisujące $\Delta T^n(t)$ mają następującą postać:

dla algorytmu PMA

$$\Delta T^n(t) = R_{th}\cdot\frac{t}{2\cdot n\cdot\tau}\left[2\cdot\frac{3\cdot\exp\left(-\frac{t}{\tau}\cdot\left(1+\frac{1}{n}\right)\right)\cdot\exp\left(\lambda\cdot\frac{t}{n}\right)-\exp\left(-\frac{t}{n\cdot\tau}\right)\cdot\exp\left(-t\cdot\lambda\cdot\left(1-\frac{1}{n}\right)\right)-\exp(-\lambda\cdot t)-\exp\left(-\frac{t}{\tau}\right)}{\exp\left(-\frac{t}{n\cdot\tau}\right)\cdot\exp\left(\lambda\cdot\frac{t}{n}\right)-1}\right] \tag{41}$$

dla algorytmu FMA

$$\Delta T^n(t) = R_{th}\cdot\left[1-\exp\left(-\frac{t}{2\cdot n\cdot\tau}\right)\right]\cdot\exp\left(-\frac{t}{\tau}\cdot\left(1-\frac{1}{2\cdot n}\right)\right)\cdot$$

$$\cdot\left(1+\exp\left(-\frac{t}{n}\cdot\left(\frac{1}{2\cdot\tau}+\lambda\right)\right)\right)\cdot\frac{\exp\left(-t\left(\frac{1}{\tau}+\lambda\right)\right)-1}{\exp\left(-\frac{t}{n}\cdot\left(\frac{1}{\tau}+\lambda\right)\right)-1} \tag{42}$$

dla algorytmu IMA rzędu 0

$$\Delta T^n(t) = R_{th}\cdot\left[1-\exp\left(-\frac{t}{n\cdot\tau}\right)\right]\cdot\frac{\exp\left(-\frac{t}{\tau}\right)\cdot\exp\left(-\lambda\cdot\frac{t}{n}\right)-\exp(-t\cdot\lambda)}{\exp\left(-\frac{t}{n\cdot\tau}\right)-\exp\left(-\lambda\cdot\frac{t}{n}\right)} \tag{43}$$

oraz dla algorytmu IMA rzędu 1

$$\Delta T^n(t) = \frac{R_{th}\cdot X(t)\cdot\exp(-t\cdot\lambda)}{1-\exp\left(-\frac{t}{n}\cdot\left(\frac{1}{\tau}-\lambda\right)\right)}\cdot\left[1-\exp\left(-t\cdot\left(\frac{1}{\tau}-\lambda\right)\right)\right] \tag{44}$$

Wyznaczając z powyższych wzorów, wartości graniczne przyrostu temperatury $\Delta T^n(t)$, dla $n \to \infty$, otrzymano dla obu algorytmów zależności o postaci identycznej z rozwiązaniem dokładnym, wyrażonym wzorem (30). A zatem, rozważane bezpamięciowe algorytmy splotowe są zbieżne dla dowolnego czasu $t \in (0; \infty)$.

4. LOKALNY BŁĄD OBCIĘCIA ALGORYTMÓW SPLOTOWYCH

Zgodnie z definicją lokalny błąd obcięcia algorytmu wielokrokowego, jest błędem występującym w bieżącym kroku obliczeń [6]. A zatem, wyznaczając lokalny błąd obcięcia w kroku n-tym zakłada się zerową wartość błędu w kroku (n-1)-szym. W związku z tym i zgodnie z ogólną postacią bezpamięciowych algorytmów splotowych opisaną wzorem (3), lokalny błąd obcięcia w kroku n-tym może wynikać tylko z niedokładnego wyznaczenia wartości S_i^n. Rozważając uproszczony przypadek jednej termicznej stałej czasowej błąd ten można wyznaczyć obliczając różnicę między rozwiązaniem analitycznym, a rozwiązaniem otrzymanym za pomocą algorytmu splotowego.

Dla algorytmu PMA lokalny błąd obcięcia ε_{TMPMA} wynika z zastosowania do obliczeń algorytmu trapezów i dany jest zależnością [6]

$$\varepsilon_{TPMA} = -\frac{1}{12} \cdot h^3 \cdot \max\left(\frac{d^2}{dt^2}[Z'(t_n - t) \cdot p(t)]\right) \tag{45}$$

gdzie $t \in \langle t_{n-1}, t_n \rangle$, natomiast funkcja $\max(x)$ oznacza największą wartość x w rozważanym przedziale.

Z kolei, jak wynika z wcześniejszych rozważań, algorytm FMA jest dokładny dla mocy stałej. Z drugiej strony, zgodnie z twierdzeniem Taylora, każdą funkcję posiadającą ciągłą pochodną w rozważanym przedziale czasu można aproksymować funkcją stałą, wiedząc, że błąd takiej aproksymacji jest proporcjonalny do pierwszej pochodnej tej funkcji. A zatem, lokalny błąd obcięcia algorytmu FMA można wyznaczyć jako różnicę między wartością S_i^n obliczoną analitycznie oraz numerycznie, dla mocy, będącej resztą pierwszego rzędu rozwinięcia w szereg Taylora rozważanej funkcji $p(t)$. Otrzymana w ten sposób zależność opisująca lokalny błąd obcięcia ε_{TFMA} algorytmu FMA ma postać [7]

$$\varepsilon_{TFMA} = R_{th} \cdot \max\left(\frac{dp(t)}{dt}\right) \cdot \left[h \cdot \left(0{,}5 \cdot \left(1 - \sqrt{\xi}\right) + \xi\right) + \tau \cdot (\xi - 1)\right] \tag{46}$$

gdzie parametr ξ opisany jest wzorem (4).

Dla algorytmu IMA rzędu 0 zależność opisująca lokalny błąd obcięcia otrzymano jako różnicę między wartością całki S_i^n obliczoną analitycznie ze wzoru (10) oraz na podstawie wzoru (15). Błąd ten opisany jest zależnością

$$\varepsilon_{TIMA0} = R_{th} \cdot \max\left(\frac{dp(t)}{dt}\right) \cdot [h \cdot + \tau \cdot (\xi - 1)] \tag{47}$$

Z kolei dla algorytmu IMA rzędu 1, lokalny błąd obcięcia opisany jest wzorem

$$\varepsilon_{TIMA1} = R_{th} \cdot \tau \cdot \max\left(\frac{d^2p(t)}{dt^2}\right) \cdot [h \cdot (3 \cdot \xi + 1) - 2 \cdot \tau\,(\xi - 1)] \tag{48}$$

Jak widać z postaci wzorów (45)–(48), lokalny błąd obcięcia rozważanych algorytmów splotowych jest funkcją długości kroku obliczeń, parametrów modelu termicznego oraz mocy wydzielanej w elemencie. Przykładowo, dla mocy opisanej funkcją liniową o postaci $p(t) = b \cdot t$, lokalne błędy obcięcia algorytmów PMA, FMA oraz IMA rzędu 0 i 1 opisane są odpowiednio zależnościami (49)–(52).

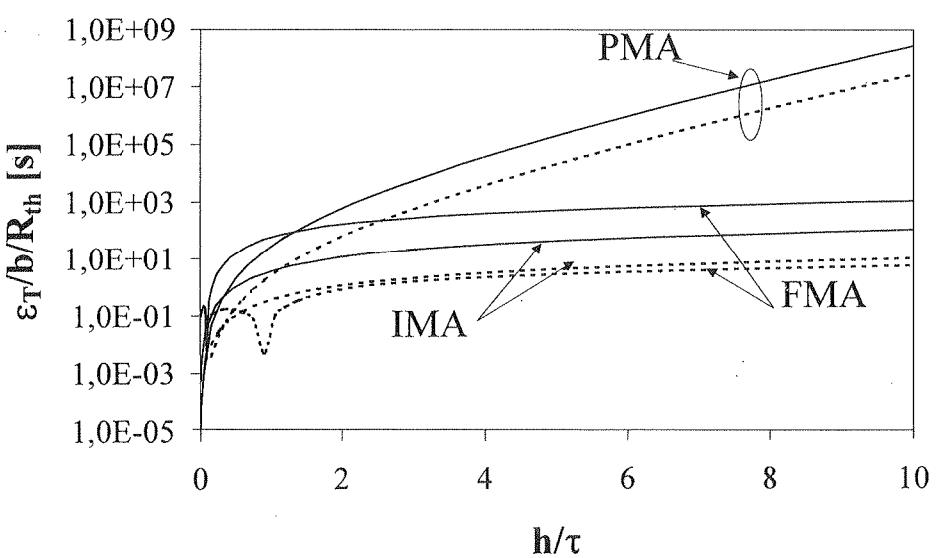
$$\varepsilon_{TPMA} = \frac{1}{12} \cdot \frac{R_{th} \cdot h^3 \cdot b}{\tau^2} \cdot \exp\left(\frac{h}{\tau}\right) \cdot \left(2 + \frac{t_n}{\tau}\right) \tag{49}$$

$$\varepsilon_{TFMA} = R_{th} \cdot b \cdot \left[h \cdot \left(0{,}5 \cdot \left(1 - \sqrt{\xi}\right) + \xi\right) + \tau \cdot (\xi - 1)\right] \tag{50}$$

$$\varepsilon_{TIMA0} = R_{th} \cdot b \cdot [h \cdot + \tau \cdot (\xi - 1)] \tag{51}$$

$$\varepsilon_{TIMA1} = 0 \tag{52}$$

Na rys. 2 przedstawiono zależność lokalnych błędów obcięcia rozważanych algorytmów, dla mocy opisanej funkcją liniową, w funkcji ilorazu h/τ przy dwóch wybranych



Rys. 2. Zależność ilorazu lokalnych błędów obcięcia algorytmów PMA, FMA oraz IMA od wartości h/τ

wartościach termicznej stałej czasowej $\tau = 1$ s (linie ciągłe) i $\tau = 10$ s (linie kreskowe). Dodatkowo, dla algorytmu PMA przyjęto $t_n/\tau = 10$. Zgodnie z wynikami przedstawionymi w pracy [7], wartość lokalnego błędu obcięcia algorytmu PMA rośnie wraz ze wzrostem wartości ilorazu t_n/τ.

Jak widać na rys. 2, w przypadku mocy opisanej funkcją liniową, dla małych, w stosunku do termicznej stałej czasowej, wartości kroku czasowego, lokalny błąd obcięcia algorytmu PMA jest najmniejszy spośród rozważanych algorytmów splotowych. Jednak przy wzroście wartości ilorazu h/τ błąd algorytmu PMA w relacji do błędu algorytmów FMA i IMA rośnie. Po upływie kilku termicznych stałych czasowych, błąd algorytmów FMA i IMA jest o kilka rzędów mniejszy od algorytmu PMA. A zatem, stosując w obliczeniach przyrostu temperatury ΔT długości kroku znacznie większe od termicznej stałej czasowej τ, przy wykorzystaniu algorytmu PMA w każdym kroku obliczeń lokalny błąd obcięcia jest znacznie większy niż przy zastosowania algorytmu FMA lub IMA.

5. OBSZARY STABILNOŚCI BEZWZGLĘDNEJ

W poprzednim rozdziale wykazano, że wszystkie rozważane w pracy algorytmy splotowe są zbieżne i stabilne dla wartości kroku obliczeń h dążącej do zera. W praktyce, obliczenia prowadzone są przy skończonej wartości kroku czasowego, dlatego istotne jest określenie zakresu dopuszczalnych zmian wartości h, dla których algorytm będzie stabilny. Zgodnie z definicją podaną w rozdziale 3, algorytm jest stabilny, jeżeli po wykonaniu dowolnie dużej ilości kroków obliczeń, błąd całkowity algorytmu ma skończoną wartość. Aby warunek ten był spełniony wystarczy, że lokalny błąd obcięcia algorytmu jest skończony oraz nie narasta z kroku na krok wartość błędu całkowitego.

Wzory opisujące zależność lokalnego błędu obcięcia rozważanych w pracy algorytmów splotowych od długości kroku przedstawiono w rozdziale 4. W celu zbadania dla jakich wartości kroku obliczeń błąd całkowity algorytmu nie narasta z kroku na krok, wyznaczono wzory opisujące błąd całkowity rozważanych algorytmów w kroku n-tym i $(n+1)$ pierwszym, a następnie zbadano, dla jakich wartości h iloraz tych błędów jest nie większy od jedności.

W przypadku, gdy moc opisana jest funkcją stałą, lokalny błąd obcięcia bezpamięciowych algorytmów splotowych będzie miał taką samą wartość ε_T w każdym kroku czasowym. Stąd błąd całkowity (δ_n w kroku n-tym będzie sumą n-elementowego ciągu geometrycznego, którego pierwszym wyrazem jest ε_T a ilorazem ξ. A zatem, wartość błędu całkowitego można wyznaczyć ze wzoru

$$\delta_n = \varepsilon_T \cdot \frac{1-\xi^n}{1-\xi} \tag{53}$$

Po wykonaniu dostatecznie dużej ilości kroków obliczeń ($n \to \infty$), całkowity błąd algorytmu splotowego opisany jest zależnością

$$\delta_n = \frac{\varepsilon_T}{1-\xi} \tag{54}$$

Jak widać, dla skończonej wartości lokalnego błędu obcięcia ε_T, błąd całkowity jest skończony. Dla ustalonej wartości ε_T, wartość δ_n maleje wraz ze wzrostem długości kroku obliczeń, ponieważ jednocześnie rośnie ξ.

Jak wynika z rozważań przedstawionych w rozdziale 4, dla mocy o postaci danej wzorem (29), lokalny błąd obcięcia wszystkich rozważanych w pracy algorytmów rośnie z kroku na krok m-krotnie, gdzie m wyrażone jest wzorem

$$m = \exp(-\lambda \cdot h) \tag{55}$$

A zatem, błąd całkowity w kroku n-tym jest opisany wzorem

$$\delta_n = \delta_1 \cdot \frac{1-m^n \cdot \xi^n}{1-m \cdot \xi} \tag{56}$$

gdzie δ_1 oznacza błąd wyznaczenia wartości splotu w pierwszym kroku obliczeń.

Jak widać, prawa strona wyrażenia (56) dla n dążącego do nieskończoności jest skończona, gdy iloczyn m oraz ξ jest mniejszy od 1 oraz δ_1 jest skończone.

Pierwszy z wymienionych warunków wymaga, aby m było mniejsze od odwrotności ξ, co oznacza, że $\mathrm{Re}(-\lambda)$ musi być mniejsze od odwrotności termicznej stałej czasowej, a w przypadku występowania wielu stałych czasowych, od odwrotności najdłuższej termicznej stałej czasowej. Sprawdzenie drugiego warunku wymaga zbadania wartości błędu δ_1. Dla rozpatrywanej mocy, błąd ten opisany jest zależnościami:
dla algorytmu PMA:

$$\delta_1 = R_{th} \cdot \exp(-\lambda \cdot h) \cdot \left\{ \frac{h}{2 \cdot \tau} \cdot \left[1 + \exp\left(-\frac{h}{\tau}\right) \cdot \exp(\lambda \cdot h) \right] + \right.$$

$$\left. + \frac{1}{1-\lambda \cdot \tau} \cdot \left[1 - \exp\left(-\frac{h}{\tau}\right) \cdot \exp(\lambda \cdot h) \right] \right\} \tag{57}$$

dla algorytmu FMA:

$$\delta_1 = R_{th} \cdot \left[\exp(-\lambda \cdot h) \cdot \left\{ \left[1 + \exp\left(-\frac{h}{2 \cdot \tau}\right) \right] - \frac{1}{1-\lambda \cdot \tau} \cdot \left[1 - \exp\left(-\frac{h}{\tau}\right) \cdot \exp(\lambda \cdot h) \right] \right\} + \right.$$

$$\left. + \exp\left(-\frac{h}{2 \cdot \tau}\right) - \exp\left(-\frac{h}{\tau}\right) \right] \tag{58}$$

dla algorytmu IMA rzędu 0:

$$\delta_1 = -\frac{R_{th}}{1-\lambda \cdot \tau} \cdot \exp(-2 \cdot \lambda \cdot h) \cdot \left\{ \lambda \cdot \tau + \exp\left(-\frac{h}{\tau}\right) \cdot [1 - \lambda \cdot \tau - \exp(\lambda \cdot h)] \right\} \tag{59}$$

oraz dla algorytmu IMA rzędu 1:

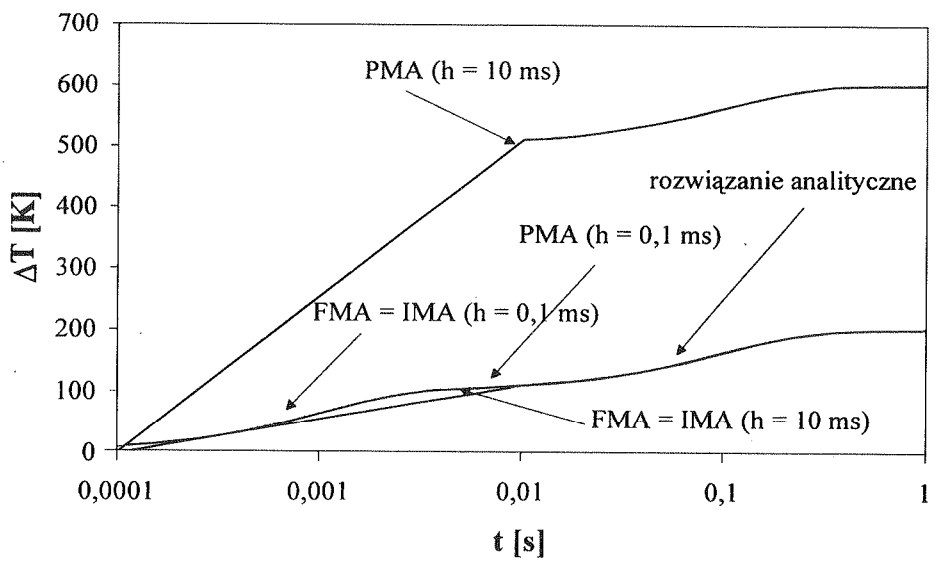
$$\delta_1 = R_{th} \cdot \exp(-\lambda \cdot h) \cdot \tau \cdot \left\{ -\frac{\lambda \cdot \tau}{1-\lambda \cdot \tau} - \frac{1-\exp\left(-\frac{h}{\tau}\right)}{h} \cdot [1-\exp(\lambda \cdot h)] \right\} \tag{60}$$

Jak widać zależności (57–60) przyjmują skończone wartości, gdy $\lambda \neq \frac{1}{\tau}$ oraz $\lambda < \infty$.

A zatem obszary stabilności bezwzględnej rozpatrywanych w pracy algorytmów splotowych są identyczne i obejmują obszar, w którym spełnione są łącznie 3 nierówności: $\lambda > -\frac{1}{\tau}$, $\lambda \neq \frac{1}{\tau}$ oraz $\lambda < \infty$. Warunki te są zawsze spełnione dla występujących w praktyce przebiegów mocy.

6. WYNIKI OBLICZEŃ I WNIOSKI

W celu zilustrowania rozważań teoretycznych, zamieszczonych w rozdziałach 3–5, przeprowadzono szereg obliczeń wartości splotu uzyskanych za pomocą algorytmów PMA oraz FMA i zweryfikowano ich dokładność za pomocą rozwiązań analitycznych.

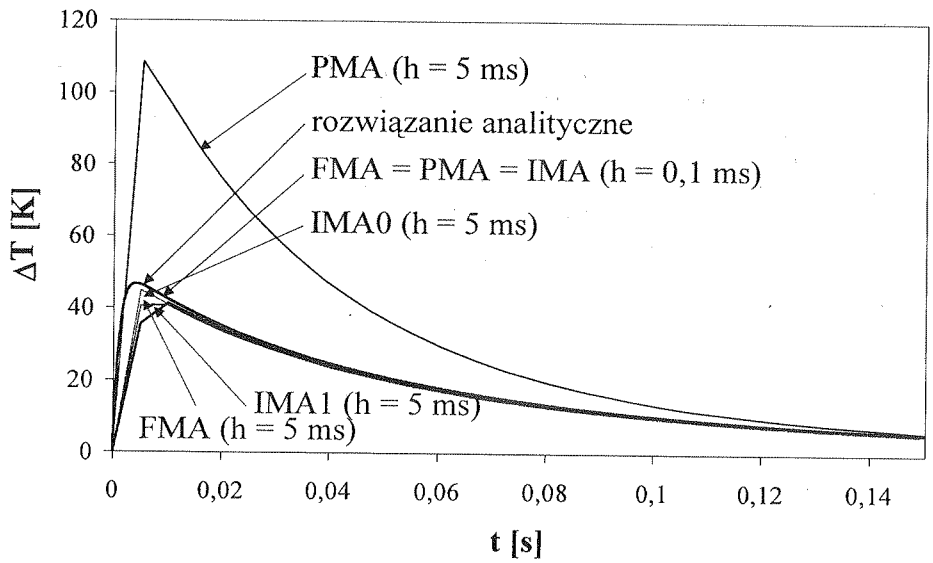


Rys. 3. Wyniki obliczeń przyrostu temperatury za pomocą algorytmu PMA, FMA oraz IMA

Przykładowo, na rys. 3 przedstawiono wyniki obliczeń przebiegu przyrostu temperatury wnętrza elementu pobudzonego przebiegiem mocy $p(t) = 2 \cdot 1(t)$ W, otrzymane z rozwiązania analitycznego oraz za pomocą algorytmów PMA i FMA oraz IMA dla

dwóch wartości kroku obliczeń. W obliczeniach przyjęto następujące wartości parametrów przejściowej impedancji termicznej $Z(t)$: $R_{th} = 100$ K/W, $a_1 = 0{,}5$, $a_2 = 0{,}5$, $\tau_1 = 1$ ms, $\tau_2 = 0{,}1$ s, $N = 2$. Uzyskano dobrą zgodność wyników otrzymanych analitycznie z wynikami wyznaczonymi algorytmem FMA oraz IMA dla obu zastosowanych wartości kroku obliczeń. Dla algorytmu PMA uzyskano także dobrą zgodność wyników obliczeń z rozwiązaniem analitycznym, ale tylko dla kroku obliczeń krótszego od termicznej stałej czasowej τ_1. Wraz ze wzrostem wartości kroku obliczeń zgodność ta ulega pogorszeniu i dla $h = 10$ ms różnica między rozwiązaniem analitycznym, a numerycznym przekraczała 200%.

Z kolei na rys. 4 przedstawiono wyniki obliczeń przebiegu nadwyżki temperatury wnętrza elementu pobudzonego mocą o przebiegu opisanym wzorem (29), w którym $\lambda = 30$, otrzymane za pomocą rozważanych w pracy algorytmów splotowych oraz z rozwiązania analitycznego. Zastosowano identyczne wartości parametrów $Z(t)$ jak w przykładzie poprzednim.



Rys. 4. Wyniki obliczeń przyrostu temperatury za pomocą algorytmu PMA, FMA oraz IMA

Jak widać, uzyskano dobrą zgodność wyników obliczeń za pomocą wszystkich rozważanych algorytmów i rozwiązania analitycznego przy kroku czasowym równym 0,1 ms, natomiast dla kroku równego 10 ms wyniki uzyskane za pomocą algorytmu PMA znacznie różnią się od wyników obliczeń analitycznych oraz od wyników otrzymanych przy wykorzystaniu algorytmów FMA i IMA.

Przedstawione wyniki obliczeń potwierdzają słuszność przytoczonych w pracy rozważań teoretycznych, z których wynika, że algorytmy FMA oraz IMA, w przeciwieństwie do PMA, umożliwiają uzyskanie zadawalająco dokładnych wyników analiz termicznych przy zastosowaniu kroku obliczeń znacznie dłuższego od najkrótszej termicznej stałej czasowej. Jest to szczególnie istotne, ponieważ w rzeczywistym

elemencie poszczególne termiczne stałe czasowe mogą różnić się nawet o kilka rzędów wielkości, a w związku z tym zastosowanie w analizie termicznej algorytmu PMA, spowoduje nieuzasadniony wzrost czasu trwania obliczeń. A zatem, chociaż wszystkie rozważane algorytmy są zbieżne i stabilne, to do analizy termicznej rzeczywistych elementów półprzewodnikowych należy stosować algorytm FMA lub IMA.

LITERATURA

1. J. Zarębski: *Wielkosygnałowa analiza oddziaływań elektrotermicznych w układach półprzewodnikowych.* Rozprawa doktorska. Politechnika Gdańska, Gdańsk, 1986.
2. A. Nowakowski: *Badanie procesów termicznych w przyrządach półprzewodnikowych.* Zesz. Nauk. Polit. Gdańskiej, Elektronika LX, nr 389, 1984.
3. W.J. Stepowicz, W. Janke: *Modele termiczne elementów półprzewodnikowych.* Zeszyty Naukowe Politechniki Gdańskiej, nr 348, Elektronika LVI, 1982, s. 3.
4. J. Zarębski: *Modelowanie, symulacja i pomiary przebiegów elektrotermicznych w elementach półprzewodnikowych i układach elektronicznych.* Prace Naukowe Wyższej Szkoły Morskiej w Gdyni, Gdynia, 1996.
5. J. Zarębski: *The Time Dependence Calculations of the Junction Temperature of Semiconductor Devices.* Int. Journal of Microcomputer Applications, Vol. 13, No. 2, June 1994, p. 62.
6. W. Janke, G. Blakiewicz: *Semi analytical recursive algorithms for convolution calculations.* IEE Proc.- Circuits Devices Syst., Vol. 142, No. 2, 1995.
7. K. Górecki, J. Zarębski: *Problem stabilności i zbieżności wybranych algorytmów splotowych do analizy termicznej układów elektronicznych.* V Konferencja Naukowo-Techniczna Zastosowanie Komputerów w Elektrotechnice ZKwE'2000, Poznań-Kiekrz 2000, t. 2, s. 383.
8. L.O. Chua, P.M. Lin: *Komputerowa analiza układów elektronicznych.* Warszawa, WNT, 1981.
9. Poradnik inżyniera. Matematyka. Praca zbiorowa, Warszawa, WNT, 1987.

K. GÓRECKI, J. ZARĘBSKI

INVESTIGATION OF THE PROPERTIES OF THE SELECTED CONVOLUTION ALGORITHMS FOR THE THERMAL ANALYSIS OF ELECTRONIC CIRCUITS

Summary

In the paper the analysis of the basic properties of the selected memoryless convolution algorithms is presented. The stability and convergence of these algorithms was investigated by the use of power in the form of two testing functions: the exponential function and Heaviside jump. The dependencies describing the value of the local truncation error and the areas of the absolute stability of the considered algorithms are elaborated as well. The theoretical considerations were illustrated by the numerical examples.

Key words: analysis of electronic circuits, convolution algorithms

SPIS TREŚCI

K. Olejarczyk: Analiza stabilności modulatorów Delta-Sigma z lokalnymi pętlami sprzężenia zwrotnego 461
H. Selvaraj, P. Sapiecha: Miary jakości syntezy logicznej w strukturach FPGA 479
J. Siuzdak, R. Nowak, T. Czarnecki: Homodynowy układ detekcji słabych sygnałów optycznych do współpracy z monochromatorami 493
K. Surowiak, D. Bochenek, S.V. Gavrilyachenko, M.F. Kupriyanov: Stabilność częstotliwości rezonansowej rezonatorów z ceramiki piezoelektrycznej a ich struktura domenowa 511
K. Wiatr, E. Jamro: Implementacja algorytmu konwolucji 2D w układach specjalizowanych VLSI oraz w procesorach ogólnego przeznaczenia i sygnałowych 553
M. Solecki, Z. Felendzer: Metody modelowania sprzężenia podłożowego w układach scalonych . . 589
J. Zarębski, K. Górecki: Wielosygnałowy dynamiczny makromodel tranzystora Darlingtona mocy do analizy elektrotermicznych stanów przejściowych 615
K. Górecki, J. Zarębski: Badanie właściwości wybranych algorytmów splotowych do analizy termicznej układów elektronicznych 633
Informacje dla Autorów 652
Roczny spis treści 655
Lista recenzentów 659

CONTENTS

K. Olejarczyk: Analysis of stability of multiple-feedback Delta-Sigma modulators with leakage . . 641
H. Selvaraj, P. Sapiecha: An efficiency measure of FPGA based logic synthesis tools 479
J. Siuzdak, R. Nowak, T. Czarnecki: Homodyne system for week optical signal detection 493
K. Surowiak, D. Bochenek, S.V. Gavrilyachenko, M.F. Kupriyanov: Stability of resonance frequency of the piezoceramic resonators and their domain structure 511
K. Wiatr, E. Jamro: Implementation of 2D convolution algorithms in dedicated VLSI devices and generalpurpose processors and DSPs 553
M. Solecki, Z. Felendzer: Modeling methods of substrate coupling in integrated circuits 589
J. Zarębski, K. Górecki: A dynamic power Darlington transistor macromatel for the analysis of electrothermal transients 615
K. Górecki, J. Zarębski: Investibation of the properties of the selected convolution algorithms for the thermal analysis of electronic circuits 633
Information for the Authors 652
Year contents of Electromics and Telecommunications Quarterly 655
The List of the reviewers 659

INFORMACJE DLA AUTORÓW

Redakcja przyjmuje do publikowania prace oryginalne, przeglądowe i monograficzne wchodzące w zakres szeroko pojętej elektroniki. Ponieważ KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJI jest czaspismem Komitetu Elektroniki i Telekomunikacji Polskiej Akademii Nauk, w związku z tym na jego łamach znajdują się prace naukowe dotyczące podstaw teoretycznych i zastosowań z zakresu elektroniki, telekomunikacji, mikroelektroniki, optoelektroniki, radiotechniki i elektroniki medycznej.
Artykuły powinno charakteryzować oryginalne ujęcie zagadnienia, własna klasyfikacja, krytyczna ocena (teorii lub metod), omówienie aktualnego stanu, lub postępu danej gałęzi techniki oraz omówienie perspektyw rozwojowych.
Artykuły publikowane w innych czasopismach nie mogą być kierowane do druku w Kwartalniku Elektroniki i Telekomunikacji w drugiej kolejności zgłoszenia.
Objętość artykułu nie powinna przekraczać 30 stron po około 1800 znaków na stronie.

Wymagania podstawowe: Artykuły należy nadsyłać w maszynopisie pisnaym jednostronnie lub na wyraźnym czarno-białym wydruku komputerowym, w 2 egzemplarzach, w języku polskim lub angielskim wybranym przez autora. Tekst artykułu musi być poprzedzony tytułem pracy, imieniem i nazwiskiem Autora wraz z podaniem miejsca jego pracy. Wszystkie strony muszą mieć numerację ciągłą.

Sposób pisania tekstu: Tekst powinien być pisany bez używania wyróżnień, a w szczególności nie dopuszcza się spacjowania, podkreślania i pisania tekstu dużymi literami z wyjątkiem wyrazów, które umownie pisze się dużymi literami (np. FORTRAN). Proponowane wyróżnienia Autor może zaznaczyć w maszynopisie zwykłym ołówkiem na pomocą przyjętych znaków adiustacyjnych np. podkreślenie linią przerywaną oznacza spacjowanie (rozstrzelenie), podkreślenie linią ciągłą — pogrubienie, podkreślenie wężykiem — kursywa. Tekst powinien być napisany z podwójnym odstępem między wierszami, tytuły i podtytuły małymi literami. Marginesy z każdej strony powinny mieć około 35 mm. Przy podziale pracy na rozdziały i podrozdziały cyfrowe ich oznaczenia nie powinny być większe niż III stopnia (np. 4.1.1).

Sposób pisania tablic: Tablice powinny być pisane na oddzielnych stronach. Tytuły rubryk pionowych i poziomych powinny być napisane małymi literami z podwójnym odstępem między wierszami. Przypisy (notki) dotyczące tablic należy pisać bezpośrednio pod tablicą. Tablice należy numerować kolejno liczbami arabskimi, u góry każdej tablicy podać tytuł. Tablice umieścić na końcu maszynopisu. Przyjmowane są tablice algorytmów i programy na wydrukach komputerowych. W tym przypadku zachowany jest ich oryginalny układ.

Sposób pisania wzorców matematycznych: Rozmieszczenie znaków, cyfr, liter i odstępów powinno być zbliżone do rozmieszczenia elementów druku. Wskaźniki i wykładniki potęg powinny być napisane wyraźnie i być prawidłowo obniżone lub podwyższone w stosunku do linii wiersza podstawowego. Znaki nad literami i cyframi, całkami i in. symbolami (strzałki, linie, kropki, daszki) powinny być umieszczone dokładnie nad tymi elementami, do których się odnoszą. Numery wzorów cyframi arabskimi powinny być kolejne i umieszczone w nawiasach okrągłych z prawej strony. Nazwy jednostek, symbole literowe i graficzne powinny być zgodne z wytycznymi IEE (International Electronical Commision) oraz ISO (International Organization of Standarization).

Powołania: Powołania na publikacje powinny być umieszczone na ostatnich stronach tekstu pod tytułem „Bibliografia", opatrzone numeracją kolejną bez nawiasów. Numeracja ta powinna być zgodna z odnośnikami w tekście artykułu. Przykłady opisu publikacji:
— periodycznej: F. Valdoni: A new milimetre wave satelite. E.T.T. 1990, vol. 2, no 5, p. 553
— nieperiodycznej: K. Andersen: A resource allocation framework. XVI International Symposium, Stockholm (Sweden), may 1991, paper A 2.4
— książki: Y.P. Tvidis: Operation and modeling of the MOS transistors. New York, Mc Graw-Hill, 1987, pp. 141–148

Materiały ilustracyjne: Rysunki powinny być wykonane wyraźnie, na papierze gładkim lub milimetrowym w formacie nie mniejszym niż 9×12 cm. Mogą być także w postaci wydruku komputerowego. Fotografie lub diapozytywy przyjmowane są raczej czarno-białe w formacie nie przekraczającym 10×15 cm. Na marginesie każdego rysunku i na odwrocie fotografii powinno być napisane ołówkiem imię i nazwisko Autora oraz skrót tytułu artykułu, do którego są przeznaczone. Spis podpisów pod rysunki i fotografie powinien być umieszczony na oddzielnej stronie.

Streszczenia: Do każdego artykułu musi być dołączone obszerne — do 60 wierzy — streszczenie z tytułem artykułu. Streszczenia (Summary) muszą być w języku polskim i angielskim. Streszczenie powinno wyjaśniać główny cel pracy, wskazywać korzyści i ograniczenia, możliwe zastosowania i zalecenia dla dalszego rozwoju danej gałęzi techniki. Pod streszczeniami powinny być podane słowa kluczowe.

Autorowi przysługuje bezpłatnie 20 odbitek artykułu. Dodatkowe egzemplarze odbitek, lub cały zeszyt Autor może zamówić u wydawcy na własny koszt.

Autora obowiązuje korekta autorska, którą powinien wykonać w ciągu 3 dni od daty otrzymania tekstu z Redakcji zwrócić osobiście, lub listownie pod adresem Redakcji. Korekta powinna być naniesiona na przekazanych Autorowi szpaltach na marginesach ew. na osobnym arkuszu w przypadku uzupełnień tekstu większych niż dwa wiersze. W przypadku nie zwrócenia korekty w terminie, korektę przeprowadza Redakcja Techniczna Wydawcy.

Redakcja prosi Autorów o powiadomienie ją o zmianie miejsca pracy i adresu prywatnego.

INFORMATION FOR THE AUTORS OF K.E.T.

The KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJI (K.E.T.) — ELECTRONICS AND TELECOMMUNICATION QUARTERLY is a journal of the Committee for Electronics and Telecommunication of the Polish Academy of Sciences and is a direct continuation for 46 years of the quarterly ROZPRAWY ELEKTROTECHNICZNE, that has been published also under the auspices of the Polish Academy of Sciences.

Its pages contain scientific articles concerned with theory and application of electronics, telecommunucation, microelectronics, optoelectronics, radioengineering and medical electronics.

Accordingly, its pages contain scientific articles, that are original works both contributory and concerned with comparative analyses of methods or systems synthetic reviews of a given field, state-of-art discussions of a given field, and presentations of developments ina given technology.

Regular papers are accepted for publication in Polish or English, but title, the estensive summary and figure caption must be bilingual. Bilingual summary must shown the significance of the results shown in the paper, emphasizing limitations and advantages, possible applications and recommendation for further works.

K.E.T. is meant for specialists and students with above subjects.

FORMAL PREPARATION OF MANUSCRIPT

Texts

Texts to be submitted to K.E.T. may be typed on one side of the paper only with double spacing between the lines and a 4 cm margin or using of wordprocessing with text formatters based on MS DOS, APPLE/MAC such as Microsoft word. In any case the text fotmatter must be specified on the Author/s, their Affiliations and relevant addresses.

All text pages should be numbered. Regular papers should have of no more than 30 pages.

Printing types and symblos

For letteral symbols, units and graphical symbols, the rtecommendation of IEE (International ElEctrotechnical Commission) and ISO (International Organisation of Standarisation) must be followed.

- In case of manuscripts conventionally typed (i.e. without wordprocessing), letters to be printed in italic should be underlined; letter to be printed in bold should be double interlined.
- Formulae and footnotes should be numbered in accordance with the quotation order followed in the text.
- The serial number of each formulae should be written at right side, between round brackets.
- Tje serial number of footnotes should be written between round brackets and upper.
- Mathematical detials, anciliary to the main discussion of the paper may be included in one or more appendixes, that are printed after conclusions and before references.

In manuscript containing lot of synmbols, a glossary may be included after introduction.

References

References are usually gathered at the end of the text and numbered according to the order of quotation in the text. The serial numbers should be written without square brackets (i.e. 1, 2...)

Illustrations

The originals of illustrations may be either drawings or photographs (black or colour). Each copy of paper must include two complete sets of illustrations.

All rights reserved. No part of the publication may be reproduced, stored in a retrival system or transmitted, in any form or by any means: electronic, photocopying, recording or otherwise, without the prior permission of the copyright owner — Redaction K.E.T.

Subscription for external subscribers:

The promotional subscription price in 2000 is $ 90 including postage for institutions. Subscriptions should be sent to the publisher, Polish Scientific Publishers PWN LTd., Journal Division, Miodowa 10, 00-250 Warsaw, POLAND, fax. (48) (22) 826 09 50, (48) (22) 826 71 63, with a copy by fast to Electronics and Telecommunications Quarterly 00-665 Warsaw, ul. Nowowiejska 15/19, p. 470. Subscription is occupated after showing a cheque for transfer documents. Our bank account is as follows: Bank account: PBK VIII O/Warszawa nr 11101037-1052-2700-1-43. At subscripber's request this journal will be air mailed at additional postage to 50% dross proce to European countries and 65% overseas.

Subscription orders available through the local press distriution or through the Foreign Trade Enterprise ARS Polona, 00-068 Warszawa, Krakowskie Przedmieście 7, Poland.

Ruch S.A. fulfils foreign customers orders, starting from any issue in the calendary year: (48) (22) 620 120 39, fax: (48) (22) 620 17 62.

LISTA RECENZENTÓW KWARTALNIKA ELEKTRONIKI I TELEKOMUNIKACJI

Dr inż. *Jarosław Arabas*
Politechnika Warszawska,
Instytut Systemów Elektronicznych

Dr hab. inż. Grzegorz Ciesielski
Politechnika Łódzka
Instytut Elektrotechniki Teoretycznej,
Metrologii i Materiałoznawstwa

Prof. dr hab. inż. *Janusz A. Dobrowolski*
Politechnika Warszawska,
Instytut Systemów Elektronicznych

Dr inż. *Jarosław Domaszewicy*
Politechnika Warszawska,
Instytut Telekomunikacji

Dr inż. *Przemysław Dymarski*
Politechnika Warszawska,
Instytut Telekomunikacji

Dr inż. *Zbigniew Gajo*
Politechnika Warszawska,
Instytut Systemów Elektronicznych

Prof. dr hab. inż. *Bogdan Galwas*
Politechnika Warszawsk,
Instytut Mikroelektroniki i Optoelektroniki

Mgr inż. *Marcin Golański*
Politechnika Warszawska
Instytut Telekomunikacji

Dr inż. *Konrad Hejn*
Politechnika Warszawska
Instytut Systemów Elektronicznych

Dr hab. inż. *Ewa Hermanowicz*
Politechnika Gdańska
Wydział Elektroniki i Telekomunikacji
i Informatyki

Prof. dr hab. inż. *Ryszard Jachowicz*
Politechnika Warszawska,
Instytut Systemów Elektronicznych

Dr inż. *Andrzej Jakubiak*
Politechnika Warszawska,
Instytut Telekomunikacji

Prof. dr hab. *Andrzej Jakubowski*
Politechnika Warszawska
Instytut Mikroelektroniki
i Optoelektroniki

Dr inż. *Stanisław Jankowski*
Politechnika Warszawskaż
Instytut Systemów Elektronicznych

Dr. hab. inż. *Swietlana Januszkiewicz*
Politechnika Szczecińska,
Wydział Informatyki

Dr inż. *Krzysztof Jasiński*
Politechnika Warszawska,
Instytut Telekomunikacji

Prof. dr hab. inż. *Janusz Jaworski*
Politechnika Lubelska,
Katedra Automatyki

Dr hab. inż. *Wojciech Kabaciński*
Politechnika Poznańska,
Instytut Elektroniki i Telekomunikacji

Prof. dr hab. inż. *Zbigniew Kaczkowski*
Polska Akademia Nauk,
Instytut Fizyki, Warszawa

Dr inż. *Ryszard Kisiel*
Politechnika Warszawska,
Instytut Mikroelektroniki i Optoelektroniki

Dr inż. *Tomasz Kosiło*
Politechnika Warszawska,
Instytut Radioelektroniki

Mgr inż. *Henryk Kowalski*
Politechnika Warszawska,
Instytut Informatyki

Prof. dr hab. inż. *Andrzej Kraśniewski*
Politechnika Warszawska,
Instytut Telekomunikacji

Prof. dr hab. inż. *Małgorzata Kujawińska*
Politechnika Warszawska,
Instytut Mikromechaniki i Fotoniki

Dr inż. *Sławomir Kukliński*
Politechnika Warszawska
Instytut Telekomunikacji

Dr hab. inż. *Wiesław Kuźmicz*
Politechnika Warszawska
Instytut Mikroelektroniki
i Optoelektroniki

Prof. dr hab. inż. *Juliusz Lech Kulikowski*
Polska Akademia Nauk, Warszawa
Instytut Biocybernetyki i Inżynierii Biomedycznej

Prof. dr hab. inż. *Zygmunt Kuśmierek*
Politechnika Łódzka, Instytut Elektrotechniki
Teoretycznej, Meteorologii i Materiałoznawstwa

Dr inż. *Lech Lewandowski*
Politechnika Warszawska
Instytut Systemów Elektronicznych

Prof. dr hab. inż. *Tadeusz Łuba*
Politechnika Warszawska,
Instytut Telekomunikacji

Mgr inż. *Grzegorz Mazur*
Politechnika Warszawska,
Instytut Informatyki

Prof. dr hab. inż. *Marian Miłek*
Politechnika Zielonogórska,
Instytut Metrologii Elektrycznej

Prof. dr hab. inż. *Tadeusz Morawski*
Politechnika Warszawska,
Instytutu Radioelektroniki

Prof. dr hab. inż. *Andrzej Napieralski*
Politechnika Łódzka,
Instytutu Elektroniki

Prof. dr *Wincenty Pajewski*
Polska Akademia Nauk
Instytut Podstawowych Problemów Techniki

Dr hab. inż. *Zdzisław Papir*
Akademia Górniczo-Hutnicza, Kraków
Katedra Telekomunikacji

Dr inż. *Andrzej Paszkiewicz*
Politechnika Warszawska,
Instytut Telekomunikacji

Dr hab. inż. *Jan Petykiewicz*
Politechnika Warszawska,
Instytut Fizyki

Dr. in. *Andrzej Pfitzner*
Politechnika Warszawska,
Instytut Mikroeletroniki i Optoelektroniki

Mgr inż. *Mariusz Rawski*
Politechnika Warszawska
Instytut Telekomunikacji

Dr inż. *Mirosław Rojewski*
Politechnika Gdańska,
Instytut Systemów Informacyjnych

Prof. nadzw. dr hab. *Stanisław Rosłoniec*
Politechnika Warszawska
Instytut Radioelektroniki

Prof. dr hab. inż. *Jerzy Siuzdak*
Politechnika Warszawska,
Instytut Telekomunikacji

Prof. nadzw. dr hab. *Władysław Skarbek*
Politechnika Warszawska
Instytut Radioelektroniki

Mgr inż. *Krzysztof Skowroński*
Politechnika Warszawska,
Instytut Systemów Elektronicznych

Prof. dr hab. inż. *Władysław Soluch*
Instytut Technologii Materiałów
Elektroniocznych

Prof. dr hab. inż. *Marek Stabrowski*
Politechnika Warszawska,
Instytutu Elektrotechniki Teoretycznej
i Miernictwa Elektryczneo

Dr hab. inż. *Witold J. Stepowicz*
Politechnika Gdańska
Wydział Elektroniki, Telekomunikacji
i Informatyki

Dr inż. *Jerzy Szabatin*
Politechnika Warszawska,
Instytut Systemów Elektronicznych

Dr inż. *Jacek Szymanowski*
Politechnika Warszawska,
Instytut Automatyki i Informatyki Stosowanej

Dr inż. *Lech Śliwa*
Politechnika Warszawska,
Instytut Podstaw Elekroniki

Prof. dr hab. inż. *Leszek Trybus*
Politechnika Rzeszowska,
Katedra Automatyki i Informatyki

Prof. dr hab. inż. *Aleksander Urbaś*
Politechnika Warszawska,
Instytut Systemów Elektronicznych

Prof. dr hab. inż. *Jacek Wojciechowski*
Politechnika Warszawska,
Instytut Systemów Elektronicznych

Prof. nzw.,
dr hab. inż. *Andrzej Wojtkiewicz*
Politechnika Warszawska,
Instytut Systemów Elektronicznych

ERRATA

In the article L. Tomawski, M. Mańka „Resonance phenomena in chosen II and III order circuits" issued in Electronic and Telecommunications Quarterly **46**, 3, 2000, pp. 42, 44 errors appeared in Figures 1, 2a and 3a. We apologise and enclose the correct schemes of the circuits.

Jest tak:

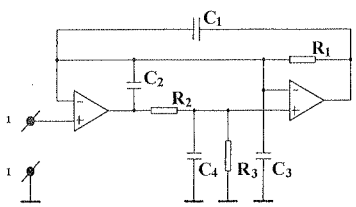


Fig. 1

Powinno być:

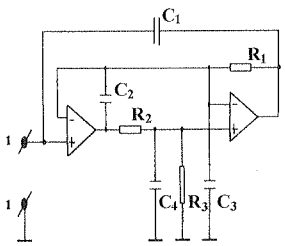


Fig. 1

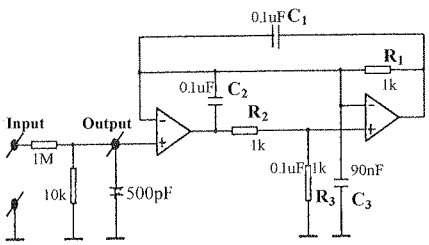


Fig. 2a

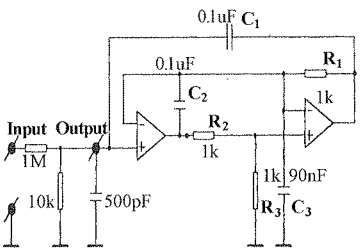


Fig. 2a

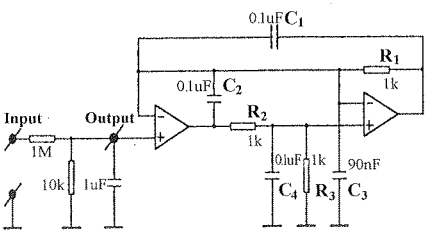


Fig. 3a

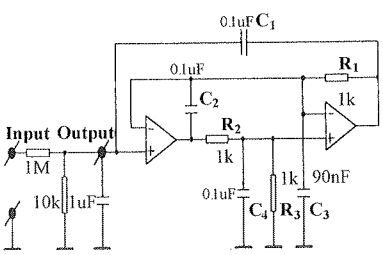


Fig. 3a