

POLSKA AKADEMIA NAUK
KOMITET ELEKTRONIKI I TELEKOMUNIKACJI

h
yjnych,
KWARTALNIK
ELEKTRONIKI I TELEKOMUNIKACJI

ELECTRONICS AND
TELECOMMUNICATIONS
QUARTERLY

TOM 49 — ZESZYT 3

WARSZAWA 2003

RADA REDAKCYJNA

Przewodniczący

Prof. dr hab. inż. STEFAN HAHN
czł. rzecz. 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 HAŁAŚ, prof. dr inż. WŁADYSŁAW MAJEWSKI, prof. dr hab. inż. JÓZEF MODELSKI, prof. dr inż. JERZY OSIOWSKI, prof. dr hab. inż. EDWARD SEDEK, 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

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

Ark. wyd. 13,25	Ark. druk. 10,625	Podpisano do druku w czerwcu 2003 r.
Papier offset. kl. III 80 g. B-1		Druk ukończono w lipcu 2003 r.

Skład, druk i oprawa: Warszawska Drukarnia Naukowa PAN
00-656 Warszawa, ul. Śniadeckich 8
Tel./fax: 628-87-77

Szanowni Autorzy

„Kwartalnik Elektroniki i Telekomunikacji” — Electronics and Telecommunications Quarterly jest kontynuatorem tradycji powstałego 49 lat temu kwartalnika p.t. „Rozprawy Elektrotechniczne”.

Kwartalnik jest czasopismem naukowym Komitetu Elektroniki i Telekomunikacji Polskiej Akademii Nauk, 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, oproelektroniki, 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 specjalistów, co zapewnia że publikacje te są uznawane jako autorski dorobek naukowy. Opublikowane w kwartalniku wyniki 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 ok. 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

K.
E.
J.

P.

K.

K.

M.
Info

K. F.
E. J.
J. K.
P. R.
K. W.

K. G.
M. G.
Inform

SPIS TREŚCI

K. Koziel: Efektywna analiza tolerancji filtrów aktywnych G_m -C czasu ciągłego	271
E. Jamro, K. Wiatr: Implementacja operacji konwolucji o stałych współczynnikach w układach FPGA	295
J. Kowalski: Analiza i projektowanie sieci neuronowych komórkowych realizowanych w technice układów pracujących w trybie prądowym	317
P. Rusek, K. Wiatr: Rekonfigurowalny kwantyzator wektorowy do kodowania obrazów w czasie rzeczywistym	355
K. Weiss, B. Gniewińska, A. Lisowiec: Zagadnienia ustalania częstotliwości drgań generatorów z rezonatorami kwarcowym pobudzonymi do drgań polem równoległym	373
K. Górecki: Modelowanie cewki z rdzeniem ferrytowym w programie SPICE z uwzględnieniem samonagrzewania	389
M. Gajer: System translacji automatycznej oparty na uogólnionych przykładach translacyjnych	405
Informacje dla Autorów (w jęz. pol.)	429

CONTENTS

K. Koziel: Efficient tolerance analysis of continuous-time G_m -C filters	271
E. Jamro, K. Wiatr: Convolution with constant coefficient implemented in FPGA chips	295
J. Kowalski: Analysis and designing of cellular neural networks based on current-mode CMOS circuits	317
P. Russek, K. Wiatr: Reconfigurable vector quantizer for image compression in real-time systems	355
K. Weiss, B. Gniewińska, A. Lisowiec: The problems of resonant frequency adjusting of the oscillators utilizing lateral field excited quartz resonators	373
K. Górecki: SPICE — aided modelling of coils with the ferrite core with selfheating taken into account	389
M. Gajer: The example based machine translation system	405
Information for the Authors (w jęz. ang.)	432

En

on
tran
ban
this
appr
para
ce e
oper
I
mos

Efficient Tolerance Analysis of Continuous-Time G_m -C Filters

ŚLAWOMIR KOZIEŁ

*Faculty of Electronics, Telecommunications and Informatics,
Gdańsk, University of Technology 80-952 Gdańsk, Poland
e-mail: koziel@ue.eti.pg.gda.pl.*

*Otrzymano 2002.10.02
Autoryzowano 2002.11.25*

In the paper, an efficient approach to tolerance analysis of arbitrary G_m -C filters is investigated. We present a general structure of G_m -C filter of arbitrary order as well as its matrix description. Using this approach, we derive formulas that allow us to calculate the change of filter transfer function caused by the deviation of filter element values from their nominal values. At the expense of some approximations we are able, with these formulas, to determine transfer function deviation with low computational cost, which makes them suitable to perform statistical analysis of arbitrary G_m -C filters. It can also be used to analyse the influence of parasitics on filter characteristics. Several application examples including Worst Case analysis of high-order G_m -C filters are presented and discussed.

Keywords: G_m -C filters, tolerance analysis, statistical analysis.

1. INTRODUCTION

Recently an increasing interest in the design of continuous-time (CT) filters based on the transconductance-capacitor (G_m -C) technique has been observed [1], [2]. The transconductors and operational transconductance amplifiers (OTAs) offer a higher bandwidth than their voltage-mode counterparts, i.e., voltage operational amplifiers. Due to this, high frequency integrated filters are mostly realised as the G_m -C ones [9]. In this approach the basic building block is a simple CT integrator for which the key performance parameter is the phase shift at its unity-gain frequency [3]-[15]. Moreover, transconductance elements can be easily tuned electronically [6], [7] and have a better suitability for operating in reduced supply environment.

In the synthesis of G_m -C filters one wants to achieve several different objectives. The most important goal is, of course, that the filter structure should realise the desired transfer

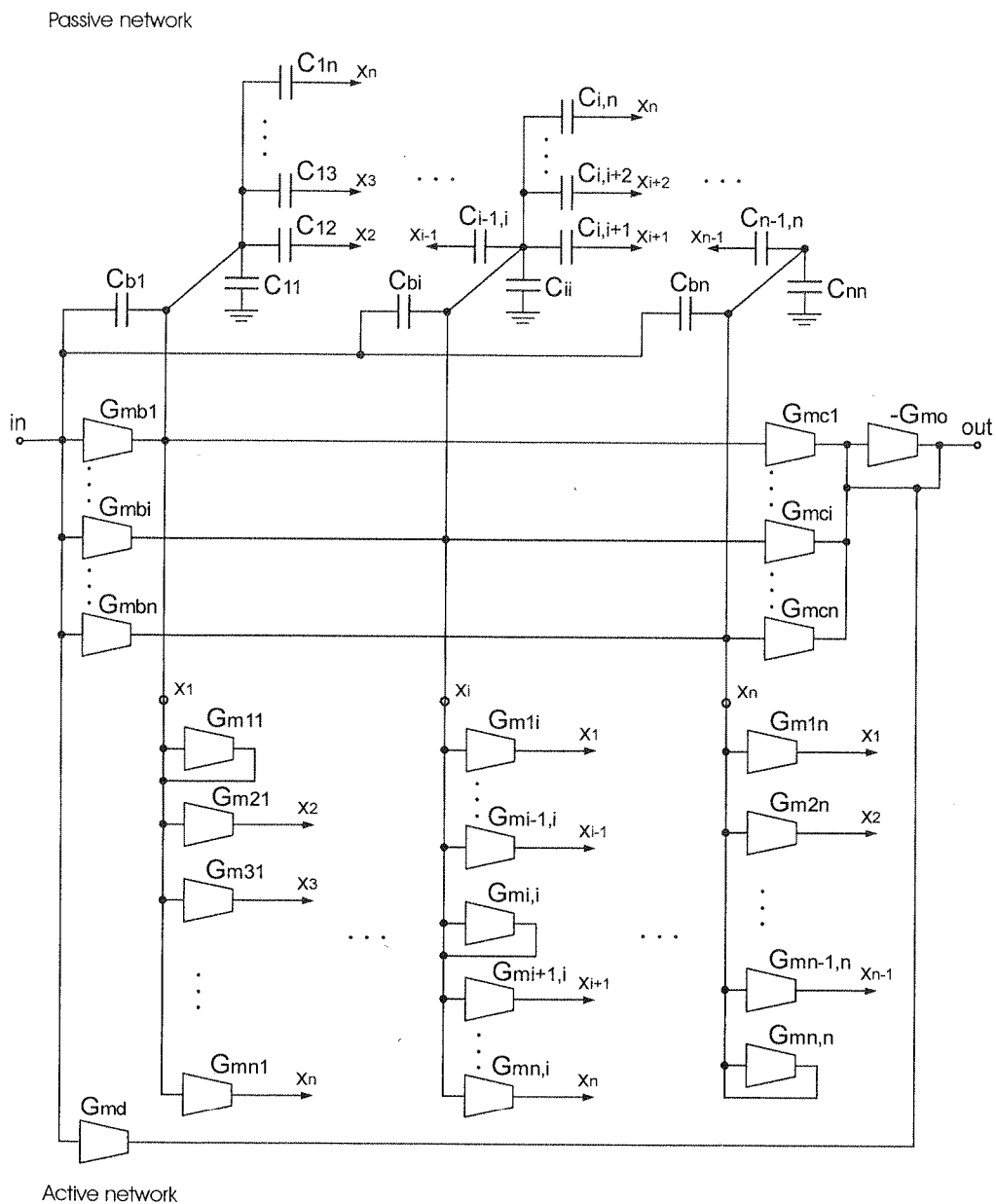
function $H(s)$. Moreover, the filter should contain a small number of components, since it results in reduction of noise, parasitic effects and power consumption. The structures containing only grounded capacitors are also more attractive, because grounded capacitors can absorb parasitic capacitances and need smaller chip areas than floating ones [2]. Eventually, we want filter structures which have large dynamic range and low sensitivity.

It is important for filter design purposes to develop efficient tools for performing statistical analysis of G_m -C filters. Monte Carlo and Worst Case analyses can provide us with useful information concerning filter properties. Specifications such as standard and maximum deviation of a filter transfer function seem to be much more reliable quantities than standard (total) sensitivity measures (e.g. [19]) to say how deviations of filter element values influence filter properties. However, statistical analysis is a very time consuming task, especially for high-order filters. The aim of this work is to develop an efficient method of tolerance analysis of G_m -C filters. The novelty of the approach follows by utilising the matrix description of general structure of G_m -C filter described in recent papers (cf. [22], [23]). At the expense of limiting the deviations of filter element values, we get a very fast procedure which is characterised by low computational complexity and applicability to any structure of G_m -C filter of arbitrary order.

The paper is organised as follows. In Section 2, a general structure of G_m -C filter is presented. We discuss a matrix description of a general G_m -C filter based on a separate treatment of passive and active network of the circuit. In Section 3 we develop explicit formulas that allow us to calculate deviation of filter transfer function given deviations of its element values. At this stage the main equation describing the filter in Laplace transform domain is extended by adding auxiliary matrices containing deviations of filter element values. In the next step, linear approximation of extended equation is performed, which leads to the required formulas. A number of examples including comparison of the results provided by the approximate and exact method are presented in Section 4. Section 5 concludes the work.

2. GENERAL STRUCTURE OF G_m -C FILTER

Consider a general structure of a voltage-mode G_m -C filter shown in Fig.1. The current-mode counterpart can be obtained by inverting all transconductors and interchanging input and output of the filter [17],[18],[21]. The structure in Fig.1 contains n internal nodes denoted as x_i , $i = 1, \dots, n$, n input transconductors G_{mbi} , an output summer consisting of transconductors G_{mci} and $-G_{mo}$ as well as a feedforward transconductor G_{md} . All transconductors form *active network*, while input capacitors C_{bi} , $i = 1, \dots, n$ and capacitors C_{ij} , $1 \leq i \leq j \leq n$ form *passive network*. It is easily seen that any G_m -C filter is a particular case of the general structure in Fig.1.

Fig. 1. General structure of voltage-mode G_m -C filterRys. 1. Ogólna struktura filtru G_m -C pracującego w trybie napięciowym

In the sequel we shall denote the voltage at the i -th node x_i also by x_i . A general structure of the voltage-mode G_m -C filter in Fig. 1 can be described by the following matrix equations:

$$\begin{aligned} sT_C X &= GB + B^T u_i \\ u_o &= CX + Du_i \end{aligned} \quad (1)$$

where u_i , u_o are the input and output voltages, respectively, and

$$T_C = \begin{bmatrix} C_{b1} + \sum_{j=1}^n C_{1j} & -C_{12} & \dots & -C_{1n} \\ -C_{12} & C_{b2} + \sum_{j=1}^n C_{2j} & \dots & -C_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ -C_{1n} & -C_{2n} & \dots & C_{bn} + \sum_{j=1}^n C_{nj} \end{bmatrix},$$

$$G = \begin{bmatrix} G_{m11} & G_{m12} & \dots & G_{m1n} \\ G_{m12} & G_{m22} & \dots & G_{m2n} \\ \vdots & \vdots & \ddots & \vdots \\ G_{mn1} & G_{mn2} & \dots & G_{mn n} \end{bmatrix}, \quad X = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix},$$

$$B = [G_{mb1} + sC_{b1} \quad \dots \quad G_{mbn} + sC_{bn}], \quad (2)$$

$$C = [c_1 \quad \dots \quad c_n], \quad c_i = G_{mci} / G_{mo}, \quad i = 1, 2, \dots, n$$

$$D = d, \quad d = G_{md} / G_{mo}$$

Note that T_C is a symmetrical matrix, i.e. $T_C^T = T_C$. On the basis of (1) we can calculate the filter transfer function:

$$H^{(0)}(s) = \frac{u_o(s)}{u_i(s)} = C(sT_C - G)^{-1} B^T + D \quad (3)$$

where index (0) refers to the fact that no filter element deviations have been taken into account. Now, let us denote adjoint matrix of $sT_C - G$ as $\tilde{A}$ where

$$\tilde{A}(s) = \text{adj}(sT_C - G) = [\tilde{A}_{ij}(s)]_{i,j=1}^n \quad (4)$$

This allows us to rewrite transfer function in the form:

$$H^{(0)}(s) = \frac{1}{\det(sT_C - G)} \sum_{i,j=1}^n c_i (G_{mbj} + sC_{bj}) \tilde{A}_{ij}(s) + d \quad (5)$$

Note that many filter structures have only one input transconductor (i.e. no input signal distribution), a trivial output summer (i.e. one of the internal nodes is the output of the filter), and no input capacitors. This means that $B = [0 \dots 0 \ G_{mbk} \ 0 \dots 0]$, $C = [0 \dots 0 \ 1 \ 0 \dots 0]$ - 1 at l -th position and $C_{bi} = 0$ for $i = 0, 1, \dots, n$. In such case, expression (5) reduces to the form:

$$H^{(0)}(s) = \frac{G_{mbk} \tilde{A}_{lk}(s)}{\det(sT_C - G)} \quad (6)$$

Similar expression can be written for slightly more general case, with no input capacitors, input signal distribution (i.e. $B = [G_{mb1} \dots G_{mbn}]$), and a trivial output summer ($C = [0 \dots 0 \ 1 \ 0 \dots 0]$ - 1 at l -th position). Then, we have

$$H^{(0)}(s) = \frac{1}{\det(sT_C - G)} \sum_{i=1}^n G_{mbi} \tilde{A}_{li}(s) \quad (7)$$

On the basis of the above expressions one can easily calculate the transmittance of any particular structure of G_m - C filter. In a similar way one can treat current, mode structures as well as state-space filters (see [22] for more details). It follows that the order n_H of transfer function of general filter structure in Fig. 1 is not necessarily equal to the number of internal nodes. It can be shown that it essentially depends on the passive network. In particular, we have that n_H is not larger than the rank of the corresponding matrix T_C .

3. TOLERANCE ANALYSIS OF G_m - C FILTERS

In this section we shall calculate the change of filter transfer function caused by the deviation of filter element values from their nominal values. We work within the setting of matrix description established in the previous section. Thus, we will consider the following deviation matrices:

$$T_{\Delta C} = \begin{bmatrix} \Delta C_{b1} + \sum_{j=1}^n \Delta C_{1j} & -\Delta C_{12} & \dots & -\Delta C_{1n} \\ -\Delta C_{12} & \Delta C_{b2} + \sum_{j=1}^n \Delta C_{2j} & \dots & -\Delta C_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ -\Delta C_{1n} & -\Delta C_{2n} & \dots & \Delta C_{bn} + \sum_{j=1}^n \Delta C_{nj} \end{bmatrix},$$

$$G_{\Delta} = \begin{bmatrix} \Delta G_{m11} & \Delta G_{m12} & \dots & \Delta G_{m1n} \\ \Delta G_{m21} & \Delta G_{m22} & \dots & \Delta G_{m2n} \\ \vdots & \vdots & \ddots & \vdots \\ \Delta G_{mn1} & \Delta G_{mn2} & \dots & \Delta G_{mnn} \end{bmatrix}, \quad \begin{aligned} B_{\Delta} &= [\Delta G_{mb1} + s\Delta C_{b1} \dots \Delta G_{mbn} + s\Delta C_{bn}], \quad (8) \\ C_{\Delta} &= [\Delta c_1 \dots \Delta c_n], \\ D_{\Delta} &= \Delta d, \end{aligned}$$

where ΔG_{mij} , $i, j=1, \dots, n$, ΔG_{mbi} , $i=1, \dots, n$, ΔC_{ij} , $i, j=1, \dots, n$, $i \leq j$, ΔC_{bi} , $i=1, \dots, n$, Δc_i , $i=1, \dots, n$, Δd , denote deviation of transconductance values, capacitance values and summation coefficients, respectively. Obviously, we have $\Delta c_i = (G_{mo} \cdot \Delta G_{mci} - \Delta G_{mo} \cdot G_{mci})/G_{mo}^2$, $i=1, \dots, n$, and $\Delta d = (G_{mo} \cdot \Delta G_{md} - \Delta G_{mo} \cdot G_{md})/G_{mo}^2$. Taking into account (8), we modify the original equations (1) describing the filter so that they take the following form:

$$\begin{aligned} s(T_C + T_{\Delta C})X &= (G + G_{\Delta})X + (B + B_{\Delta})^T u_i, \\ u_o &= (C + C_{\Delta})X + (D + D_{\Delta})u_i \end{aligned} \quad (9)$$

Using the first equation in (9) we can calculate the vector of node voltages X as:

$$\begin{aligned} X &= [(sT_C - G) + (sT_{\Delta C} - G_{\Delta})]^{-1} (B + B_{\Delta})^T u_i \\ &= [(sT_C - G)(I + (sT_C - G)^{-1}(sT_{\Delta C} - G_{\Delta}))]^{-1} (B + B_{\Delta})^T u_i \\ &= [I + (sT_C - G)^{-1}(sT_{\Delta C} - G_{\Delta})]^{-1} (sT_C - G)^{-1} (B + B_{\Delta})^T u_i \end{aligned} \quad (10)$$

Here, I stands for the identity matrix. For the rest of the paper we shall assume that element deviations are small. This means, in particular, that the norm of matrix $(sT_C - G)^{-1}(sT_{\Delta C} - G_{\Delta})$ is small, so we can use the following approximation

$$(I + A)^{-1} \approx I - A, \quad (11)$$

which holds for any matrix A such that $\|A\| \ll 1$. Using (11) we obtain

$$X \approx [I - (sT_C - G)^{-1}(sT_{\Delta C} - G_{\Delta})](sT_C - G)^{-1} (B + B_{\Delta})^T u_i \quad (12)$$

Hence, the filter transfer function at the presence of element deviations is given by the following formula

$$H(s) = (C + C_{\Delta})[I - (sT_C - G)^{-1}(sT_{\Delta C} - G_{\Delta})](sT_C - G)^{-1} (B + B_{\Delta})^T + D + D_{\Delta}. \quad (13)$$

By neglecting all higher order terms with respect to deviation we get

$$\begin{aligned} H(s) &= C(sT_C - G)^{-1} B^T + C(sT_C - G)^{-1} B_{\Delta}^T + C_{\Delta} (sT_C - G)^{-1} B^T \\ &\quad + D + D_{\Delta} - C(sT_C - G)^{-1} (sT_{\Delta C} - G_{\Delta})(sT_C - G)^{-1} B^T \end{aligned} \quad (14)$$

Introducing matrices C_L and C_R defined as follows

$$C_L = C(sT_C - G)^{-1}, \quad C_R = (sT_C - G)^{-1}B^T, \quad (15)$$

we can rewrite (14) as

$$H(s) = C_L (B^T + B_\Delta^T) + C_\Delta C_R + D + D_\Delta - C_L (sT_{\Delta C} - G_\Delta) C_R. \quad (16)$$

Now, using (3) and (16), we can calculate the transfer function deviation $\Delta H(s) = H(s) - H^{(0)}(s)$:

$$\Delta H(s) = C_L B_\Delta^T + C_\Delta C_R - C_L (sT_{\Delta C} - G_\Delta) C_R + D_\Delta, \quad (17)$$

the amplitude deviation of the transfer function $\Delta |H(s)| = |H(s)| - |H^{(0)}(s)|$:

$$\Delta |H(s)| = |C_L (B^T + B_\Delta^T) + C_\Delta C_R - C_L (sT_{\Delta C} - G_\Delta) C_R + D + D_\Delta| - |C_L B_\Delta^T + D|, \quad (18)$$

or phase deviation of the transfer function $\Delta(\arg H(s)) = \arg H(s) - \arg H^{(0)}(s)$

$$\Delta(\arg H(s)) = \arg(C_L(B^T + B_\Delta^T) + C_\Delta C_R - C_L(sT_{\Delta C} - G_\Delta)C_R + D + D_\Delta) - \arg(C_L B_\Delta^T + D). \quad (19)$$

Obviously, formula (17) is nothing else but the differential of $H(s)$, i.e.

$$\Delta H(s) = \sum_{i=1}^N \frac{\partial H(s)}{\partial z_i} \Delta z_i \quad (20)$$

where z_i denote filter elements (i.e. capacitors and transconductors). Expression (17) is just a very convenient way to calculate it.

Note that if we assume trivial output summer ($C = [0 \dots 0 \ 1 \ 0 \dots 0]$ - 1 at l -th position), then we have $C_\Delta = 0$ and $D_\Delta = 0$ and (17), (18) and (19) simplify to

$$\Delta H(s) = C_L [B_\Delta^T - (sT_{\Delta C} - G_\Delta) C_R], \quad (21)$$

$$\Delta |H(s)| = |C_L (B + B_\Delta^T) - C_L (sT_{\Delta C} - G_\Delta) C_R + D| - |C_L B_\Delta^T + D|, \quad (22)$$

$$\Delta(\arg H(s)) = \arg(C_L(B^T + B_\Delta^T) - C_L(sT_{\Delta C} - G_\Delta)C_R + D) - \arg(C_L B_\Delta^T + D). \quad (23)$$

It is seen that the above formulas are very simple and general, since they allow us to analyse an arbitrary G_m -C filter structure. Due to the fact that they use a matrix formulation, it is easy to implement computer software to evaluate these formulas in a very efficient way. In particular, one can perform a very fast Monte Carlo and/or Worst Case analysis. Note that matrices C_L and C_R have to be calculated only once for a given filter structure. Then, the evaluation of formulas (17)-(19) (or (21)-(23)) is nothing else but a simple matrix (or vector) multiplication, no matter how we change the deviation of the filter element values.

There is no need to invert the matrix $s(\mathbf{T}_C + \mathbf{T}_{\Delta C}) - (\mathbf{G} + \mathbf{G}_{\Delta})$ while changing these deviations. Due to this fact a computational cost of determining transfer function deviation using approximate formulas is only $O(n^2)$. For example, an analysis of 8th order filter is about five times faster than the same analysis being done using the exact method.

In the case of low order filters it is possible to perform hand tolerance analysis. For illustration purposes consider a second-order G_m -C filter shown in Fig.2. We shall assume for simplicity that nominal values of all filter transconductors are the same (up to the sign), i.e. $G_{b1}=G_{21}=-G_{12}=-G_{22}=g$, while element deviations will be denoted as ΔC_1 , ΔC_2 , ΔG_{b1} , ΔG_{21} , ΔG_{12} , ΔG_{22} . The matrices describing this filter are

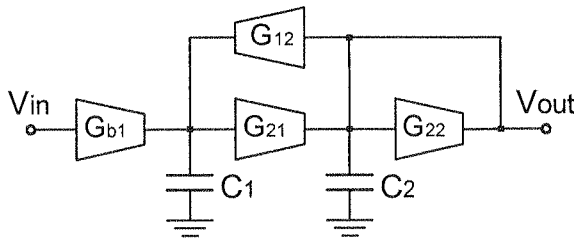


Fig.2. Diagram of the second-order low-pass G_m -C filter

Rys.2. Schemat dolnoprzepustowego filtru G_m -C drugiego rzędu

$$\mathbf{T}_C = \begin{bmatrix} C_1 & 0 \\ 0 & C_2 \end{bmatrix}, \mathbf{G} = \begin{bmatrix} 0 & -g \\ g & -g \end{bmatrix}, \mathbf{B} = [g \quad 0], \mathbf{C} = [0 \quad 1], \mathbf{D} = 0. \quad (24)$$

Using (24) one can easily calculate matrices $\mathbf{C}_L$ and $\mathbf{C}_R$ that are

$$\mathbf{C}_L = \frac{1}{D(s)} [g \quad sC_1], \quad \mathbf{C}_R = \frac{1}{D(s)} \begin{bmatrix} sC_2 g + g^2 \\ g^2 \end{bmatrix}, \quad (25)$$

with

$$D(s) = C_1 C_2 s^2 + C_1 g s + g^2. \quad (26)$$

Moreover, it follows from (3) that

$$\mathbf{H}^{(0)}(s) = \frac{g^2}{C_1 C_2 s^2 + C_1 g s + g^2}. \quad (27)$$

Now, using (21) we can calculate $\Delta \mathbf{H}(s) = \mathbf{H}(s) - \mathbf{H}^{(0)}(s)$, which is

$$\Delta \mathbf{H}(s) = \frac{1}{D(s)} [g \quad sC_1] \left(\begin{bmatrix} \Delta G_{b1} \\ 0 \end{bmatrix} - \frac{1}{D(s)} \begin{bmatrix} s\Delta C_1 & -\Delta G_{12} \\ \Delta G_{21} & s\Delta C_2 - \Delta G_{22} \end{bmatrix} \begin{bmatrix} sC_2 g + g^2 \\ g^2 \end{bmatrix} \right) \quad (28)$$

Hence, we get

$$\Delta H(s) = \frac{g}{D^2(s)} [-s(sC_2 + g)\Delta C_1 + s^2 C_1 g \Delta C_2 + g^2 \Delta G_{12} - sC_1(sC_2 + g)\Delta G_{21} + sC_1 g \Delta G_{22} + D(s)\Delta G_{b1}]. \quad (29)$$

In a similar way one can calculate explicit formulas for $\Delta |H(s)|$ and $\Delta(\arg H(s))$.

4. APPLICATION EXAMPLES

In this section we consider several application examples of a tolerance analysis method introduced in the previous section. Throughout this section we shall use normalised element values and normalised frequency. All the results of Monte Carlo simulations have been obtained on the basis of 5000 random samples. The standard deviation of amplitude and phase, used in the sequel, are defined as

$$S_{|H(j\omega)|} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (|H_i(j\omega)| - |H^{(0)}(j\omega)|)^2} \quad (30)$$

$$S_{\arg H(j\omega)} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (\arg H_i(j\omega) - \arg H^{(0)}(j\omega))^2} \quad (31)$$

where $|H^{(0)}(j\omega)|$ ($\arg H^{(0)}(j\omega)$) is the nominal amplitude (phase) and $|H_i(j\omega)|$ ($\arg H_i(j\omega)$) is the actual amplitude (phase) for i -th random sample. We have assumed a gaussian distribution of the deviations of filter element values while calculating standard deviations (30), (31) and a flat distribution while calculating maximum deviations of amplitude and phase. In figures containing the results of the statistical simulations solid lines show the results achieved according to approximate formulas (22) and (23); points indicate the results obtained according to the exact formulas based on equation (9). Deviations of amplitude and phase are calculated in dB and degrees, respectively.

As the first example, consider again a second-order filter in Fig.2. We assume that this filter implements Butterworth transfer function. Then, the normalised element values are: $C_1=1.4142$, $C_2=0.7071$, $G_{b1}=G_{21}=1.0$, $G_{12}=G_{22}=-1.0$. Figs.3 and 4 show the modulus and phase of the nominal transfer function of the filter, respectively. Figs.5,6 (Figs.7,8) show the results of Monte Carlo simulation (Worst Case analysis) of amplitude and phase of the transfer function. We have assumed 1%, 2%, 5% and 10% deviations for all filter elements. Figs.5 and 6 show standard deviation of amplitude and phase, respectively. Figs.7 and 8 show maximum deviation of amplitude and phase, respectively. One can infer from the data in Figs.5–8 that formulas (22), (23) provide us with a very good approximation. Standard deviation calculated by means of both methods is almost identical even for the deviations in filter element values equal to 10%. The results are slightly worse for

maximum deviations, however, the error is negligible for deviations of element values smaller or equal to 5%.

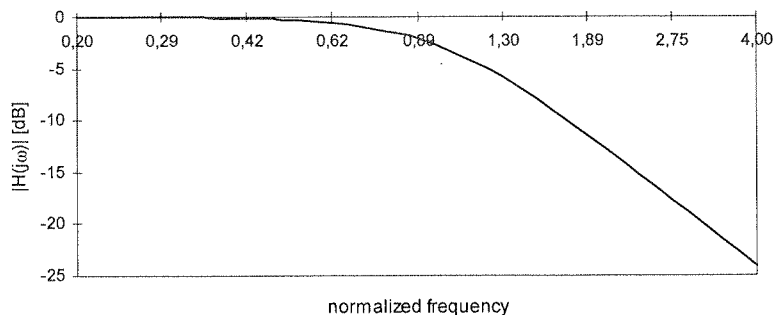


Fig. 3. Nominal amplitude characteristic $|H(j\omega)|$ of 2nd order Butterworth filter in Fig. 2

Rys. 3. Nominalna charakterystyka amplitudowa $|H(j\omega)|$ filtru Butterwortha 2-go rzędu z Rys. 2

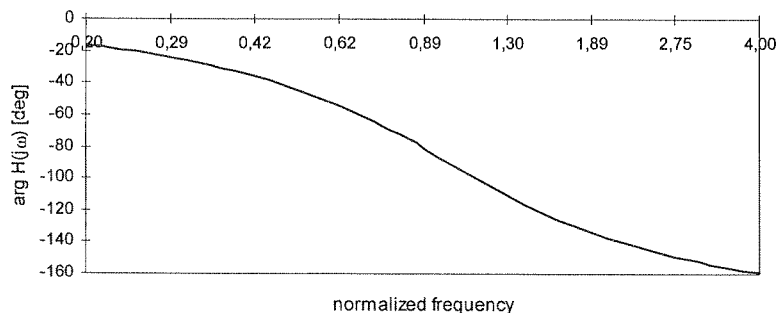


Fig. 4. Nominal phase characteristic $\arg[H(j\omega)]$ of 2nd order Butterworth filter in Fig. 2

Rys. 4. Nominalna charakterystyka fazowa $\arg[H(j\omega)]$ filtru Butterwortha 2-go rzędu z Rys. 2

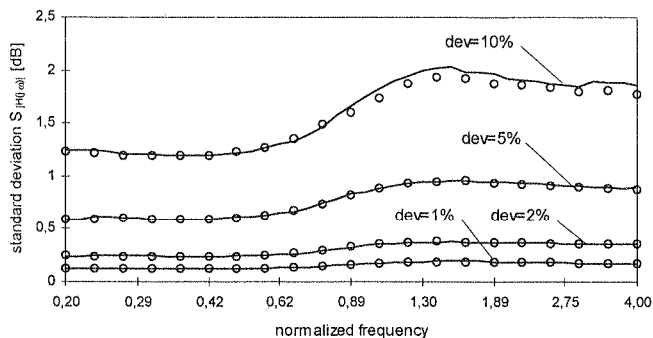


Fig. 5. Standard deviation of amplitude $S_{|H(j\omega)|}$ versus normalised frequency for 2nd order Butterworth filter in Fig. 2 with a deviation in filter element values as a parameter; approximate values denoted as solid line and exact values as circles; gaussian distribution in element values was assumed

Rys. 5. Odchylenie standardowe amplitudy $S_{|H(j\omega)|}$ w funkcji znormalizowanej częstotliwości dla filtru Butterwortha 2-go rzędu z Rys. 2 z odchyłką wartości elementów jako parametrem; wartości przybliżone oznaczono linią ciągłą, wartości dokładne oznaczono punktami; przyjęto gaussowski rozkład odchyłek wartości elementów

values

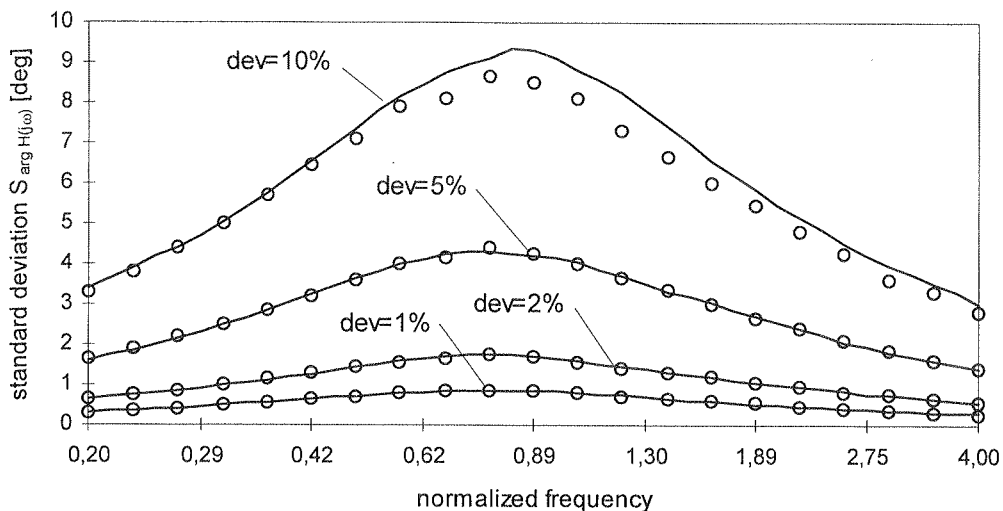


Fig. 6. Standard deviation of phase $S_{arg H(j\omega)}$ versus normalised frequency for 2nd order Butterworth filter in Fig. 2 with a deviation in filter element values as a parameter; approximate values denoted as solid line and exact values as circles; gaussian distribution in element values was assumed

Rys. 6. Odchylenie standardowe fazy $S_{arg H(j\omega)}$ w funkcji znormalizowanej częstotliwości dla filtru Butterwortha 2-go rzędu z Rys. 2 z odchyłką wartości elementów jako parametrem; wartości przybliżone oznaczono linią ciągłą, wartości dokładne oznaczono punktami; przyjęto gaussowski rozkład odchyłek wartości elementów

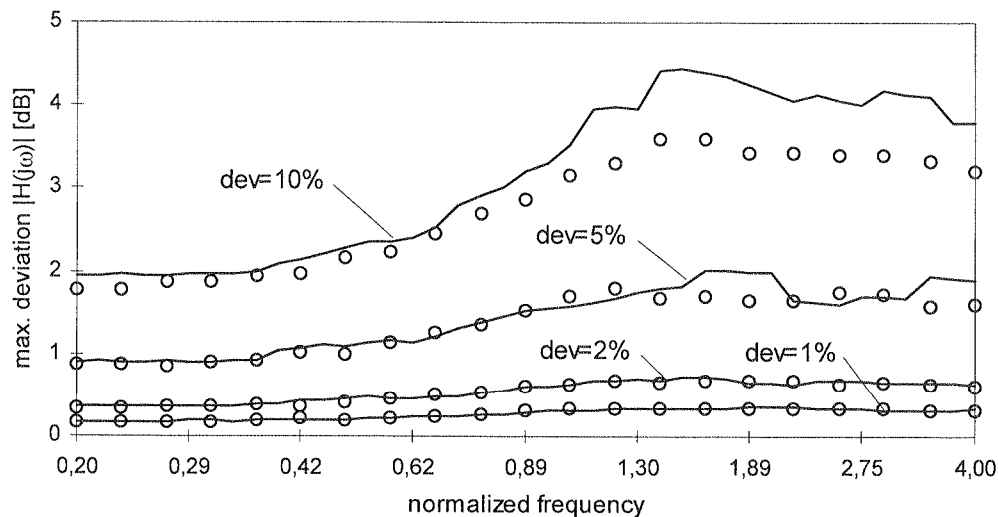


Fig. 7. Maximum deviation of amplitude versus normalised frequency for 2nd order Butterworth filter in Fig. 2 with a deviation in filter element values as a parameter; approximate values denoted as solid line and exact values as circles; flat distribution in element values was assumed

Rys. 7. Odchylenie maksymalne amplitudy w funkcji znormalizowanej częstotliwości dla filtru Butterwortha 2-go rzędu z Rys. 2 z odchyłką wartości elementów jako parametrem; wartości przybliżone oznaczono linią ciągłą, wartości dokładne oznaczono punktami; przyjęto jednostajny rozkład odchyłek wartości elementów

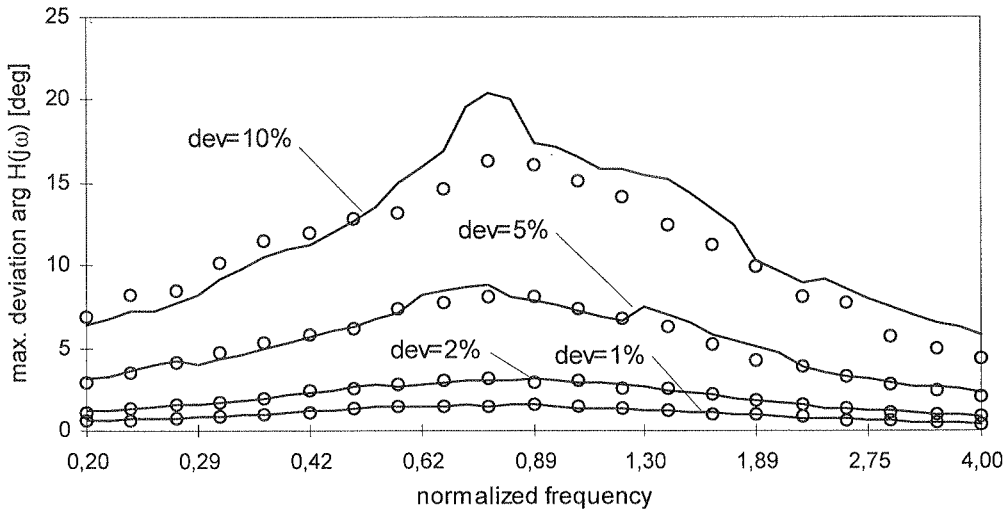


Fig. 8. Maximum deviation of phase versus normalised frequency for 2nd order Butterworth filter in Fig. 2 with a deviation in filter element values as a parameter; approximate values denoted as solid line and exact values as circles; flat distribution in element values was assumed

Rys. 8. Odchylenie maksymalne amplitudy w funkcji znormalizowanej częstotliwości dla filtra Butterwortha 2-go rzędu z Rys. 2 z odchyłką wartości elementów jako parametrem; wartości przybliżone oznaczono linią ciągłą, wartości dokładne oznaczono punktami; przyjęto jednostajny rozkład odchyłek wartości elementów

Another example is a 3rd order elliptic low-pass filter in leap-frog (LF) structure with a floating capacitor [24] shown in Fig.9. We assume that the filter implements transfer function with minimum pass-band attenuation $A_1=0.9$, maximum stop-band attenuation $A_2=0.16$ and selectivity $\omega_s/\omega_p=1.2$ [24]. The normalised filter element values in this case are: $C_1=1.1653$, $C_2=0.6241$, $C_3=0.4710$, $C_4=0.9428$, $G_{b1}=G_{21}=G_{32}=1.0$, $G_{12}=G_{23}=G_{33}=-1.0$.

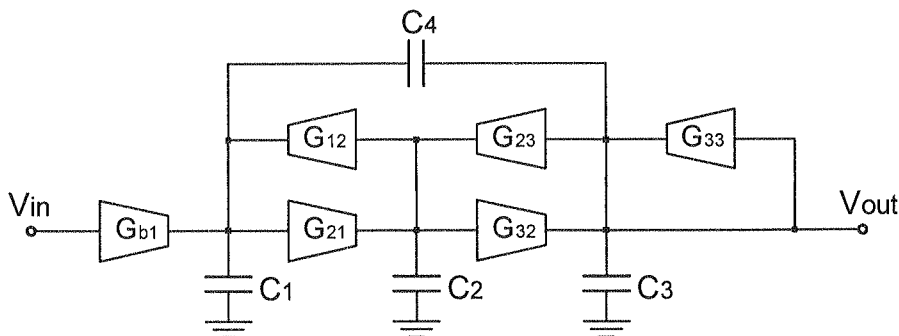


Fig. 9. A third order low-pass elliptic filter in LF structure with a floating capacitor

Rys. 9. Dolnoprzepustowy filtr eliptyczny trzeciego rzędu w strukturze LF z kondensatorem nieziemionym

Figs. 10 and 11 show the modulus and phase of the nominal transfer function of the filter, respectively. Figs. 12, 13 (Figs. 14, 15) show the results of Monte Carlo simulation (Worst Case analysis) of amplitude and phase of the transfer function. We have assumed 1%, 2%, 5% and 10% deviations for all filter elements. Figs. 12 and 13 show standard deviation of amplitude and phase, respectively. Figs. 14 and 15 show maximum deviation of amplitude and phase, respectively. It follows from the data in Figs. 12–15 that formulas (22), (23) give a very good approximation for frequencies which are not too close to the frequency of the transmission zero of the filter. Standard deviation calculated by means of both methods is almost the same even for the deviation in filter element values equal to 10%. The results for maximum deviation are slightly worse. However, for frequencies close to the frequency of the transmission zero, approximation is not that good, especially for deviations of element values equal to or higher than 5%. We can account for this fact by recalling that $|H(j\omega)|$ decreases to minus infinity (in dB scale) as the frequency approaches zero of the transfer function. Thus, even very small error produced by the approximate formula (22) in a linear scale, gives considerable error in a logarithmic scale. Similarly, the transfer function has a considerable phase jump at its zero, which means that the results of maximum phase deviation are not reliable for frequencies close to this frequency point and for large deviation of filter element values.

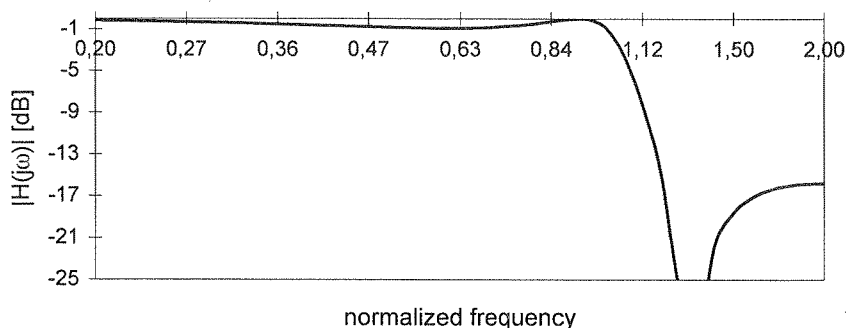


Fig. 10. Nominal amplitude characteristic $|H(j\omega)|$ of 3rd order elliptic filter in Fig. 9

Rys. 10. Nominalna charakterystyka amplitudowa $|H(j\omega)|$ filtru eliptycznego 3-go rzędu z Rys. 9

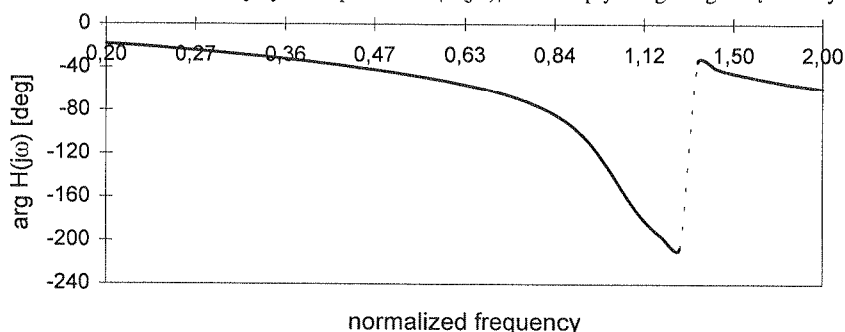


Fig. 11. Nominal phase characteristic $\arg[H(j\omega)]$ of 3rd order elliptic filter in Fig. 9

Rys. 11. Nominalna charakterystyka fazowa $\arg[H(j\omega)]$ filtru eliptycznego 3-go rzędu z Rys. 9

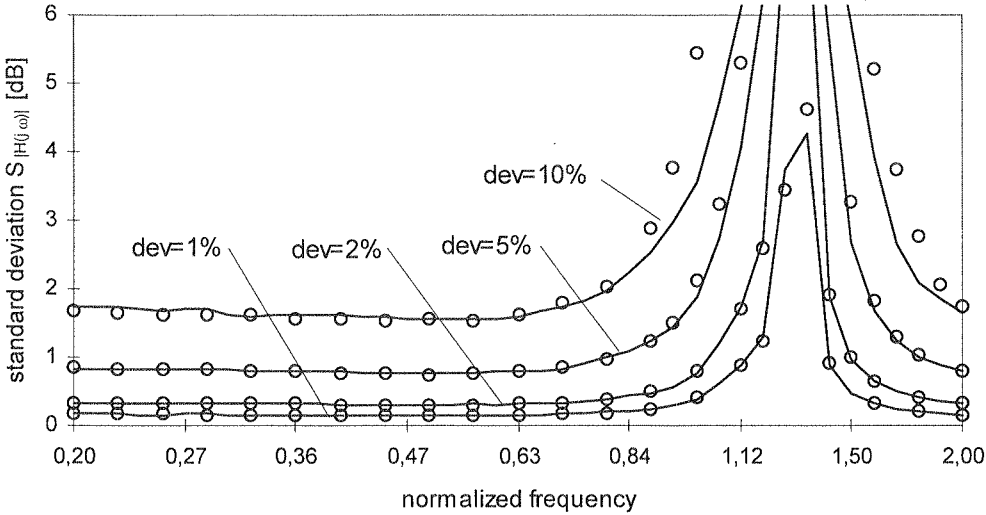


Fig. 12. Standard deviation of amplitude $S_{|H(j\omega)|}$ versus normalised frequency for 3rd order elliptic filter in Fig. 9 with a deviation in filtr element value as a parameter; approximate values denoted as solid line and exact values as circles; gaussian distribution in element values was assumed

Rys. 12. Odchylenie standardowe amplitudy $S_{|H(j\omega)|}$ w funkcji znormalizowanej częstotliwości dla filtra eliptycznego 3-go rzędu z Rys. 9 z odchyłką wartości elementów jako parametrem; wartości przybliżone oznaczono linią ciągłą, wartości dokładne oznaczono punktami; przyjęto gaussowski rozkład odchyłek wartości elementów

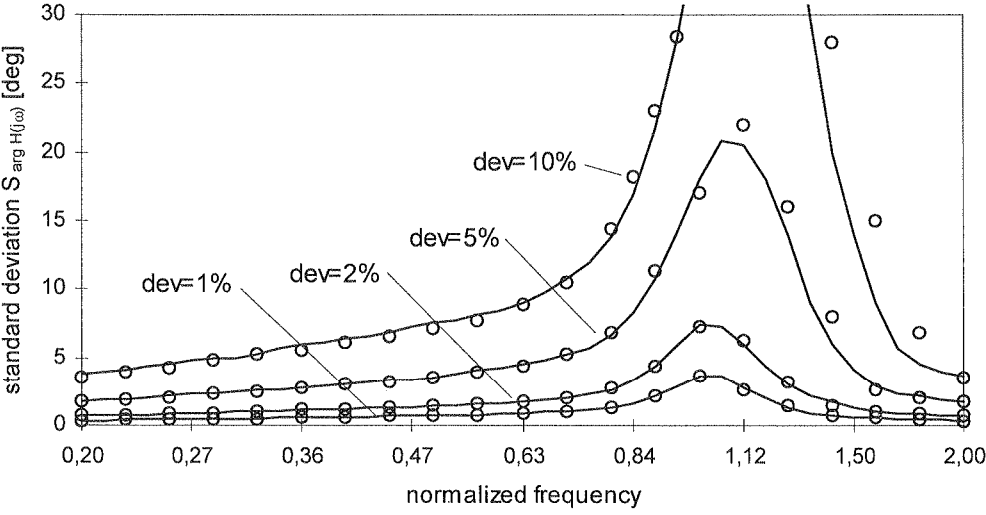


Fig. 13. Standard deviation of phase $S_{arg H(j\omega)}$ versus normalised frequency for 3rd order elliptic filter in Fig. 9 with a deviation in filtr element values as a parameter; approximate values denoted as solid line and exact values as circles; gaussian distribution in element values was assumed

Rys. 13. Odchylenie standardowe fazy $S_{arg H(j\omega)}$ w funkcji znormalizowanej częstotliwości dla filtra eliptycznego 3-go rzędu z Rys. 9 z odchyłką wartości elementów jako parametrem; wartości przybliżone oznaczono linią ciągłą, wartości dokładne oznaczono punktami; przyjęto gaussowski rozkład odchyłek wartości elementów

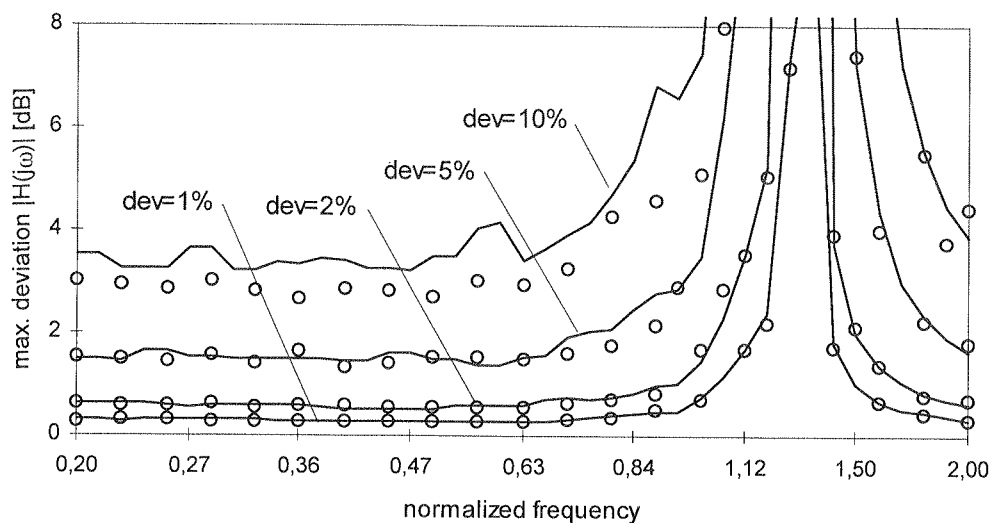


Fig. 9

Fig. 14. Maximum deviation of amplitude versus normalised frequency for 3rd order elliptic filter in Fig. 9 with a deviation in filter element values as a parameter; approximate values denoted as solid line and exact values as circles; flat distribution in element values was assumed

ltru
naczo
mentów

Rys. 14. Odchylenie maksymalne amplitudy w funkcji znormalizowanej częstotliwości dla filtru eliptycznego 3-go rzędu z Rys. 9 z odchyłką wartości elementów jako parametrem; wartości przybliżone oznaczono linią ciągłą, wartości dokładne oznaczono punktami; przyjęto jednostajny rozkład odchyłek wartości elementów

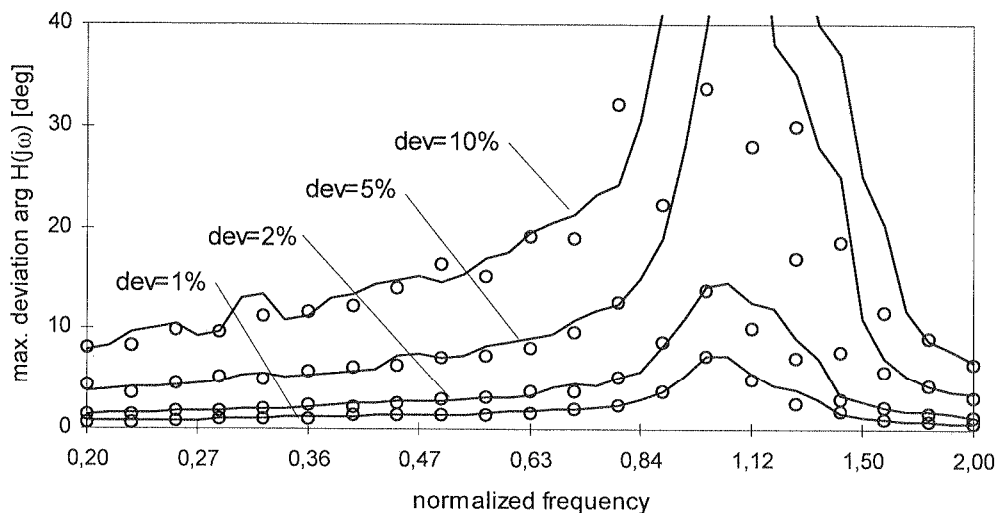


Fig. 9

Fig. 15. Maximum deviation of phase versus normalised frequency for 3rd order elliptic filter in Fig. 9 with a deviation in filter element values as a parameter; approximate values denoted as solid line and exact values as circles; flat distribution in element values was assumed

tu
naczo
mentów

Rys. 15. Odchylenie maksymalne fazy w funkcji znormalizowanej częstotliwości dla filtru eliptycznego 3-go rzędu z Rys. 9 z odchyłką wartości elementów jako parametrem; wartości przybliżone oznaczono linią ciągłą, wartości dokładne oznaczono punktami; przyjęto jednostajny rozkład odchyłek wartości elementów

Now, consider the all-pole low-pass LF filter that realizes 8th order Butterworth transfer function. A diagram of a general structure of all-pole low-pass IFLF filter is shown in Fig. 16. Here, we have $n=8$. As before, we use normalised values of elements and normalised frequency. We assume that $G_{b1} = G_{i,i-1} = -G_{j8} = 1, i=2,...,n, j=1,...,n$, i.e. all transconductors are the same (up to the sign). Normalised values of capacitances are: $C_1=1.560, C_2=1.827, C_3=1.729, C_4=1.526, C_5=1.259, C_6=0.937, C_7=0.578, C_8=0.195$. Normalised value of corner frequency equals 1.

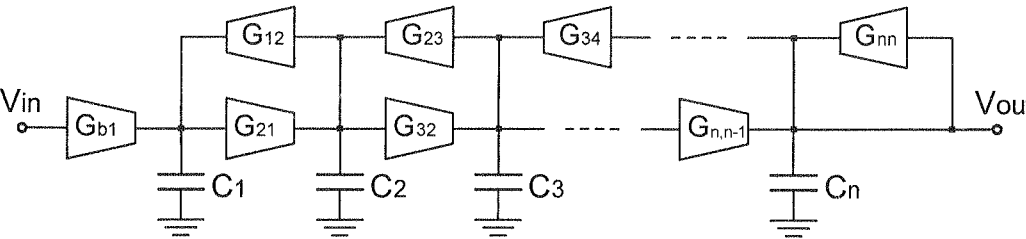


Fig. 16. Diagram of a n -th order all-pole low-pass LF filter structure

Rys. 16. Schemat filtru dolnoprzepustowego rzędu n w strukturze LF bez zer transmisyjnych

Figs.17 and 18 show the modulus and phase of the nominal transfer function of the filter, respectively. Figs.19,20 (Figs.21,22) show the results of Monte Carlo simulation (Worst Case analysis) of amplitude and phase of the transfer function. We have assumed 1%, 2%, 3% and 5% deviations for all filter elements. Figs.19 and 20 show standard deviation of amplitude and phase, respectively. Figs.21 and 22 show maximum deviation of amplitude and phase, respectively. Also in this case we can observe a very good agreement between the results obtained by means of the approximate and exact formulas.

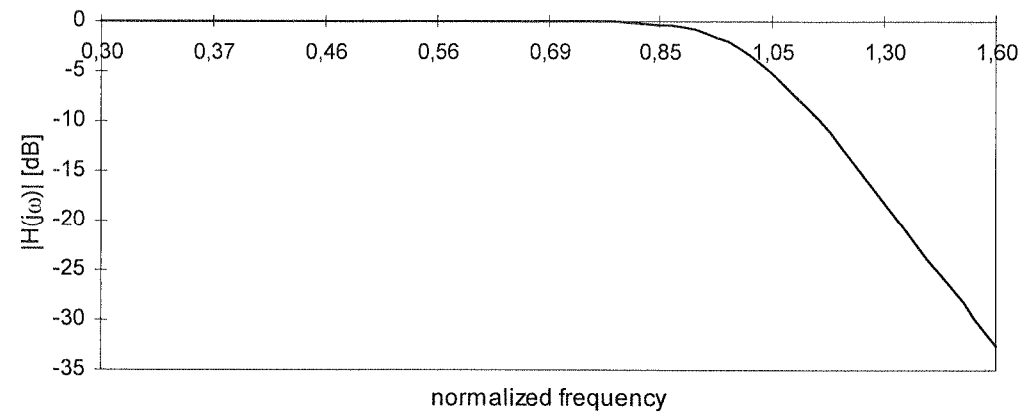


Fig. 17. Nominal amplitude characteristic $|H(j\omega)|$ of 8th order Butterworth LF filter in Fig. 16

Rys. 17. Nominalna charakterystyka amplitudowa $|H(j\omega)|$ filtru Butterwortha 8-go rzędu z Rys. 16

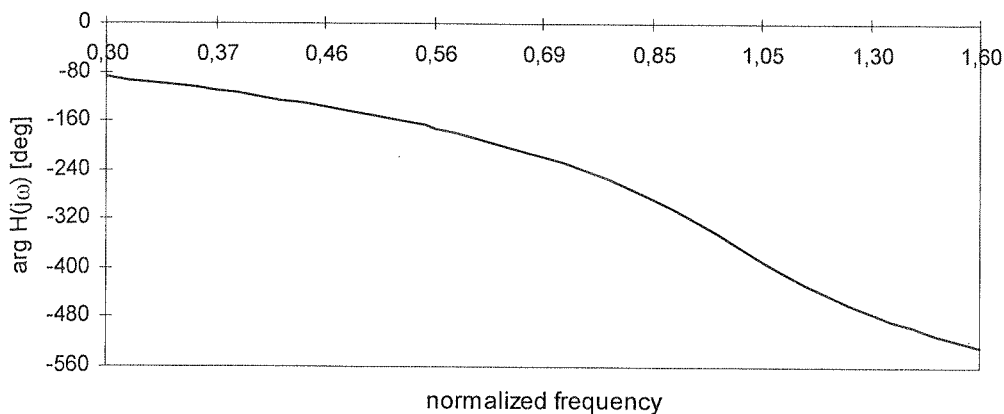


Fig. 18. Nominal phase characteristic $\arg[H(j\omega)]$ of 8th order Butterworth LF filter in Fig. 16

Rys. 18. Nominalna charakterystyka fazowa $\arg[H(j\omega)]$ filtru Butterwortha 8-go rzędu z Rys. 16

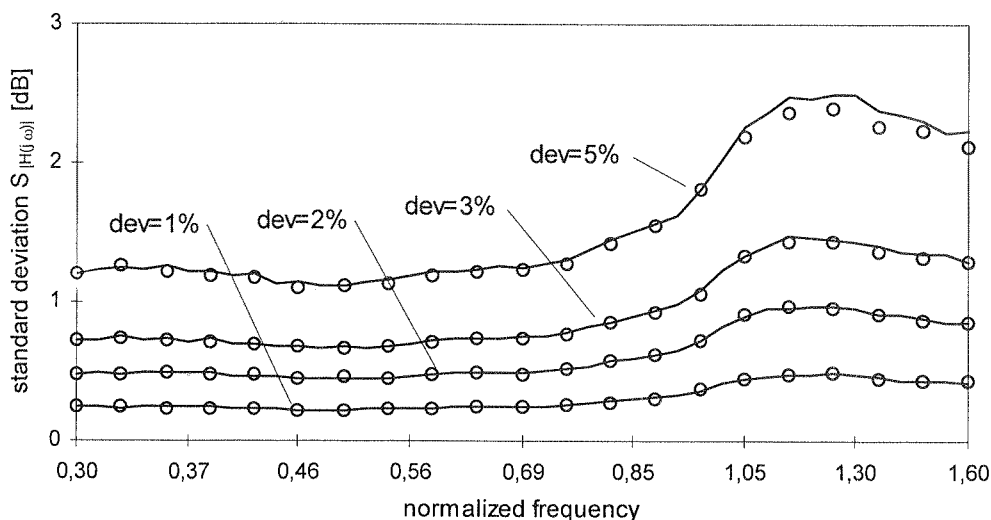


Fig. 19. Standard deviation of amplitude $S_{|H(j\omega)|}$ versus normalised frequency for 8th order Butterworth LF filter in Fig. 16 with a deviation in filter element values as a parameter; approximate values denoted as solid line and exact values as circles; gaussian distribution in element values was assumed

Rys. 19. Odchylenie standardowe amplitudy $S_{|H(j\omega)|}$ w funkcji znormalizowanej częstotliwości dla filtru Butterwortha 8-go rzędu z Rys. 16 z odchyłką wartości elementów jako parametrem; wartości przybliżone oznaczono linią ciągłą, wartości dokładne oznaczono punktami; przyjęto gaussowski rozkład odchyłek wartości elementów

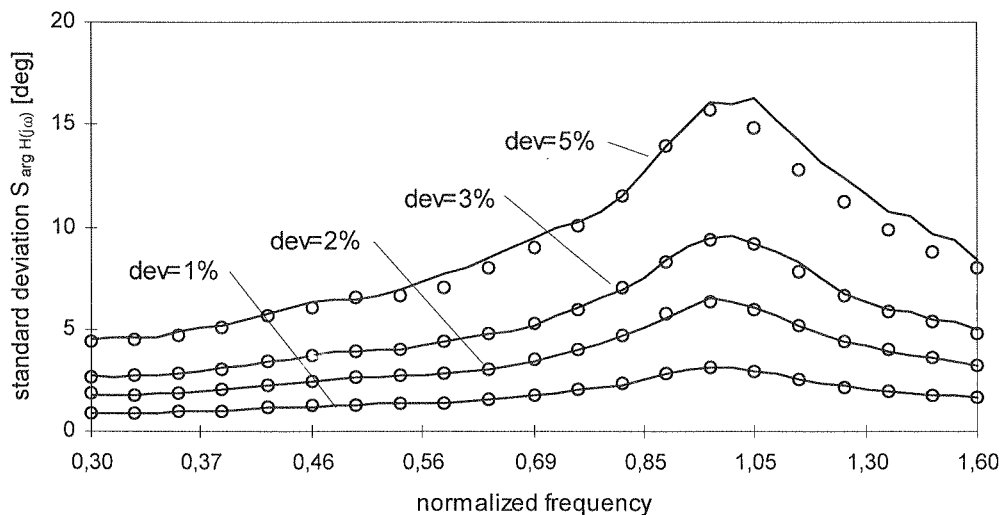


Fig. 20. Standard deviation of phase $S_{\arg H(j\omega)}$ versus normalised frequency for 8th order Butterworth LF filter in Fig. 16 with a deviation in filter element values as a parameter; approximate values denoted as solid line and exact values as circles; gaussian distribution in element values was assumed

Rys. 20. Odchylenie standardowe fazy $S_{\arg H(j\omega)}$ w funkcji znormalizowanej częstotliwości dla filtru Butterwortha 8-go rzędu z Rys. 16 z odchyłką wartości elementów jako parametrem; wartości przybliżone oznaczono linią ciągłą, wartości dokładne oznaczono punktami; przyjęto gaussowski rozkład odchyłek wartości elementów

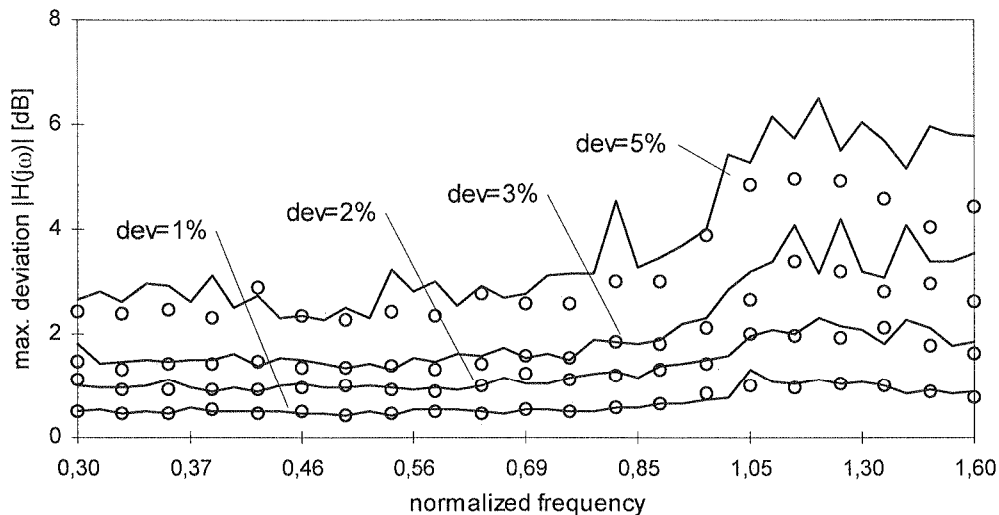


Fig. 21. Maximum deviation of amplitude $|H(j\omega)|$ versus normalised frequency for 8th order Butterworth LF filter in Fig. 16 with a deviation in filter element values as a parameter; approximate values denoted as solid line and exact values as circles; flat distribution in element values was assumed

Rys. 21. Odchylenie maksymalne amplitudy w funkcji znormalizowanej częstotliwości dla filtru Butterwortha 8-go rzędu z Rys. 16 z odchyłką wartości elementów jako parametrem; wartości przybliżone oznaczono linią ciągłą, wartości dokładne oznaczono punktami; przyjęto jednostajny rozkład odchyłek wartości elementów

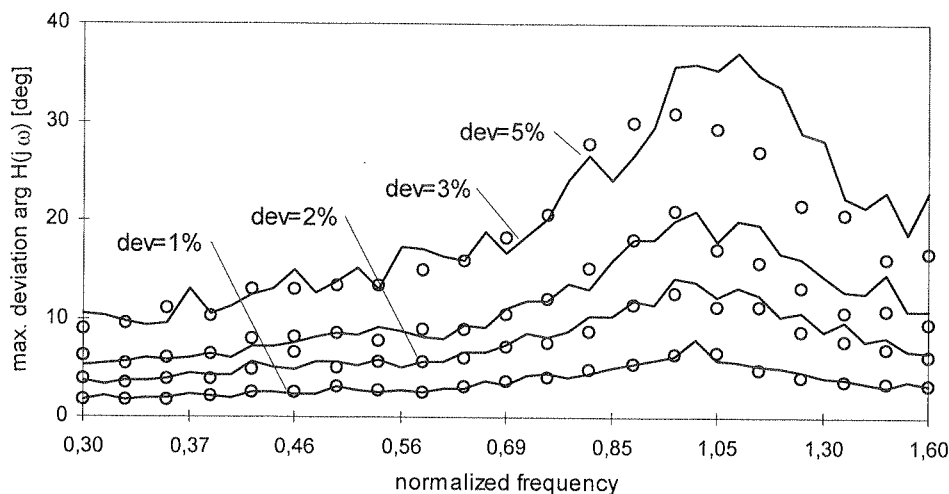


Fig. 22. Maximum deviation of phase versus normalised frequency for 8th order Butterworth LF filter in Fig. 16 with a deviation in filter element values as a parameter; approximate values denoted as solid line and exact values as circles; flat distribution in element values was assumed

Rys. 22. Odchylenie maksymalnej fazy w funkcji znormalizowanej częstotliwości dla filtra Butterwortha 8-go rzędu z Rys. 16 z odchyłką wartości elementów jako parametrem; wartości przybliżone oznaczono linią ciągłą, wartości dokładne oznaczono punktami; przyjęto jednostajny rozkład odchyłek wartości elementów

Since the proposed method is very fast, it can be used to make a comparison of different filter structures. As an example we shall compare different structures of all-pole canonical low-pass G_m -C filters that realize the same transfer function. Fig.23 shows a general structure of n -th order all-pole canonical low-pass G_m -C filter.

If the feedback signal at intrinsic node x_i in Fig.23 is taken from the node x_{i+1} for $i=1,2,\dots,n-1$ and the node x_n has the inner feedback loop then we obtain so-called leap-frog (LF) structure. If all feedback signals at nodes x_i , $i=1,2,\dots,n$ are taken from the node x_n then we get so-called inverse follow-the-leader (IFLF) structure. It is well known that IFLF and LF filters can be considered as two 'extreme' canonical structures of all-pole filters [23]. In general, the feedback signal emerging at node x_i can be taken from node x_j with $j=i+1,\dots,n$. This yields the total number $(n-1)!$ of different canonical multiple-loop integrator structures of order n .

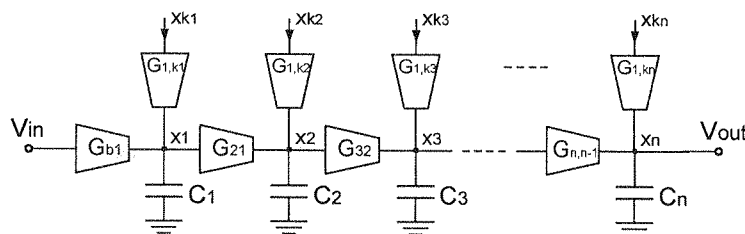


Fig. 23. A general structure of n -th order all-pole canonical low-pass G_m -C filter

Rys. 23. Ogólna struktura kanonicznego filtra dolnoprzepustowego G_m -C rzędu n bez zer transmisyjnych

Table 1

Standard deviation of amplitude for 5-th order all-pole low-pass Butterworth filters *

Odchylenie standardowe amplitudy dla dolnoprzepustowych filtrów Butterwortha 5-go rzędu

Filter structure	$\omega=0.5$			$\omega=1.0$			$\omega=2.0$		
	G+C dev.	C dev.	G dev.	G+C dev.	C dev.	G dev.	G+C dev.	C dev.	G dev.
2 3 4 5 5	0.19	0.05	0.18	0.29	0.19	0.22	0.32	0.22	0.23
2 3 5 5 5	0.19	0.06	0.18	0.30	0.20	0.22	0.34	0.22	0.26
2 4 4 5 5	0.19	0.06	0.18	0.42	0.28	0.32	0.33	0.22	0.24
2 4 5 5 5	0.19	0.06	0.18	0.49	0.28	0.40	0.36	0.23	0.27
2 5 4 5 5	0.16	0.07	0.14	0.48	0.30	0.36	0.33	0.23	0.24
2 5 5 5 5	0.16	0.07	0.14	0.51	0.31	0.40	0.36	0.23	0.28
3 3 4 5 5	0.22	0.09	0.19	0.40	0.23	0.31	0.32	0.22	0.23
3 3 5 5 5	0.21	0.10	0.19	0.40	0.25	0.33	0.34	0.22	0.26
3 4 4 5 5	0.28	0.09	0.26	0.55	0.35	0.43	0.34	0.23	0.24
3 4 5 5 5	0.30	0.10	0.27	0.66	0.34	0.56	0.36	0.23	0.28
3 5 4 5 5	0.19	0.09	0.17	0.60	0.37	0.47	0.33	0.22	0.24
3 5 5 5 5	0.21	0.10	0.19	0.62	0.37	0.52	0.36	0.23	0.29
4 3 4 5 5	0.23	0.10	0.21	0.34	0.19	0.27	0.32	0.22	0.24
4 3 5 5 5	0.23	0.11	0.21	0.39	0.22	0.32	0.34	0.22	0.26
4 4 4 5 5	0.28	0.10	0.26	0.57	0.33	0.46	0.33	0.23	0.25
4 4 5 5 5	0.29	0.11	0.27	0.71	0.34	0.62	0.36	0.23	0.28
4 5 4 5 5	0.21	0.10	0.19	0.60	0.36	0.49	0.33	0.23	0.24
4 5 5 5 5	0.22	0.11	0.20	0.67	0.37	0.58	0.37	0.23	0.29
5 3 4 5 5	0.26	0.11	0.24	0.33	0.17	0.28	0.32	0.23	0.23
5 3 5 5 5	0.26	0.12	0.23	0.33	0.20	0.26	0.34	0.22	0.26
5 4 4 5 5	0.31	0.12	0.30	0.54	0.32	0.43	0.34	0.23	0.24
5 4 5 5 5	0.32	0.12	0.29	0.66	0.32	0.59	0.37	0.23	0.28
5 5 4 5 5	0.23	0.11	0.21	0.63	0.36	0.49	0.33	0.23	0.24
5 5 5 5 5	0.24	0.12	0.21	0.64	0.36	0.53	0.36	0.23	0.29

* - G+C dev. means that both transconductors and capacitors were deviated, G dev. (resp. C dev.) refers to the situation when only transconductors (respectively: capacitors) were deviated.

Filter structure
2 3 4
2 3 5
2 4 4
2 4 5
2 5 4
2 5 5
3 3 4
3 3 5
3 4 4
3 4 5
3 5 4
3 5 5
4 3 4
4 3 5
4 4 4
4 4 5
4 5 4
4 5 5
5 3 4
5 3 5
5 4 4
5 4 5
5 5 4
5 5 5

* - G+C dev. means that both transconductors and capacitors were deviated, G dev. (resp. C dev.) refers to the situation when only transconductors (respectively: capacitors) were deviated.

sequ
the f
struc

Table 1

Table 2

Maximum deviation of amplitude for 5-th order all-pole low-pass Butterworth filters*
 Odchylenie maksymalne amplitudy dla dolnoprzepustowych filtrów Butterwortha 5-go rzędu

Filter structure	$\omega=0.5$			$\omega=1.0$			$\omega=2.0$		
	G+C dev.	C dev.	G dev.	G+C dev.	C dev.	G dev.	G+C dev.	C dev.	G dev.
2 3 4 5 5	0.36	0.08	0.35	0.56	0.29	0.42	0.70	0.39	0.45
2 3 5 5 5	0.37	0.09	0.35	0.61	0.31	0.45	0.75	0.40	0.48
2 4 4 5 5	0.36	0.09	0.32	0.81	0.46	0.59	0.65	0.40	0.48
2 4 5 5 5	0.38	0.09	0.35	0.96	0.47	0.77	0.68	0.43	0.51
2 5 4 5 5	0.30	0.09	0.24	0.97	0.52	0.67	0.67	0.37	0.47
2 5 5 5 5	0.29	0.09	0.25	0.94	0.53	0.81	0.71	0.39	0.52
3 3 4 5 5	0.46	0.16	0.33	0.76	0.42	0.62	0.84	0.42	0.46
3 3 5 5 5	0.43	0.17	0.41	0.88	0.41	0.62	0.74	0.42	0.46
3 4 4 5 5	0.64	0.17	0.52	0.16	0.62	0.83	0.76	0.41	0.46
3 4 5 5 5	0.61	0.19	0.54	1.35	0.60	1.11	0.73	0.44	0.52
3 5 4 5 5	0.37	0.17	0.35	1.19	0.58	0.89	0.68	0.40	0.50
3 5 5 5 5	0.48	0.17	0.34	1.31	0.61	1.00	0.76	0.42	0.57
4 3 4 5 5	0.46	0.15	0.44	0.67	0.29	0.51	0.70	0.39	0.45
4 3 5 5 5	0.53	0.17	0.40	0.79	0.38	0.66	0.74	0.39	0.46
4 4 4 5 5	0.54	0.16	0.50	1.12	0.55	0.93	0.68	0.43	0.45
4 4 5 5 5	0.60	0.19	0.52	1.54	0.57	1.27	0.83	0.40	0.59
4 5 4 5 5	0.41	0.15	0.35	1.21	0.58	0.95	0.68	0.40	0.46
4 5 5 5 5	0.43	0.19	0.36	1.40	0.60	1.20	0.75	0.39	0.57
5 3 4 5 5	0.61	0.18	0.44	0.67	0.26	0.53	0.65	0.41	0.44
5 3 5 5 5	0.53	0.18	0.44	0.65	0.32	0.51	0.66	0.42	0.56
5 4 4 5 5	0.60	0.17	0.58	1.14	0.59	0.76	0.65	0.40	0.54
5 4 5 5 5	0.68	0.19	0.62	1.42	0.50	1.22	0.72	0.42	0.60
5 5 4 5 5	0.51	0.17	0.38	1.27	0.57	0.93	0.63	0.42	0.48
5 5 5 5 5	0.53	0.19	0.39	1.32	0.60	1.12	0.76	0.40	0.52

* – G+C dev. means that both transconductors and capacitors were deviated, G dev. (resp. C dev.) refers to the situation when only transconductors (respectively: capacitors) were deviated.

To simplify further description, we introduce the following notation: let the sequence $j_1, j_2, \dots, j_n$ denote the canonical all-pole low-pass structure of n -th order, for which the feedback signal emerging at node x_i is taken from node x_{j_i} . For example, n -th order LF structure will be denoted as 2,3,4,..., n , while n -th order IFLF structure as $n, n, \dots, n$.

Here, we consider 5-th order Butterworth filters. According to the discussion above, there are 24 different canonical structures that realize this transfer function. They can be implemented with the same value of transconductances for all filter transconductors (see [23] for more details). Table 1 and 2 show the standard and maximum deviation of amplitude, respectively, for all considered filter structures, obtained for 5000 random samples. 1% deviation in capacitance elements, transconductance elements or all filter elements has been assumed (with gaussian distribution for standard deviation and with flat distribution for maximum deviation). The simulations have been performed for normalised frequency equal to 0.5, 1.0 and 2.0. It follows that for low frequency ($\omega=0.5$) transconductance element deviations determine the deviation of transfer function of the filter. For higher frequencies ($\omega=2.0$) the contributions of transconductance and capacitance element deviations becomes roughly the same. It is also seen that transfer function deviation is larger for IFLF filter than for LF one. Note that there is a kind of monotonic dependence between the filter structure and transfer function deviation. In particular, the more the filter is similar to LF one, the smaller transfer function deviation it exhibits. The similarity is 'measured' by the number of feedback transconductors located in the same place as in LF filter (see also [25] for more detailed discussion on this subject).

5. CONCLUSIONS

In the paper, an efficient procedure for calculating deviation of filter transfer function given deviations of filter element values has been developed. The procedure has been used to perform statistical analysis of G_m -C filters. It follows that accuracy of the results obtained by means of the method is very good, provided that the deviations of filter element values are not larger than 5–10% (depending on the considered filter). A number of examples that illustrate the proposed approach have been presented including statistical analysis of family of filters realizing the same all-pole low-pass transfer function of 5th order. The computational complexity of the proposed procedure is $O(n^2)$, while the complexity of the exact method is $O(n^3)$ since it involves inverting $n(n$ matrix during each iteration. It is also worth noticing that the proposed method is versatile, since it can be applied to any G_m -C filter structure using the same general formulas (and, as a consequence, the same software packages).

6. ACKNOWLEDGEMENT

The author would like to thank the anonymous reviewer for his comments and remarks which were helpful in improving the quality of the paper.

7. REFERENCES

1. R. Schaumann, M.S. Ghausi, K.R. Laker: *Design of Analog Filters, Passive, Active RC, and Switched Capacitor*. Englewood Cliff, NJ: Prentice-Hall, 1990.

2. T.
3. C.
L.
4. R.
an
5. E.
Ty
6. M.
ce
7. G.
D.
8. G.
IE
9. B.
10. J.
fil
11. R.
fr
12. Y.
19
13. S.
Th
14. J.
th
15. J.
fo
PE
16. M.
N.
17. G.
Th
18. G.
vo
19. J.
vo
20. R.
S.
21. S.
St
vo
22. S.
B.
PE
23. S.
F.
24. S.
G.
25. S.
Pr
Pr

2. T. Deliyannis, Y. Sun, J.K. Fidler: *Continuous-time active filter design*. CRC Press, USA, 1999.
3. C. Toumazou, J. Lidgey, D. Haigh, Eds.: *Analogue IC Design – The Current-Mode Approach*. London, U.K.: IEE, Apr. 1990.
4. R.L. Geiger, E. Sánchez-Sinencio: *Active filter design using operational transconductance amplifiers: A tutorial*. IEEE Circuit and Devices Mag., March 1985, vol. 1, pp. 20–32.
5. E. Sánchez-Sinencio, R.L. Geiger, H. Nevarez-Lozano: *Generation of Continuous-Time Two Integrator Loop OTA Filter Structures*. IEEE Trans. Circuits Syst., August 1988, vol. 35, pp. 936–946.
6. M.A. Tan, R. Schaumann: *Design of a General Biquadratic Filter Section with Only Transconductances and Grounded Capacitors*. IEEE Trans. Circuits Syst.-II, April 1988, vol. 35, pp. 478–480.
7. G. Groenewold: *Optimal Dynamic Range Integrated Continuous-Time Filters*. Delft, The Netherlands: Delft Univ. Press, 1992.
8. G. Groenewold: *The Design of High Dynamic Range Continuous-Time Integratable Bandpass Filters*. IEEE Trans. Circuits and Syst., August 1991, vol. 38, No.8, pp. 838–852.
9. B. Nauta: *Analog CMOS filters for very high frequencies*. Kluwer Academic Publishers, 1993.
10. J. Silva-Martinez, M.S.J. Steyaert, W. Sansen: *A 10.7MHz 68-dB CMOS continuous-time filter with on-chip automating tuning*. IEEE J. Solid-State Circuits, 1992, vol. 27, No. 12, pp. 1843–1853.
11. R. Allini, A. Baschirotto, R. Castello: *Tuneable BiCMOS continuous-time filter for high-frequency applications*. IEEE J. Solid-State Circuits, 1992, vol. 27(12), pp. 1905–1915.
12. Y.P. Tsividis: *Integrated continuous-time filter design – An overview*. IEEE J. Solid-State Circuits, March 1994, vol. 29, pp. 166–176.
13. S. Szczepański, J. Jakusz, R. Schaumann: *A linear CMOS OTA for VHF applications*. IEEE Trans. Circuits Syst.-II, March 1997, vol.44, pp. 174–187.
14. J. Mahattanakul, C. Toumazou: *Current-Mode Versus Voltage-Mode G_m -C Biquad Filters: What the Theory Says*. IEEE Trans. Circuits Syst.-II, February 1998, vol. 45, pp. 173–186.
15. J. Glinianowicz, J. Jakusz, S. Szczepański, Y. Sun: *High-frequency two-input CMOS OTA for continuous-time filter applications*. IEE Proc.-Circuits Dev. Syst., February 2000, vol.147, No.1, pp.13–18.
16. M.S. Ghausi, K.R. Laker: *Modern Filter Design: Active RC and Switched Capacitor*. Englewood Cliffs, NJ, Prentice-Hall, Inc., 1981.
17. G.W. Roberts, A.S. Sedra: *A general class of current amplifier-based biquadratic filter circuits*. IEEE Trans. Circuits Syst.-I, April 1992, vol. 38, pp. 257–263.
18. G.W. Roberts, A.S. Sedra: *All current-mode frequency selective circuits*. Electron. Lett., June 1989, vol. 25, pp. 759–761.
19. J.D. Schoeffler: *The synthesis of minimum sensitivity networks*. IEEE Trans. Circuit Theory, 1964, vol. 11, pp. 271–276.
20. R. Mackay, A.S. Sedra: *Generation of low-sensitivity state-space active filters*. IEEE Trans. Circuit Syst., October 1980, vol. CAS-27, pp. 863–870.
21. S. Koziel, S. Szczepański: *Dynamic Range Comparison of Voltage-Mode and Current-Mode State-Space G_m -C Biquad Filters in Reciprocal Structures*. Proc. IEEE Int. Conf. Elect. Circuits Syst., 2001, vol. II, pp. 819–822.
22. S. Koziel, S. Szczepański, R. Schaumann: *General Approach to Continuous-Time G_m -C Filters Based on Matrix Description*. Proc. Int. Symp. Circuits, Syst., ISCAS, Phoenix, AZ, USA, 2002, vol.IV, pp. 647–650.
23. S. Koziel, S. Szczepański, R. Schaumann: *General Approach to Continuous-Time G_m -C Filters*. To appear, 2003, Int. J. Circuit Theory and Applications.
24. S. Koziel, S. Szczepański: *Structure Generation and Performance Comparison of Canonical Elliptic G_m -C Filters*. Proc. Int. Conf. Electron. Circuits, Syst., ICECS, vol. I, pp. 157–160, Dubrovnik, Croatia, 2002.
25. S. Koziel, S. Szczepański: *Sensitivity Properties of All-Pole Canonical Low-Pass G_m -C Filters*. Proc. IEEE Int. Conf. Circuits Syst. for Communications, ICCSC'2002, St. Petersburg, Russia, 2002, pp. 54–57.

S. KOZIEŁ

EFEKTYWNA ANALIZA TOLERANCJI FILTRÓW AKTYWNYCH G_m -C CZASU CIĄGŁEGO

Streszczenie

W pracy przedstawiono efektywną metodę analizy tolerancji dla filtrów aktywnych G_m -C czasu ciągłego. Zaprezentowano ogólny opis macierzowy filtrów G_m -C, który obejmuje wszystkie możliwe realizacje filtrów w rozważanej klasie elementów. Zaletą tego opisu jest między innymi to, że wszystkie macierze występujące w równaniach mogą być tworzone bezpośrednio przez wgląd w schemat układu.

W oparciu o przedstawiony model wyprowadzono formuły umożliwiające wyznaczanie tolerancji funkcji przenoszenia filtru (zarówno modułu jak i fazy) przy zadanych odchyłkach wartości elementów filtru od wartości nominalnych. Zastosowano pewne przybliżenie, które przy założeniu małych odchyłek wartości elementów filtru umożliwia obliczanie odchyłki funkcji przenoszenia bez konieczności odwracania macierzy głównej układu równań opisujących filtr przy każdorazowej zmianie wartości odchyłek elementów. Powoduje to zmniejszenie asymptotycznej złożoności obliczeniowej całej procedury z $O(n^3)$ do $O(n^2)$.

Przedstawioną metodę wykorzystano do przeprowadzenia analizy statystycznej filtrów G_m -C. W szczególności, dokonano analizy Monte Carlo i analizy najgorszego przypadku (Worst Case) dla wybranych struktur filtrów różnych rzędów. Zaprezentowano również wyniki analizy porównawczej rodziny filtrów realizujących tą samą dolnoprzepustową funkcję przenoszenia piątego rzędu.

Przedstawione wyniki eksperymentów pokazują, że wzory przybliżone dają bardzo dobre rezultaty przy odchyłkach wartości elementów układu nie przekraczających 5–10% od wartości nominalnej w zależności od badanej struktury filtru, przy czym uzyskano kilkukrotne zmniejszenie czasu obliczeń (w porównaniu do metody dokładnej) już dla filtrów o średniej złożoności (6–8 rzędu). Należy również zaznaczyć, że opisana procedura jest uniwersalna i pozwala analizować dowolne struktury filtrów G_m -C za pomocą tego samego programu komputerowego. Może ona zatem być wykorzystana w systemach automatycznego projektowania i analizy filtrów tej klasy.

Słowa kluczowe: filtry G_m -C, analiza tolerancji, analiza statystyczna

Implementacja operacji konwolucji o stałych współczynnikach w układach FPGA

ERNEST JAMRO, KAZIMIERZ WIATR

*Katedra Elektroniki, Akademia Górniczo-Hutnicza
30-059 Kraków, al. Mickiewicza 30,
wiatr/jamro@uci.agh.edu.pl*

*Otrzymano 2002.07.10
Autoryzowano 2002.11.20*

W artykule omówiono różne architektury układu konwolwera zoptymalizowane pod kątem implementacji w układach programowalnych FPGA. Współczynniki konwolucji są stałe, jednakże układ FPGA może być szybko przeprogramowany co pozwala na zmianę tych współczynników. Mnożenie jest najbardziej skomplikowaną operacją wykonywaną podczas obliczania operacji konwolucji dlatego w pierwszej części artykułu zostaną omówione układy mnożące. Niemniej rozbieżność operacji mnożenia wewnątrz układu konwolwera pozwala na lepszą optymalizację operacji konwolucji, dlatego w artykule porównano dwie architektury: układ konwolwera zbudowany przy użyciu pamięci LUT (ang. *Look-Up Table*) i nazywany LC (ang. *LUT-based Convolver*) oraz układ konwolwera zbudowany jako suma mnożeń wykonywanych przy użyciu układów mnożących opartych o pamięć LUT i nazywanych LM (ang. *LUT based Multiplier*). Ponadto omówiono alternatywną technikę: konwolwer oparty na (równoległej) arytmetyce rozproszonej (ang. *Parallel*) i nazywaną DAC (ang. *Distributed Arithmetic Convolver*). Głównym celem tego artykułu jest przedstawienie nowej architektury układu konwolwera wykorzystującego nieregularną arytmetykę rozproszoną IDAC (ang. *Irregular Distributed Arithmetic Convolver*), która to w porównaniu z architekturą DAC jest nieregularna, a przez to pozwala na lepszą optymalizację układu konwolwera. Wszystkie architektury konwolwera omówione w tym artykule mogą być automatycznie generowane przez autorskie narzędzie AuToCon.

Słowa kluczowe: procesory specjalizowane, architektury elementów obliczeniowych, arytmetyka rozproszona, programowalne układy logiczne

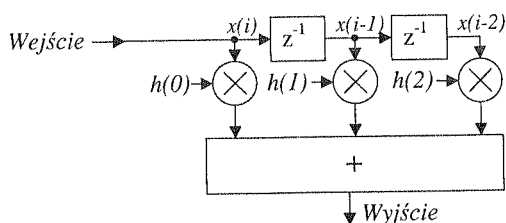
1. WSTĘP

Operacja konwolucji, która jest wykonywana podczas np. filtracji FIR może być wyrażona jako arytmetyczna suma mnożeń:

$$y(i) = \sum_{k=0}^{N-1} h(k) \cdot x(i-k) \quad (1)$$

gdzie: $y(i)$, $x(i)$ oraz $h(i)$ oznaczają odpowiednio odpowiedź i wejście w czasie i oraz współczynnik konwolucji.

Operacja konwolucji jest z reguły wykonywana przy użyciu układów mnożących a następnie układu dodającego. Przykład układu konwolwera jest pokazany na rys. 1.



Rys. 1. Układ konwolwera dla okna o rozmiarze 3×1

Fig. 1. Convolution block with window 3×1

Mnożenie jest najbardziej skomplikowaną operacją wykonywaną podczas operacji konwolucji i parę różnych technik zostało opracowanych w celu optymalizacji tej operacji. Po pierwsze współczynniki mnożenia są bardzo często stałe dlatego wartości współczynników mnożenia mogą być „wbudowane” w układ mnożący, to rozwiązanie jest dalej oznaczane jako KCM (ang. *Constant Coefficient Multiplier*). Układ mnożący KCM zajmuje w zależności od wartości współczynnika średnio 17÷23%, i maksymalnie 29÷41% obszaru zajmowanego przez w pełni funkcyjny układ mnożący nazywany dalej Variable Coefficient Multiplier (VCM) [1]. Konsekwentnie układ mnożący KCM powinien być użyty kiedy współczynnik mnożenia jest stały.

Alternatywnie, kiedy współczynniki mnożenia zmieniają się rzadko, układ programowalny FPGA (ang. *Field Programmable Gate Arrays*) może być przeprogramowany (w całości lub też tylko jego część) w celu zmiany wartości współczynników. Układy FPGA mogą być (częściowo) przeprogramowane w czasie rzędu paru milisekund, niemniej układ mnożący musi być powtórnie zaprojektowany i zaimplementowany, co jest dużo bardziej czasochłonne niż samo przeprogramowanie układu FPGA. Dlatego przeprogramowanie układu FPGA w celu zmiany współczynnika mnożenia może być praktycznie użyte tylko w przypadku, gdy liczba dozwolonych wartości współczynników mnożenia jest ograniczona. W tym przypadku, każdy współczynnik mnożenia (lub też zestaw współczynników mnożenia w przypadku konwolucji) powinien być uprzednio zaimplementowany a następnie odpowiednio użyty do przeprogramowania układu FPGA.

Układy FPGA w celu implementacji logiki używają pamięci LUT, dlatego bezpośrednim sposobem implementacji operacji mnożenia wydaje się użycie układu mnożącego LM z pamięcią LUT (ang. *LUT-based Multiplier*) [2, 3], gdzie wartość współczynnika mnożenia jest zakodowana przez zawartość pamięci LUT.

(1)

i oraz

ających

s. 1.

Alternatywną metodą implementacji operacji konwolucji jest zastosowanie układu konwolwera w oparciu o arytmetykę rozproszoną DAC (ang. *Distributed Arithmetic Convolver*) [4, 5]. Układ DAC wykorzystuje pamięć LUT podobnie jak LM, niemniej operacja konwolucji jest wykonywana w sekwencji bitowej tj. tak jakby wejście układu konwolwera miało szerokość tylko jednego bitu, a następnie poszczególne wyniki cząstkowe układów 1-bitowych są ze sobą odpowiednio sumowane. W konsekwencji wszystkie wejścia do pamięci LUT mają taką samą wagę bitową, dzięki czemu szerokość magistrali danych pamięci LUT jest mniejsza w porównaniu z układem LM. Układ DAC ma regularną strukturę, taką samą dla różnych wag bitów wejściowych, a to powoduje, że swoboda grupowania wejść do pamięci LUT jest ograniczona. Pozbawione tej wady jest nowatorskie rozwiązanie nieregularnej arytmetyki rozproszonej IDAC (ang. *Irregular Distributed Arithmetic Convolver*), dla którego struktura układu DAC różni się dla różnych wag bitów wejściowych.

2. UKŁAD MNOŻĄCY O STAŁYM WSPÓŁCZYNNIKU (KCM)

2.1. UKŁAD MNOŻĄCY BEZMNOŻNY (MM)

operacji

operacji.

współczyn-

t dalej

KCM

9÷41%

variable

en być

d pro-

nowany

Układy

emniej

st dużo

progra-

tycznie

nia jest

współczyn-

towany

pośred-

ego LM

zynnika

Układ mnożący KCM jest często implementowany przy użyciu układu bezmnożnego MM (ang. *Multiplierless Multiplier*), dla którego używa się tylko przesunięć bitowych oraz układów dodających. Schemat takiego układu mnożącego jest określany przez niezerowe bity współczynnika mnożącego. Na przykład wartość wejściowa A pomnożona przez współczynnik $B = 14 = 1110_2$ jest implementowana jako $(A \ll 1) + (A \ll 2) + (A \ll 3)$, gdzie ' $\ll$ ' oznacza przesunięcie bitowe w lewo. Należy podkreślić, że koszt (zajmowany obszar układu scalonego) układu mnożącego zależy bezpośrednio od wartości współczynnika mnożącego – liczba jedynek w jego binarnej reprezentacji powinna być jak najmniejsza.

Postać kanoniczna cyfry ze znakiem (CSD)

W reprezentacji CSD (ang. *Canonic Signed Digit*) [6, 7] uzyskuje się redukcję obszaru zajmowanego przez układ mnożący poprzez redukcję jedynek w reprezentacji liczby. W reprezentacji CSD w porównaniu z reprezentacją binarną wprowadzona została dodatkowa cyfra $\bar{1}$ odpowiadająca wartości -1 i symbolizująca operację odejmowania (zamiast dodawania). Dla przykładu mnożenie przez $B = 15 = 1111_2 = 1000 \bar{1}$ może być zrealizowane jako odejmowanie $(A \ll 4) - A$.

Warto zwrócić uwagę, że standardowy algorytm konwersji do kodu CSD szacuje koszt operacji dodawania i odejmowania jako taki sam. Jednakże w przypadku odejmowania bity najmniej znaczące (ang. LSB) przesuniętego w prawo odjemnika nie mogą być bezpośrednio kopiowane na wyjście jak to ma miejsce w przypadku dodawania. Dlatego zaproponowano zmodyfikowany algorytm konwersji [3], dla którego symbol $\bar{1}$ jest wstawiany tylko wtedy gdy dzięki temu wstawieniu liczba niezerowych bitów CSD zostanie zredukowana. W niniejszym artykule używa się więc zmodyfikowanego algorytmu konwersji.

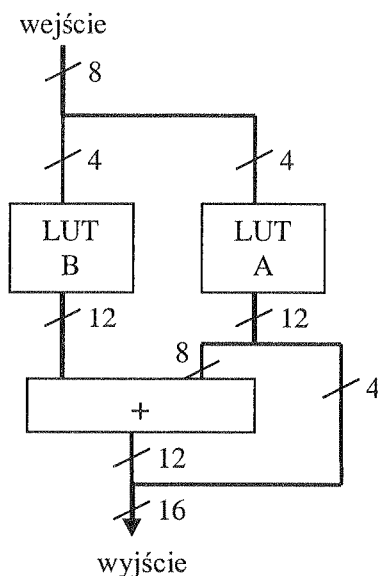
Współdzielenie wspólnej podstruktury (SS)

Inną metodą redukcji obszaru zajmowanego przez układ mnożący jest dzielnie wspólnej podstruktury SS (ang. *Sub-structure Sharing*) [8]. Na przykład, mnożenie przez liczbę $27 = 11011_2$ może być zrealizowane z użyciem dodatkowej zmiennej pomocniczej tmp , $tmp = a + (a \ll 1)$, jako $27 \cdot a = tmp + (tmp \ll 3)$.

Operacja SS może być również przeprowadzona na reprezentacji CSD, dlatego kombinacja SS i CSD została również uwzględniona podczas implementacji algorytmu optymalizacji generowanego układu.

2.2. MNOŻENIE Z UŻYCIEM PAMIĘCI LUT

Obliczenie dowolnej ograniczonej wartościowo funkcji możliwe jest z wykorzystaniem pamięci typu LUT, na której wejście adresowe podane są wartości argumentów a na wyjściu otrzymuje się wynik funkcji. Zastosowanie pamięci LUT daje najszybszą w teorii implementację funkcji jako, że nie jest wymagane zastosowanie żadnych układów arytmetycznych. Niestety zastosowanie w praktyce pojedynczej pamięci LUT dla mnożenia jest możliwe tylko dla małych wartości argumentu, ponieważ wielkość pamięci wzrasta wykładniczo wraz z szerokością argumentu. Na przykład, dla L -bitowego argumentu i K -bitowego współczynnika mnożenia, wymagana jest pamięć o głębokości 2^L i szerokości $(L+K)$ bitów. Dlatego praktycznym rozwiązaniem jest zastosowanie kombinacji małych pamięci LUT i układów dodających, poprzez rozbicie argumentu wejściowego, użycie małych pamięci LUT i następnie odpowiednie dodanie wyjść pamięci LUT [2, 3]. Przykład układu mnożącego jest pokazany na rys. 2.



Rys. 2. Układ mnożący LM dla szerokości argumentu $L=8$ i szerokości współczynnika mnożącego $K=8$

Fig. 2. Look-up table based multiplication for 8-bit wide input and coefficient

Zawartość pamięci LUT dla mnożenia $Y=A \cdot B$ może być obliczona bezpośrednio z mnożenia tak jak to jest pokazane na przykładzie w tab. 1.

Tabela 1

Zawartość komórek pamięci LUT ($y_5 - y_0$) dla różnych wartości argumentu podanego na magistralę adresową; współczynnik mnożenia wynosi 19; szerokość adresu – szerokość magistrali adresowej dla poszczególnych komórek pamięci.

The contents of the LUT memory ($y_5 - y_0$) for different address values and the coefficient equal 19.
Address width – the width of address bus for each memory cell

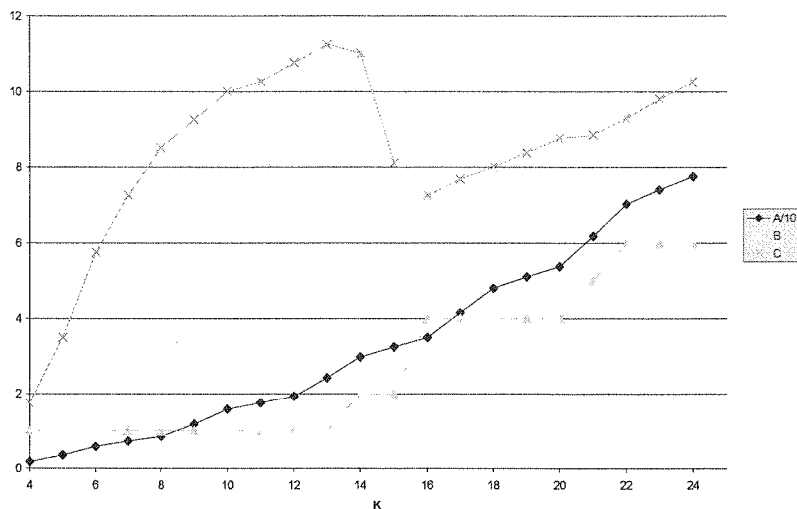
Adress	Wartość	y_5	y_4	y_3	y_2	y_1	y_0
0	0	0	0	0	0	0	0
1	19	0	1	0	0	1	1
2	38	1	0	0	1	1	0
3	57	1	1	1	0	0	1
Szerokość adresu		1	1	2	2	2	1

Warto zwrócić uwagę, że zawartość komórek pamięci LUT zależy tylko od linii adresowych o wadze mniejszej lub równej wadze komórki wyjściowej. Dlatego w przedstawionym przykładzie w tab. 1, komórka pamięci y_0 zależy tylko od linii adresowej a_0 ; y_1 zależy od a_0 i a_1 ; itd. Maksymalna szerokość magistrali jest jednak ograniczona od góry przez szerokość argumentu, dlatego szerokość magistrali adresowej komórki y_i w ogólnym przypadku wynosi $MAX(i+1, n)$, gdzie n – szerokość magistrali adresowej pamięci LUT. W konsekwencji $n-1$ najmniej znaczących bitów pamięci wymaga mniejszej szerokości magistrali adresowej a przez to i mniejszych rozmiarów pamięci, co powoduje znaczącą redukcję rozmiaru pamięci i będzie oznaczane dalej jako LAWR (ang. *LSBs Address Width Reduction*).

Dodatkową redukcję szerokości magistrali adresowej można zaobserwować dla komórek pamięci, dla których zawartość pamięci nie zależy od pewnej linii adresowej. Osiągnięte dzięki tej metodzie oszczędności będą dalej w skrócie nazywane jako DAWR (ang. *Don't care Address Width Reduction*). DAWR jest z reguły obserwowany dla najbardziej znaczących bitów pamięci LUT, co potwierdza przykład w tab. 1, dla którego oszczędności DAWR występują dla komórek pamięci y_5 i y_4 . Dodatkowe oszczędności można osiągnąć dzięki dzieleniu wspólnych komórek pamięci MS (ang. *Memory Sharing*), jeżeli te komórki mają taką samą zawartość. Na przykładzie tab. 1 można zauważyć, że komórki pamięci y_0 i y_4 są takie same, dlatego tylko jedna z nich może być zastosowana. Warto w tym miejscu podkreślić, że redukcję szerokości magistrali adresowej DAWR oraz dzielenie wspólnej pamięci MS nie można uogólnić i zależą one silnie od wartości współczynnika mnożenia i szerokości magistrali adresowej. Ponadto, wymagany jest zaawansowany algorytm przeszukujący i implementujący wspomniane techniki. W przedstawionych rozważaniach zastosowano algorytm porównujący wszystkie możliwe kombinacje komórek i linii adresowych.

W układzie mnożącym rozbić dużej pamięci LUT na mniejsze powinno być przeprowadzone z uwzględnieniem relacji koszt-rozmiar pamięci oraz kosztów układów dodających. Układy XC4000 firmy Xilinx mają wbudowane pamięci 16×1 oraz 32×1 , jednakże obszar zajmowany przez pamięci 32×1 jest dwa razy większy niż dla pamięci 16×1 , dlatego w praktyce rekomendowane są przede wszystkim pamięci 16×1 [1].

Układy Virtex firmy Xilinx oferują jeszcze bogatszy wybór pamięci, oprócz pamięci 16×1 i 32×1 wprowadzono w układzie Virtex dodatkowe duże bloki pamięci BSR (ang. *BlockSelectRAM*) o rozmiarze 4kb i różnej szerokości magistrali danych: $4k \times 1$, $2k \times 2$, $1k \times 4$, 512×8 , 256×16 [9]. Powierzchnia krzemu zajmowana przez pojedynczy BSR jest porównywalna do powierzchni zajmowanej przez 16 komórek Virtex CLB (może pomieścić 64 pamięci 16×1 RAM). Jednakże rzeczywisty koszt tych pamięci może zależeć od wolnych zasobów układu FPGA, np. układ ma już zajęte wszystkie CLB ale wolne zasoby BSR, dlatego rzeczywistą zależność kosztów tych pamięci należy dostosować do aktualnego projektu. Rys. 3 przedstawia ekwiwalentny koszt pamięci BSR (dodatkowy koszt układu po zastąpieniu pamięci BSR pamięciami 16×1 i 32×1 i podzieleniu przez liczbę zastąpionych BSR) dla różnych szerokości wejścia i współczynnika mnożenia. Na podstawie tego wykresu można odczytać, że pojedyncza pamięć BSR może być zastąpiona przez średnio 8-11 komórek VCLB (około 32-44 pamięci 16×1 RAM). Warto w tym miejscu podkreślić, że dane na rys. 3 są podane przy założeniu, że wykorzystany jest tylko jeden port pamięci BSR, jednakże istnieje możliwość wykorzystania drugiego portu tej samej pamięci BSR, co może prowadzić do podwojenia kosztu ekwiwalentnego pamięci BSR.



Rys 3. Powierzchnia układu LM dla różnej szerokości magistrali wejściowej i współczynnika mnożenia K. A) powierzchnia podawana dla Virtex CLB (1 VCLB= 2 XC4000 CLB) w skali 1:10 dla LM z użyciem tylko pamięci 16×1 i 32×1 ; B) liczba użytych bloków BSR 256×16 ; C) ekwiwalentny koszt (podawany dla VCLB) jednego bloku BSR w porównaniu z rozwiązaniem przy użyciu tylko pamięci 16×1 i 32×1

Fig. 3. Area of the LM block for different input and coefficient width K. A) Area for Virtex in CLB (1VCLB = 2 XC4000 CLB) scaled 1:10 and using only 16×1 and 32×1 RAMs. B) number of Block SelectRAMs (BSR) 156×16 . C) equivalent cost (in VCLB) of a BSR in comparison to the solution A.

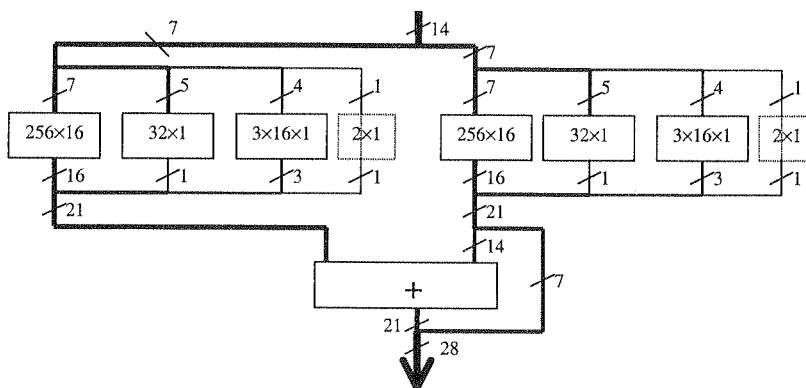
mo być
układów
z 32×1 ,
pamięci
1].

pamięci
ci BSR
i: $4k \times 1$,
edynczy
ex CLB
pamięci
wszystkie
pamięci
y koszt
nięciami
erokości
dczytać,
VCLB
dane na
pamięci
pamięci

Bloki pamięci BSR są relatywnie duże i dlatego trudno jest znaleźć architekturę układu mnożącego, dla której są one całkowicie wykorzystane, a stopień wykorzystania bloków pamięci bardzo wpływa na ich ekwiwalentny koszt. Konsekwentnie, dla małej szerokości magistrali wejściowej i współczynnika mnożenia K ($K < 8$) ekwiwalentny koszt BSR jest raczej mały (zob. rys. 3), co nie zachęca do użycia tych pamięci. Dla $K \leq 13$ ekwiwalentny koszt BSR wzrasta. Natomiast dla $K > 13$ są użyte dwa lub więcej bloki BSR, co wymaga użycia dodatkowych układów arytmetycznych oraz ponownie pojawia się problem pełnego wykorzystania pamięci BSR (występuje efekt ziarnistości dużych bloków BSR), przez co zmniejsza się ekwiwalentny koszt pamięci BSR. Dla $K = 16$, liczba użytych bloków BSR wzrasta gwałtownie, jako że bloki BSR są grupowane w pary stanowiące w sumie dużą pamięć 256×32 , co również zmniejsza ekwiwalentny koszt BSR. Dalsze zwiększanie K powoduje wzrost ekwiwalentnego kosztu, wydaje się jednak, że wartość maksymalna 11 nie zostanie przekroczona.

W tym artykule w celu znalezienia optymalnej architektury układu mnożącego wykorzystano technikę przeszukiwania wszystkich możliwych kombinacji (ang. *exhausted search*). Wszystkie możliwe i sensowne kombinacje pamięci BSR, 32×1 i 16×1 oraz układów dodających są rozważane i najlepszy układ zostaje wybrany. W celu wizualizacji rozważanych architektur na rys. 4 przedstawiono przykład układu mnożącego, w którym użyto kombinacje różnych pamięci RAM i układu dodającego. W przykładzie tym (jak i we wszystkich rozważaniach w tym podrozdziale) nie uwzględniono możliwej optymalizacji MS i DAWR, ze względu na ich ścisłe powiązanie z wartością współczynnika mnożenia, dlatego po uwzględnieniu tych optymalizacji rys. 3 może ulec dalszej komplikacji.

Warto zwrócić uwagę, że autorski program AuToCon nie zakłada jakichkolwiek związków pomiędzy kosztami modułów pamięci i układów dodających, dlatego moduły pamięci mogą być dowolnie zdefiniowane i program może być użyty dla różnych rodzin układów FPGA lub nawet dla projektów ASIC. Parametrami wejściowymi jest koszt układów dodających, rozmiar i koszt modułów pamięci, koszt rejestrów D (potrzebny dla architektury potokowej). Wyjściem programu jest plik tekstowy VHDL, który po syntezie może być zaimplementowany do dowolnego układu FPGA.



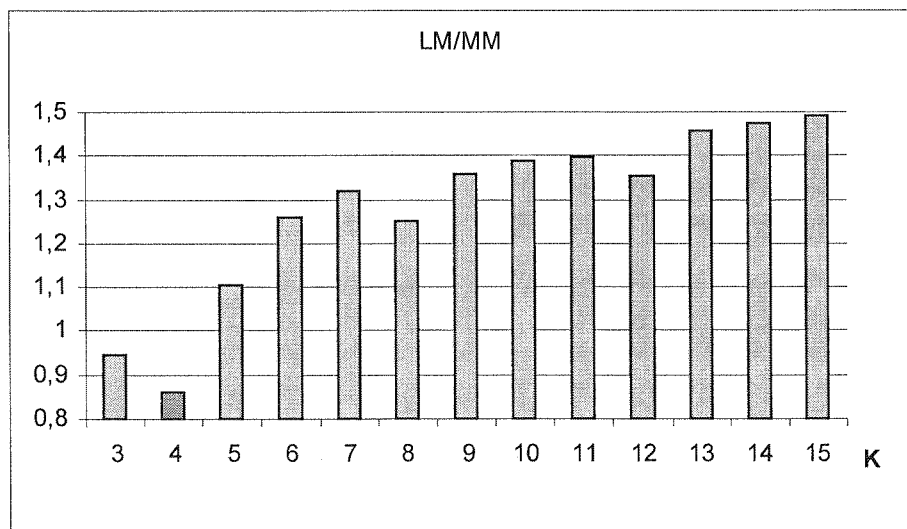
Rys. 4. Układ mnożący LM dla wejść i $K = 14$

Fig. 4. A LM block for input and coefficient width equal 14

2.3. PORÓWNANIE UKŁADÓW MNOŻĄCYCH KCM

2.3.1. Powierzchnia

W przedstawionych rozważaniach omówiono dwie różne techniki implementacji KCM: LM oraz MM. Dlatego powstaje pytanie, która z nich jest bardziej efektywna. Rys. 5 przedstawia ich wzajemne zależności uwzględniając wartości średnie zajmowanej powierzchni. Analizując rys. 5 można zauważyć, że dla XC4000 układ LM jest preferowany dla szerokości $K \geq 4$, a dla $K \geq 5$ najlepszy rezultat jest z reguły osiągnięty dla układu MM. Warto podkreślić, że dla rosnącego K przewaga architektury MM nad LM wzrasta, ze względu na lepsze wyniki optymalizacji takimi metodami jak CSD i SS.

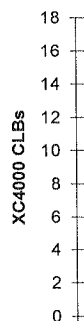


Rys. 5. Relacja pomiędzy stosunkiem średnich powierzchni układu XC4000 zajętych przez LM / MM. Wynik dla różnej szerokości K danej wejściowej (zakres danej wejściowej $0 \div 2^K - 1$) i współczynnika mnożenia (zakres: $1 \div 2^K - 1$)

Fig. 5. Relation between average area of XC4000 occupied by the LM vs. MM. Results for the different input width K (input range $0 \div 2^K - 1$) and coefficient values $1 \div 2^K - 1$

Kolejnym aspektem wzajemnych porównań jest badanie relacji ilościowych jak dużą redukcję powierzchni można osiągnąć dzięki zastosowaniu technik optymalizacji MS i DAWR dla LM. Według wyników doświadczalnych uzyskuje się średnio zmniejszenie powierzchni układu o $5 \div 20\%$ w zależności od szerokości wejścia K , przy czym optymalizacja ta jest bardziej skuteczna dla małych pamięci LUT.

W
W rze
liwoś
pracy u
zastos
układy
zastos
Jednak
cut-set
(pamię
tury (n
określe
zajmow
liczby
Jednak
taktow
częstot
wzrost
podkre
optym
tenden
i dlate

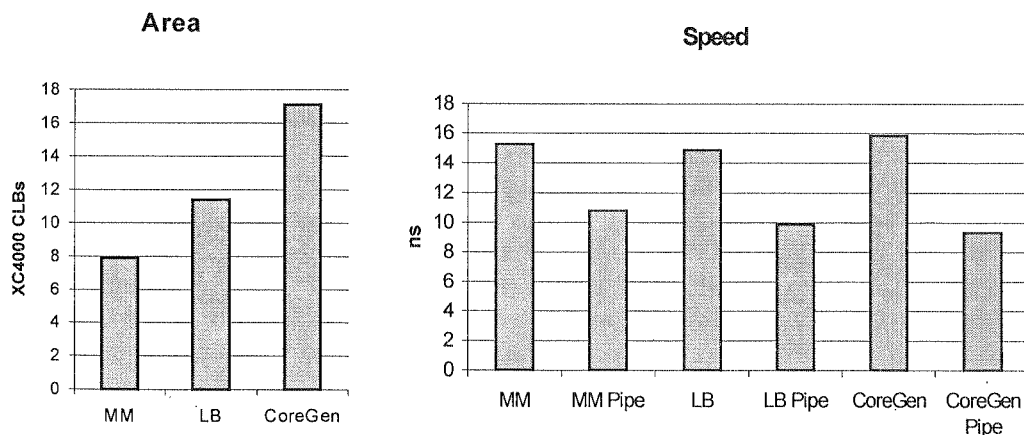


Rys. 6.
dla MM

Fig. 6.
Geno

2.3.2. Częstotliwość pracy

W rozdziale 2.3.1 rozważana była tylko powierzchnia zajmowanego układu. W rzeczywistym układzie należy jednak rozważyć zarówno powierzchnię jak i częstotliwość pracy układu (w szczególności stosunek wartości powierzchni do częstotliwości pracy układu). W konsekwencji, aby zwiększyć przepustowość układu mnożącego należy zastosować architekturę potokową, tym bardziej, że po każdym elemencie logicznym układy FPGA zawierają wbudowane rejestry typu D. Dlatego w pierwszym przybliżeniu zastosowanie architektury potokowej nie wymaga dodatkowej powierzchni układu FPGA. Jednakże, niektóre sygnały nie wymagają logiki, dlatego zgodnie z zasadą tnij-wstaw (ang. *cut-set*) [7] rejestry potokowe muszą być czasami użyte bez odpowiadającej im logiki (pamięci LUT, układu dodającego). W konsekwencji dla całkowicie potokowej architektury (rejestry występują po każdej logice), powierzchnia zajmowana przez układ jest określona raczej przez liczbę rejestrów niż liczbę użytej logiki. Dodatkowa powierzchnia zajmowana przez układ potokowy wzrasta około 0-50% i redukuje się do zera dla mniejszej liczby etapów potokowości gdy np. rejestry są implementowane co dwa poziomy logiki. Jednakże redukcja liczby etapów potokowych powoduje zmniejszenie częstotliwości taktowania układu. Z drugiej strony, architektura potokowa powoduje znaczne zwiększenie częstotliwości pracy układu i dlatego w większości przypadków można zaakceptować wzrost powierzchni układu potokowego, szczególnie wtedy gdy nie jest on duży. Warto podkreślić, że koszt rejestrów został również brany pod uwagę podczas poszukiwania optymalnej architektury układu mnożącego. Dla przykładu, technika optymalizacji SS ma tendencję do implementacji dużej liczby rejestrów w przeciwieństwie do techniki CSD i dlatego metoda CSD jest skuteczniejsza dla architektury potokowej.



Rys. 6. Średnia zajmowana powierzchnia układu mnożącego (bez potokowości) i okres zegara z/bez potokowości dla MM, LB i Core Generator [10]. Wynik implementacji dla układu XC4000E-1, 8-bitowej danej bez znaku oraz losowo wybranym współczynnikom mnożenia: 41, 108, 132, 190, 225

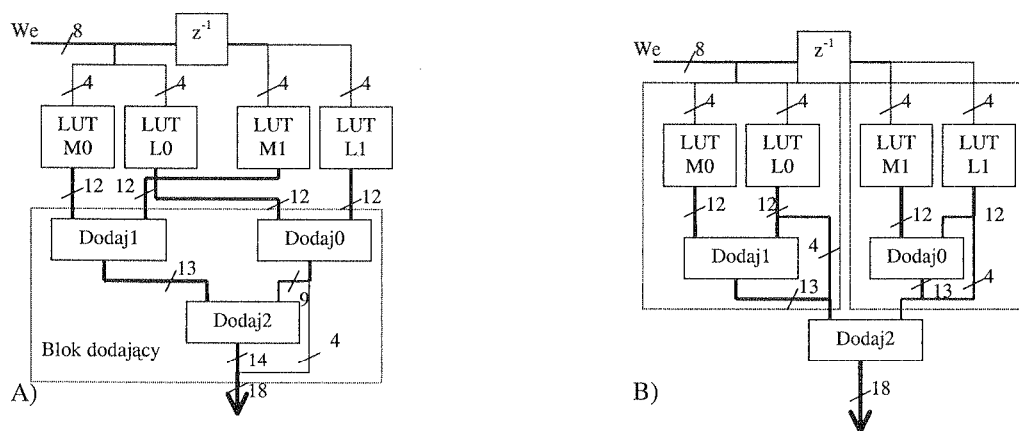
Fig. 6. Average area without pipelining and system period without and with pipelining for the MM, LM and Core Generator [10] multiplier. Implementation results for XC4000E-1 and for the 8-bit unsigned input and 5 randomly chosen coefficients: 41, 108, 132, 190, 225

Wykresy na rys. 6 przedstawiają średnią powierzchnię i okres zegara dla układów MM i LM. Widać, że układy MM są bardziej efektywne niż LM i charakteryzują się tylko nieznacznym spadkiem częstotliwości pracy spowodowanej propagacją przeniesienia w dedykowanym układzie dodającym. Spadek szybkości w niektórych przypadkach może powodować, że przy porównywalnych powierzchniach układu, bardziej optymalną architekturą jest układ LM. Wykres na rys. 6 przedstawia również wynik implementacji dla komercyjnego programu Core Generator [10]. Uwidacznia on, że rozwiązanie układów MM i LM opracowane w ramach referowanych w tym artykule badań przewyższa rozwiązanie komercyjne.

3. KONWOLWER Z UŻYCIEM PAMIĘCI LUT (LC)

Struktura układu konwolwera LC o stałym współczynniku mnożenia z użyciem pamięci LUT (ang. *LUT-based Convolver*) jest podobna do struktury układu dodającego wyniki mnożeń otrzymanych w układach mnożących LM. Jednakże w celu optymalizacji układów dodających dla układu LC wszystkie operacje dodawania są wykonywane wewnątrz jednego bloku dodającego. W konsekwencji operacje dodawania wykonywane wewnątrz każdego układu mnożącego LM powinny być przeniesione do jednego bloku dodającego, tak więc operacja mnożenia LM nie jest rozpatrywana jako oddzielna operacja.

Rys. 7 ilustruje oszczędności otrzymane dzięki użyciu układu LC zamiast sumy mnożeń LM.



Rys. 7. Struktura układu konwolwera dla $Y = A \cdot B_0 + z^{-1} \cdot A \cdot B_1$ i dla szerokości danej wejściowej i współczynników konwolucji równej 8-bitów; A) LC, B) suma LM

Fig. 7. The structure of the convolver for $Y = A \cdot B_0 + z^{-1} \cdot A \cdot B_1$ for input and coefficient width $K=8$ A) LC, B) sum of LM

ów MM
ię tylko
iesienia
ch może
ymalną
tacji dla
układów
ewyższa

Rozpatrzmy teraz oszczędności otrzymane dzięki zastosowaniu układu LC zamiast sumy LM dla szerokości magistrali adresowej pamięci LUT równej s oraz szerokości wyjścia pamięci LUT równej w (w w pierwszym przybliżeniu jest równe sumie szerokości współczynnika mnożenia oraz współczynnika s) dla przykładu podanego na rys. 7. Dla układu sumy mnożeń LM szerokość układów dodających *Dodaj0* i *Dodaj1* na rys. 7B wynosi w przybliżeniu w -bitów. Ostatni układ dodający *Dodaj2* ma szerokość około $w+s$. Podsumowując całkowita liczba 1-bitowych bloków dodających (pół-sumatorów lub pełnych-sumatorów) dla układu sumy LM jest równa:

$$w_{LM} = 3 \cdot w + s \quad (2)$$

Dla układu konwolucji LC natomiast użyto trzech układów dodających o szerokości około w -bitów, co w sumie daje całkowitą liczbę sumatorów 1-bitowych równą:

$$w_{LC} = 3 \cdot w \quad (3)$$

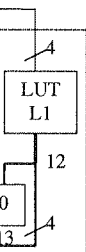
Podsumowując, dzięki zastosowaniu układu LC zamiast sumy LM uzyskano oszczędność rzędu:

$$p = \frac{w_{LM} - w_{LC}}{w_{LC}} = \frac{s}{3 \cdot w} \quad (4)$$

życiem
ającego
malizacji
nywane
nywane
o bloku
peracja.
t sumy

Ponadto dzięki użyciu układu LC zredukowano maksymalną szerokość układu dodającego (a dzięki temu skrócono czas propagacji) z $w+s$ do w .

Rys. 7 pokazuje bardzo prosty przykład układu konwolwera dla okna konwolucji równego 2×1 , w rzeczywistych układach sieć układów dodających jest dużo bardziej skomplikowana. W konsekwencji znalezienie optymalnego drzewa układów dodających, dla którego każdy układ dodający ma tylko dwa wejścia a całkowita liczba użytych sumatorów jest najmniejsza, jest zadaniem wykraczającym poza ramy tego artykułu i zostało szeroko przestudiowane w [11, 12]. W tym miejscu należy jednak dodać, że problem optymalizacji drzewa dodającego nie da się rozwiązać w sposób intuicyjny i należy użyć zaawansowanych technik optymalizacji takich jak: przeszukiwanie wyczerpujące, algorytm zachłanny, algorytm genetyczny czy też symulowane wyżarzanie. W wyniku doświadczeń zaleca się użycia algorytmu zachłannego w przypadku kiedy czas generacji optymalnego drzewa dodającego ma duże znaczenie, w przeciwnym wypadku zaleca się użycie algorytmu symulowanego wyżarzania, który daje dodatkową redukcję zajmowanej powierzchni układu dodającego o około 10÷20% w porównaniu z algorytmem zachłannym [11].



wej

K= 8

Program generujący układ konwolucji LC używa zaawansowanych technik optymalizacji drzewa dodającego ale wymaga również użycia zaawansowanych technik optymalizacji pamięci LUT, zwłaszcza jeżeli mogą być użyte różne rodzaje pamięci, jak to ma miejsce dla układu mnożącego LM pokazanego na rys. 4. Dla układu LM, który został omówiony w rozdziale 2.2 liczba możliwych kombinacji pamięci LUT i układów dodających jest stosunkowo mała, dlatego możliwe było użycie algorytmu przeszukiwania

wyczerpującego. W konsekwencji dla układu LM optymalizowany był koszt pamięci, układów dodających jak i przerzutników (w przypadku architektury potokowej). Niestety w przypadku układu konwolwera LC algorytm przeszukiwania wyczerpującego nie może być wykorzystany ze względu na dużo większą liczbę możliwych kombinacji. Dlatego dla układu LC użyto algorytmu lokalnego przeszukiwania wyczerpującego, tzn. każdy układ mnożący jest optymalizowany osobno przy użyciu przeszukiwania wyczerpującego, tak jak to ma miejsce dla układu LM. Następnie wszystkie układy dodające są grupowane wewnątrz jednego bloku dodającego i tam powtórnie optymalizowane wszystkie razem przez techniki opisane w [11, 12].

Oprócz technik optymalizacji opisanych powyżej układ konwolucji LC jest optymalizowany przez algorytmy, które są typowe tylko dla układu konwolwera. Techniki te zostaną opisane krótko poniżej.

Optymalizacja podobnych współczynników (SCO)

Filtry FIR (o skończonej odpowiedzi impulsowej) są bardzo często implementowane jako filtry o liniowym przesunięciu fazowym, dla których współczynniki filtru są symetryczne. Wykorzystując tą symetrię liczba użytych układów mnożących zmniejsza się o połowę, poprzez lustrzane odbicie biegu sygnału wejściowego w punkcie symetrii współczynników konwolucji oraz zastosowanie dodatkowych układów dodających sumujących wejścia o tym samym współczynniku konwolucji przed wykonaniem wspólnej operacji mnożenia [7]. Niemniej różne rodzaje symetrii lub też asymetryczne mogą być użyte, szczególnie w przypadku konwolucji dwu-wymiarowej [13]. Dlatego program optymalizujący AuToCon [14] porównuje wszystkie współczynniki konwolucji i grupuje je w grupy o podobnych wartościach. Współczynniki w jednej grupie mogą być przesunięte bitowo i mogą mieć przeciwne znaki, dlatego wejścia odpowiadające współczynnikom w jednej grupie są odpowiednio przesuwane bitowo a następnie dodawane lub odejmowane od siebie. Ostatecznie dla jednej grupy współczynników użyty jest tylko jeden układ mnożący, dzięki czemu liczba układów mnożących ulega redukcji.

Dla przykładu, dla filtru:

$$H(z) = H_1(z) + 5 \cdot z^{-i} - 5 \cdot z^{-j} - 10 \cdot z^{-k} + 20 \cdot z^{-l} \quad (5)$$

następujące współczynniki są przyporządkowane jednemu układowi mnożącemu:

$$A_5 = z^{-1} - z^{-j} - 2 \cdot z^{-k} + 4 \cdot z^{-l} \quad (6)$$

i w rezultacie ostateczny wynik jest następujący:

$$H(z) = H_1(z) + 5 \cdot A_5 \quad (7)$$

W podanym przykładzie liczba układów mnożących zmniejszyła się o 3.

Optymalizacja rejestrów potokowych.

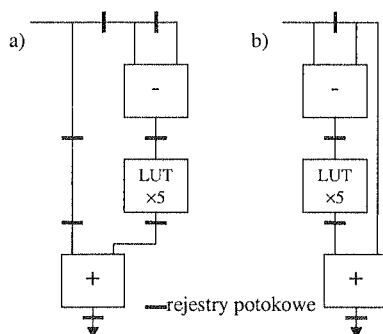
Program AuToCon generuje układ konwolwera o skomplikowanej strukturze potokowej, dla którego dodatkowy parametr potokowości p definiuje maksymalną liczbę elementów logicznych pomiędzy dwoma rejestrami potokowymi. Rys. 8a przedstawia

przykład
potokow
potrzeb
nadmia
presur
zmniejs
implem
podkre
we ukł
implem

UL
Distrib
osobn
tematy

gdzie
i-tej c

przykład układu konwolwera, dla którego nie zastosowano optymalizacji liczby rejestrów potokowych. Taki układ bardzo często wymaga dodatkowych rejestrów potokowych potrzebnych do kompensacji różnicy opóźnień potokowych. W celu likwidacji tych nadmiarowych rejestrów zaimplementowano dodatkową optymalizację, której celem jest przesunięcie miejsca pobierania sygnału do bloków arytmetycznych w taki sposób aby zmniejszyć liczbę rejestrów potokowych. Podobny algorytm optymalizacji został zaimplementowany w [8]. Przykładowy wynik optymalizacji jest pokazany na rys. 8b. Warto podkreślić, że dzięki zastosowaniu optymalizacji potokowej całkowite opóźnienie potokowe układu konwolwera często ulega redukcji. Optymalizacja potokowa została zaimplementowana dla każdego układu konwolucji opisanego w tym artykule.



Rys. 8. Przykład implementacji konwolucji $(2 + 5 \cdot z^{-1} - 5 \cdot z^{-2})$ dla parametru potkowości $p=1$ oraz

a) bez b) z optymalizacją potokową

Fig. 8. An example of the convolver $(2 + 5 \cdot z^{-1} - 5 \cdot z^{-2})$ for pipeline parameter $p=1$ and

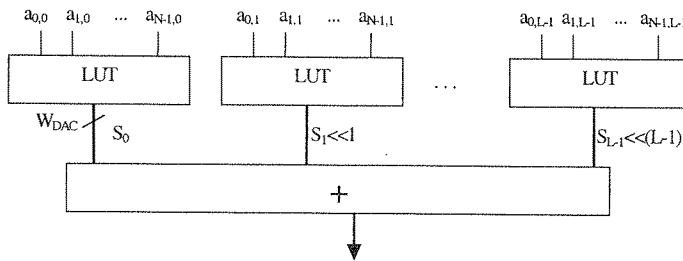
a) without b) with pipelining optimisation

4. KONWOLWER Z UŻYCIEM ARYTMETYKI ROZPROSZONEJ (DAC)

Układ konwolwera DAC zbudowany z użyciem arytmetyki rozproszonej (ang. *Distributed Arithmetic Convolver*) [4, 5] wykonuje operacje mnożenia i dodawania osobno dla każdego bitu danej wejściowej wykorzystując następującą tożsamość matematyczną:

$$\sum_{i=0}^{N-1} w_i \cdot a_i = \sum_{j=0}^{L-1} 2^j \cdot \sum_{i=0}^{N-1} w_i \cdot a_{i,j} \quad (8)$$

gdzie: N — rozmiar konwolucji, L — szerokość danej wejściowej w bitach, $a_{i,j}$ — j -ty bit i -tej danej wejściowej.



Rys. 9. Schemat blokowy układu konwolwera w arytmetyce rozproszonej

Fig. 9. Block diagram of Distributed Arithmetic Convolver

Porównując układ DAC z układem LC można zauważyć że szerokość magistrali danych pamięci LUT dla układu LC wynosi w przybliżeniu:

$$W_{LC} = K + N \quad (9)$$

Dla układu DAC szerokość ta jest równa:

$$W_{DAC} = K + \lceil \log_2(N + 1) \rceil \quad (10)$$

gdzie: W_{LC} , W_{DAC} — szerokość magistrali danych pamięci LUT odpowiednio układu LC i DAC, N — szerokość magistrali adresowej pamięci LUT, K — szerokość współczynników konwolucji (przy założeniu takiej samej szerokości dla wszystkich współczynników).

Porównując równania (9) i (10) można zauważyć, że szerokość magistrali danych pamięci LUT jest dużo mniejsza dla układu DAC niż dla układu LC, ponieważ bity wejściowe do pamięci LUT dla układu DAC mają taką samą wagę bitową. Mniejsza szerokość magistrali danych pamięci LUT pociąga za sobą duże oszczędności ponieważ wymagana jest nie tylko mniejsza pamięć, ale również mniejsza szerokość układów dodających. W konsekwencji układ DAC jest preferowany w stosunku do układu LC.

Przykładowy schemat układu DAC jest pokazany na rys. 9. Podobnie jak dla układu LM czy też LC, rozmiar pamięci LUT rośnie wykładniczo wraz ze wzrostem szerokości adresowej, dlatego pojedyncza pamięć LUT jest dzielona na dwie lub więcej niezależnych pamięci LUT a następnie wyjścia tych pamięci są razem ze sobą zsumowane. Podobnie jak dla układu LC operacja dodawania iloczynów cząstkowych dla podzielonej pamięci LUT nie powinna przebiegać osobno dla każdego bitu. Wszystkie operacje dodawania powinny być włączone w jeden blok dodający co pozwala na lepszą optymalizację układów dodających.

Podział pamięci LUT na mniejsze pamięci powinien przebiegać zgodnie z relacją kosztów różnych rodzajów pamięci LUT i układów dodających, podobnie jak to miało miejsce dla układu mnożącego LM. Układ DAC zakłada, że struktura układu jest taka sama dla różnych wag bitów wejściowych, dzięki czemu proces optymalizacji układu DAC jest prostszy, ale odbija się to ujemnie na powierzchni zajmowanej przez układ. Dla przykładu, konwolucja o rozmiarze 3×3 wymaga wykonania 9 operacji mnożenia, czyli 9-bitowej

pamięć
magist
podzie
Rozwi
LUT-b
wykon
W kon
nych w
oraz 8
DAC v

Ar
układu
strukt
podłąc
magist
użyć
zaprop
Arithm
szerok
odpow
oraz uk
wpływ
W
zgodni
rzędzi
LUT,
podkre
wejści
algory
wybran
podkre
rodzaje
współ
to sze
najmł
współ
wyższ
zapisan

pamięci LUT dla każdego bitu wejścia. W układach FPGA z reguły optymalną szerokością magistrali adresowej pamięci LUT jest 4. W konsekwencji 9-bitowa pamięć LUT musi być podzielona na 4+4+1 lub też na 4+5, co pociąga za sobą pewną nadmiarową logikę. Rozwiązaniem tego problemu może być użycie rozwiązania hybrydowego DAC i LM (ang. *LUT-based Hybrid Convolver — LHC*) [15] dla którego część operacji mnożenia jest wykonana wykorzystując operację DAC a część wykonując standardowe mnożenie LM. W konsekwencji dla powyższego przykładu konwolucji 3×3 , 8 mnożeń jest wykonywanych w układzie DAC a 1 mnożenie w układzie LM. W konsekwencji dla konwolucji 3×3 oraz 8-bitowej danej i 8-bitowych współczynników mnożenia obszar zajmowany przez DAC wynosi 266 komórek CLB a dla układu LHC 209 komórek CLB układu XC4000 [15].

5. KONWOLWER Z UŻYCIEM NIEREGULARNEJ ARYTMETYKI ROZPROSZONEJ (IDAC)

Architektura układu DAC zakłada regularną strukturę układu. tzn. że struktura układu jest taka sama dla różnych wag bitowych danej wejściowej. Niemniej regularność struktury DAC nie musi być zachowana, co więcej do tej samej pamięci LUT mogą być podłączone wejścia o różnej wadze bitowej, o ile pociągnie to za sobą mniejszą szerokość magistrali danych pamięci LUT. W konsekwencji zamiast architektury DAC powinno się użyć kombinacji układów DAC i LC. Ta nowa architektura układu konwolwera zaproponowana w tym artykule została nazwana: IDAC (ang. *Irregular Distributed Arithmetic Convolver*). Algorytm optymalizujący układ IDAC powinien redukować szerokość magistrali danych i adresowej pamięci LUT oraz szerokość układów dodających odpowiednio w stosunku do relacji kosztów pomiędzy poszczególnymi blokami pamięci oraz układami dodającymi, a waga bitu wejściowego powinna być tylko parametrem, który wpływa na szerokość magistrali danych pamięci LUT.

W celu optymalizacji układu IDAC zaproponowany będzie algorytm zachłanny, który zgodnie z nazwą optymalizuje tylko rozwiązanie cząstkowe, tj. optymalizuje w pierwszym rzędzie szerokość magistrali adresowej oraz danych wejściowych do pojedynczej pamięci LUT, pozostałe wejścia są optymalizowane w następnym kroku optymalizacji. Warto podkreślić, że w tym algorytmie każdy bit danej wejściowej jest traktowany jako osobne wejście. Algorytm optymalizujący jest podany w Listingu 1. Głównym zadaniem tego algorytmu jest nie tylko optymalne dobranie wejść do rozważanej pamięci LUT ale również wybranie optymalnego rozmiaru (rodzaju) użytej pamięci LUT. W tym miejscu należy podkreślić, że podobnie jak dla układu LM, układ IDAC umożliwia użycie różnych rodzajów pamięci LUT (porównaj rys. 4). Przed rozpoczęciem procesu optymalizacji współczynniki konwolucji są przesuwane w prawo tak aby były one nieparzyste, redukuje to szerokość magistrali danych pamięci LUT — dla współczynników parzystych najmłodsze bity są równe zero i nie niosą żadnej informacji. Wejścia dla których współczynnik mnożenia został przesunięty w prawo są traktowane dalej tak jakby miały wyższą wagę bitową odpowiednio do liczby przesunięć. Jeżeli dana wejściowa jest zapisana w kodzie uzupełnień do dwóch, to wszystkie bity wejściowe oprócz najbardziej

znaczącego są traktowane tak jakby dana wejściowa była zawsze dodatnia, natomiast dla bitu najbardziej znaczącego współczynnik konwolucji jest mnożony przez -1 .

Algorytm podany w Listingu 1 stosuje następujący priorytet parametrów wejścia oraz współczynnika mnożenia: Najpierw algorytm znajduje wolne wejście o najniższej wadze bitowej s_i . Jeżeli krok S1 daje więcej niż jedno rozwiązanie to w kroku drugim S2 dla bitów o tej samej najniższej wadze bitowej wybranych w kroku S1, wybierane jest wejście, dla którego współczynnik mnożenia ma najmniejszą szerokość. Jeżeli poprzednie dwa kroki nie dały jednoznacznej odpowiedzi w kroku trzecim S3 wybierane jest to wejście, dla którego znak współczynnika mnożenia (zawsze dodatni lub też kodowany w kodzie uzupełnień do dwóch) jest taki sam jak znak pamięci LUT dla poprzednio przypisanych wejść.

Listing 1. Algorytm wyboru najlepszego rozwiązania cząstkowego dla architektury IDAC
Listing 1. The algorithm choosing the best partial solution for the IDAC

Warunki początkowe:

$c_{best} = \alpha$

$width = 1$

Początek pętli:

S1: Wybierz wolne wejście (wejścia) o najmniejszej wadze bitowej s_i .

S2: Dla wejść wybranych w punkcie S1 znajdź wejście, dla którego współczynnik mnożenia ma najmniejszą szerokość w_i .

S3: Dla wejść z punktu S3, wybierz wejście o tym samym znaku co znak pamięci LUT.

S4: Oblicz średni koszt c_a przypadający na jedno wejście pamięci LUT (rozważ również poprzednio wybrane wejścia).

S5: Jeżeli $c_a < c_{best}$ wtedy $c_{best} = c_a$ (został znaleziony lepszy układ).

S6: $width = width + 1$.

S7: Jeżeli $width > max_width$ wtedy wybierz układ o najlepszym koszcie c_{best} w przeciwnym wypadku idź na początek pętli.

gdzie:

$width$ — szerokość magistrali adresowej rozważanej pamięci IDAC LUT,

max_width — maksymalna szerokość magistrali adresowej dla rozważanych pamięci lub też liczba wolnych wejść (bitów wejściowych) jeżeli mniejsza,

s_i — waga wejścia, s^i = waga bitu wejściowego + przesunięcie w prawo współczynnika konwolucji (wykonywane w celu uzyskania nieparzystego współczynnika),

w_i — szerokość współczynnika mnożenia (po przesunięciu w prawo),

c_a — średni koszt na wejście do pamięci LUT, $c_a = \frac{Cost_memory + Cost_adder}{width}$,

$Cost_memory$ — koszt pamięci LUT, dla którego szerokość magistrali adresowej wynosi $width$, a szerokość magistrali danych jest otrzymana na podstawie zakresu danej wyjściowej pamięci LUT (suma współczynników mnożenia uwzględniająca dodatkowe przesunięcie bitowego w przypadku gdy wejścia mają różną wagę bitową),

$Cost_adder$ — koszt układu dodającego o szerokości równej szerokości magistrali danych pamięci LUT + 1,

c_{best} — najmniejszy koszt przypadający na jeden bit danej wejściowej, odpowiada on najlepszemu układowi, który został do tej pory znaleziony.

Kroki S1-S3 w podanym algorytmie znajdują najlepsze wejście, które jest następnie podłączane do pamięci IDAC LUT. Kroki S4-S7 znajdują natomiast optymalną szerokość magistrali adresowej pamięci LUT. Krok S4 oblicza średni koszt pamięci i układu dodającego przypadający na jedno wejście. W tym kroku założono, że średnia szerokość układu dodającego jest równa szerokości magistrali danych pamięci LUT + 1. Rzeczywista szerokość układu dodającego zależy od połączeń układów dodających i jest znana dopiero w etapie końcowym po skomplikowanym procesie optymalizacji drzewa dodającego. Krok S5 wybiera najlepszy znaleziony układ, tzn. układ o najmniejszym średnim koszcie na jedno wejście. Kroki S6 i S7 służą do kontroli pętli oraz do zakończenia procesu wyboru najlepszych wejść oraz pamięci LUT. Powyższy algorytm jest powtarzany do momentu, w którym zostaną przypisane wszystkie wolne wejścia. Przykład implementacji tego algorytmu jest podany w rozdz. 6.

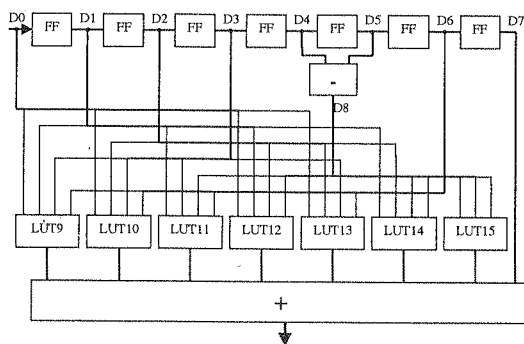
Algorytm podany w Listingu 1 jest oryginalnym opracowaniem autorskim oraz rozwiązuje problem, który nie był to tej pory nigdy jeszcze rozważany. Jednakże lepsze algorytmy optymalizujące mogą być tu przytoczone, jak wiadomo algorytm zachłanny należy do grupy algorytmów szybkich i łatwych do implementacji jednakże rezultat tego algorytmu z reguły odbiega od optymalnego. Dlatego następnym krokiem powinno być skonstruowanie algorytmu pozwalającego na rozważenie problemu w sposób globalny, np. użycie algorytmu symulowanego wyżarzania.

6. WYNIK IMPLEMENTACJI

W celu porównania wyników implementacji dla układów DAC i IDAC zostanie zaimplementowany następujący (przykładowo wybrany) filtr FIR:

$$H(z) = 59 + 183 \cdot z^{-1} + 162 \cdot z^{-2} - 7 \cdot z^{-3} - 48 \cdot z^{-4} + 12 \cdot z^{-5} + 9 \cdot z^{-6} + 2 \cdot z^{-7} \quad (11)$$

Schemat blokowy tego filtru przy użyciu architektury IDAC jest przedstawiony na rys. 10.



Rys. 10. Schemat blokowy układu IDAC (FF- rejestry opóźniające)

Fig. 10. Example of a block diagram of the IDAC (FF- flip-flops)

Zanim zostanie zaimplementowany algorytm zachłanny podany w Listing 1, grupowane są podobne współczynniki czyli takie, które są przesunięte bitowo lub/i zanegowane. Dla rozważanego filtru podobne współczynniki to $-48 \cdot z^{-4} + 12 \cdot z^{-5}$, dlatego wejścia $D4$ i $D5$ na rys. 10 stowarzyszone z tymi współczynnikami tworzą nowe wejście $D8 = D5 \text{ — } (D4 \ll 2)$, a współczynnik mnożenia stowarzyszony z wejściem $D8$ wynosi 12. Następnie każdy współczynnik jest przesuwany w prawo w celu otrzymania współczynnika nieparzystego. Dla przykładu współczynnik 162 jest przesuwany w prawo o 1-bit i ostatecznie otrzymuje się współczynnik 81. Ostateczne wartości współczynników po przesunięciu są podane w tab. 2. w rzędzie 2.

Zgodnie z algorytmem podanym w Listingu 1 najpierw wybrane jest wejście $D6$ bit 0 (współczynnik równy 9), potem wejście $D3$ bit 0, potem wejście $D0$ bit 0 i ostatecznie wejście D bit 0. Zgodnie z podanym kosztem poszczególnych bloków pamięci najniższy średni koszt został uzyskany dla 4-wejściowej pamięci RAM16x1. Dlatego następne wejścia są grupowane do następnej pamięci IDAC LUT. Warto podkreślić, że wejście $D7$ dla którego współczynnik mnożenia jest potęgą liczby 2 jest podłączony bezpośrednio (bez wykonywania operacji mnożenia) do bloku dodającego. Dokładna kolejność łączenia poszczególnych wejść do następnych pamięci LUT jest podana w tab. 2.

Tabela 2

Przypisanie różnych wejść do pamięci IDAC LUT (zob. rys. 10)

IDA-LUT assignment for different inputs bits (see Fig. 10)

Wej.	D0	D1	D2	D3	D6	D7	D8
Wsp.	59	183	81	-7	9	1	3
Bit 0	9,2	9,3		9,1	9,0	B	
Bit 1	10,2	11,0	10,3	10,1	10,0	E	
Bit 2	12,0	12, 2	12,1	11,3	11,2	Z	11,1
Bit 3	13,2	14,0	13,3	13,1	13,0	P	12,3
Bit 4			14,2			O	14,1
Bit 5						Ś	14,3
Bit 6						R	15,0
Bit 7						E	15,1
Bit 8						D	15,2

Wyniki implementacji dla filtru podanego w równaniu (11) oraz przy użyciu różnych architektur są podane w tab. 3. Core Generator [10] — program zalecany przez producenta układów FPGA — Xilinx, generuje układ konwolucji używając architektury (równoleg-

grupowa-
wane. Dla
4 i D5 na
($D4 < 2$),
nie każdy
rzystego.
trzymuje
a podane
ejście D6
statecznie
najniższy
następne
ejście D7
dnio (bez
łączenia

łego) DAC. Program autorski AuToCon generuje układ konwolwera używając architektury IDAC, ponadto możliwe jest ustalenie parametru potokowości p , w tab. 3 znajdują się dwie opcje: $p=\alpha$ — brak potokowości oraz $p=1$ — rejestry potokowe są wstawiane po każdej pamięci LUT oraz po każdym układzie dodającym. Analizując tab. 3 można zauważyć, że wyniki implementacji dla IDAC są zdecydowanie lepsze niż dla DAC, co potwierdza tezę postawioną w niniejszym artykule.

Architektura układu IDAC jest nieregularna dlatego początkowo może się wydawać, że minimalny okres zegara dla tej architektury musi być większy niż dla regularnej struktury DAC. Jednakże wyniki doświadczalne pokazują, że opóźnienia w układach IDAC i DAC są porównywalne, co więcej układ IDAC jest nieznacznie szybszy. Wynik ten można wytłumaczyć faktem, że układ zajmujący mniejszą liczbę CLB może być łatwiej umiejscowiony w układzie FPGA oraz że zajmuje mniej zasobów połączeniowych, dzięki czemu jest on szybszy.

Tabela 3

Wyniki implementacji dla układu XC4005PC84-09

Implementation results for XC4005PC84-09

Układ	# 16x1 LUTs	# FFs	T [ns]
1) Core Generator	169	205	9.22
2) AuToCon $p=\alpha$	125	28	28.4
3) AuToCon $p=1$	126	179	8.7

Tabela 2

D8
3
11,1
12,3
14,1
14,3
15,0
15,1
15,2

7. PODSUMOWANIE

Niniejszy artykuł opisuje implementacje operacji konwolucji w układach FPGA oraz nowe rozwiązanie układu konwolwera: nieregularną architekturę rozproszoną IDAC. Układ IDAC różni się od architektury regularnej DAC tym, że pozwala na to aby układ DAC był różny dla różnych wag bitu wejściowego. Umożliwienie nieregularności układu DAC umożliwia lepszą optymalizację układu IDAC, co z kolei pociąga za sobą znaczną redukcję obszaru zajmowanego przez układ, co potwierdzają wyniki doświadczalne.

Zaproponowano również algorytm optymalizujący architekturę IDAC. Zaproponowanym algorytmem jest algorytm zachłanny, który najprawdopodobniej może być ulepszony poprzez zaproponowanie bardziej rozbudowanego kryterium wyboru. Ponadto w przyszłych pracach należy opracować bardziej zaawansowany algorytm optymalizujący, np. symulowane wyżarzanie. Co więcej ulepszony algorytm powinien rozważać redukcję szerokości magistrali adresowej pamięci LUT oraz inne metody opisane dla układu LM [3]. Dla przykładu można zauważyć, że dla pamięci LUT11 na rys. 10 tylko jedna linia adresowa jest wymagana dla najmniej znaczącego bitu wyjściowego pamięci.

u różnych
roducenta
ównoleg-

W rozdziale 2.1 opisano układy bezmnożne MM, dla których wynik implementacji podany w rozdziale 2.3 jest często lepszy niż dla układów mnożących LM (opartych na pamięciach LUT). Podobnie układ bezmnożny konwolwera może być alternatywą dla układu IDAC, co więcej podobnie jak dla układów MM i LM, wybór pomiędzy różnymi rodzajami układu konwolwera może zależeć od wartości współczynników mnożenia, szerokości danej wejściowej czy też relacji kosztów pomiędzy różnymi elementami podstawowymi układu FPGA. Dlatego w przyszłości należy dla ściśle podanych parametrów konwolwera porównać koszty różnych architektur oraz wybrać najlepszą z nich.

Architektura IDAC jest zaimplementowana w autorskim programie AuToCon napisanym w języku C++ oraz w języku opisu sprzętu VHDL, który automatycznie generuje opis w języku VHDL zoptymalizowanego układu konwolwera dla podanych parametrów wejściowych takich jak rozmiar okna konwolucji, szerokość danej wejściowej, wartości współczynników konwolucji, itp.

8. BIBLIOGRAFIA

1. K. Wiatr, E. Jamro: *Implementation of Multipliers in FPGA Structures*. Proc. of the IEEE Int. Symposium on Quality Electronic Design, San Jose, California, 26-28 March 2001, pp. 415-420.
2. A. R. Omond: *Computer Arithmetic Systems. Algorithms Architecture and Implementations*. Prentice Hall, 1994
3. K. Wiatr, E. Jamro: *Constant Coefficient Multiplication in FPGA Structures*. Proc. of the IEEE Int. Conf. Euromicro, Maastricht, Netherlands, Sep. 5-7, 2000, Vol. I, pp. 252-259.
4. C. S. Burrus: *Digital filter structure described by arithmetic*. IEEE Transaction on Circuits and Systems, 1977, pp. 674-680.
5. T. T. Do, C. Reuter, P. Pirsch: *Alternative approaches implementing high-performance FIR filters on lookup table-based FPGAs: A comparison*. SPIE Conference on Configurable Computing and Applications, Boston, Massachusetts, 2-3 Nov. 1998, pp. 248-254.
6. H. Garner: *Number Systems and Arithmetic*. Advances in Computing, vol. 6, 1965, pp. 131-194.
7. P. Pirsch: *Architectures for Digital Signal Processing*. Chichester UK, Wiley, 1998.
8. R. I. Hartley: *Subexpression Sharing in Filters Using Canonic Signed Digit Multipliers*. IEEE Transactions on Circuits and Systems II — Analog and Digital Signal Processing, vol. 43, no. 10, Oct. 1996.
9. Xilinx Co.: *The Programmable Logic Data Book*, 1999.
10. Xilinx Co.: *Core Generator*, Foundation 2.1i Software Packet, 1999.
11. E. Jamro, K. Wiatr: *Genetic Programming in FPGA Implementation of Addition as a Part of the Convolution*. Proc. of the IEEE Int. Conf. Digital System Design, Warszawa, Poland, 4-6 Sep. 2001, pp. 466-473.
12. E. Jamro, K. Wiatr: *Implementation of convolution operation on general purpose processors*. Proceedings of the Euromicro Conf. Warszawa, Poland, 4-6 Sep. 2001, pp. 410-417.
13. W. S. Lu: *Two-Dimensional Digital Filters*. Marcel Dekker, New York, 1992.
14. E. Jamro: *Parameterised automated generation of convolvers implemented in FPGAs*. Ph.D. Thesis, AGH Technical University, Kraków, Poland, June 2001.
15. K. Wiatr, E. Jamro: *Implementation of image data convolutions operations in FPGA reconfigurable structures for real-time vision systems*. International IEEE Conference on Information Technology: Coding and Computing, Nevada 2000, pp. 152-157.

E. JAMRO, K. WIATR

CONVOLUTION WITH CONSTANT COEFFICIENT IMPLEMENTED IN FPGA CHIPS

Summary

This paper reviews different architectural solutions for calculating constant coefficient convolution operation in FPGAs. At first, different architectures of multipliers are approached. Disregarding the multiplier entity allows for further circuit optimisations, therefore Look-Up-Table (LUT) based Convolver (LC) versus the sum of the LUT-based Multipliers (LM) is described. Further, an alternative technique - (Parallel) Distributed Arithmetic Convolver (DAC) is approached. The key issue of this paper is, however, a novel architectural solution: Irregular Distributed Arithmetic Convolver (IDAC) which, in comparison to the DAC, has an irregular form, and therefore allows for better circuit optimisation. All architectural solutions described hereby can be automatically generated by the Automated Tool for generation Convolver in FPGAs (AuToCon).

Keywords: specialised processors, computing elements architecture, distributed arithmetic, programmable logic device

Ana
rea

Ko
pra
ora
ele
pro
2.4
ran
uk

Sta

Roz
klasy sy
gać pra
charakte
układów
napięcia
mocy v

Analiza i projektowanie sieci neuronowych komórkowych realizowanych w technice układów pracujących w trybie prądowym

JACEK KOWALSKI

*Instytut Elektroniki, Politechnika Łódzka
ul. Stefanowskiego 18/22. 90-924 Łódź
e-mail: jackowal@ck-sg.p.lodz.pl*

*Otrzymano 2002.03.27
Autoryzowano 2002.10.24*

W artykule przedstawiono analizę i propozycję metody projektowania Sieci Neuronowych Komórkowych (SNK) realizowanych z zastosowaniem układów CMOS pracujących w trybie prądowym. Analizę przeprowadzono wykorzystując model tranzystora MOS Shichmana-Hodgesa oraz model Sakurai'a-Newtona. Wyjaśniono sposób projektowania na przykładach SNK detektora elementów spójnych. Rezultaty projektowania potwierdzono symulacjami komputerowymi programem SPICE z wykorzystaniem parametrów modelu tranzystora MOS technologii MIETEC 2.4 μm oraz technologii CMOS 0.25 μm . Przedstawiono model układowy komórki SNK „full range” oraz wyniki jego symulacji i pomiarów. Ponadto zaproponowano zmodyfikowany model układowy komórki SNK „full range”, którego zaletą jest dużo lepsza liniowość.

Słowa kluczowe: sieci neuronowe komórkowe, układy CMOS VLSI, ASIC, przetwarzanie informacji obrazowej

1. WSTĘP

Rozwój sztucznych Sieci Neuronowych Komórkowych (SNK), należących do klasy systemów równoległego, rozproszonego przetwarzania informacji nie może przebiegać prawidłowo w oderwaniu od realizacji praktycznych, zwłaszcza że ich cechą charakterystyczną jest dobrze określona regularna architektura. Postęp w technologii układów scalonych MOS z elementami o bardzo krótkich kanałach zmusza do zmniejszania napięcia zasilającego [14], przede wszystkim w celu ograniczenia gęstości wydzielanej mocy w strukturze krzemowej oraz ze względu na możliwość niszczących przebieg

w strukturach tranzystorów MOS technologii submikronowych. To powoduje zmniejszenie zakresu dynamiki sygnałów napięciowych przetwarzanych analogowo w układzie oraz maksymalnych osiągalnych sygnałów wyjściowych. Jednym z nowszych rozwiązań układowych analogowego przetwarzania sygnałów, które jest pozbawione wad rozwiązań tradycyjnych - w tym przede wszystkim opartych na wykorzystaniu wzmacniaczy operacyjnych napięciowych i wzmacniaczy transkonduktancyjnych [3, 15, 33] - jest idea przetwarzania sygnałów prądowych w miejsce dotychczasowego przetwarzania sygnałów napięciowych. Dotychczasowe osiągnięcia w zakresie realizacji takich układów dają podstawy do przypuszczenia, że technika prądowa analogowego przetwarzania sygnałów (ang. current-mode approach) staje się ważną alternatywą projektowania układów scalonych. Technika ta znalazła zastosowanie również do realizacji sieci neuronowych komórkowych [4-6, 9, 13, 19-32, 34-38].

2. MODEL KOMÓRKI SNK CHUA'Y PRACUJĄCEJ W TRYBIE PRĄDOWYM

Równanie stanu komórki sieci neuronowej komórkowej w wersji napięciowej [2, 7, 17] jest wyrażone następująco:

$$C \frac{dV_{xij}}{dt} = -\frac{V_{xij}}{R_x} + \sum_{C(k,l) \in N_r(i,j)} A(i, j; k, l) V_{ykl} + \sum_{C(k,l) \in N_r(i,j)} B(i, j; k, l) V_{ukl} + I \quad (1)$$

przy czym zmienną stanu jest tu napięcie V_{xij} na kondensatorze modelu obwodowego. Jeśli dokonamy w tym równaniu następujące podstawienie formalne $i_{xij} = V_{xij} / R_x$ to otrzymamy równanie:

$$R_x C \frac{di_{xij}}{dt} = -i_{xij} + \sum_{C(k,l) \in N_r(i,j)} A(i, j; k, l) V_{ykl} + \sum_{C(k,l) \in N_r(i,j)} B(i, j; k, l) V_{ukl} + I \quad (2)$$

Równanie (2) można potraktować jako równanie stanu dla prądu i_{xij} płynącego przez rezystor liniowy R_x połączony równolegle z pojemnością w modelu obwodowym komórki. Ponieważ jednak zmienna prądowa rezystora nie może być zmienną stanu w sensie przyjętym w klasycznej analizie układów dynamicznych, zatem równanie (2) jest również równaniem stanu tylko w sensie formalnym. Jest ono jednak równaniem dynamiki układu równoważnym równaniu (1), ponieważ podstawienie użyte przy transformacji tych równań jest przekształceniem liniowym.

W podobny sposób przekształca się równanie określające zmienną wyjściową. W modelu napięciowym równanie to przedstawia się następująco:

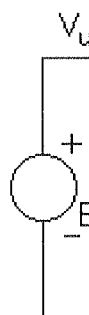
$$V_{yij} = 0.5 (|V_{xij} + 1| - |V_{xij} - 1|) \quad (3)$$

Stosując
przekształ-

przy cz
Uwzględ
można c

Poniewa
orygina
komórk
1 przed
prądowe
w ramac
(tw. m

przy cz



Dopływ
komóre
suma pr
wyrażo

Stosując podobne przekształcenie liniowe dla prądu wyjściowego $i_{yij} = V_{yij}/R_y$ otrzymujemy przekształcone równanie wyjściowe:

$$i_{yij} = 0.5 (|i_{xij} + I_Q| - |i_{xij} - I_Q|) \quad (4)$$

przy czym $I_Q = IV/R_y = IV/R_x$.

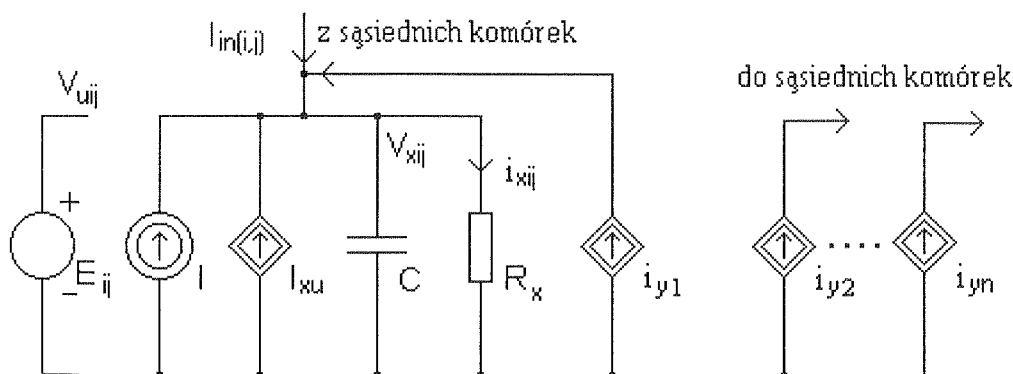
Uwzględniając powyższe przekształcenia, równanie dynamiki komórki w wersji prądowej można ostatecznie zapisać następująco:

$$R_x C \frac{dj_{xij}}{dt} = -i_{xij} + \sum_{C(k,l) \in N_r(i,j)} A_I(i,j;k,l) i_{ykl} + \sum_{C(k,l) \in N_r(i,j)} B_G(i,j;k,l) V_{ukl} + I \quad (5)$$

Ponieważ równanie (2) zostało otrzymane drogą przekształceń liniowych równania oryginalnego (1), zatem własności dynamiczne, znane z teorii sieci neuronowych komórkowych [2], są w odniesieniu do zmiennej prądowej i_{xij} analogiczne. Na rys. 1 przedstawiono model obwodowy odpowiadający równaniu (5). W modelu tym źródła prądowe $I_{xu}(i,j;k,l)$ są sterowane napięciami wejściowymi V_{ukl} wszystkich komórek w ramach sąsiedztwa N_r , podobnie jak to ma miejsce w przypadku modelu napięciowego (tzw. modelu Chua'y). Wartości tych prądów są określone zależnościami:

$$I_{xu}(i,j;k,l) = \sum_{C(k,l) \in N_r(i,j)} B_G(i,j;k,l) V_{ukl} \quad (6)$$

przy czym szablon B_G jest równy szablonowi B modelu napięciowego Chua'y.



Rys. 1. Model obwodowy komórki w koncepcji prądowej

Fig. 1. Circuit representation of current-mode CNN cell

Dopływający do węzła sumacyjnego prąd $I_{in}(i,j)$ jest równy sumie prądów wyjściowych komórek sąsiedztwa z wyłączeniem własnego prądu wyjściowego komórki (tzn. jest to suma prądów wyjściowych komórek należących do otoczenia danej komórki). Prąd ten jest wyrażony zależnością:

$$I_{in}(i,j) = \sum A_I(i,j;k,l) i_{ykl}, C(k,l) \in N_r(i,j), k \neq i, l \neq j \quad (7)$$

przy czym $A_I = A R_x$, gdzie A jest szablonem w modelu napięciowym Chua'y. Źródła prądowe i_{yn} dla $n = 1, 2, \dots, M$ (M jest liczbą komórek sąsiedztwa) reprezentują prądy wyjściowe komórek, określone wzorem:

$$i_{yn} = A_I(k,l;i,j) i_{yij} \quad (8)$$

przy czym źródło prądowe i_{y1} reprezentuje własne sprzężenie zwrotne.

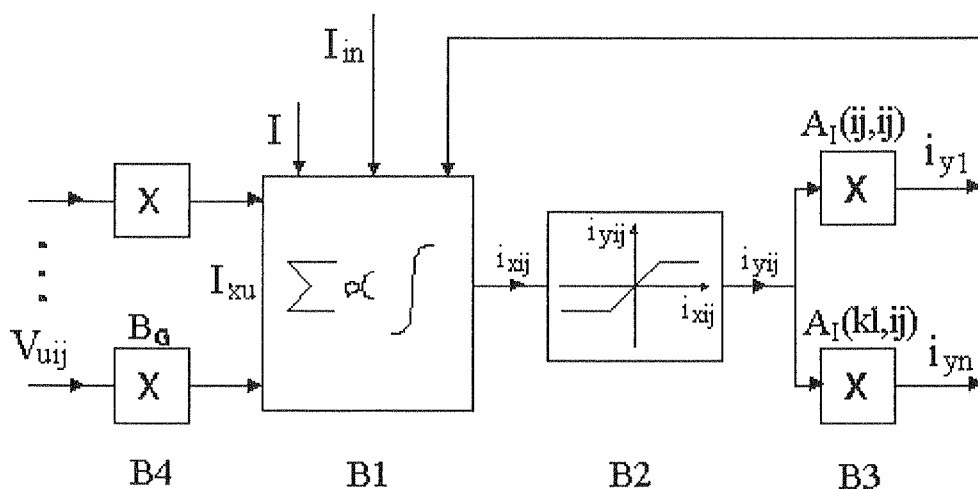
Wprowadzając następujące ograniczenia na wartość początkową prądu $i_{xij}(0)$ oraz napięcie wyjściowe V_{ukl} :

$$|i_{xij}(0)| \leq I_Q \text{ i } |V_{ukl}| \leq IV \quad (9)$$

można (analogicznie jak w przypadku modelu napięciowego Chua'y) określić maksymalną wartość prądu rezystora R_x (prądu „stanu” i_{xij}), której żadna komórka sieci nie przekroczy w dowolnej chwili $t > 0$:

$$i_{max} = I_Q + |I| + I_Q \sum_{C(k,l) \in N_r(i,j)} |A_I(j,j;k,l)| + IV \sum_{C(k,l) \in N_r(i,j)} |B_G(i,j;k,l)| \quad (10)$$

Zależność (10) określa maksymalny zakres dynamiki zmian sygnałów prądowych w sieci, który trzeba zapewnić dla swobodnej odpowiedzi czasowej. Zakres ten jest w każdym przypadku zależny od przeznaczenia danej sieci do realizacji konkretnych zadań przetwarzania obrazów. Na rys. 2 przedstawiono schemat funkcjonalny blokowy [19]



Rys. 2. Schemat funkcjonalny blokowy komórki SNK w wersji prądowej

Fig. 2. Functional block scheme of current-mode CNN cell

(7)
Źródła
prądy

(8)
napięcie

(9)
ymalna
ekroczy

(10)

w sieci,
każdym
zadań
wy [19]

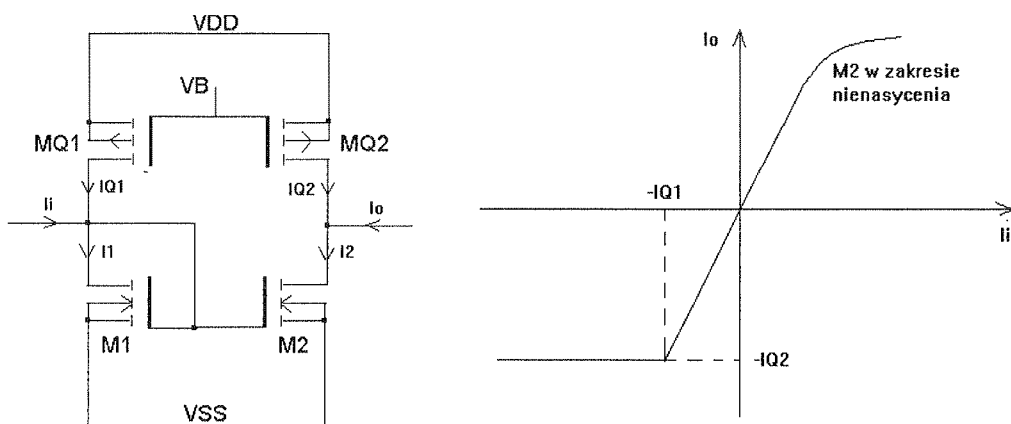
i_{y1}

i_{yn}

realizujący równanie dynamiki komórki. Sygnałem wyjściowym bloku sumowania algebraicznego i całkowania $B1$ jest prąd i_{xij} będący zmienną „stanu”. W bloku $B2$ prąd ten jest poddawany nieliniowej transformacji dając na wyjściu prąd i_{yij} . Z kolei ten prąd jest skalowany w blokach $B3$ przez wymnożenie przez wartości szablonu A_i , w wyniku czego na każdym wyjściu otrzymujemy odpowiedni prąd i_{yn} dołączany do wejść sumacyjnych komórek sąsiednich. W blokach $B4$ następuje skalowanie napięć wejściowych V_{ukl} (przez przemnożenie przez wartości szablonu B_G) z komórek sąsiednich dając w wyniku prądy I_{xu} . Bloki $B1$, $B2$ i $B3$ są konwerterami prądowo-prądowymi, natomiast bloki $B4$ są konwerterami napięciowo-prądowymi. Schemat blokowy z rys. 2 stanowi podstawę opracowania realizacji VLSI sieci neuronowej komórkowej w wersji prądowej.

2.1. REALIZACJA KOMÓRKI SNK W TECHNOLOGII CMOS

Podstawowym układem przetwornika analogowego prądowo-prądowego jest proste zwierciadło prądowe obciążone źródłami prądowymi [7, 16, 19], jak na rys. 3. Sygnałem wejściowym układu jest prąd I_i dopływający do węzła tranzystora $M1$ w konfiguracji „diodowej”, natomiast sygnałem wyjściowym jest prąd I_o wpływający do węzła wyjściowego. Przyjmujemy, że prądami dodatnimi są prądy dopływające do węzłów wejściowego i wyjściowego.



Rys. 3. Podstawowy układ CMOS przetwornika analogowego prądowo-prądowego

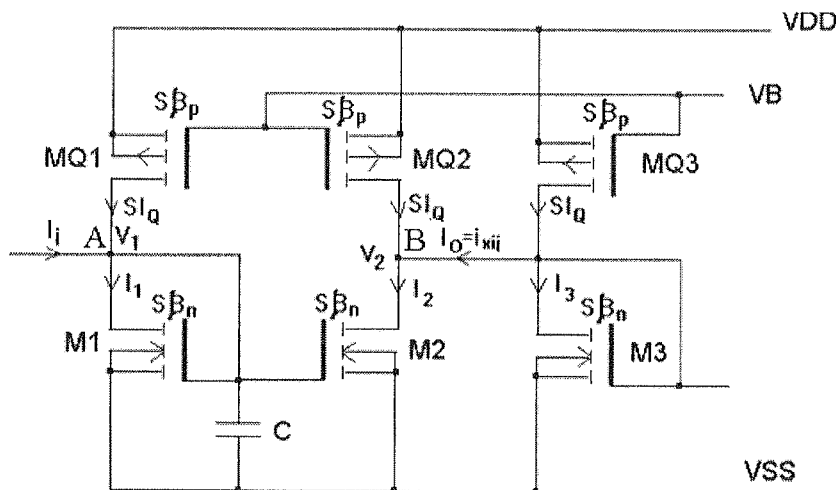
Fig. 3. A basic CMOS circuit of analog current to current converter

Zależnie od stosunku wymiarów tranzystorów lewej i prawej strony układu, charakterystyka statyczna przejściowa układu ma zmienne nachylenie i zakres części liniowej. Jeżeli prąd wejściowy ma wartość ujemną (wypływa z węzła wejściowego) a jego wartość bezwzględna rośnie od zera, to początkowo wzrasta od zera ujemny prąd wyjściowy (wypływa z węzła wyjściowego). Gdy wartość bezwzględna prądu wejściowego zostanie zrównana z wartością prądu wypływającego z tranzystora $MQ1$ to przez

tranzystor $M1$ nie płynie już żaden prąd, a tym samym jego odbicie zwierciadlane w postaci prądu I_2 jest równe zero. Z węzła wyjściowego wypływa stały prąd tranzystora I_{Q2} . Charakterystyka przejściowa układu dla prądów mniejszych od $-I_{Q1}$ ulega nasyceniu na poziomie $-I_{Q2}$. Dla dodatnich prądów wejściowych początkowo charakterystyka układu jest linią prostą, tak długo jak długo tranzystor $M2$ pozostaje w nasyceniu. Nachylenie charakterystyki w części liniowej zależy od wymiarów geometrycznych tranzystorów, może być zatem dobierane łatwo w bardzo szerokim zakresie. Opisany układ przetwornika stanowi podstawowy blok realizacyjny komórki. Przez dołączenie kondensatora równoległe do tranzystora w połączeniu „diodowym” otrzymujemy układ realizacji całkowania sygnałów dopływających do wejścia. Dwa układy połączone kaskadowo stanowią blok realizacji funkcji wyjściowej, natomiast pojedynczy układ o odpowiednio dobranym nachyleniu charakterystyki w części liniowej służy do realizacji elementów szablonu sieci.

2.2. PROJEKTOWANIE UKŁADU VLSI KOMÓRKI SNK

Podstawą sformułowania procedury projektowania układu VLSI komórki danej sieci jest analiza elektryczna pracy podstawowego, wyżej opisanego konwertera prądowo-prądowego. Poniżej przeprowadza się taką analizę [19] biorąc za punkt wyjścia odpowiednio dobrane modele tranzystorów MOS. Przyjęto model Shichmana-Hodgesa odpowiadający modelowi *Level 1* oraz alternatywnie model Sakurai'a-Newtona odpowiadający modelowi *Level 6* pakietu programu *SPICE3 ICAP4*. Wybór tych modeli jest uzasadniony ich względną prostotą pozwalającą na otrzymanie wzorów projektowych przy jednoczesnej wystarczającej do obliczeń przybliżonych dokładności odwzorowania charakterystyk elementów z długim i krótkim kanałem. Analizę przeprowadza się dla układu integratora z obciążeniem przedstawionym na rys. 4, przy czym jej celem jest określenie



Rys. 4. Integrator z obciążeniem rzeczywistym

Fig. 4. Integrator with real load

postaci
ra I_{Q2} .
eniu na
układu
chylene
storów,
wornika
wnoleg-
owania
ia blok
brany
u sieci.

i danej
ądowo-
wyjścia
odgesa
ona od-
deli jest
ch przy
a chara-
układu
reślenie

wymiarów wszystkich tranzystorów komórki, tak aby ich punkty pracy nie wychodziły z zakresu nasycenia. Wybór integratora wynika z tego, że jest najbardziej krytycznym elementem komórki, ponieważ musi zapewnić swobodną ewolucję w czasie zmiennej prądowej „stanu” podczas odpowiedzi sieci na zadane wymuszenie początkowe. Wymiary tranzystorów pozostałych bloków komórki pozostają w prostej proporcji do wymiarów integratora i są określone szablonem SNK.

2.3. ANALIZA DLA PRZYPADKU MODELU TRANZYSTORA MOS LEVEL 1

W dostępnej literaturze i praktyce projektowania układów VLSI przyjmuje się najczęściej, że zakres stosowalności modelu *Level 1* tranzystora MOS jest od dołu ograniczony długością kanału nie mniejszą niż $3\ \mu\text{m}$. Dla istniejących dostępnych technologii CMOS model ten może być zatem użyty, zwłaszcza do obliczeń przybliżonych. Za punkt wyjścia analizy przyjmuje się, że podstawowe wymiary tranzystorów NMOS i PMOS wyrażone poprzez parametry β_n i β_p odnoszą się do układu realizującego funkcję wyjściową z nasyceniem na poziomie prądu I_Q .

1. Pierwszym punktem analizy jest określenie współczynnika skali S dobrego tak, aby zmienna „stanu” i_{xij} miała zapewnioną swobodę zmian w czasie w zakresie:

$$-i_{\max} \leq i_{xij} \leq i_{\max} \quad (11)$$

przy czym wartość $i_{\max}$ jest dla danego szablonu sieci określona wzorem (10). Zmienna ta może przyjmować tylko wartości z przedziału

$$-S I_Q \leq i_{xij} \leq S I_Q \quad (12)$$

co wynika z ograniczeń wydajności prądowych tranzystorów $MQ2$ i $MQ3$ dla przypadku $I_2=0$ i $I_3=0$. Współczynnik S jest zatem określony nierównościami:

$$S \geq i_{\max}/I_Q \quad (13)$$

Wymiary tranzystorów integratora będą zatem S razy większe od wymiarów tranzystorów bloku realizującego funkcję wyjściową.

2. Tranzystor $MQ1$. Trzeba tak dobrać jego polaryzację V_B , aby uzyskać wymagany prąd $S I_Q$ tranzystora.

Z równania prądu tranzystora $MQ1$: $S I_Q = 0.5 S \beta_p (V_{GSMQ1} - V_{Tp})^2$, otrzymuje się zależność na

napięcie źródło-bramka $V_{SGMQ1} = |V_{Tp}| + \sqrt{\frac{2 I_Q}{\beta_p}}$. Uwzględniając związek: $V_B = V_{DD} - V_{SGMQ1}$

napięcie polaryzacji V_B jest zatem równe:

$$V_B = V_{DD} - |V_{Tp}| - \sqrt{\frac{2 I_Q}{\beta_p}} \quad (14)$$

Napięcia: dren-źródło i bramka-źródło tranzystora $M1$ są sobie równe ($M1$ ma zwarty dren z bramką), zatem: $V_{DSM1} = V_{Tn} + \sqrt{\frac{2I_1}{S\beta_n}}$, zaś potencjał drenu tego tranzystora względem punktu wspólnego (najczęściej ziemi) jest równy $V_1 = V_{DSM1} + V_{SS} = V_{Tn} + V_{SS} + \sqrt{\frac{2I_1}{S\beta_n}}$. Wartość maksymalna prądu $I_{1max} = I_{2max}$ jest równa $I_{1max} = I_{2max} = SI_Q + i_{xijmax} = SI_Q + i_{max}$. Zatem maksymalna wartość potencjału V_1 jest równa:

$$V_{1max} = V_{Tn} + V_{SS} + \sqrt{\frac{2(SI_Q + i_{max})}{S\beta_n}} = V_{Tn} + V_{SS} + \sqrt{\frac{2(I_Q + \frac{i_{max}}{S})}{\beta_n}} \quad (15)$$

Warunek pracy tranzystora $MQ1$ w nasyceniu jest określony nierównością: $V_{SDMQ1} > V_{SGMQ1} - |V_{Tp}|$. Minimalne napięcie źródło-dren tranzystora $MQ1$ wynosi $V_{SDMQ1min} = V_{DD} - V_{1max}$

$= V_{DD} - V_{Tn} - V_{SS} - \sqrt{\frac{2(I_Q + \frac{i_{max}}{S})}{\beta_n}}$. Ostatecznie następujący warunek musi być spełniony, aby tranzystor $MQ1$ pracował w nasyceniu:

$$V_{DD} - V_{SS} - V_{Tn} > \sqrt{\frac{2(I_Q + \frac{i_{max}}{S})}{\beta_n}} + \sqrt{\frac{2I_Q}{\beta_p}} \quad (16)$$

Warunek (16) zapewnia pracę tranzystora $MQ1$ w nasyceniu w całym zakresie zmian prądu i_{xij} określonym wzorem (11).

3. Dla tranzystora $M2$ w zakresie nasycenia: $V_{DSM2} > V_{GSM2} - V_{Tn}$, ale ponieważ napięcie dren-źródło tranzystora $M1$ jest równe napięciu bramka-źródło tranzystora $M2$, to $V_{GSM2} = V_{DSM1} = V_{Tn} + \sqrt{\frac{2I_1}{S\beta_n}}$. W przypadku najgorszym do węzła wyjściowego B dopływa prąd I_O równy wartości SI_Q , zatem maksymalny prąd drenu tranzystora $M2$ jest równy $I_{2max} = 2SI_Q$, zatem maksymalne napięcie bramka-źródło tranzystora $M2$ wynosi:

$$V_{GSM2max} = V_{Tn} + \sqrt{\frac{4SI_Q}{S\beta_n}} = V_{Tn} + \sqrt{\frac{4I_Q}{\beta_n}} \quad (17)$$

Dla takiego przypadku prąd drenu $I_3 = 0$, co oznacza że tranzystor $M3$ jest odcięty, a ponieważ pracuje w konfiguracji „diodowej” więc jego napięcie dren-źródło jest mniejsze od napięcia progowego, zatem $V_{DSM2} \leq V_{Tn}$. Uwzględniając to otrzymujemy: $V_{Tn} \geq V_{DSM2}$

$> V_{GSM2} - V_{Tn} = V_{Tn} + \sqrt{\frac{4I_Q}{\beta_n}} - V_{Tn}$, czyli $V_{Tn} > \sqrt{\frac{4I_Q}{\beta_n}}$. Ostatecznie warunkiem zapewniającym pracę tranzystora $M2$ w nasyceniu jest:

$$\beta > \frac{4I_Q}{V_{Tn}^2} \quad (18)$$

4. Dla tranzystora $MQ2$ warunek pracy w nasyceniu jest podobny: $V_{SDMQ2} > V_{SGMQ2} - |V_{Tp}|$. Przypadek najgorszy wystąpi wtedy, gdy napięcie dren-źródło tranzystora $M2$ osiągnie wartość największą. Ponieważ dren tego tranzystora jest połączony z drenem tranzystora $M3$ w konfiguracji „diodowej”, zatem tranzystor $M3$ będzie określał sytuację wystąpienia maksymalnej wartości napięcia dren-źródło. Będzie to miało miejsce wtedy, gdy przez tranzystor $M2$ nie płynie żaden prąd ($I_2=0$), a jednocześnie przez $M3$ płynie suma prądów górnych źródeł, tzn. $I_3 = 2 S I_Q$. Mamy zatem:

$$V_{SDM2max} = V_{Tn} + \sqrt{\frac{2I_{3max}}{S\beta_n}} = V_{Tn} + \sqrt{\frac{4I_Q}{\beta_n}} \quad (19)$$

Wartość największa potencjału V_2 jest równa:

$$V_{2max} = V_{SDM2max} - V_{SS} = V_{Tn} + \sqrt{\frac{4I_Q}{\beta_n}} - V_{SS} \quad (20)$$

W związku z powyższym najmniejsza wartość napięcia źródło-dren tranzystora $MQ2$ będzie wynosiła:

$$V_{SDMQ2min} = V_{DD} - V_{2max} = V_{DD} - V_{SS} - V_{Tn} - \sqrt{\frac{4I_Q}{\beta_n}} \quad (21)$$

Ponieważ $V_{SGMQ2} = V_{SGMQ1} = |V_{Tp}| + \sqrt{\frac{2I_Q}{\beta_p}}$, zatem otrzymujemy:

$$V_{DD} - V_{SS} - V_{Tn} - \sqrt{\frac{4I_Q}{\beta_n}} > |V_{Tp}| + \sqrt{\frac{2I_Q}{\beta_p}} - |V_{Tp}| \quad (22)$$

lub po redukcji:

$$V_{DD} - V_{SS} - V_{Tn} > \sqrt{\frac{4I_Q}{\beta_n}} + \sqrt{\frac{2I_Q}{\beta_p}} \quad (23)$$

Wzory (16) i (23) można zastąpić jednym warunkiem:

$$d = \max [(i_{\max}/S), I_Q], \quad V_{DD} - V_{SS} - V_{Tn} > \sqrt{\frac{2(I_Q + d)}{\beta_n}} + \sqrt{\frac{2I_Q}{\beta_p}} \quad (24)$$

którego spełnienie zapewnia pracę tranzystorów MQ1 i MQ2 w nasyceniu.

2.4. PRZYKŁAD OBLICZENIOWY PROJEKTOWANIA KOMÓRKI SNK Z UŻYCIEM MODELU TRANZYSTORA MOS LEVEL 1

Zostanie zaprojektowany układ komórki sieci realizującej detekcję elementów spójnych [3, 8, 18] w obrazie. Przykład ten wybrano ze względu prostotę jego szablonu, który składa się tylko z trzech elementów macierzy sprzężenia zwrotnego $A_I = [1 \ 2 \ -1]$. Pozostałe elementy są równe 0. Przyjmijmy następujące parametry technologii CMOS $3\mu\text{m}$: $V_{Tn} = 0.9\text{V}$, $KP_n = 57\text{E-}6\text{ A/V}^2$, $V_{Tp} = -0.9\text{V}$, $KP_p = 17\text{E-}6\text{ A/V}^2$. Załóżmy napięcie zasilające $V_{DD} = +5\text{V}$, $V_{SS} = -5\text{V}$ oraz prąd $I_Q = 10\mu\text{A}$.

Procedura projektowania.

1. Obliczamy wartość maksymalną prądu stanu $i_{\max}$ ze wzoru (10), przy założeniu że $I_Q = 10\mu\text{A}$. Wartość ta wynosi: $i_{\max} = 10\mu\text{A} + 10\mu\text{A} (1 + 2 + 1) = 10\mu\text{A} + 40\mu\text{A} = 50\mu\text{A}$. Współczynnik skali S jest zatem równy $S \geq i_{\max} / I_Q = 5$.

2. Ze wzoru (18) określamy parametr tranzystora podstawowego NMOS: $\beta_n > 4 * 10\mu\text{A} / (0.9\text{V})^2 = 49.38\mu\text{A/V}^2$. Przyjmijmy, że $\beta_n = \beta_{n\min}$. Niech długość kanału tego tranzystora ma wartość minimalną w danej technologii CMOS, czyli $3\mu\text{m}$, wtedy szerokość W_n jest równa: $W_n = (\beta_n L_n) / KP_n = (49.38\mu\text{A/V}^2 * 3\mu\text{m}) / 57\mu\text{A/V}^2 = 2.6\mu\text{m}$. Otrzymaną wartość szerokości W_n zaokrąglamy do $W_n = 3\mu\text{m}$ ze względu na to, że szerokość kanału tranzystora MOS nie może być mniejsza od minimalnego wymiaru technologii $\lambda = 3\mu\text{m}$. Ponadto zaokrąglenie to czyni nierówność (18) bardziej ostrą.

3. Przyjmijmy minimalne wymiary tranzystorów podstawowych PMOS, tzn. $W_p = L_p = 3\mu\text{m}$. Obliczamy współczynnik $\beta_p = KP_p (W_p / L_p) = 17\mu\text{A/V}^2 (3\mu\text{m} / 3\mu\text{m}) = 17\mu\text{A/V}^2$. Obliczamy współczynnik $d = \max [(i_{\max}/S), I_Q] = \max (50\mu\text{A}/5, 10\mu\text{A}) = 10\mu\text{A}$.

Sprawdzamy warunek (24): $V_{DD} - V_{SS} - V_{Tn} > \sqrt{\frac{2(I_Q + d)}{\beta_n}} + \sqrt{\frac{2I_Q}{\beta_p}}$ $5\text{V} + 5\text{V} - 0.9\text{V} = 9.1\text{V} >$

$$\sqrt{\frac{4 * 10\mu\text{A}}{49.38\mu\text{A/V}^2}} + \sqrt{\frac{2 * 10\mu\text{A}}{17\mu\text{A/V}^2}} = 0.9\text{V} + 1.08\text{V} = 1.98\text{V} - \text{warunek spełniony.}$$

4. Obliczamy napięcie polaryzacji źródeł prądowych PMOS, $V_B = V_{DD} - V_{Tp} - \sqrt{\frac{2I_Q}{\beta_p}}$.

$$= 5\text{V} - 0.9\text{V} - \sqrt{\frac{2 * 10\mu\text{A}}{17\mu\text{A/V}^2}} = 4.1\text{V} - 1.085\text{V} = 3.015\text{V}.$$

5. Przeliczenie warunków początkowych. Formalność zmiennej „stanu” i_{xij} polega między innymi na tym, że nie można zadać prądu początkowego w rezystancji R_x , ponieważ nie jest to element magazynujący energię. Dlatego warunek początkowy należy przeliczyć

na pocz
jest rów
Z komb

Dla ro
odpowi
(oblicz

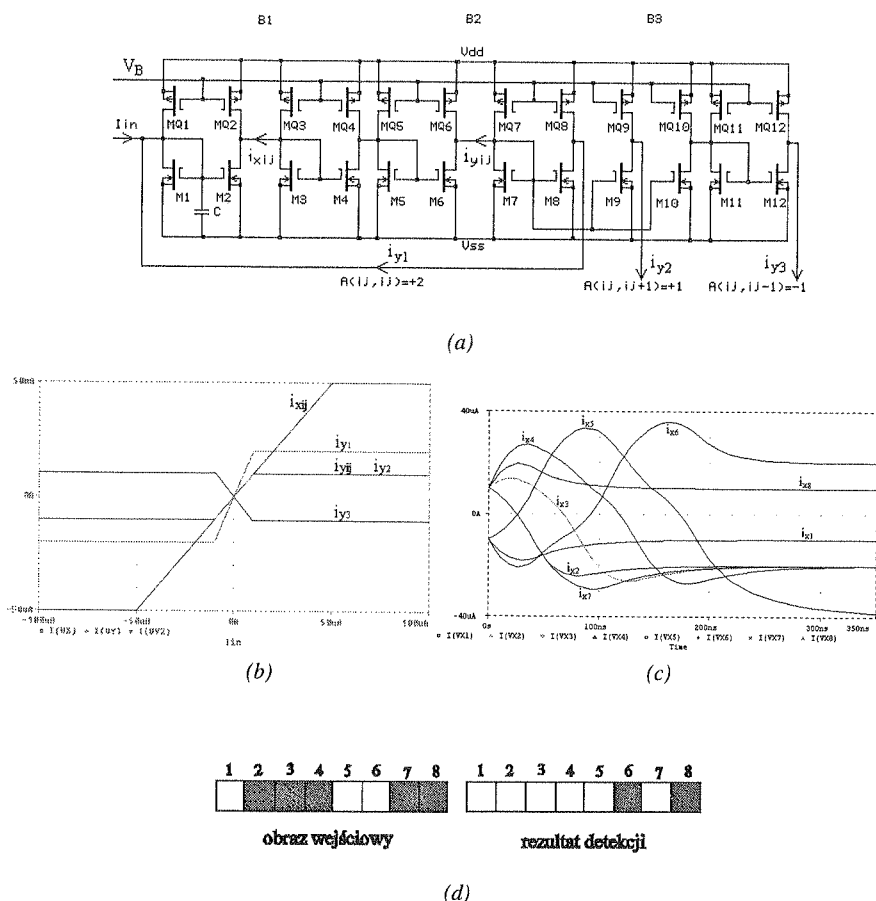
Rys. 5.
komórk

Fig. 5.
CNN co

na początkowe napięcie $V_1(0)$ kondensatora C . Prąd w chwili początkowej tranzystora $M2$ jest równy: $SI_Q + i_{xij}(0) = 0.5 S \beta_n (V_{GSM2} - V_{Tn})^2$, zaś potencjał $V_1(0) = V_{SS} + V_{GSM2}(0)$. Z kombinacji tych dwóch zależności otrzymujemy:

$$V_1(0) = V_{Tn} + V_{SS} + \sqrt{\frac{2(SI_Q + i_{xij}(0))}{S\beta_n}} \quad (25)$$

Dla rozważanego przykładu zmianom prądu i_{xij} w zakresie od $-50 \mu A$ do $+50 \mu A$ odpowiadają zmiany napięcia kondensatora C względem masy V_1 od $-4.1V$ do $-3.2V$ (obliczenia ze wzoru (25)).



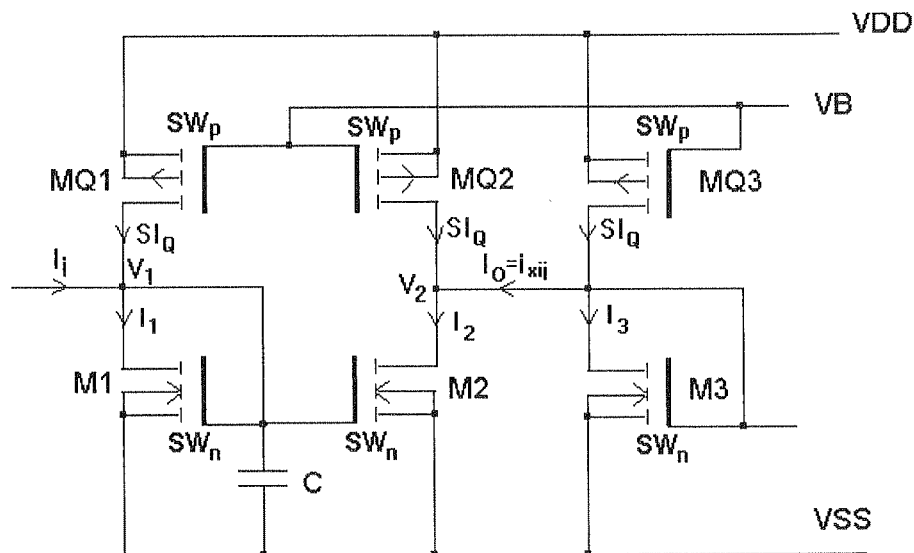
Rys. 5. Schemat komórki w wersji prądowej sieci detektora elementów spójnych (a) wyniki symulacji DC komórki programem *PSPICE* (b), wyniki symulacji 8-komórkowej SNK detektora elementów spójnych (MOSFET model *Level 1*) (c), obraz poddany przetwarzaniu i jego wynik (d)

Fig. 5. Scheme of current-mode connected component detecting CNN cell (a), *PSPICE* DC simulation results of CNN cell (MOSFET model *Level 1*) (b), simulation results of 8-cells connected component detecting CNN (c), input and output image (d)

Ostatecznie wymiary tranzystorów są następujące: wszystkie kanały mają długość minimalną $3\text{ }\mu\text{m}$, natomiast szerokości odpowiednio: $W_p = 3\text{ }\mu\text{m}$, $W_n = 3\text{ }\mu\text{m}$ (dla elementu podstawowego) oraz $S \times W_p = 15\text{ }\mu\text{m}$, $S \times W_n = 15\text{ }\mu\text{m}$, $2 \times W_n = 6\text{ }\mu\text{m}$, $2 \times W_p = 6\text{ }\mu\text{m}$. Na rys. 5 przedstawiono schemat komórki sieci komórkowej jednowymiarowej detektora elementów spójnych do zliczania wzdłuż wierszy obrazu elementów spójnych oraz wyniki symulacji programem *PSPICE v.5.2* firmy *MicroSim*. Wyniki przetwarzania detektora elementów spójnych są poprawne, gdyż SNK wykryła 2 elementy spójne i zapisała je w postaci sekwencji pikseli przesuniętych do prawej krawędzi, co obrazują rys. 5c i rys. 5d.

2.5. ANALIZA DLA PRZYPADKU MODELU TRANZYSTORA MOS LEVEL 6

Model Sakurai'a-Newtona [1], mający swoją implementację w programie *SPICE3* pakietu *ICAP4* firmy *Intusoft*, jest polecany do celów obliczeń przybliżonych układów z tranzystorami MOS o długościach kanałów równych i mniejszych od $1\text{ }\mu\text{m}$. Zaletą tego modelu jest to, że dla bardzo krótkiego kanału (długości submikrometrowe) opis prądu drenu w zakresie nasycenia zależy liniowo od napięcia bramki, co znacznie upraszcza obliczenia i analizę układów. Pamiętać jednak trzeba, że istotną wadą takich tranzystorów jest duże nachylenie charakterystyk w zakresie nasycenia, tzn. niewielka rezystancja wyjściowa, co ma istotny wpływ na prądy i napięcia szczytkowe w układach analogowych. Za punkt wyjścia analizy przyjmuje się, że podstawowe wymiary tranzystorów NMOS i PMOS odnoszą się do układu realizującego funkcję wyjściową z nasyceniem na poziomie prądu I_0 . Analizę przeprowadza się dla układu integratora z obciążeniem przedstawionym na rys. 6.



Rys. 6. Integrator z obciążeniem do analizy z modelem MOSFET Level 6

Fig. 6. Integrator with real load used for analysis with MOSFET model Level 6

długość
elementu
= 6 μm .
detektora
az wyniki
detektora
apisała je
i rys. 5d.

SPICE3
układów
n. Zaletą
metrowe)
znacznie
ną wadą
enia, tzn.
czątkowe
stawowe
o funkcję
ła układu

DD

1. Pierwszy punkt analizy jest analogiczny jak w poprzednim przypadku analizy dla modelu tranzystora MOSFET *Level 1*.

2. Tranzystor *MQ1*. Trzeba tak dobrać polaryzację V_B , aby uzyskać wymagany prąd SI_Q tranzystora.

Z równania prądu tranzystora *MQ1*: $SI_Q = S \frac{W_p}{L_p} B_p (V_{GSMQ1} - |V_{Tp}|)^{n_p}$, otrzymuje się zależność określającą napięcie źródło-bramka: $V_{SGMQ1} = |V_{Tp}| + \left(\frac{I_Q L_p}{W_p B_p} \right)^{\frac{1}{n_p}}$. Napięcie polaryzacji V_B jest zatem równe:

$$V_B = V_{DD} - |V_{Tp}| - \left(\frac{I_Q L_p}{W_p B_p} \right)^{\frac{1}{n_p}} \quad (26)$$

Napięcie dren-źródło tranzystora *M1* jest równe: $V_{DSM1} = V_{Tn} + \left(\frac{I_1 L_n}{SW_n B_n} \right)^{\frac{1}{n_n}}$, zaś potencjał drenu tego tranzystora względem ziemi: $V_1 = V_{DSM1} + V_{ss} = V_{ss} + V_{Tn} + \left(\frac{I_1 L_n}{SW_n B_n} \right)^{\frac{1}{n_n}}$. Wartość maksymalną prądu $I_{1max} = I_{2max}$ jest równa $I_{1max} = I_{2max} = SI_Q + i_{xijmax} = SI_Q + i_{max}$. Zatem maksymalna wartość potencjału V_1 jest równa:

$$V_{1max} = V_{ss} + V_{Tn} + \left\{ \frac{(SI_Q + i_{max}) L_n}{SW_n B_n} \right\}^{\frac{1}{n_n}} \quad (27)$$

Warunek pracy tranzystora *MQ1* w nasyceniu jest określony zależnością: $V_{SDMQ1} > K_p (V_{SGMQ1} - |V_{Tp}|)^{m_p}$. Minimalne napięcie źródło-dren tranzystora *MQ1* wynosi $V_{SDMQ1min} = V_{DD} - V_{1max} = V_{DD} - V_{ss} - V_{Tn} - \left\{ \frac{(SI_Q + i_{max}) L_n}{SW_n B_n} \right\}^{\frac{1}{n_n}}$. Ostatecznie następujący warunek musi być spełniony, aby tranzystor *MQ1* pracował w nasyceniu:

$$V_{DD} - V_{ss} - V_{Tn} > \left\{ \frac{(I_Q + \frac{i_{max}}{S}) L_n}{W_n B_n} \right\}^{\frac{1}{n_n}} + K_p \left(\frac{I_Q L_p}{W_p B_p} \right)^{\frac{m_p}{n_p}} \quad (28)$$

Warunek (28) zapewnia pracę tranzystora *MQ1* w nasyceniu w całym zakresie zmian prądu i_{xij} , określonym wzorem (11).

3. Tranzystor *M2*. W zakresie nasycenia: $V_{DSM2} > K_n (V_{GSM2} - V_{Tn})^{m_n}$, ale ponieważ napięcie dren-źródło tranzystora *M1* jest równe napięciu bramka-źródło tranzystora *M2*, zatem: $V_{GSM2} = V_{DSM1} = V_{Tn} + \left(\frac{I_1 L_n}{SW_n B_n} \right)^{\frac{1}{n_n}}$. W przypadku najgorszym do węzła wyjściowego dopływa prąd I_O równy wartości SI_Q , zatem maksymalny prąd drenu tranzystora *M2* jest równy $I_{2max} = 2SI_Q$, zatem maksymalne napięcie bramka-źródło tranzystora *M2* wynosi:

$$V_{GSM2max} = V_{Tn} + \left(\frac{2I_Q L_n}{W_n B_n} \right)^{\frac{1}{n_n}} \quad (29)$$

Dla takiego przypadku prąd drenu $I_3 = 0$, co oznacza że tranzystor $M3$ jest odcięty, a ponieważ pracuje w konfiguracji „diodowej”, więc jego napięcie dren-źródło jest mniejsze od napięcia progowego, zatem $V_{DSM2} \leq V_{Tn}$. Uwzględniając to otrzymujemy:

$$V_{Tn} \geq V_{DSM2} > K_n (V_{GSM2} - V_{Tn})^{m_n} = K_n \left\{ V_{Tn} + \left(\frac{2I_Q L_n}{W_n B_n} \right)^{\frac{1}{n_n}} - V_{Tn} \right\}^{m_n}, \text{ czyli } V_{Tn} > K_n \left(\frac{2I_Q L_n}{W_n B_n} \right)^{\frac{1}{n_n}}.$$

Ostatecznie warunkiem zapewniającym pracę tranzystora $M2$ w nasyceniu jest:

$$\frac{L_n}{W_n} < \frac{B_n}{2I_Q} \left(\frac{V_{Tn}}{K_n} \right)^{\frac{n_n}{m_n}} \quad (30)$$

4. Tranzystor $MQ2$. Warunek wyjściowy pracy w nasyceniu jest podobny: $V_{SDMQ2} > K_p (V_{SGMQ2} - |V_{Tp}|)^{m_p}$. Przypadek najgorszy wystąpi wtedy, gdy napięcie dren-źródło tranzystora $M2$ osiągnie wartość największą. Ponieważ dren tego tranzystora jest połączony z drenem tranzystora $M3$ w konfiguracji „diodowej”, zatem tranzystor $M3$ będzie określał moment wystąpienia maksymalnej wartości napięcia dren-źródło. Będzie to miało miejsce wtedy, gdy przez tranzystor $M2$ nie płynie żaden prąd ($I_2=0$), a jednocześnie przez $M3$ płynie suma prądów górnych źródeł, tzn. $I_3 = 2 I_{I_Q}$. Mamy zatem:

$$V_{DSM2max} = V_{Tn} + \left(\frac{2SI_Q L_n}{SW_n B_n} \right)^{\frac{1}{n_n}} = V_{Tn} + \left(\frac{I_3 L_n}{SW_n B_n} \right)^{\frac{1}{n_n}} = V_{Tn} + \left(\frac{2I_Q L_n}{W_n B_n} \right)^{\frac{1}{n_n}} \quad (31)$$

Wartość największa potencjału V_2 jest równa:

$$V_{2max} = V_{DSM2max} + V_{SS} = V_{Tn} + \left(\frac{2I_Q L_n}{W_n B_n} \right)^{\frac{1}{n_n}} + V_{SS} \quad (32)$$

A zatem najmniejsza wartość napięcia źródło-dren tranzystora $MQ2$ będzie wynosiła:

$$V_{SDMQ2} = V_{DD} - V_{2max} = V_{DD} - V_{SS} - V_{Tn} - \left(\frac{2I_Q L_n}{W_n B_n} \right)^{\frac{1}{n_n}} \quad (33)$$

Ponieważ $V_{SDMQ2} = V_{SDMQ1} = |V_{Tp}| + \left(\frac{I_Q L_p}{W_p B_p} \right)^{\frac{1}{n_p}}$, zatem otrzymujemy:

$$V_{DD} - V_{SS} - V_{Tn} - \left(\frac{2I_Q L_n}{W_n B_n} \right)^{\frac{1}{n_n}} > K_p \left\{ |V_{Tp}| + \left(\frac{I_Q L_p}{W_p B_p} \right)^{\frac{1}{n_p}} - |V_{Tp}| \right\}^{m_p} \quad (34)$$

lub po redukcji:

(29)

$$V_{DD} - V_{SS} - V_{Tn} > \left(\frac{2I_Q L_n}{W_n B_n} \right)^{\frac{1}{n_n}} + K_p \left(\frac{I_Q L_p}{W_p B_p} \right)^{\frac{m_n}{n_p}} \quad (35)$$

...t odcięty,
...ródło jest
...ujemy:

Wzory (28) i (35) można zastąpić jednym warunkiem:

$$\left(\frac{2I_Q L_n}{W_n B_n} \right)^{\frac{1}{n_n}}.$$

$$a = \max [(i_{max}/S, I_Q], \quad V_{DD} - V_{SS} - V_{Tn} > \left(\frac{(I_Q + a)L_n}{W_n B_n} \right)^{\frac{1}{n_n}} + K_p \left(\frac{I_Q L_p}{W_p B_p} \right)^{\frac{m_p}{n_p}} \quad (36)$$

którego spełnienie zapewnia pracę tranzystorów *MQ1* i *MQ2* w nasyceniu.

t:

2.6. PRZYKŁAD OBLICZENIOWY DLA PRZYPADKU MODELU TRANZYSTORA MOS LEVEL 6

(30)

... napięcie
... tranzystora
... zystor *M3*
... Będzie to
... nocześnie

Według powyższego algorytmu obliczono wymiary tranzystorów komórki SNK detektora elementów spójnych przyjmując parametry technologii CMOS 0.25 μm [1] jak w tabeli 1. Przyjęto napięcia zasilające $V_{SS}=0$ i $V_{DD}=+3V$, prąd $I_Q=10\mu A$ oraz założono, że długości kanałów wszystkich tranzystorów są równe 0.25 μm. Schemat ideowy komórki SNK detektora elementów spójnych jest przedstawiony na rys. 5a. Otrzymane wymiary zestawiono w tabeli 2. Na rys. 7 przedstawiono wyniki symulacji komputerowej *SPICE/ICAP4* jednowymiarowej SNK detektora elementów spójnych otrzymane przy zastosowaniu tzw. modelu *Level 6* z parametrami jak w tabeli 1. Porównując rys. 7 z rys. 5c można zauważyć, że wyniki przetwarzania są identyczne oraz SNK zrealizowana w technologii submikronowej MOSFETs 0.25μm będzie szybsza.

(31)

Tabela 1

Parametry technologii CMOS 0.25 μm użyte do projektu
0.25μm CMOS technology parameters used for designing

(32)

ynosiła:

(33)

Parametr ICAP4	Parametr we wzorach	NMOS	PMOS
KC	B	4.97E-5	1.12E-5
NC	n	1.0484	1.3649
KV	K	0.835	1.054
NV	m	0.620	0.740
VT	V_T	0.860	-0.872

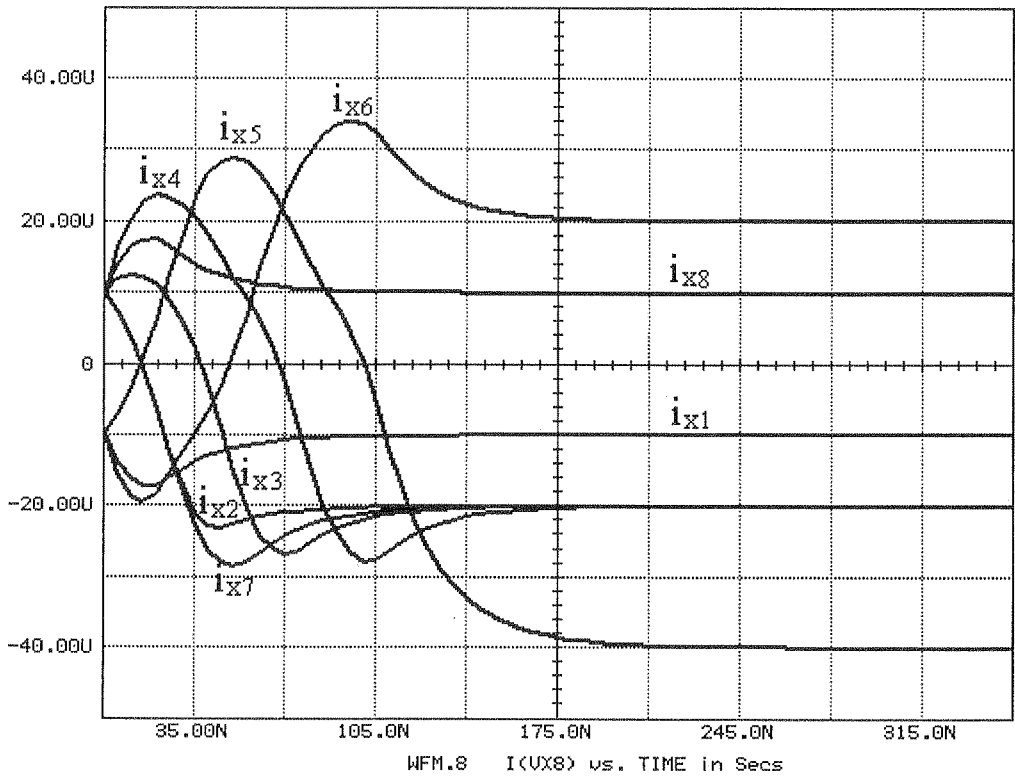
(34)

Tabela 2

Zaprojektowane wymiary tranzystorów komórki SNK detektora elementów spójnych

Designed transistors dimensions of connected componnt detecting CNN cell

	M1,M2	M3,M4	M5,M6	M7	M8	M9-M12
W[μm]	1.25	1.25	0.25	0.25	0.5	0.25
	MQ1,MQ2	MQ3,MQ4	MQ5,MQ6	MQ7	MQ8	MQ9-MQ12
W[μm]	1.25	1.25	0.25	0.25	0.5	0.25



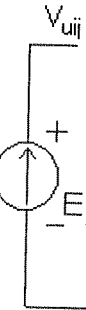
Rys. 7. Wyniki symulacji SPICE3 pakietem ICAP4 sieci detektora elementów spójnych z zastosowaniem modelu Level 6 MOSFET 0.25μm

Fig. 7. SPICE3 ICAP4 simulation results of connected component detecting CNN with MOSFET model Level 6 of 0.25μm technology

3. MODEL KOMÓRKI SNK „FULL RANGE” [9]

Model komórki w wersji prądowej prezentowany do tej pory jest równoważny modelowi podstawowemu, przedstawionemu w pracy [2]. Jego implementacja fizyczna

jest na c
łatwe. P
szablon
podstaw
skompli
gdy kor
różnych
Pow
względ
rozwiąz
prądowa
element
jak na r



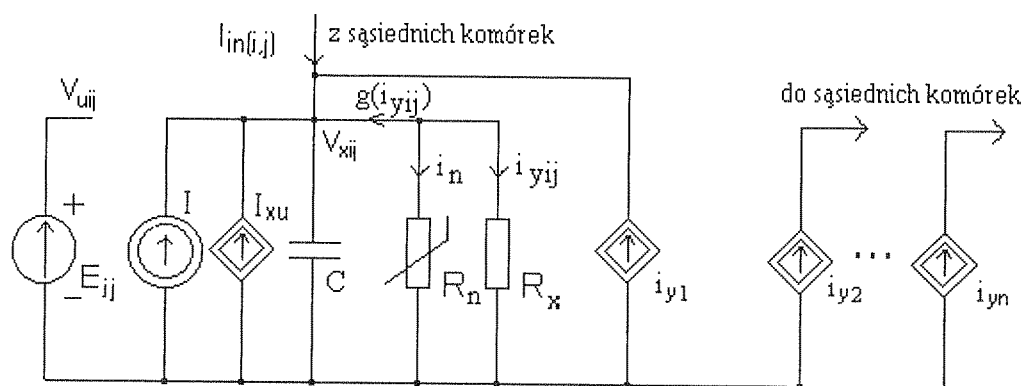
Mode

gdzie fun

zaś współ

jest na ogół dość skomplikowana jak również spełnienie wymagań poprawnej pracy nie jest łatwe. Ponadto model ten nie zawiera elementów pozwalających na zmianę parametrów szablonu, czyli nie ma zdolności do reprogramowania swoich funkcji. Może być zatem podstawą projektowania tylko takich sieci, których funkcje są ustalone i na ogół niezbyt skomplikowane. Funkcjonalność sieci komórkowych zwiększa się istotnie w przypadku, gdy komórki można w ograniczony sposób adaptować do różnych zadań lub sekwencji różnych zadań z dziedziny przetwarzania obrazów.

Powyższe aspekty spowodowały konieczność poszukiwania nowych i lepszych pod względem funkcjonalności rozwiązań układowych komórek SNK. Jednym z takich rozwiązań jest zaproponowany w pracy [9] tzw. model "full range". Jest to nowa wersja prądowa komórki (rys. 1), w której do liniowego rezystora R_x dołączony jest równolegle element rezystancyjny nieliniowy z ograniczeniem prądowym na zadanym poziomie tak jak na rys. 8.



Rys. 8. Model obwodowy komórki w koncepcji prądowej „full range”

Fig. 8. Circuit representation of current-mode „full-range” CNN cell

Model matematyczny komórki „full range” opisany jest następującym równaniem:

$$R_x C(di_{yij}/dt) = g(i_{yij}) + I + \sum_{C(k,l) \in Nr(i,j)} [A_I(ij,kl)i_{ykl} + B_G(ij,kl)v_{ukl}] \quad (37)$$

gdzie funkcja $g(i_{yij})$ jest określona następująco:

$$g(i_{yij}) = \begin{cases} -h(i_{yij} + I_Q) + I_Q, & i_{yij} < -I_Q \\ -i_{yij}, & -I_Q \leq i_{yij} \leq I_Q \\ -h(i_{yij} - I_Q) - I_Q, & i_{yij} > I_Q \end{cases} \quad (38)$$

zaś współczynnik h przyjmuje odpowiednio duże wartości.

Funkcję $g(i_{yij})$ można rozdzielić na dwa następujące składniki:

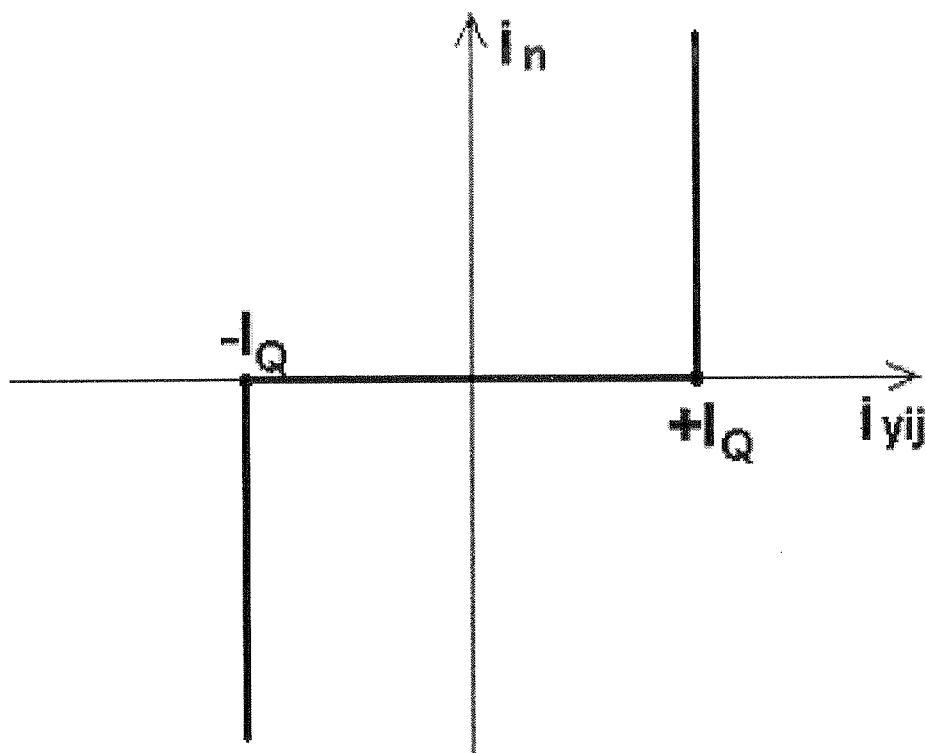
$$g(i_{yij}) = -i_{yij} - \begin{cases} (h-1)(i_{yij} + I_Q), & i_{yij} < -I_Q \\ 0, & -I_Q \leq i_{yij} \leq I_Q \\ (h-1)(i_{yij} - I_Q), & i_{yij} > I_Q \end{cases} \quad (39)$$

Pierwszy składnik wzoru (39) bez znaku minus określa prąd i_n opornika nieliniowego R_n uzależniony od prądu „stanu” i_{yij} z rys. 8, mianowicie :

$$i_n(i_{yij}) = \begin{cases} (h-1)(i_{yij} + I_Q), & i_{yij} < -I_Q \\ 0, & -I_Q \leq i_{yij} \leq I_Q \\ (h-1)(i_{yij} - I_Q), & i_{yij} > I_Q \end{cases} \quad (40)$$

Wzór (39) można przepisać w prostej postaci:

$$g(i_{yij}) = -[i_n(i_{yij}) + i_{yij}] \quad (41)$$



Rys. 9. Charakterystyka opornika ograniczającego R_n w modelu „full range”

Fig. 9. Characteristic of nonlinear limiting resistor R_n in „full-range” CNN model

Charak
W ten
względ
modelu
z chwi
wykorz
ramow
szego s
sekwen

gdzie k
przetw
z grada
przy u
binaryz
obiekto

in
obraz

Wy
stawo
komór
tranzj
twarza
„diod
a jego
pływaj
zmieni
sumow
(tranzj
wyjści

wego R_n

(39)

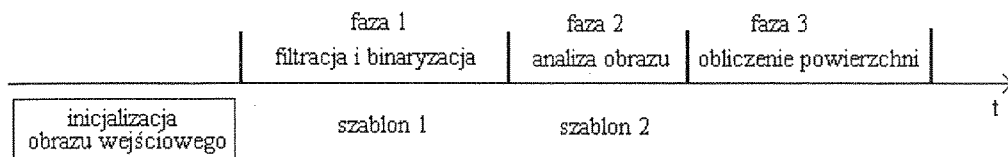
Charakterystyka nieliniowego opornika ograniczającego R_n jest przedstawiona na rys. 9. W ten sposób ogranicza się do zadanego przedziału zmiany prądu stanu $i_{xij} = i_{yij}$, co ze względu na tę równość pozwala wyeliminować przekształtnik $i_{yij} = f(i_{xij})$. Zaletą takiego modelu jest prosta budowa układowa oraz to, że stan ustalony sieci jest ściśle określony z chwilą osiągnięcia przez zmienną stanu i_{xij} poziomu ograniczenia. Pozwala to na wykorzystanie komórki w pracy sekwencyjnej z wieloma szablonami w trybie reprogramowalnym. Po ustaleniu się stanu przejściowego z pierwszej sekwencji (efekt pierwszego szablonu) szablon sieci może być skokowo zmieniony i rozpocznie się następna sekwencja przetwarzania informacji w sieci bez naruszenia warunku:

(40)

$$|i_{xij}(t_k)| \leq I_Q, \quad (42)$$

(41)

gdzie k – numer sekwencji przetwarzania, t_k – moment przełączenia na k -tą sekwencję przetwarzania. Przykładem może być tu sieć do estymacji powierzchni obiektów obrazu z gradacją szarości, której sekwencja pracy składać się musi z trzech faz wykonywanych przy użyciu dwóch szablonów, rys. 10. W pierwszej fazie sieć powinna wykonać binaryzację i filtrację szumów obrazu, w drugiej i trzeciej zaś dokonać identyfikacji obiektów i policzenia ich pikseli, czyli powierzchni obiektów.

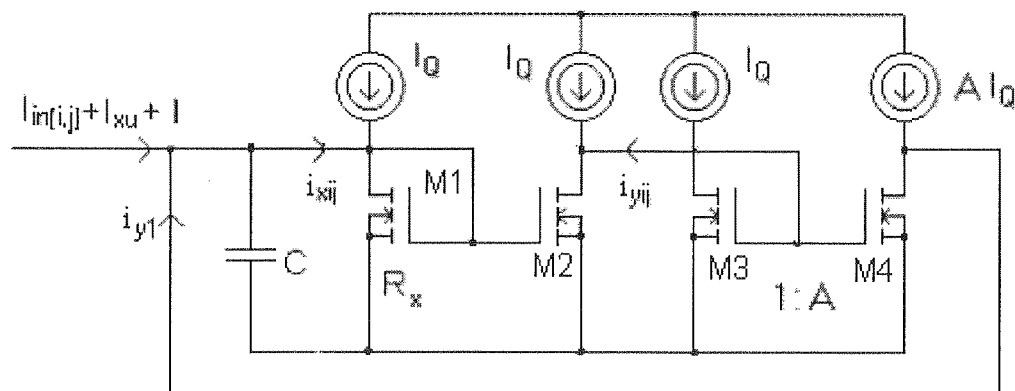


Rys. 10. Sekwencje pracy SNK do obliczania powierzchni obiektu obrazu

Fig. 10. Time diagram of the operation of CNN image object area estimator

yij

Wykazuje się pełną równoważność proponowanego modelu z modelem podstawowym z punktu widzenia wyjścia i_{yij} . Na rys. 11 przedstawiono układową realizację komórki „full range”, gdzie opornik R_x jest realizowany jako połączenie „diodowe” tranzystora MOS $M1$. Jednakże w zastosowaniach wymagających liniowości przetwarzania informacji w stanach przejściowych nieliniowa rezystancja R_x połączenia „diodowego” tranzystora $M1$ jest poważnym ograniczeniem. Zmienną stanu jest tu prąd i_{xij} , a jego maksymalny poziom nie powinien przekraczać wartości I_Q dla prądów dopływających i wartości $-I_Q$ dla prądów odpływających. W takim samym zakresie powinien zmieniać się prąd i_{yij} , zatem pierwszy wzmacniacz prądowy (tranzystory $M1$ i $M2$) realizuje sumowanie, całkowanie i generację sygnału wyjściowego. Drugi wzmacniacz prądowy (tranzystory $M3$ i $M4$) realizuje funkcję wyjścia oraz służy do multiplikacji sygnału wyjściowego przez element centralny szablonu sprzężenia zwrotnego sieci.



Rys. 11. Model układowy komórki „full range”

Fig. 11. „Full range” CMOS current-mode cell

Model matematyczny komórki jest opisany równaniem podobnym do (37):

$$\tau \frac{di_{yij}}{dt} = g(i_{yij}) + i_{y1} + I_{in}(i,j) + I_{xu} + I \quad (43)$$

przy czym stała czasowa zależy od parametrów tranzystora $M1$ i prądu przezeń płynącego, czyli nie jest stała w trakcie przetwarzania informacji w SNK. Może to prowadzić do znacznych ubytków informacji w SNK o niektórych szablonach — szczególnie wtedy, gdy informację wyjściową stanowią stany przejściowe w sieci. Jednakże dla dużej grupy szablonów sieci SNK zjawisko nieliniowości stałej czasowej nie stanowi większego problemu i model z rysunku 11 może być stosowany.

Jako przykład analizy wybrano 32-komórkową SNK o jednowymiarowym schemacie połączeń międzykomórkowych służącą do filtracji zakłóceń szumowych i binaryzacji obrazu. Do realizacji obydwu zadań wybrany został model sieci bez sterowania, natomiast szablon sprzężenia zwrotnego użytej sieci ma następującą postać:

$$A = [1 \ 4 \ 1], \quad B = 0, \quad I = 0$$

Przedstawiony szablon jest podobny do szablonu sieci $A=[1 \ 2 \ 1]$, $B=0$, $I=0$. Charakter przetwarzania informacji zawartej w linii jest dla obydwu szablonów taki sam i sprowadza się do operacji wielokrotnego uśredniania, przy czym zapewnienie warunku $A_{00} > 1$ [2] gwarantuje uzyskanie w stanie stabilnym sieci binarnych sygnałów wyjściowych wszystkich komórek. Zwiększenie wartości współczynnika A_{00} prowadzi do zbliżania się wyniku przetwarzania do wyniku, jaki zostałby uzyskany przez zastosowanie zwykłego progowania, przy czym maleje zdolność sieci do tłumienia szumu addytywnego. Wybór wartości $A_{00}=4$ stanowi wynik kompromisu między zachowaniem zdolności sieci do redukcji poziomu zakłóceń a możliwością poprawnej segmentacji obrazów niezaszumionych (dla

wartości v
krawędzi

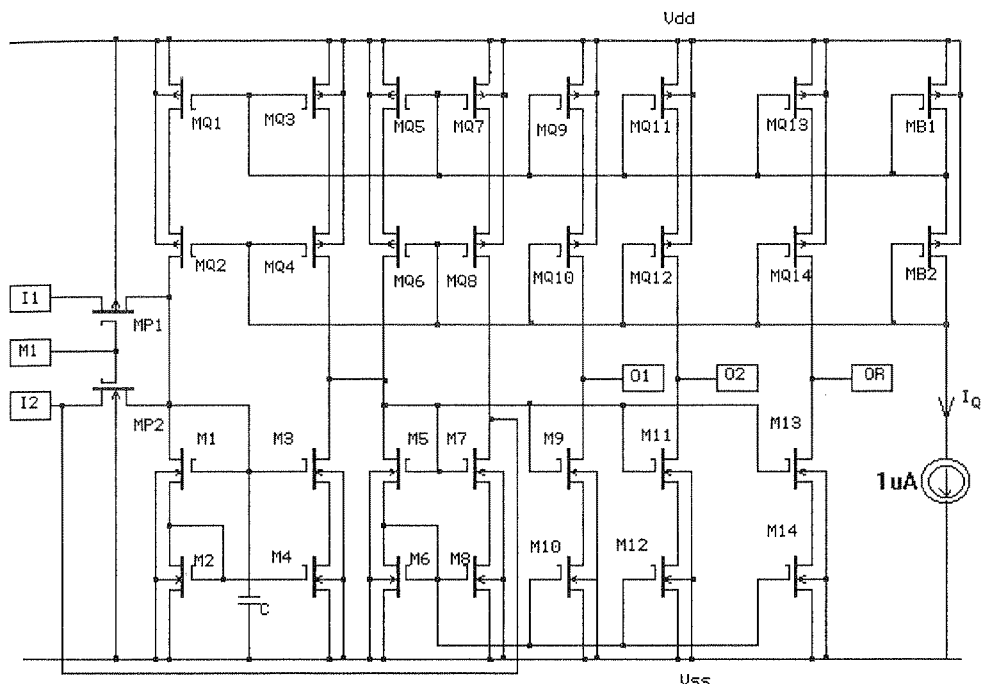
Podsta
prądowy
tranzystor
powierz
wykorzys
prądowy
większą



Sche
szablonu
wartości
na wejśc
sygnałów
oddziały
zwrotne
sygnałów

wartości wagi $A_{00}=2$ zmiana rozmiaru obiektów obrazu wynikająca z "rozmywania" jego krawędzi w trakcie trwania procesów przejściowych jest bardziej prawdopodobna).

Podstawowym elementem opisywanej sieci jest komórka wykorzystująca model prądowy "full range". Wybór tej metody realizacji cechuje się minimalną liczbą tranzystorów wymaganych dla realizacji układowej komórki i w konsekwencji, minimalną powierzchnią krzemu zajmowaną przez komórkę. Podstawowym elementem układowym wykorzystanym w komórce jest wzmacniacz prądowy z kaskodowymi zwierciadłami prądowymi, wybranymi ze względu na ich dużą rezystancję wyjściową, a tym samym większą odporność na stałe napięcia szcztkowe.



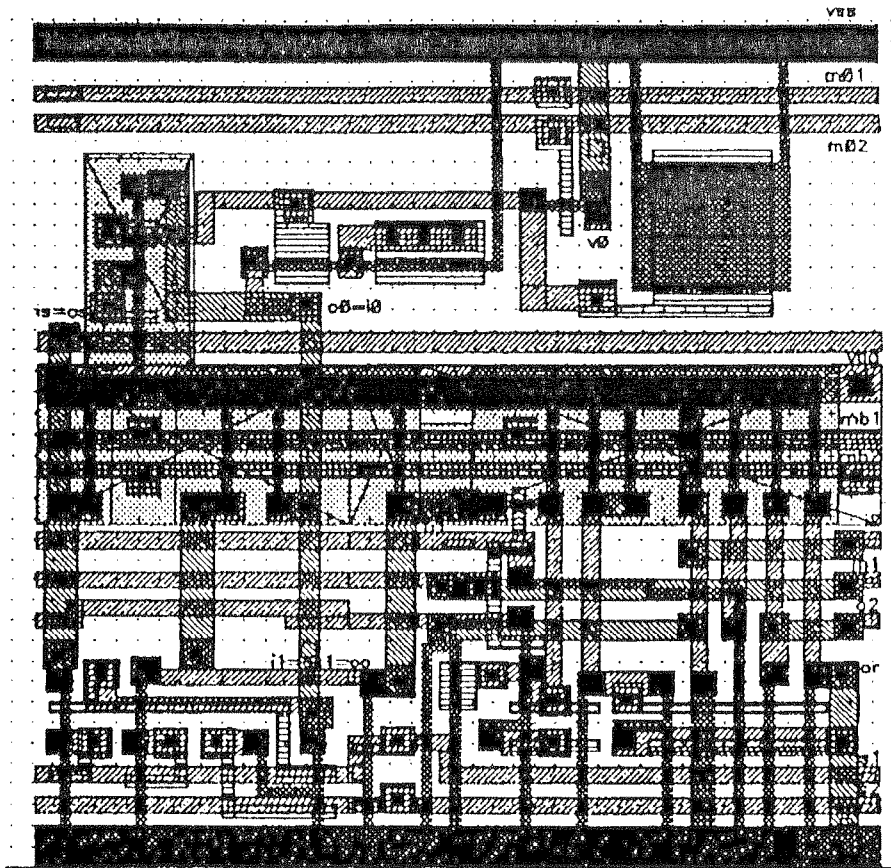
Rys. 12. Schemat ideowy komórki SNK do filtracji szumów i binaryzacji obrazu [12]

Fig. 12. Scheme of the cell of CNN for noise removal and image binaryzation [12]

Schemat ideowy komórki został przedstawiony na rys. 12. Z uwagi na stałość szablonu sieci, w przedstawionym schemacie nie występują sygnały przeprogramowujące wartości wag. Sygnałami sterującymi komórką są sygnały wyjściowe komórek sąsiednich na wejściu $I2$ oraz sygnał wyjściowy komórki sprzężenia zwrotnego własnego. Wartości sygnałów sterujących są ważone przez współczynniki o wartościach równych jedności dla oddziaływań międzykomórkowych i wartości wynoszącej cztery dla własnego sprzężenia zwrotnego. Istotną cechą przyjętego modelu komórki jest dokonywanie operacji ważenia sygnałów sterujących na wyjściu a nie na wejściu komórki, co oznacza, że każda komórka

ma dwa różne wyjścia prądowe $O1$ i $O2$. Sygnały sterujące komórką są sumowane w węźle oznaczonym symbolem $I2$, podlegając całkowaniu w obwodzie zawierającym pojemność C (tworzą ją pojemności wejściowe tranzystorów MOS). Odcinkowo-liniowa charakterystyka wyjściowa komórki realizowana jest w układzie złożonym z tranzystorów $M1-M8$, $MQ1-MQ8$. Stan początkowy komórki jest wprowadzany linią oznaczoną symbolem $I1$, a o trybie pracy sieci - przetwarzania lub ładowania informacji początkowej - decyduje sygnał sterujący oznaczony symbolem $M1$, przełączający klucze $MP1$ i $MP2$. Obok wyjść ważonych prądów wyjściowych (wyznaczanych w gałęziach z tranzystorami $M9$ i $M10$ dla wyjścia prądowego $O1$ i $M11$, $M12$ dla wyjścia $O2$), każda z komórek jest dodatkowo wyposażona w wyjście oznaczone symbolem OR , przeznaczone do wyprowadzania informacji z SNK.

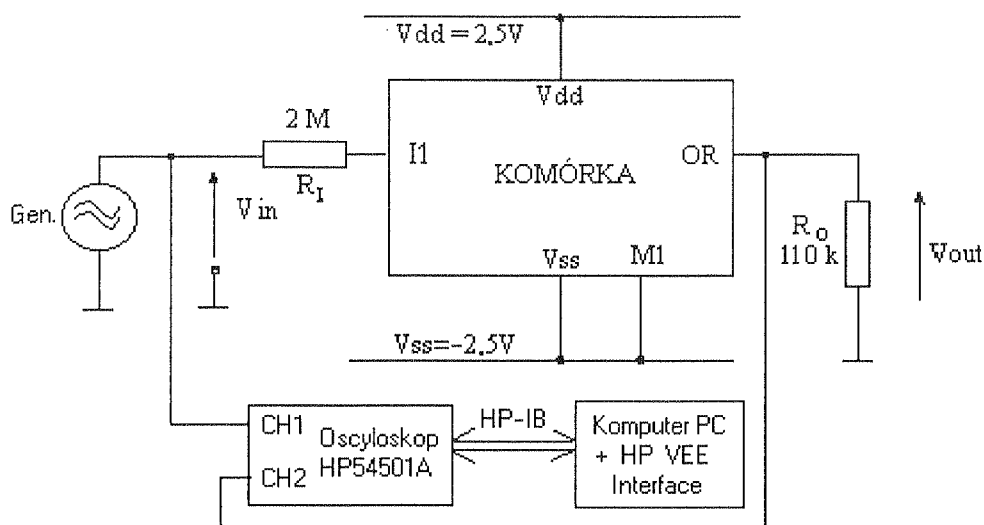
Opisana SNK została wykonana jako część układu scalonego wykonanego w technologii MIETEC CMOS $2.4\mu\text{m}$ [12] reprogramowalnej sieci neuronowej komórkowej dla celów specjalizowanej analizy obrazów. Topografię komórki „full range” SNK do filtracji szumów i binaryzacji obrazu przedstawiono na rys. 13.



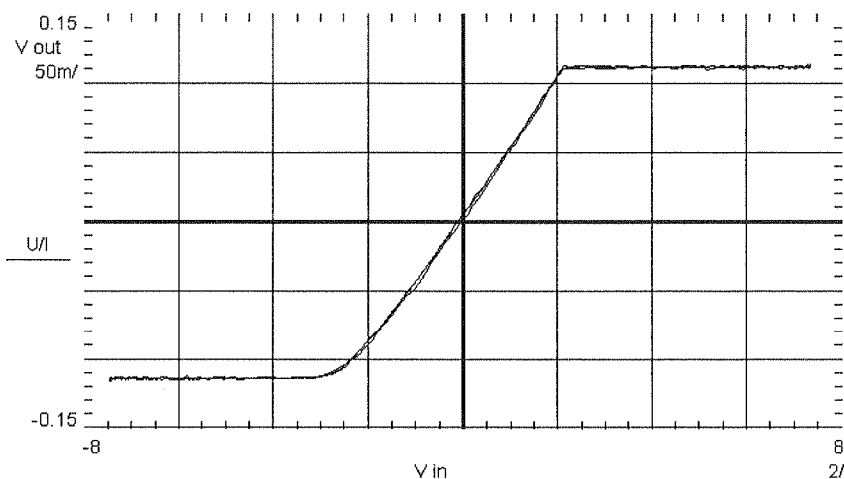
Rys. 13. Topografia komórki SNK do filtracji szumów i binaryzacji obrazu [12]

Fig. 13. Layout of the cell of CNN for noise removal and image binaryzation [12]

Przeprowadzono następujące pomiary pojedynczej komórki o schemacie z rys. 12: pomiar charakterystyki przejściowej komórki z wyłączonym sprzężeniem zwrotnym oraz pomiar charakterystyki przejściowej komórki z włączonym sprzężeniem zwrotnym.



(a)

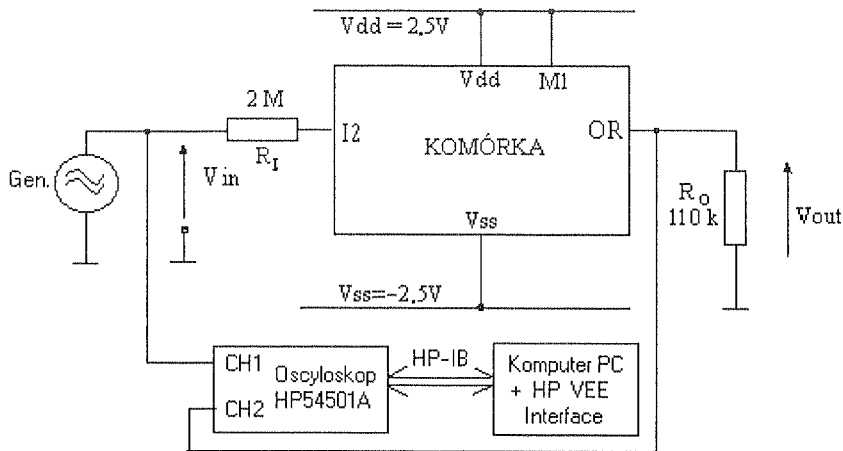


(b)

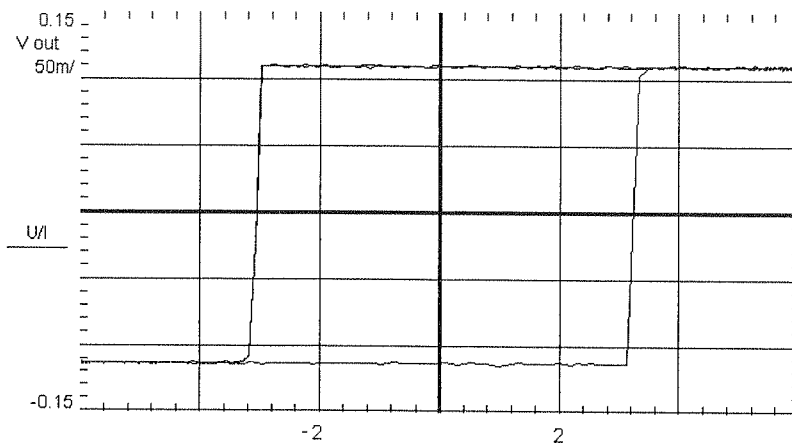
Rys. 14. Schemat układu pomiarowego (a) i zmierzona charakterystyka funkcji wyjścia komórki (b)

Fig. 14. Scheme of measurement set-up (a) and measured characteristic of cell output function (b)

Układ pomiarowy nieliniowej charakterystyki przejściowej komórki bez sprzężenia zwrotnego oraz wynik pomiaru przedstawione zostały odpowiednio na rys. 14a i rys. 14b. Układ pomiarowy nieliniowej charakterystyki przejściowej komórki z włączonym sprzężeniem zwrotnym oraz wynik pomiaru przedstawione zostały odpowiednio na rys. 15a i rys. 15b.



(a)



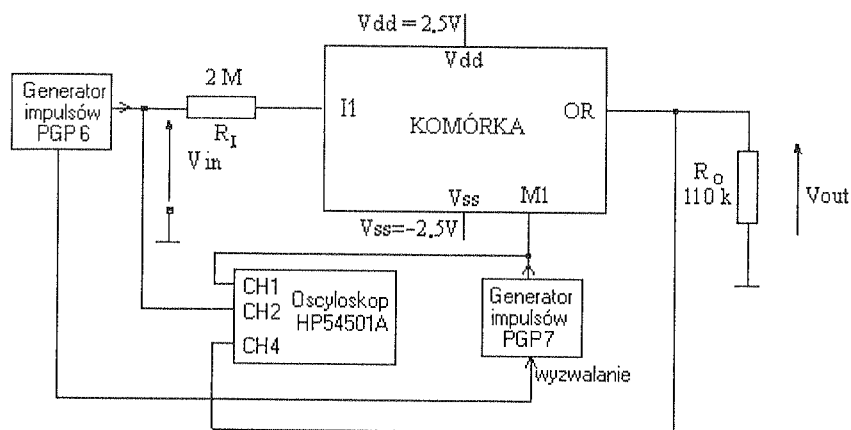
(b)

Rys. 15. Układ pomiarowy (a) i uzyskana charakterystyka, pozwalająca na oszacowanie współczynnika sprzężenia zwrotnego komórki (b)

Fig. 15. Measurement set-up (a) and measured characteristic, which allows for estimating value of cell feedback coefficient

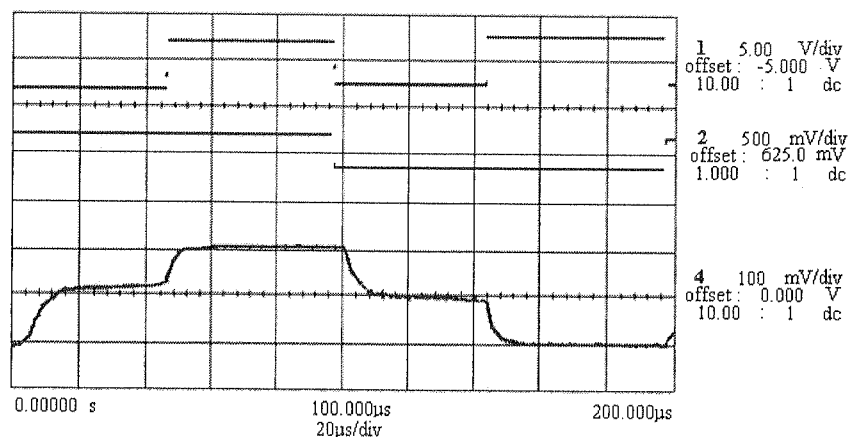
tego oraz
miarowy
raz wynik

Kolejnym testem była obserwacja oscyloskopowa stanów przejściowych w pojedynczej komórce prądowej "full range" o schemacie ideowym z rys. 12. Schemat pomiarowy oraz rezultaty tych pomiarów przedstawia rys. 16.



(a)

Δp running



(b)

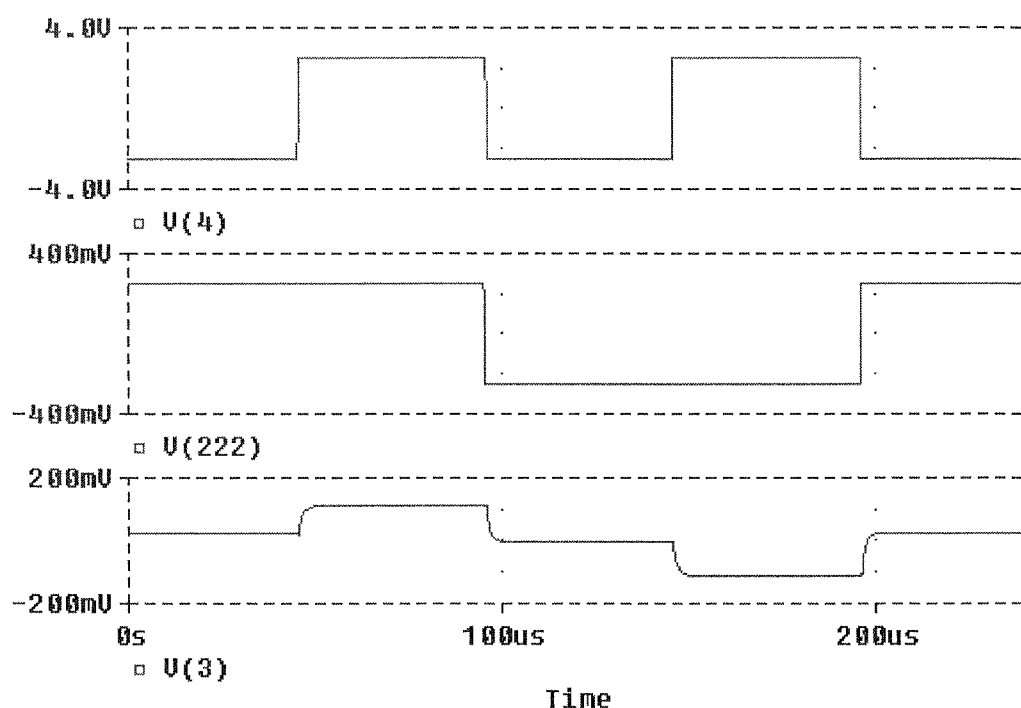
zynnika

e of cell

Rys. 16. Pomiary oscyloskopowe stanów przejściowych w komórce prądowej „full range”, schemat pomiarowy (a) i rezultaty pomiaru (b)

Fig. 16. Oscilloscope observations of transient responses in current-mode „full range” CNN cell, scheme of measurement set-up (a) and measurement results (b)

Na rys. 16b przebieg 1 prezentuje sygnał na wejściu *M1* komórki (rys. 12) sterujący rodzajem jej pracy, mianowicie niski poziom ustawia komórkę w stan wprowadzania informacji (załadowanie obrazu wejściowego), zaś wysoki stan inicjuje przetwarzanie informacji. Przebieg 2 jest sygnałem napięciowym, który pobudza przez opornik R_i wejście *I1* komórki inicjując warunki początkowe w komórce (obraz wejściowy). Przebieg 4 jest sygnałem napięciowym na wyjściu *OR* komórki obciążonym rezystancją $110\text{k}\Omega$ i odzwierciedla on przebieg czasowy zmiennej stanu w komórce. Analizując te przebiegi można stwierdzić, że komórka binaryzuje obraz wejściowy, tzn. spełnia zaprojektowane zadanie. Jeżeli jest inicjowany słaby dodatni sygnał wejściowy (obraz wejściowy) to wynikiem przetwarzania jest wysoki binarny stan na wyjściu komórki, zaś jeżeli jest inicjowany słaby ujemny sygnał wejściowy to wynikiem przetwarzania jest niski binarny stan na wyjściu komórki. Analogiczne wyniki symulacji komputerowej przedstawia rys. 17.



Rys. 17. Wyniki symulacji stanów przejściowych w komórce prądowej „full range”

Fig. 17. Simulation results of transient responses in current-mode „full range” CNN cell

4. ZMODYFIKOWANY MODEL KOMÓRKI SNK „FULL RANGE” [20, 21]

Przedstawiony w punkcie 3 model układowy komórki SNK posiada szereg wad, które uniemożliwiają jego zastosowanie w przypadku SNK wymagających niezmienności stałej czasowej procesów przejściowych oraz ograniczenia dynamiki zmiennej stanu

w zakr
połącze
wadą j
Nielini
wać z
zastos
nich p
szablon
W
"full ra
dynam
zasadn
Prz
i zwrac

V_C

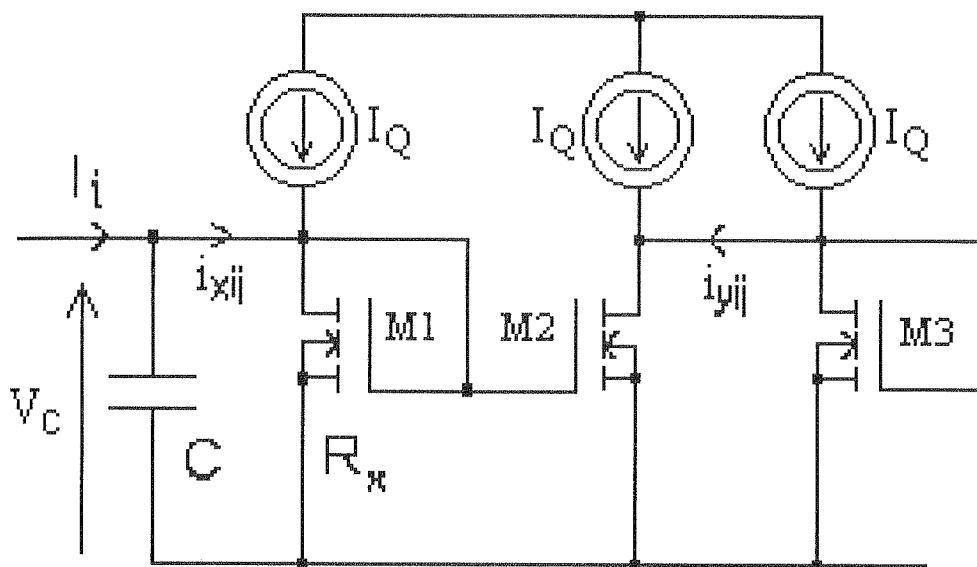
1. Wid
napięci
nie od
a jedyn
od I_Q .
zmienn
która j

sterujący
wzrostania
warzanie
wejście
eg 4 jest
odzwier-
gi można
zadanie.
ynikiem
ny słaby
wyjściu

w zakresie $-I_Q \leq i_{xij} \leq +I_Q$. Ponieważ w tym modelu (rys. 11) opornik R_x jest realizowany jako połączenie „diodowe” tranzystora MOS, powoduje to nieliniowość stałej czasowej komórki. Drugą wadą jest brak ograniczenia prądu i_{xij} od strony jego wartości dodatnich na poziomie $+I_Q$. Nieliniowość wynikająca z zależności stałej czasowej od prądu i_{xij} może powodować zniekształcenia lub utratę informacji w stanach przejściowych SNK, jeżeli w danym zastosowaniu taka jest istotna, zaś brak ograniczenia prądu stanu i_{xij} od strony wartości dodatnich powoduje, że nie jest spełniony warunek (42) przy pracy sekwencyjnej SNK z wieloma szablonami.

W niniejszym punkcie postawiono sobie za cel próbę optymalizacji układowej komórki „full range” z punktu widzenia niezmienności jej stałej czasowej i dokładnego ograniczenia dynamiki zmiennej stanu komórki w zakresie $-I_Q \leq i_{xij} \leq +I_Q$. Poniżej przedstawiono zasadnicze wyniki tej optymalizacji.

Przeanalizujemy działanie układu komórki „full range” posługując się rysunkiem 18 i zwracając uwagę na jego wady.



Rys. 18. Model komórki do analizy wad rozwiązania układowego „full range”

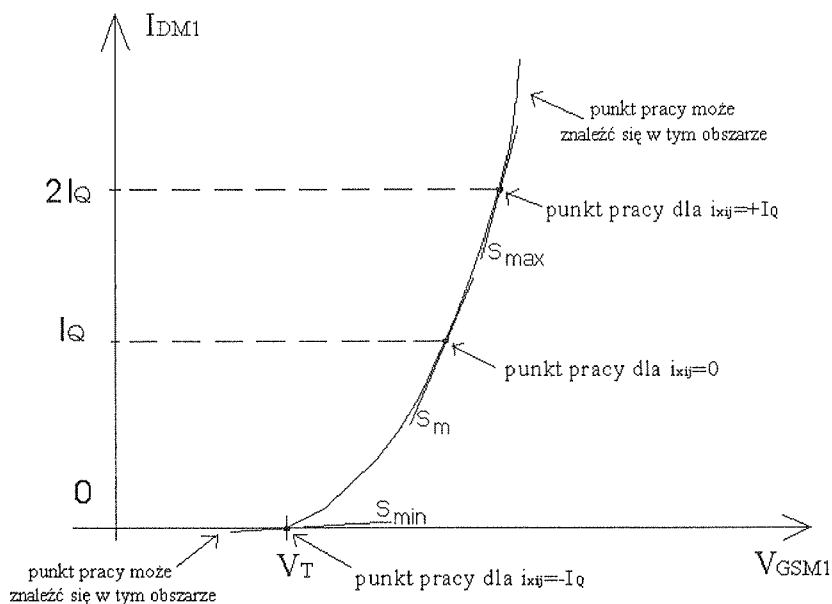
Fig. 18. CNN cell model used for analysis of drawbacks of „full range” cell realization

1. Widać z rys. 18, że prąd i_{xij} odzwierciedla jednoznacznie stan integratora dla każdego napięcia stanu V_C . Niech $I_i > I_Q$. Wtedy w stanie ustalonym $i_{xij} > I_Q$, ale $i_{yij} = I_Q$. Zatem prąd i_{yij} nie odwzorowuje wzajemnie jednoznacznie stanu integratora dla każdego prądu i_{xij} a jedynie w przedziale $-I_Q \leq i_{xij} \leq +I_Q$. Wiadomo jednak, że w SNK prąd i_{xij} może być większy od I_Q . Ten brak wzajemnie jednoznacznej relacji w całym przedziale pomiędzy faktyczną zmienną stanu, którą jest napięcie na kondensatorze C a zmienną wyjściową i_{yij} jest wadą, która jest szczególnie istotna w tych zastosowaniach sieci, w których ważny jest proces

przetwarzania informacji w stanach przejściowych a nie tylko wynik przetwarzania w stanie ustalonym. Przykładem takiej SNK jest opisana w dalszej części tego rozdziału sieć służąca do estymacji powierzchni obiektów w obrazie (patrz rys. 23). Konsekwencją braku ograniczenia prądu i_{xij} od strony wartości dodatnich na poziomie $+I_Q$ jest także brak spełnienia warunku (42) przy pracy sekwencyjnej SNK z wieloma szablonami.

2. Rezystor R_x jest silnie nieliniowy, ponieważ jest nim tranzystor $M1$ w tzw. połączeniu "diodowym" z polaryzacją I_Q , rys. 19. Punkt pracy może znaleźć się w obszarze, gdzie

$V_{GSM1} < V_T$ jak i gdzie $V_{GSM1} > V_T + \sqrt{\frac{4I_Q}{\beta_{M1}}}$. Zdefiniujemy następująco nieliniowość N charakterystyki rezystora R_x :



Rys. 19. Charakterystyka tranzystora MOS w połączeniu „diodowym” jako rezystora R_x

Fig. 19. Characteristic of diode connected MOS transistor working as resistor R_x

$$N = 100\% (S_{max} - S_{min})/S_m \quad (44)$$

przy czym S_{max} jest maksymalnym nachyleniem charakterystyki, S_{min} jest minimalnym nachyleniem charakterystyki natomiast S_m jest środkowym nachyleniem. Obliczymy ten współczynnik stosując model *Level 1* tranzystora MOS. Nachylenie charakterystyki jest równe $g_m = \beta_{M1} (V_{GSM1} - V_T) = \sqrt{2I_D \beta_{M1}}$. Ograniczymy zakres zmienności prądu do $0 \leq I_D \leq 2 I_Q$, zatem nieliniowość według wzoru (44):

$$N = 100\% [g_m(2I_Q) - g_m(0)]/g_m(I_Q) = 100\% \sqrt{\frac{4I_Q}{2I_Q}} \cong 1.41 \times 100\% = 141\%$$

Tak duża
zwłaszcza
sach prz
powierz
proporcj
długość.
wyjściow
sieci na
Aby roz
1. zapev
pojemno
faktyczn
2. należ
 $\leq i_{xij} \leq +$
„full ran
szablon
3. należ
realizacj
Proponu
 i_{xij} oraz
relacja
z polary
paramet

$\ln[i_{ij}] + 1$

Równan
niż (37)

Tak duża nieliniowość opornika R_x dla wielu realizacji może być nie do przyjęcia, zwłaszcza w tych zastosowaniach, w których interesuje nas informacja zawarta w procesach przejściowych. Przykładem takiego zastosowania jest SNK służąca do estymacji powierzchni obiektów w obrazie, gdzie impuls, którego czas trwania powinien być proporcjonalny do długości obiektu w linii obrazu, propagując przez SNK skraca swoją długość. Zjawisko to powoduje brak liniowej zależności pomiędzy czasem trwania impulsu wyjściowego SNK a długością obiektu w linii obrazowej oraz zmienniczość odpowiedzi sieci na położenie obiektu w obrazie.

1. zapewnić, aby zachodziła wzajemnie jednoznaczna relacja pomiędzy napięciem na pojemności C a prądem stanu i_{xij} , ponieważ prąd „stanu” i_{xij} powinien odzwierciedlać faktyczną zmienną stanu jaką jest napięcie na kondensatorze C ,
2. należy ograniczyć w układach CMOS zakres zmian zmiennej stanu i_{xij} do przedziału $-I_Q \leq i_{xij} \leq +I_Q$, aby spełnić dokładnie wszystkie wymagania założeń modelu matematycznego „full range” i zapewnić spełnienie warunku (42) przy pracy sekwencyjnej SNK z wieloma szablonami,

3. należy zmniejszyć nieliniowość N do wielkości, która jest dopuszczalna dla danej realizacji SNK tak, aby uzyskać minimalne błędy odpowiedzi SNK.

Proponuje się rozwiązanie przedstawione na rysunku 20, w którym prąd stanu integratora i_{xij} oraz i_{yij} jest ograniczony do zakresu $-I_Q \leq i_{xij} \leq +I_Q$ i istnieje wzajemnie jednoznaczna relacja pomiędzy V_C a i_{yij} . Rezystor R_x tworzy połączenie diodowe tranzystora $M5$ z polaryzacją prądową mI_Q , rys. 20. Łatwo wykazać, że nieliniowość rezystora nie zależy od parametrów tranzystora ani doboru I_Q a jedynie od wartości współczynnika m :



Rys. 20. Zmodyfikowana komórka „full range”

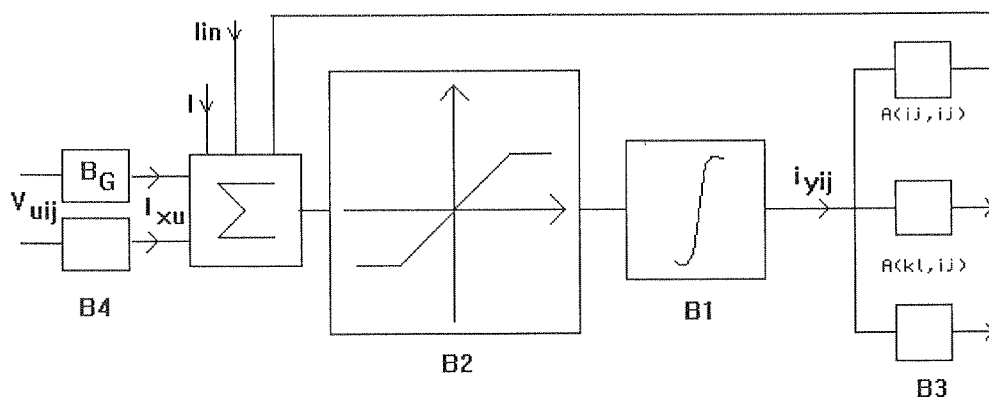
Fig. 20. Modified „full range” CNN cell

$$N(m) = 100\% (\sqrt{m+1} - \sqrt{m-1})/\sqrt{m} \quad (45)$$

Równanie matematyczne stanu tak zmodyfikowanej komórki "full range" jest nieco inne niż (37) i jest opisane wzorem (46).

$$-g[\tau(di_{yij}/dt) + i_{yij}] = I + \sum_{C(k,l) \in Nr(i,l)} [A_c(ij, kl)i_{ykl} + B_G(ij, kl)v_{ukl}] \quad (46)$$

gdzie funkcja g jest opisana wzorem (38). Na rys. 21 przedstawiono schemat funkcjonalny blokowy realizujący równanie dynamiki zmodyfikowanej komórki "full range". W sumatorze Σ są sumowane: prąd polaryzacji I , prądy z wyjść komórek sąsiednich I_{in} oraz prądy I_{xu} z przetworników transkonduktancyjnych B_G związanych z szablonem B . Sygnał wyjściowy tego bloku jest poddawany nieliniowej transformacji w bloku $B2$. Z kolei prąd po tej transformacji jest poddany całkowaniu w integratorze strasnym $B1$, a następnie skalowany w blokach $B3$ przez wymnożenie przez wartości szablonu A , w wyniku czego na każdym wyjściu otrzymujemy odpowiedni prąd i_{yij} dołączany do wejść sumacyjnych komórek sąsiednich. W blokach $B4$ następuje skalowanie napięć wejściowych V_{uij} (przez przemnożenie przez wartości szablonu B_G) z komórek sąsiednich dając w wyniku prądy I_{xu} . Bloki $B1$, $B2$ i $B3$ są konwerterami prądowo-prądowymi, natomiast bloki $B4$ są konwerterami napięciowo-prądowymi.



Rys. 21. Schemat funkcjonalny blokowy zmodyfikowanej komórki „full range”

Fig. 21. Functional scheme of modified „full range” CNN cell

Na rysunku 22 przedstawiono porównanie wykresów odpowiedzi czasowych komórki "full range" z rys. 11 oraz komórki zmodyfikowanej "full range" z rys. 20 z wybranym współczynnikiem $m=5$. Wyraźnie jest widoczne polepszenie liniowości odpowiedzi przejściowej dla przypadku zmodyfikowanej komórki "full range", co świadczy o zmniejszeniu zależności stałej czasowej od zmian sygnałów. Dla przypadku komórki "full range" stała czasowa znacząco zwiększa się w miarę zmniejszania wartości prądu stanu. Dla potwierdzenia dobrych własności liniowych zmodyfikowanej komórki "full range" wykonano symulacje w programie *HSPICE* na komputerze *SUN workstation SPARC 10* jednowymiarowej *SNK* przeznaczonej do estymacji powierzchni obiektów w obrazie złożonej ze 100 komórek ($N=100$), której struktura jest przedstawiona na rys. 23 [10, 11, 12, 20-23]. Zadaniem takiej sieci jest generacja impulsów o czasie trwania proporcjonalnym do długości obiektów w obrazie. Impulsy wyjściowe z tej *SNK* są pobierane z wyjścia komórki o numerze N . Dodatkowym wymaganiem jest niezależność czasu trwania impulsu od lokalizacji obiektu o takiej samej długości w obrazie. Symulacje

SNK
zastos
"full r
wanej
26 prz
propo
przypa
komór
i sieci
długo
zmod
odpow
optym
do au

100

-8

-10

Rys.

Fig. 22

jonalny
W suma-
prądy I_{xu}
yjściowy
d po tej
alowany
każdem
komórek
ez prze-
prądy I_{xu}
i B4 są

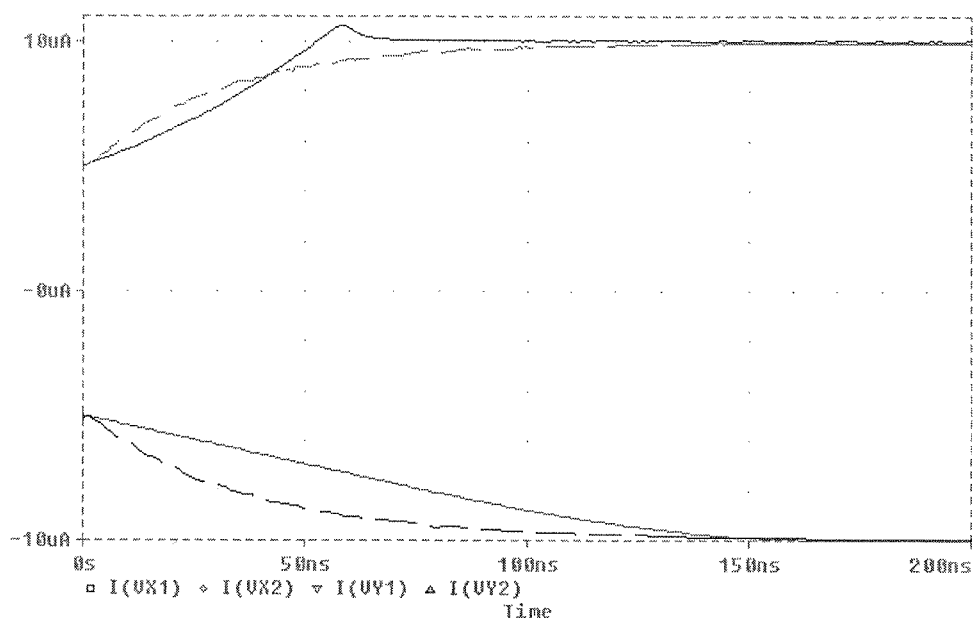
→
(, i, j)

→
(, i, j)

→
B3

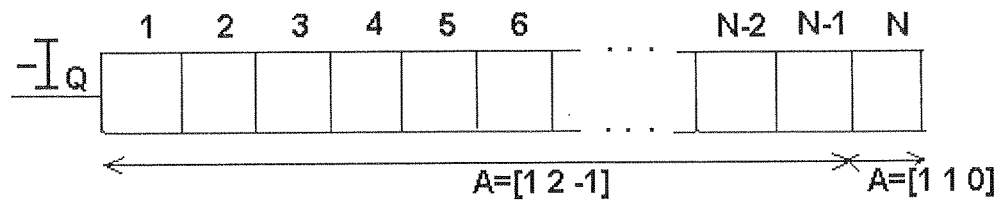
asowych
z rys. 20
niowości
nge", co
przypadku
wartości
komórki
rkstation
obiektyw
na rys. 23
e trwania
SNK są
zależność
umulacje

SNK do estymacji powierzchni obiektów w obrazie wykonano dla dwóch przypadków zastosowanej do realizacji sieci komórki. Mianowicie dla pierwszego przypadku komórki "full range", której schemat przedstawia rys. 24 oraz dla drugiego przypadku zmodyfikowanej komórki "full range" o schemacie ideowym zaprezentowanym na rys. 25. Rysunek 26 przedstawia wyniki symulacji czasu trwania impulsu na końcu łańcucha sieci, który jest proporcjonalny do długości obiektu liczonej w liczbach pikseli (np. czarnych) tylko dla przypadku SNK do estymacji obiektów w obrazie zbudowanej w oparciu o zmodyfikowaną komórkę "full range". Przedstawiono wyniki dla sieci złożonej z komórek "full range" i sieci złożonej z komórek zmodyfikowanych. Liniowa zależność czasu trwania impulsu od długości obiektu jest dowodem dużej liniowości sieci zbudowanej z użyciem komórek zmodyfikowanych "full range". Podobne badania potwierdziły również niezmienniczość odpowiedzi na położenie obiektu w łańcuchu 100 komórek zmodyfikowanych. Wyniki optymalizacji stanowiły bazę wyjściową do opracowania projektu układu scalonego SNK do automatycznej estymacji powierzchni obiektów w obrazie [12].



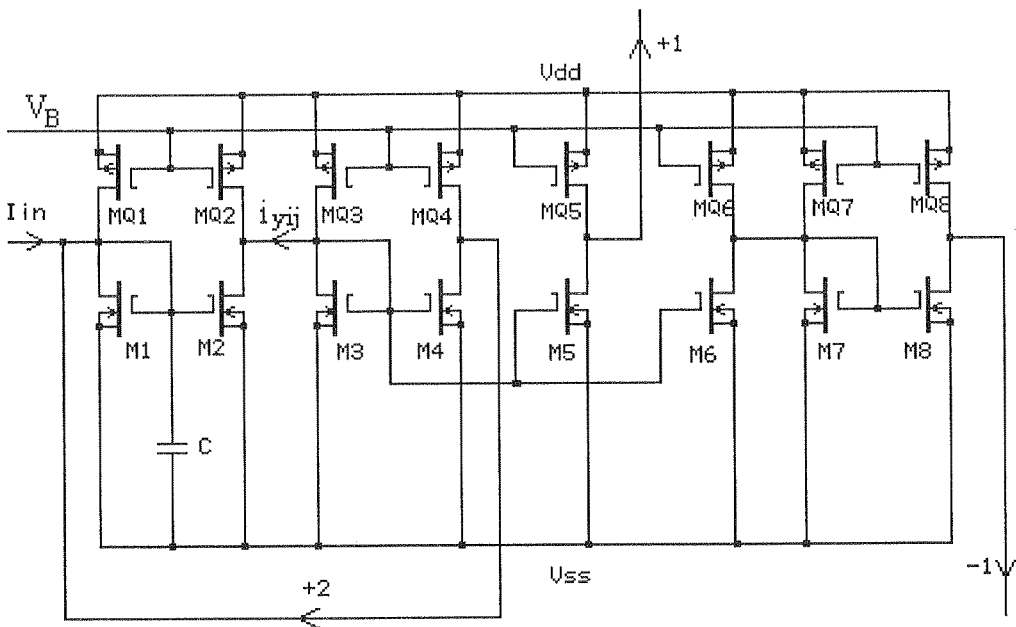
Rys. 22. Odpowiedzi czasowe komórki "full range" (linia ciągła) i komórki zmodyfikowanej "full range" (linia przerywana)

Fig. 22. Transient response of "full range" CNN cell (solid line) and transient response of modified "full range" CNN cell (dashed line)



Rys. 23. Struktura SNK do estymacji powierzchni obiektów w obrazie

Fig. 23. Structure of CNN image object area estimator



Rys. 24. Schemat ideowy komórki "full range" zastosowanej do symulacji SNK estymatora powierzchni obiektów w obrazie

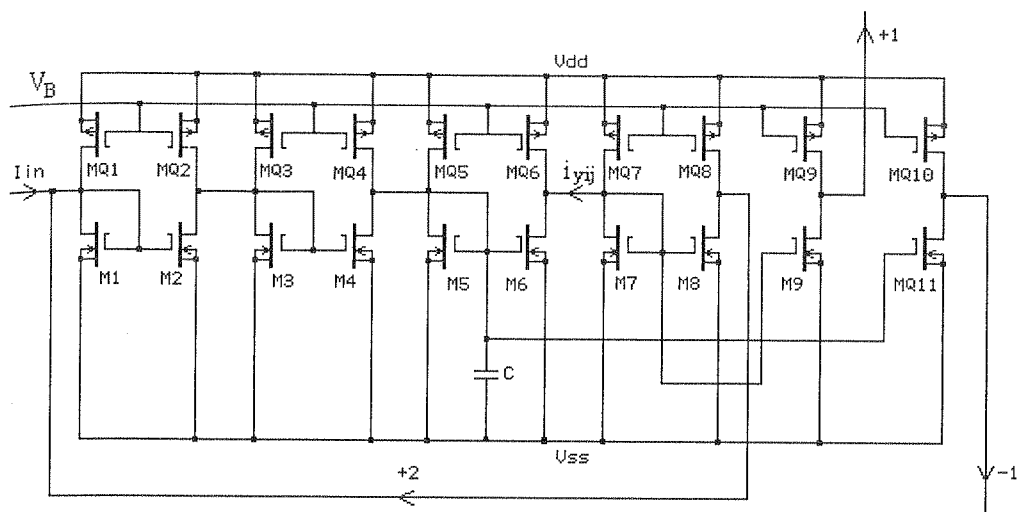
Fig. 24. Scheme of the "full range" CNN image object area estimator cell



Rys. 25

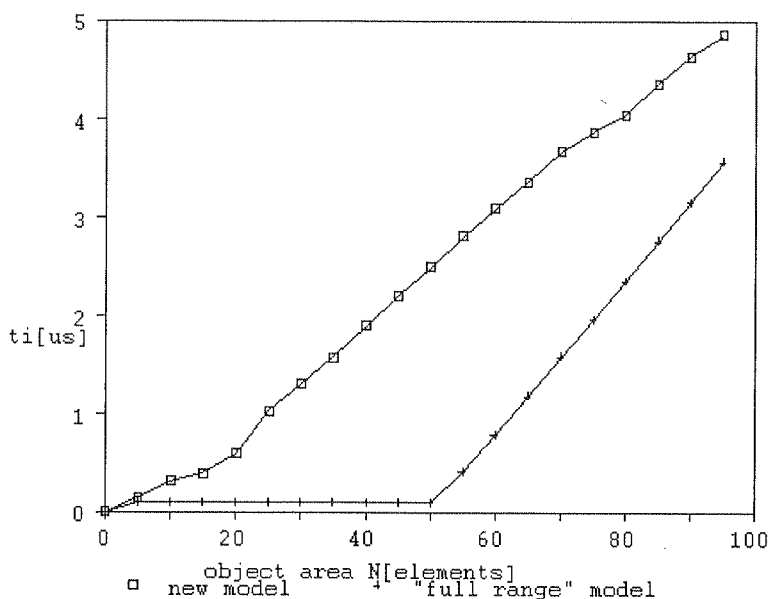
Rys. 26

Fig. 26.
estim



Rys. 25. Schemat ideowy zmodyfikowanej komórki "full range" zastosowanej do symulacji SNK estymatora powierzchni obiektów w obrazie

Fig. 25. Scheme of the modified "full range" CNN image object area estimator cell



Rys. 26. Wykres czasu trwania impulsu w funkcji długości obiektu, SNK z komórką "full range" (krzyżyki), SNK ze zmodyfikowaną komórką "full range" (kwadraciki)

Fig. 26. A plot of a pulse duration versus a connected component length in CNN system for image object area estimation — computer simulation results. The squares — modified model, the crosses — original "full range" model

5. PODSUMOWANIE

Przedstawione w artykule wyniki badań potwierdzają przydatność układów pracujących w trybie prądowym do realizacji sprzętowej sieci neuronowych komórkowych. Komórka neuronowa pracująca w trybie prądowym zajmuje mniejszą powierzchnię płytki krzemowej w porównaniu z komórką neuronową pracującą w trybie napięciowym. Szczególnie dużą oszczędnością powierzchni krzemu charakteryzują się realizacje sieci komórkowych neuronowych wykorzystujące model prądowy „full range”. W tym przypadku, gdy nie zależy na dobrej liniowości stałej czasowej możemy uzyskać największą oszczędność powierzchni krzemu, gdyż liczba tranzystorów przypadająca na pojedynczą komórkę jest mała oraz nie trzeba stosować tranzystorów o szerokich kanałach w stopniu integratora, aby zapewniać swobodną dynamikę zmian zmiennej stanu. Ponieważ w przypadku realizacji VLSI sieci neuronowych komórkowych liczba komórek, a co za tym idzie rozdzielczość przetwarzanego obrazu, zależy od stopnia złożoności układu elektronicznego realizującego model komórki, więc przy niskim stopniu złożoności komórki pracującej w trybie prądowym można osiągać najwyższą rozdzielczość przetwarzanego obrazu za pomocą pojedynczego układu scalonego SNK.

W przypadku, gdy wymagana jest duża liniowość stałej czasowej i jeżeli sieć neuronowa komórkowa ma pracować sekwencyjnie z kilkoma szablonami przełączanymi skokowo jeden po drugim po zakończeniu przetwarzania informacji w danej sekwencji, można zastosować realizację SNK wykorzystującą zmodyfikowany model „full range”.

Ponadto realizację SNK w oparciu o technikę prądową charakteryzuje mały pobór mocy. Nadają się one do wykorzystania w układach z pojedynczym zasilaniem o obniżonym napięciu i nie wymagają symetrycznego zasilania. Aby zapewnić możliwość programowania szablonu SNK pracującej w trybie prądowym można wykorzystać układy programowalnych zwierciadeł prądowych [34] i układ programowalnego wzmacniacza prądowego [28, 30-32].

6. BIBLIOGRAFIA

1. T. Sakurai, A. R. Newton: *A Simple MOSFET Model for Circuit Analysis*. IEEE Trans. on Electron Devices, vol. 38, no. 4, April 1991, pp. 887-893.
2. L. O. Chua, L. Yang: *Cellular Neural Networks: Theory and Applications*. IEEE Trans. Circuits and Systems, CAS-35, October 1988, pp. 1257-1290.
3. J. M. Cruz, L. O. Chua: *A CNN Chip for Connected Component Detection*. IEEE Transactions on Circuits and Systems, vol. 38, no. 7, July 1991, pp.812-817.
4. R. Dominguez-Castro, S. Espejo, A. Rodriguez-Vazquez, R. Carmona: *A CNN Universal Chip in CMOS Technology*. Proc. of CNNA-94, 1994, pp.91-97.
5. S. Espejo, A. Rodriguez-Vazquez, J.L. Huertas: *Design and Testing Issues in Current-Mode Cellular Neural Networks*. Second International Workshop CNNA'92, Monachium, Niemcy, listopad 1992, pp. 169-174.
6. J.L. Huertas, A. Rodriguez-Vazquez, S. Espejo: *Analog VLSI Implementation of Cellular Neural Networks*. Second International Workshop CNNA'92, Monachium, Niemcy, listopad 1992, pp. 141-150.

7. T. 199
8. T. 1 Tra
9. A. S a Tim Pro
10. K. Net pp.
11. K. Co
12. K. S siec ukł
13. V a Cur Ana
14. K.
15. J. M. Ele
16. T. K i m
17. K. S prze
18. F. 2 Tra
19. T. K app paż
20. J. K cell
21. J. K obje ss. 4
22. J. K netw Net
23. K. Ś circ 90-9
24. K. Ś lma Zgo
25. K. Ś leve 311-
26. K. Ś real 30.0

7. T. Kacprzak, K. Ślot: *Sieci neuronowe komórkowe*. Wydawnictwo Naukowe PWN, Warszawa-Łódź 1995.
8. T. Matsumoto, O. Chua, H. Suzuki: *CNN Cloning Template: Connected Component Detector*. IEEE Transactions on Circuits and Systems, vol. 37, no. 5, May 1990, pp.633-635.
9. A. Rodriguez-Vazquez, S. Espejo, R. Dominguez-Castro, J. L. Huertas E. Sanchez-Sinencio: *Current-Mode Techniques for the Implementation of Continuous- and Discrete-Time Cellular Neural Networks*. IEEE Trans. on Circuits and Systems - II. Analog and Digital Signal Processing, vol. 40, no.3, March 1993, pp.132-146.
10. K. Ślot: *A Histogram Calculating and Connected Component Detecting Pulse-Mode Cellular Neural Network*. Proc. of European Conf. on Circuit Theory and Design 1993, Davos, Switzerland, September, 1993, pp.629-637.
11. K. Ślot: *A Cellular Neural Network for Image Objects Area Estimation*. Proc. of the Advanced Training Course: Mixed Design of VLSI Circuits, Dębe, Poland, 5-9 April 1994, pp.72-77.
12. K. Ślot: Raport merytoryczny z realizacji projektu badawczego KBN Nr 0496/S5/94/07, *Reprogramowalna sieć neuronowa komórkowa dla celów specjalizowanej analizy obrazów: projekt i realizacja fizyczna w formie układu scalonego wielkiej skali integracji*. Łódź, czerwiec 1996.
13. Varrientos J.E. Varrientos, E. Sanchez-Sinencio, J., A. Ramirez-Angulo: *Current-Mode Cellular Neural Network Implementation*. IEEE Transactions on Circuits and Systems-II: Analog and Digital Processing, vol. 40, no. 3, March 1993, pp.147-155.
14. K. Wawryn: *Układy z przełączanymi prądami*. Warszawa, WNT 1997.
15. J. M. Cruz, L. O. Chua: *High-Speed High-Density CMOS CNNs*. Memorandum No. UCB/ERL M91/28, Electronics Research Laboratory, College of Engineering, University of California, Berkeley, 19 April 1991.
16. T. Kacprzak: Raport merytoryczny z realizacji projektu badawczego KBN Nr 3 3510 92 03, *Rozwój teorii i metod projektowania sieci neuronopodobnych*. Łódź, lipiec 1994.
17. K. Ślot: *Analiza projektowa i synteza jednowarstwowych komórkowych sieci neuronopodobnych dla celów przetwarzania obrazów*. Rozprawa Doktorska, Politechnika Łódzka, Instytut Elektroniki, Łódź, 1993.
18. F. Zou, J. A. Nossek: *Stability of Cellular Neural Networks with Opposite-Sign Templates*. IEEE Transactions on Circuits and Systems, vol. 38, no. 6, June 1991, pp.675-677.
19. T. Kacprzak, J. Kowalski: *Design methodology of cellular neural networks in CMOS current-mode approach*. XVI Krajowa Konferencja Teoria Obwodów i Układy Elektroniczne, Kołobrzeg, ss. 89-95, październik 1993.
20. J. Kowalski, T. Kacprzak, K. Ślot: *A reprogrammable cell circuit for an image area estimating cellular neural network*. Workshop Mixed Design of VLSI Circuits, Dębe, Poland, pp. 88-93, kwiecień 1994.
21. J. Kowalski, T. Kacprzak, K. Ślot: *VLSI design problems of cellular neural network for image object area estimation*. XVII Krajowa Konferencja Teoria Obwodów i Układy Elektroniczne, Polanica Zdrój, ss. 471-476, październik 1994.
22. J. Kowalski, K. Ślot, T. Kacprzak: *A CMOS current-mode VLSI implementation of cellular neural network for an image objects area estimation*. Third IEEE International Workshop on Cellular Neural Networks and their Applications, CNNA-94, Rome, p.351, grudzień 1994.
23. K. Ślot, J. Kowalski, J. Pacholik: *An object area estimating cellular neural network based VLSI circuit: an overview of circuit analog part*. Workshop Mixed Design of VLSI Circuits, Kraków, Poland, pp. 90-95, maj 1995.
24. K. Ślot, J. Kowalski, J. Pacholik, P. Dębiec: *VLSI Integrated Circuit for Realizing Selected Image Processing Operations*. XVIII Krajowa Konferencja Teoria Obwodów i Układy Elektroniczne, Polana Zgorzelisko, 25-27 X 1995, ss.305-310.
25. K. Ślot, J. Kowalski, J. Pacholik, P. Dębiec: *VLSI integrated circuit for realizing selected gray level image analyses: circuits test results*. Workshop Mixed Design of VLSI Circuits, Łódź, Poland, pp. 311-316, 30.05-01.06.1996.
26. K. Ślot, J. Kowalski, J. Pacholik, P. Dębiec: *Cellular neural network based integrated circuit for realizing selected image analyses*. II Konferencja Sieci Neuronowe i ich Zastosowania, Szczyrk, ss. 477-482, 30.04-04.05.1996.

27. K. Ślot, J. Kowalski, J. Pacholik, P. Dębiec: *Cellular Neural Network Based VLSI Architecture for Image Processing*. Proc. of 4-th IEEE Conf. CNNA-96, Seville, Spain, June 24-26, 1996, pp. 249-254.
28. J. Kowalski, T. Kacprzak: *Continuously Adjustable MOSFET Current Amplifier*. Proceedings of the 1997 European Conference on Circuits Theory and Design, Budapest, Hungary, 30.08-3.09.1997, pp. 196-199.
29. T. Kacprzak, J. Kowalski: Raport naukowy z realizacji projektu badawczego KBN Nr 8 T11B 059 11, *Analiza projektowa reprogramowalnych sieci neuronowych komórkowych w technice półprzewodnikowych układów analogowych CMOS wielkiej skali integracji*. Łódź, grudzień 1997.
30. K. Ślot, J. Kowalski, P. Dębiec, C. Bolek, J. Pacholik: *Programmable Parallel Image FIR Filter Based on The Cellular Neural Network Paradigm: Circuit Test Results*. Workshop Mixed Design of VLSI Circuits, Łódź, Poland, June 18-20, 1998, pp. 369-374.
31. J. Kowalski, T. Kacprzak: *Continuously Gain Programmable CMOS Current Amplifier: Results of Circuit Measurements*. XXII National Conference on Circuit Theory and Electronic Networks - KKTOiUE, Warszawa — Stare Jabłonki, Poland, Proceedings, vol. 2, October 20-23, 1999, pp. 343-348.
32. T. Kacprzak, J. Kowalski: *Cellular Neural Network Implementation Based on Continuously Programmable Gain CMOS Current Amplifier*. Kwartalnik Elektroniki i Telekomunikacji, T. 46, Z. 3, Warszawa PWN, 2000, ss. 321-337.
33. J. Kowalski, T. Kacprzak: *Analiza i projektowanie sieci neuronowych komórkowych realizowanych w technice układowej wzmacniaczy transkonduktancyjnych jedno- i wielowejściowych*. Kwartalnik Elektroniki i Telekomunikacji, T. 47, Z. 1, Warszawa, Wydawnictwo Naukowe PWN, 2001, ss. 87-119.
34. J. Kowalski: *Układy CMOS programowalnych zwierciadeł prądowych*. Wydział Elektrotechniki i Elektroniki Politechniki Łódzkiej Elektronika — Prace Naukowe, Z. Nr 6, Łódź 2001, ss. 114-134.
35. S. Espejo, A. Rodriguez-Vazquez, R. Carmona, P. Foldesy, A. Zarandy, P. Szolgay, T. Sziranyi, T. Roska: *A 0.8 μ m CMOS Programmable Mixed-Signal Focal-Plane Array Processor with On-Chip Binary Imaging and Instructions Storage*. IEEE Int. Journal of Solid State Circuits, Vol. 32, No. 7, 1997, pp. 1013-1026.
36. A. Passio, A. Kananen, K. Halonen, V. Porra: *A 48 by 48 CNN Chip Operating with B/W Images*. IEEE International Conference on Electronics, Circuits and Systems, Lisboa, Portugal, 7-10 September, 1998, pp. 191-194.
37. S. Espejo, R. Dominguez-Castro, G. Linan, A. Rodriguez-Vazquez: *A 64 x 64 CNN Universal Chip with Analog and Digital I/O*. IEEE International Conference on Electronics, Circuits and Systems, Lisboa, Portugal, 7-10 September, 1998, pp. 203-206.
38. G. Linan, S. Espejo, R. Dominguez-Castro, A. Rodriguez-Vazquez: *The CNNUC3: an Analog I/O 64 x 64 CNN Universal Machine Chip Prototype with 7-bit Analog Accuracy*. 6-th IEEE International Workshop on Cellular Neural Networks and Their Applications, Proceedings, Catania, Italy, May 23-25, 2000, pp. 201-206.

J. KOWALSKI

ANALYSIS AND DESIGNING OF CELLULAR NEURAL NETWORKS BASED ON CURRENT-MODE CMOS CIRCUITS

Summary

In this article an analysis and designing method proposal of Cellular Neural Networks (CNN) based on current-mode CMOS circuits are presented. A detailed analytical analysis is made using Shichman-Hodges and Sakurai-Newton MOS transistor model. A designing way is explained for a connected component detector CNN example. Designing results are checked by SPICE simulations using CMOS MIETEC 2.4 μ m and CMOS 0.25 μ m technology parameters. CMOS CNN full range cell model, its simulations and measurements results have been presented. The current-mode "full-range" concept of the cell circuit allows for minimize number of MOS

transistor
the time o
applicatio
in time. In
"full rang
of the inte
simulation
range CN
image ob

Keyword.

Architecture
p. 249-254.
Proceedings
-3.09.1997,

11B 059 11,
odnikowych

Image FIR
d Design of

r: Results if
KKTOiUE,

Continuously
T. 46, Z. 3,

alizowanych
nik Elektro-
19.

ktrotechniki
14-134.

r and y, P.
Plane Array
ate Circuits,

g with B/W
tugal, 7-10

64 x 64 CNN
Circuits and

e CNNUC3:
p. 6-th IEEE
atania, Italy,

transistors and allows for application of switched templates in time sequence. However, strong nonlinearities of the time constant of an integrator stage in CMOS full range CNN cell realization exclude this circuit from CNN application where the result of processing depends on the precision of signal waveforms during the state evolution in time. Investigations directed to overcome this problem have resulted in modified version of current-mode CNN "full range" cell. The main idea of the time constant nonlinearity reduction results from increasing the bias current of the integrator stage, which is now greater than the maximal level of state variable. A comparison of the computer simulation of the CNN system operation using two models of cell is presented. The conclusion is that modified full range CNN cell model is much more linear, therefore this concept is better suited for requirements of the CNN for image object area estimation.

Keywords: Cellular Neural Networks, CMOS VLSI circuits, ASIC, image processing.

N) based on
-Hodges and
etector CNN
IOS 0.25 μm
ts have been
ber of MOS

Qu
war
i os
tery
odp
rea
sys
wie

Sło

Kom

Warto z
multime
tym, to
obrazu te
bardziej
złożonoś
możliwo
one odp

Rekonfigurowalny kwantyzator wektorowy do kodowania obrazów w czasie rzeczywistym

PAWEŁ RUSSEK, KAZIMIERZ WIATR

*Katedra Elektroniki, Akademia Górniczo-Hutnicza w Krakowie
30-059 Kraków, al. Mickiewicza 30
russek/wiatr@uci.agh.edu.pl*

*Otrzymano 2002.05.09
Autoryzowano 2002.09.11*

Artykuł opisuje implementację kwantyzacji wektorowej typu PVQ (ang. Pyramid Vector Quantization) w układach FPGA. Algorytm kodowania symboli kwantyzatora został zmodyfikowany, aby ułatwić realizację sprzętową. Zaproponowano odpowiednią architekturę sprzętową i oszacowano zasoby sprzętowe konieczne do jej realizacji. Proponowana architektura charakteryzuje się dużą mocą obliczeniową. Szacowane zasoby sprzętowe służą do określenia odpowiedniej wielkości układu reprogramowalnego do rzeczywistej realizacji. Do przykładowej realizacji posłużyła rodzina Virtex firmy Xilinx. Zaproponowano również rekonfigurowany system kompresji obrazu, w którym rekonfiguracja kwantyzatora jest wykonywana dla kontroli wielkości strumienia danych i alokacji bitów w strukturze informacji wizyjnej.

Słowa kluczowe: przetwarzanie obrazów, kompresja obrazów, systemy czasu rzeczywistego, układy programowalne

1. WSTĘP

Kompresja obrazu wymaga systemów cyfrowych o dużej mocy obliczeniowej. Warto zaznaczyć, że chociaż dzisiejsze komputery osobiste wyposażone w funkcje multimedialne są w stanie przeprowadzić dekompresję sygnału wideo w czasie rzeczywistym, to jednak ciągle brakuje im wydajności, aby wykonać kompresję standardowego obrazu telewizyjnego lub kompresję/dekompresję obrazu wysokiej rozdzielczości. Do tych bardziej zaawansowanych zadań musi być użyty dedykowany sprzęt. Ze względu na swoją złożoność takie systemy są drogie. Ponadto, z powodu znacznego kosztu i braku możliwości użycia ich do realizacji innych zadań niż te do jakich je zaprojektowano, nie są one odpowiednie do szerokiego, domowego użytku.

Autorzy uważają, że systemy reprogramowalne [1] dają możliwość rozwiązania powyższego problemu kosztów. Chociaż są one obecnie również drogie (ze względu na ciągle jeszcze małą powszechność ich stosowania), to jednak mogą stanowić platformę obliczeniową dla dowolnego problemu ponieważ nie posiadają stałej struktury połączeń bramek logicznych, a to efektywnie równoważy ich stosunkowo znaczny koszt. Architektury logiki wewnętrznej tych układów można kształtować z taką samą łatwością z jaką komputery zmieniają wykonywany program. W przeciwieństwie do architektury procesorowej nasza rekonfigurowalna architektura może tak zmieniać swoją strukturę, aby jak najlepiej odpowiadała zadaniom jakie w danej chwili wykonuje. Jednym z podejść do rekonfigurowalnego systemu jest rekonfigurowalny koprocesor. Podobnie jak klasyczny koprocesor arytmetyczny tak taki koprocesor rekonfigurowalny może wspomagać główny procesor w wykonywaniu zadań złożonych obliczeniowo. Oczywiście charakter tych zadań zależy od aktualnie załadowanej konfiguracji sprzętowej. Odpowiednia konfiguracja musi być jednak najpierw przygotowana i przechowywana w pamięci komputera gotowa do użycia.

Autorzy zdecydowali się zaproponować w FPGA realizację sprzętową algorytmu kwantyzacji wektorowej. Chociaż kwantyzacja wektorowa jest problemem złożonym obliczeniowo [5], to jednak jako bardziej wydajna od kwantyzacji skalarnej jest chętniej stosowana w kompresji obrazu. Sednem kompresji obrazu jest dopasowanie parametrów kwantyzacji do aktualnie kwantowanych danych. Parametryzacja zazwyczaj przeszkadza w realizacji architektury dedykowanej dla danego algorytmu ponieważ różne parametry mogą wymagać odmiennej architektury. Jednak nie odnosi się to do systemu, który w każdej chwili może być zrekonfigurowany tak, by odpowiadał aktualnie dobranym parametrom.

Rozważana kwantyzacja PVQ należy do klasy kwantyzatorów kratkowych. Parametry zastosowanej kratki wpływają na wielkość potrzebnych zasobów sprzętowych.

Dla danej wartości przepływności bitowej (ang. bitrate) można dobrać wiele zestawów parametrów, które ją spełniają. Zestawy te wymagają różnej wielkości zasobów sprzętowych, aby mogły zostać w nich zaimplementowane. Zestawy które wymagają więcej sprzętu zazwyczaj cechuje lepsza jakość kwantyzacji. Należy zatem znaleźć optimum, czyli takie parametry, które będą jak najlepsze przy zadanej wielkości zasobów sprzętowych. Klasycznie parametry kwantyzacji wynikają z założonego strumienia danych. W naszym podejściu dodatkowym czynnikiem jest wielkość hardware'u, w którym ma być ta kwantyzacja implementowana.

2. KWANTYZACJA WEKTOROWA PVQ

Kwantyzacja PVQ (ang. Pyramid Vector Quantization) jest odmianą kwantyzacji kratkowej, którą jako pierwszy zaproponował T.R. Fisher [4]. Ponieważ kwantyzacja PVQ cechuje się prostotą i wymaga sporo obliczeń, stosunkowo szybko zaproponowano rozwiązania sprzętowe w strukturach VLSI [7], [8], [11].

W
„piram

gdzie
nazywa

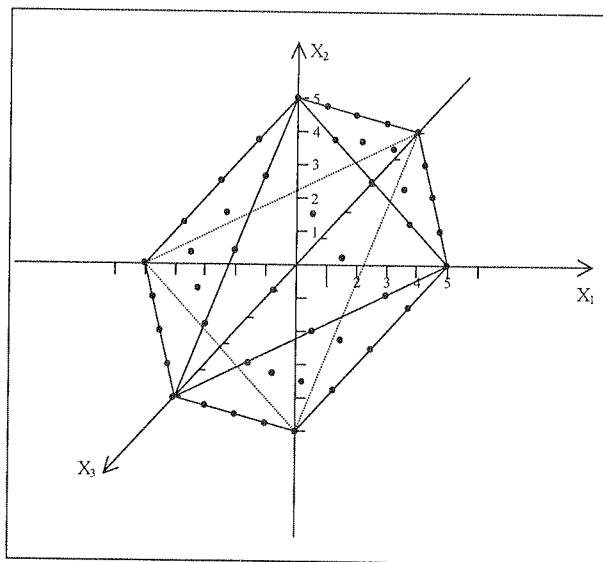
T. I
wartość
powierz
zbudow

Im v
wektora
wielkoś
poprawi
algorytm
kwantow
rzutowa
punkt kw

W kwantyzacji PVQ siatkę kwantyzacji tworzy się na powierzchni L wymiarowej „piramidy” (rys. 1). Powierzchnię tej piramidy opisuje równanie

$$\sum_{i=1}^L |x_i| = K \quad (1)$$

gdzie $X(x_1, x_2, \dots, x_L)$ jest wektorem w L wymiarowej przestrzeni, a K jest parametrem nazywanym promieniem piramidy.

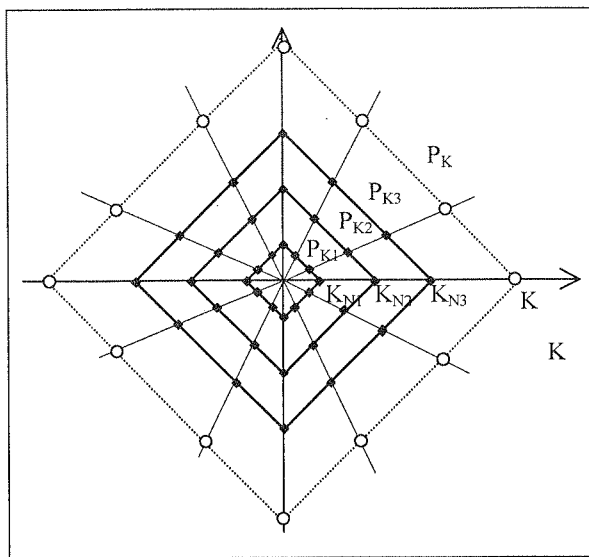


Rys. 1. Siatka kwantyzacji w 3 wymiarowej przestrzeni

Fig. 1. Quantization grid in 3-D space

T. R. Fisher pokazał, że dla źródła danych o rozkładzie Laplace'a, dla dużych wartości L , wektor $X(x_1, x_2, \dots, x_L)$ z dużym prawdopodobieństwem leży w pobliżu powierzchni piramidy. Tak zatem dla takiego źródła danych dobrym rozwiązaniem jest zbudowanie kwantyzatora w oparciu o taką piramidę.

Im większą przyjmujemy wartość L tym lepszym przybliżeniem jest dla kwantowanego wektora powierzchnia piramidy. Ze względu na ograniczenia sprzętowe (wymagania wielkości pamięci) niemożliwe jest zapewnienie wystarczająco dużej wartości L . Aby poprawić jakość kodowania dla małych L , T.R. Fisher zaproponował modyfikację algorytmu zwaną PCPVQ (ang. Product Code Pyramid Vector Quantization). PCPVQ do kwantowania używa zestawu koncentrycznych piramid (rys. 2). Kwantowany wektor jest rzutowany na powierzchnię najbliższej piramidy i wtedy poszukiwany jest odpowiedni punkt kwantyzacji na jej powierzchni.



Rys. 2. Kwantyzacja PCPVQ w 2 wymiarach

Fig. 2. PCPVQ quantization in 2-D space

Piramidy w PCPVQ mają przypisane indeksy i odpowiedni indeks jest przesyłany razem z punktem siatki. Dzięki temu kodowany jest nie tylko kierunek, ale i długość wektora.

W PCPVQ (rys. 2) wektor jest rzutowany na powierzchnię najbliższej piramidy (oznaczonej P_{K_N}), a następnie przeskalowany tak, aby znalazł się na powierzchni piramidy o promieniu K (piramida oznaczona P_K). Kodowanie na P_K przebiega w punktach, które mają całkowite współrzędne. Tak zatem im większe K przyjmujemy, tym większa będzie dokładność kwantyzacji. W celu znalezienia punktu kwantyzacji realizowany jest poniższy algorytm.

Algorytm 1.

1. Znajdź X_p - rzut wektora $X(x_1, x_2, \dots, x_N)$ na powierzchnię P_{K_N} .
2. Znajdź odpowiadający punkt na powierzchni P_K (punkt X_K) - przeskaluj X_p przez współczynnik K/K_N .
3. Znajdź y - zaokrąglaj współrzędne X_K w dół do wartości całkowitych.
4. Oblicz

$$\|y\|_1 = \sum_{i=1}^L y_i \quad (2)$$

5. Jeżeli rezultat jest mniejszy od K , to $K - \|y\|_1$ współczynników które miały największy błąd z zaokrąglenia zwiększyć o 1.
6. Punktem kwantyzacji jest punkt $Y = (y * K_N) / K$.

3. KODOWANIE PUNKTÓW SIATKI

Skwantowany w punkcie siatki wektor jest przesyłany do odbiornika jako kod binarny. Ilość punktów na siatce kwantyzatora o promieniu K w L wymiarowej przestrzeni $N(L,K)$ zakładając całkowite współrzędne punktów siatki, można obliczyć korzystając z rekurencyjnej zależności

$$N(L, K) = N(L - 1, K) + N(L, K - 1) + N(L - 1, K - 1) \quad (3)$$

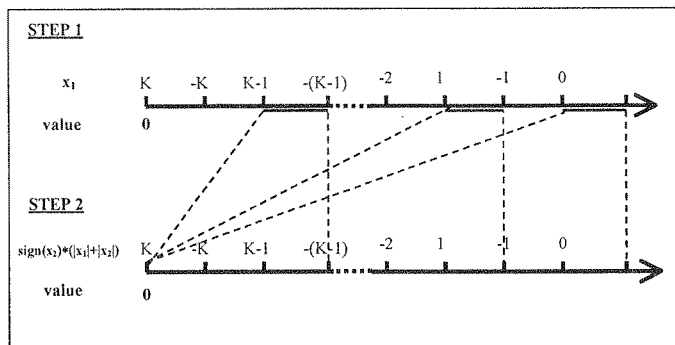
Aby kodowanie było możliwe należy zaproponować sposób przypisania punktom siatki unikalnego symbolu binarnego. T.R. Fisher zaproponował algorytm kodowania punktów siatki jako wartości binarnych z zakresu od 0 do $N(L,K)-1$. Ta poszukiwana wartość binarna może być obliczona na podstawie wartości współrzędnych kodowanego punktu siatki. Metoda numeracji zaproponowana przez Fishera nie jest oczywiście jedyną możliwą. Została ona zmodyfikowana dla uzyskania prostszej realizacji sprzętowej.

Metoda zakłada podział całego zakresu kodowania na przedziały (rys. 3). Dla $X(x_1, x_2, \dots, x_N)$ proces dzielenia rozpoczynamy od współrzędnej x_1 . Ponieważ może ona przyjąć jakąkolwiek wartość z zakresu od $-K$ do K tak zatem dzielimy przedział na $2K+1$ przedziałów. Przedziały nie mają równych długości. Jeżeli $x_1=K$ lub $x_1=-K$ to pozostałe współrzędne muszą być zerowe. Tak więc przedział dla $x_1=K$ lub $x_1=-K$ ma długość 1. Dla przykładu jeżeli $x_1=0$ to mamy $N(L-1,K)$ możliwych punktów siatki. Dalej dla $x_1=1$ szerokość przedziału jest równa $N(L-1,K-1)$, a dla $x_1=2$ mamy $N(L-1,K-2)$ itd. Następnie dzielimy uzyskane przedziały na podprzedziały biorąc pod uwagę x_2 . Zgodnie z wartością x_1 dzielimy odpowiedni przedział. Jeżeli $x_1=0$, x_2 może być dowolną wartością z zakresu $-K$ do K , a zatem możliwych jest $2K+1$ podprzedziałów. Jeżeli $x_1=1$, x_2 może być dowolną wartością z zakresu $-K+1$ do $K-1$. Podprzedziały mają różną długość wynikającą z wartości x_1 i x_2 . Dla przykładu jeżeli $x_1=x_2=0$ to szerokość odpowiedniego przedziału wynosi $N(L-2,K)$. Uogólniając dla $x_1=a$ i $x_2=b$ przedział ma szerokość $N(L-2, K-|a|-|b|)$. Dzielenia kontynuujemy aż do x_L tak, aby wszystkie rozważane przedziały miały szerokość jednostkową i odpowiadały jednemu wektorowi. Wskazywana wartość jest symbolem wektora.

Jak to pokazano na rysunku 3, pozycja podprzedziału w stosunku do początku przedziału jest zawsze tak sama. Oznacza to, że niezależnie od zdekodowanej pierwszej współrzędnej, kolejne współrzędne do dekodowania używają tych samych poziomów dzielących przedziały. Znakomicie ułatwia to realizację sprzętową i jest to podstawowa różnica w stosunku do kodowania zaproponowanego przez Fishera.

(2)

ajwiększy



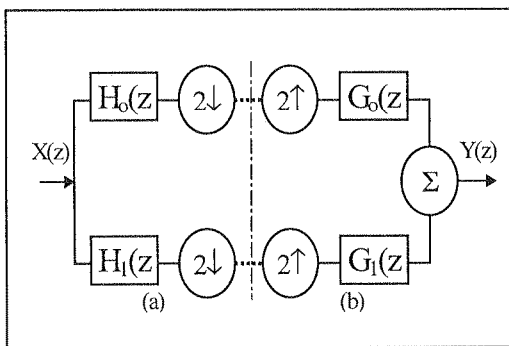
Rys. 3. Numerowanie punktów siatki poprzez dzielenie na kolejne podprzedziały

Fig. 3. Grid point numbering by successively sub-partition distribution

4. KODOWANIE PODPASMOWE

Kodowanie podpasmowe jest jedną z metod kodowania obrazów opartych na dekompozycji częstotliwościowej. Ponieważ punkty obrazu charakteryzują się dużą korelacją przestrzenną, ich dekorelacja poprzez przekształcenie częstotliwościowe pozwala osiągnąć lepsze wyniki kodowania. Z informacji wizyjnej wydzielane są jej składowe częstotliwościowe, które następnie są niezależnie kwantowane.

Aby przeprowadzić dekompozycję podpasmową używa się banku filtrów. Bank filtrów jest zestawem filtrów o wzajemnie uzupełniających się pasmach przepuszczania. Przykład prostego banku filtrów przedstawia rysunek 4. $H_0(z)$ i $H_1(z)$ są odpowiednio dolno i górnoprzepustowym filtrem analizującym, $G_0(z)$ i $G_1(z)$ są filtrami syntezującymi. Potrzeba stosowania filtrów syntezujących przy odtworzeniu zdekomponowanego sygnału wynika z zawsze istniejącego w praktyce pewnego pokrywania się pasm filtrów analizujących, co musi być kompensowane.



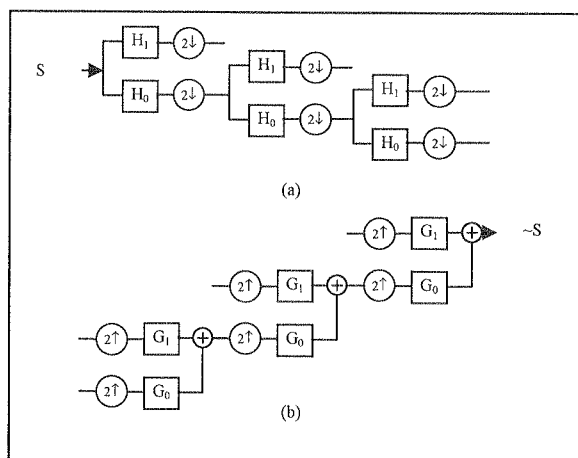
Rys. 4. Bank filtrów

Fig. 4. Filter bank

podpas
poziom
składow
dekom

W
filtracja

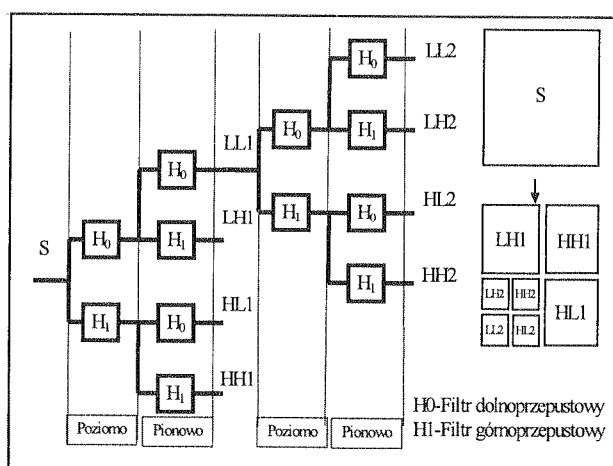
Dla potrzeb kompresji obrazu najczęściej stosowanym sposobem dekompozycji podpasmowej jest dekompozycja falkowa. Dekompozycja falkowa polega na wielopoziomowej dekompozycji podpasmowej, w której kolejnej dekompozycji poddawana jest składowa niskoczęstotliwościowa. Dekompozycja falkowa zakłada każdorazowy podział dekomponowanego pasma na połowę (rys. 5).



Rys. 5. Dekompozycja falkowa: a) filtr analizujący, b) filtr syntezujący

Fig. 5. Wavelet decomposition: a) analysis filter, b) synthesis filter

W odniesieniu do obrazów stosujemy dekompozycję dwuwymiarową w której filtracja odbywa się na przemian w kierunku pionowym i poziomym (rys. 6).



Rys. 6. Dwuwymiarowa dekompozycja falkowa

Fig. 6. 2-D wavelet decomposition

popr
czej c
tak d
chwil
częst

A
kwan
algor
kwan
PVQ
Na
turan
opera
i kom

S
czyst
potr

5. KWANTYZACJA WEKTOROWA WSPÓŁCZYNNIKÓW FALKOWYCH

Spotykany w praktyce sygnał wizyjny i natura wzroku ludzkiego powodują, że większość informacji istotnych z punktu widzenia kodowania, znajduje się w zakresie niskich częstotliwości. Zgodnie z teorią informacji, aby optymalnie gospodarować pasmem przesyłowym, alokacja bitów pasma powinna być proporcjonalna do energii składników sygnału. Tak zatem, zgodnie z powyższym stwierdzeniem, kwantyzacja powinna być dokładna dla podpasm o niższych składowych częstotliwościowych i bardziej zgrubna dla wyższych.

Kwantowane wektory były formowane z liniowo następujących po sobie współczynników obrazowych. Punkty wybierano kolumna po kolumnie przechodząc z linii do linii. Wybieranie rozpoczęło w lewym górnym rogu.

Przyjęcie za rozmiar wektora kwantyzacji kwadratów kolejnych liczb naturalnych mogło sugerować, że wektory formowano z kwadratowych obszarów na które podzielono obraz. Takie było pierwotne założenie lecz na skutek nie podzielności szerokości i wysokości ramek przez sparаметryzowany rozmiar tak przyjętego wektora i powstające na wskutek tego problemy na brzegach zrezygnowano z tego rozwiązania.

Rozmiar wektora jest w rozważanym artykule parametrem narzucanym przez wielkość sprzętu dlatego nie można go było dostosować tak, aby dzielił ramkę bez reszty.

Niezależnie od dekompozycji częstotliwościowej dalsza dekokrelacja kodowanych danych może być osiągnięta przez zastosowanie kwantyzacji wektorowej w miejsce skalarnej. Im większy wymiar wektora przyjmujemy przy tej kwantyzacji tym lepszych rezultatów można się spodziewać. W praktyce rozmiar wektora jest ograniczony przez możliwości użytego sprzętu.

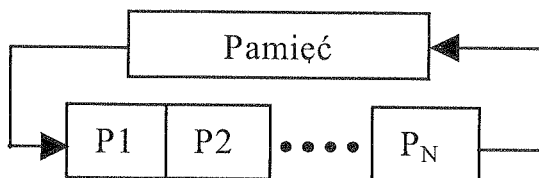
Innym problemem jest sposób alokacji bitów do poszczególnych pasm. Najprostszą metodą jest zastosowanie z góry narzuconych wartości otrzymanych na przykład w wyniku wyczerpującego przeszukiwania przeprowadzonego dla obrazów sklasyfikowanych według różnych kryteriów (na przykład według estymacji ruchu). Odpowiednie wartości przepływności bitowej mogą być przypisane podpasmom obrazów określonej klasy. Przy kodowaniu wideo może być także różna alokacja dla kodowania wewnątrz i między ramkowego (ang. inter/intra frame coding).

6. REALIZACJA SPRZĘTOWA

6.1. ARCHITEKTURA POTOKOWA

W architekturze potokowej dane są przetwarzane przez ciąg szeregowo połączonych procesorów. Każdy procesor wykonuje pewną część całego algorytmu i przekazuje częściowe wyniki do następnego procesora. Główną zaletą tej architektury jest możliwość załadowania nowych danych zaraz po opuszczeniu pierwszego procesora przez dane

poprzednie. W ten sposób dane są w pewnym sensie przetwarzane równolegle. W pojedynczej chwili czasowej N danych jest przetwarzanych w N procesorach. Chwila czasowa trwa tak długo jak czas przetwarzania najwolniejszego procesora. W układach synchronicznych chwila czasowa odpowiada częstotliwości zegara. W konsekwencji można powiedzieć, że częstotliwość wykonywania całego algorytmu jest równa częstotliwości zegara systemu.



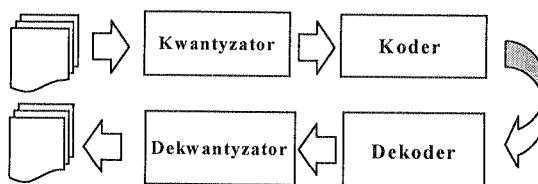
Rys. 7. Architektura potokowa

Fig. 7. Pipelined architecture

6.2. IMPLEMENTACJA SPRZĘTOWA

Algorytm PVQ może zostać podzielony na dwa etapy nazywane w tym artykule: kwantyzacja i kodowanie. Chociaż określenie kwantyzacja może być stosowane do całego algorytmu, to jednak w tym przypadku chodzi o proces znajdowania punktu na siatce kwantyzatora. Kodowanie to przypisanie kodu do punktu siatki. Proces odwrotny tzn. PVQ^{-1} podzielimy na: dekodowanie i dekwantyzację (rys. 8).

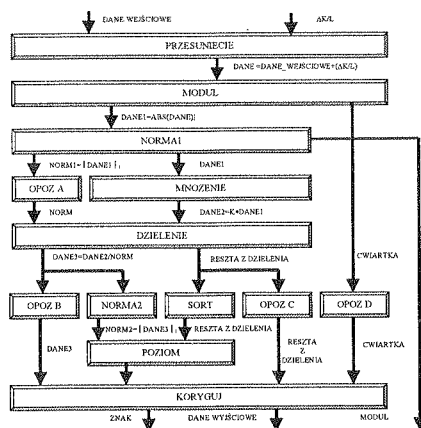
Należy podkreślić, że prezentowane na schematach bloki są same w sobie architekturami potokowymi. Aby osiągnąć maksymalną szybkość przetwarzania, podstawowymi operacjami realizowanymi przez procesory w potoku są operacje typu dodawanie i komparacja.

Rys. 8. System kwantyzacji PVQ/ PVQ^{-1} Fig. 8. PVQ/ PVQ^{-1} system quantizer

6.2.1. Kwantyzator

Schemat blokowy kwantyzatora przedstawiono na rysunku (rys. 9). Nie jest to czysto potokowe rozwiązanie, ponieważ nie było to możliwe (Algorytm 1) ze względu na potrzebę przechowywania wyników tymczasowych. Taka cecha realizowanego algorytmu

poważnie utrudnia warunki przetwarzania potokowego. Aby ominąć tą niedogodność zastosowano linie opóźniające (OPOZ). Zamiast przechowywać wyniki tymczasowe są one przesuwane wzdłuż głównego potoku przetwarzającego.



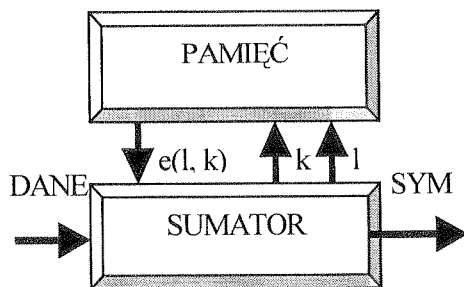
Rys. 9. Moduł kwantyzatora

Fig. 9. Quantizer module

6.2.2. Koder

Blok kodera używany jest do znalezienia odpowiadającego danemu punktowi symbolu. Zgodnie z algorytmem opisanym w rozdziale 2 odpowiedni symbol jest otrzymywany poprzez dodawanie do siebie wartości progów przechowywanych w pamięci (rys. 10). Zawartość pamięci jest wcześniej przygotowana dla danego $P(L, K)$.

Kolejno indeks współrzędnej kodowanego punktu (l) i jej wartość (k) adresują odpowiednio pamięć. Wartość $e(l, k)$ zapisana w pamięci jest dodawana do wartości w rejestrze tymczasowym. Co L współrzędnych otrzymujemy nowy symbol i kasujemy zawartość rejestru tymczasowego.



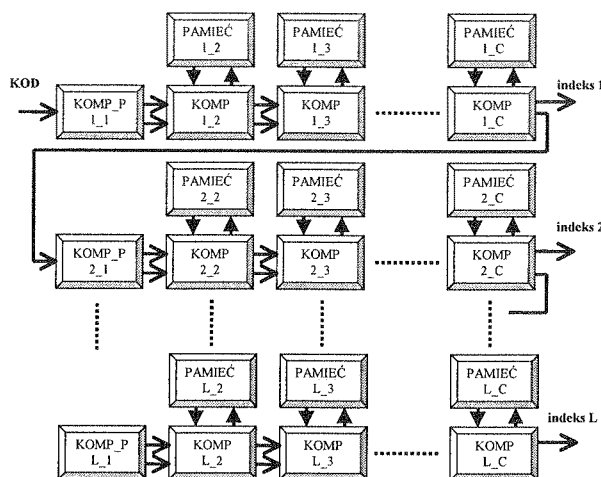
Rys. 10. Moduł kodera

Fig. 10. Coder module

6.2.3. Dekoder

Dekoder jest bardziej skomplikowany niż koder, ponieważ zachodzi konieczność porównywania wielu progów w celu jednoznacznej identyfikacji przedziału. Moduł dekodera składa się z tablicy komparatorów (KOMP) i elementów pamięciowych (MEM) (rys.11). Każdy wiersz tablicy dekoduje jedną współrzędną. Ponieważ dekodowanie współrzędnej odbywa się bit po bicie, to aby zdekodować c-bitową współrzędną należy wykonać c porównań. Dlatego wymiary tablicy dekodera wynoszą $L \times c$. Aby zdekodować pierwszy bit potrzebujemy jeden próg komparacji. Dla dekodowania drugiego bitu potrzebujemy dwóch progów komparacji. To, który próg będzie użyty do porównania, zależy od wyniku pierwszej komparacji. Ogólnie do dekodowania n-tego bitu potrzeba 2^{n-1} progów komparacji. Wybór odpowiedniego progów odbywa się na podstawie wartości n-1 zdekodowanych wcześniej bitów.

Kiedy pierwsza współrzędna jest kompletnie zdekodowana przedział zmiennosci jest przesunięty o wartość początkową zdekodowanego przedziału i w kolejnym wierszu jest dekodowana następna współrzędna. Ponieważ do kodowania użyto zmodyfikowanego algorytmu, progi komparacji dla kolejnych współrzędnych nie zależą od wartości zdekodowanych wcześniej.



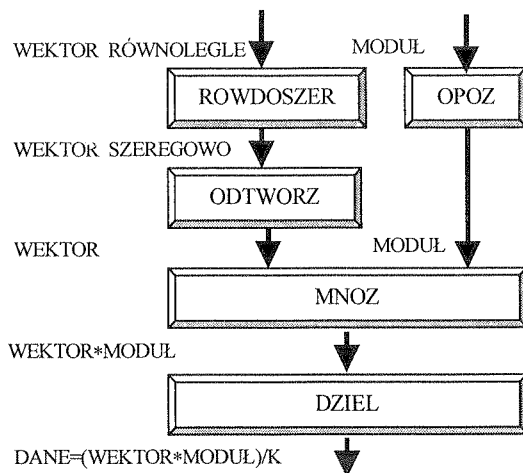
Rys. 11. Moduł dekodera

Fig. 11. Decoder module

6.2.4. Dekwantyzator

Algorytm dekwantyzacji w pewien sposób wykonuje operację odwrotną do kwantyzacji. W zasadzie polega on na odpowiednim przeskalowaniu (dzieleniu i mnożeniu) otrzymanej z dekodera wartości. Najpierw równoległa reprezentacja danych wycho-

dzących z dekodera musi być zamieniona na szeregową (blok ROWDOSZER). Blok ODTWORZ pełni funkcję dekodującą wynik otrzymywany z dekodera na właściwe współrzędne dekodowanego punktu. Bloki MNOŻ i DZIEL są użyte, aby punkt z piramidy o promieniu K przeliczyć na punkt o promieniu LEN (LEN promień piramidy PCPVQ).



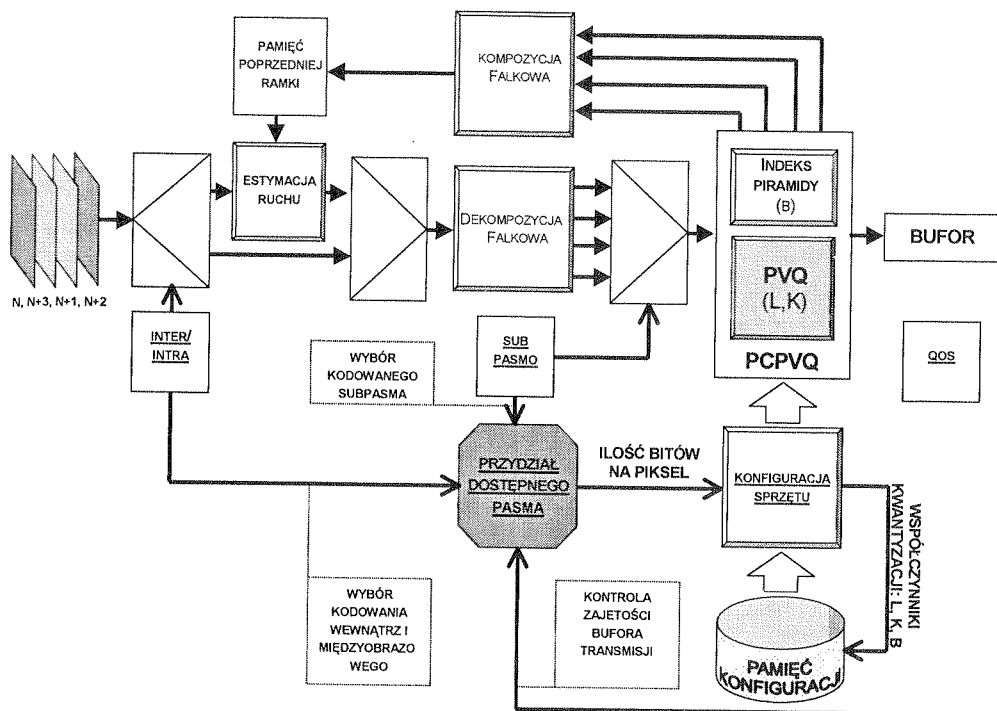
Rys. 12. Moduł dekwantyzatora

Fig. 12. De-quantizer module

7. REKONFIGUROWALNY KWANTYZATOR PVQ

Od czasu kiedy T. R. Fisher zaproponował kwantyzację PVQ podejmowano wiele prób jej zastosowania, w tym również w odniesieniu do kompresji falkowej. Ponieważ współczynniki falkowe w pewnym przybliżeniu spełniają założenia rozkładu Laplace'a [2] próby te wskazały na przydatność PVQ i PCPVQ do takiego kodowania. Choć należy przyznać, że rozkład Gaussa lepiej odpowiada współczynnikom falkowym jednak prostota algorytmu PVQ w szczególnych przypadkach (systemy małych mocy, czasu rzeczywistego) usprawiedliwia godzenie się na gorszą jakość kodowania.

Ponieważ PVQ skutecznie może być zastosowany do kodowania obrazów [9], [3], [2] poniżej przedstawiono rekonfigurowalny system kompresji obrazów (rys. 13).



Rys. 13. Rekonfigurowalny system kompresji obrazu

Fig. 13. Reconfigurable image compression system

no wiele
ponieważ
ace'a [2]
ć należy
prostota
eczywis-

Ponieważ głównym założeniem projektowanego systemu jest jego praca w czasie rzeczywistym użycie procesora ogólnego przeznaczenia jest niemożliwe i należy użyć procesor dedykowany. Z drugiej strony takie rozwiązanie stoi w sprzeczności z koniecznością parametryzacji algorytmu PVQ dla różnych gęstości bitowych, która jest kluczowa z punktu widzenia kompresji obrazu. System rekonfigurowalny zakłada zestaw dedykowanych procesorów, przy czym każdy zaprojektowano do innej wartości gęstości bitowej. Procesory mają charakter procesorów wirtualnych, ponieważ są przechowywane jako pliki konfiguracyjne gotowe do załadowania do rekonfigurowalnego układu FPGA. Taki zestaw procesorów może być bardzo duży ponieważ ich ilość nie wpływa na cenę systemu. Jedyną wadą jest jedynie to, że stosowane do kompresji parametry są z natury rozwiązaniami bardziej dyskretnymi niż w rozwiązaniu software'owym.

W proponowanym systemie poszczególne konfiguracje procesora dedykowanego są przechowywane w pamięci komputera, w którym zainstalowany jest system i ładowane poprzez sprzęg „Konfiguracja sprzętu”. Blok „Przydział dostępnego pasma” jest jednostką podejmującą na podstawie typu kodowanej ramki i rodzaju podpasma oraz zajętości bufora wyjściowego decyzje o właściwej w danej chwili gęstości bitowej kodowania.

Tabela 1

Oszacowane zasoby sprzętowe (liczba bloków CLB) potrzebne do realizacji modułu dekodera.
Zaznaczono parametry (L,K) dla układów FPGA XCV600, XCV1000, XCV3200

The estimated hardware resources (in number of CLBs) necessary for implementation of decoder block.
The parameters for targeting XCV600, XCV1000, XCV3200 FPGA chips are also shown as shaded areas

L\BITRATE	0,2		0,4		0,5		0,7		1		2		4		6	
	K	CLB	K	CLB	K	CLB	K	CLB	K	CLB	K	CLB	K	CLB	K	CLB
4									1	33	4	117	24	318	147	1030
9									3	237	7	499	36	1862	171	4863
16									4	868	12	2003	57	5737	248	15086
25	1	299	2	738	3	954	4	1568	6	2037	18	6080	83	18259	255	34096
39	2	1163	3	1508	4	2565	6	3340	9	5610	25	12429	115	37672		
49	2	1702	4	3789	5	4458	7	5590	12	10231	33	29155	154	97994		
64	2	2367	5	6291	6	7130	9	12240	15	16989	43	49676	198	166970		
81	3	4213	6	9648	8	15417	12	20299	19	34732	54	79423	247	266970		
100	3	5464	7	14125	9	2205	14	29774	23	52350	66	160247	255	385620		
121	4	11857	8	25867	11	32391	17	55845	27	75217	80	235036				
144	4	14677	10	37834	13	45687	20	78518	32	136755	94	331280				
169	5	21179	11	49842	15	62211	23	106693	38	189941	110	456033				
196	6	29196	13	67933	17	105300	27	144781	43	251800	128	867319				
225	6	34535	15	89817	20	141207	30	186737	49	330497	146	1140536				
256	7	45511	17	148305	22	178598	34	310594	56	429312	165	1472695				

Pr
dobrze
sprzęt
do uk
odnies
wielko
W
PVQ z
kompr
PCPV
Wynik
Go
standa
stosow
gdy p
JPEG

8. REALIZACJA PVQ W UKŁADACH FPGA VIRTEX

Przed przystąpieniem do właściwej implementacji w układzie programowalnym FPGA dobrze jest wstępnie ocenić konieczne do realizacji zasoby. Aby uzyskać dalsze oszczędności sprzętowe poszczególne moduły kwantyzatora, kodera, dekodera i dekwantyzatora są ładowane do układu FPGA odrębnie. Ponieważ moduł dekodera jest największy, jego wielkość jest odniesieniem dla wielkości zastosowanego układu FPGA. Rezultaty przeprowadzonej ewaluacji wielkości potrzebnego układu przedstawiono w tabeli 1.

Wytypowano zestawy parametrów (L,K) dla różnych gęstości bitowych tak, aby układ PVQ zmieścił się do układu XCV1000. Uzyskane w ten sposób parametry zastosowano do kompresji obrazu „Lena”. Zastosowano dekompozycję falkową 4 poziomową i algorytm PCPVQ. Alokacja bitów została przeprowadzona metodą poszukiwania wyczerpującego. Wyniki przedstawia tabela 3.

Gorsze wyniki kompresji PCPVQ w porównaniu z wynikami kompresji przy pomocy standardu JPEG2000 wynikają z bardzo wyrafinowanego i złożonego procesu kwantyzacji stosowanego w tej drugiej metodzie. Trzeba przyznać, że PCPVQ wypada jeszcze gorzej gdy porównać go z takimi opracowanymi metodami kompresji jak EZW i SPIHT. JPEG2000 czerpie wiele z tych metod i dlatego osiąga wspaniałe rezultaty.

Tabela 2

Wyniki rzeczywistej implementacji algorytmu PVQ w układzie XCV
1000BG560-4 (6144 CLB), częstotliwość zegara 50MHz

The results of real implementation of PVQ algorithm in Virtex XCV
1000BC560-4 chips (6144 CLB), the clock frequency at least 50 MHz

BIT RATE	CLB					
	L	K	Dequantizer	Coder	Decoder	Dequantizer
0,2	100	3	4620 75%	502 8%	5980 97%	1800 29%
A0,5	49	5	2558 41%	445 7%	4702 76%	1851 30%
0,7	49	7	2582 42%	588 9%	5777 94%	1849 30%
1	36	9	2075 33%	630 10%	54 88%	1938 31%
4	16	57	1355 22%	1329 21%	5619 91%	2049 33%
6	9	171	1207 19%	1515 24%	8077 65%	2156 35%

Tabela 3

Rezultaty kompresji obrazu „Lena” 512×512
dla poszczególnych gęstości bitowych

Results of „Lena” 512×512 greyscale image compression
for various target bitrates

BITRATE	1	0,5	0,2	0,1
BAND				
HH1	0	0	0	0
LH1	0,5	0	0	0
HL1	0,5	0,2	0	0
HH2	1	0,5	0,2	0,1
LH2	2	1	0,5	0,2
HL2	4	1	0,9	0,3
HH3	5	5	0,6	0,2
LH3	6	5	0,7	0,2
HL3	6	5	3	2
HH4	6	5	4	4
LH4	6	6	4	4
HL4	6	6	6	6
LL4	8	8	8	8
PSNR[dB]	31,53	30,24	28,86	26,01

Przede wszystkim EZW i SPIHT opierają się nie na DCT, ale na dekompozycji falkowej. Dzięki temu podlegające subiektywnej ocenie obrazy kodowane tymi metodami pozbawione są charakterystycznej dla DCT segmentacji. Ponadto EZW i SPIHT wykorzystują korelację poszczególnych podpasm i dlatego bazując na tym zjawisku oferują lepsze wyniki kompresji. Chociaż ze względów praktycznych w JPEG2000 zrezygnowano z łącznego kodowania poszczególnych podpasm osiągając w ten sposób lepszą odporność na błędy transmisji to jednak wykorzystano w tym standardzie kodowanie współczynników nie jednorazowo jako całej wartości, ale iteracyjnie ich kolejnych bitów od najbardziej do najmniej znaczącego (ang. bit-planes). Daje to możliwość skalowalności kodowania i dodatkowo powoduje powstawanie obszarów zer, które mogą być kodowane łącznie. Taka technika znacznie polepsza efektywność kodowania entropijnego.

Mimo znacznie lepszych od JPEG wyników kompresji, JPEG2000 nadal bazuje na kodowaniu entropowym co jest w niektórych zastosowaniach wadą, która daje metodzie PCPVQ przewagę. Nie jest możliwe bowiem w metodach stosujących jako ostatni etap kompresji kodowanie entropowe tak efektywnie jak w PCPVQ sterowanie gęstością generowanego strumienia danych. W PCPVQ możemy założyć gęstość strumienia przed rozpoczęciem kodowania. Cecha ta może być bardzo istotna w niektórych zastosowaniach np.: w systemach czasu rzeczywistego. Należy ponadto pamiętać, że podstawowy cel czyli realizacja kwantyzatora jako procesora sprzętowego doprowadził do szeregu założeń upraszczających. Przede wszystkim założono, że współczynniki falkowe mają rozkład Laplace'a, choć bliższe prawdzie byłoby przyjęcie rozkładu Gaussa. W przypadku przyjęcia rozkładu Gaussa jako modelu dla współczynników falkowych kwantyzować należałoby na powierzchni sfery, co byłoby znacznie bardziej skomplikowane. Ponadto narzucony wielkością zasobów sprzętowych mały rozmiar wektora dodatkowo osłabia osiąganą jakość kompresji.

Rzeczywistą implementację sprzętową wykonano w układzie XCV1000BG560-4 (tab. 2). Ponieważ współczynniki cztero-poziomowej dekompozycji falkowej obrazu o 8-bitowych pikselach mają maksymalnie 45 bitów (dla stosowanych filtrów), szerokość danych wejściowych podczas implementacji ustalono na taką wartość. Opis architektury w języku VHDL został skompilowany przy pomocy programu Synopsys FPGA Express 3.6.1. Implementację przeprowadzono przy pomocy programu Xilinx Alliance 3.1i. Wymaganie dotyczące szybkości zegara, wynoszącej 50MHz, zostało spełnione w każdym implementowanym przypadku.

9. PODSUMOWANIE

Kompresja obrazów jest obecnie bardzo ważną klasą aplikacji obliczeniowych współczesnych systemów. Szczególnie zastosowanie kompresji obrazów w czasie rzeczywistym ma duże znaczenie dla automatyki i zastosowań multimedialnych. Jednak złożoność obliczeniowa algorytmów kompresji obrazów poważnie utrudnia działalność w tym zakresie z wykorzystaniem systemów komputerowych ogólnego przeznaczenia. Dlatego dobrym rozwiązaniem jest zastosowanie do tego celu specjalizowanych procesorów sprzętowych implementowanych w układach programowalnych FPGA.

W niniejszym artykule przedstawiono wyniki przykładowych implementacji algorytmu kwantyzacji wektorowej PVQ i PCPVQ. Algorytmy te szczególnie dobrze nadają się do implementacji sprzętowej ponieważ wymagają bardzo dużo, ale prostych obliczeń. Na szczególną uwagę zasługuje zaproponowany w artykule schemat rekonfigurowalnego systemu kompresji obrazu. System ten z jednej strony spełnia wymagania zastosowanego algorytmu, z drugiej strony możliwa jest jego realizacja sprzętowa.

W tabeli 1 przedstawiono szczegółowe wyniki przykładowych implementacji. Ich wartość to przede wszystkim możliwość oszacowania rzeczywistych potrzeb wielkości koniecznego do zastosowania układu programowalnego z rodziny Virtex firmy Xilinx

(600k do 3200k bramek) w zależności od wartości przepływności bitowej (0,2 do 6). Należy jeszcze raz podkreślić, że w zaprojektowanym systemie dane obrazowe mogą być przetwarzane z częstotliwością 50MHz (efektywny czas potrzebny na przetworzenie pojedynczego piksela danych wynosi zaledwie 20ns).

10. BIBLIOGRAFIA

1. M. Barr: *A Reconfigurable Computing Primer*. Multimedia Systems Design, September 1998, pp. 44–47
2. M.E. Blain, T.R. Fisher: *A Comparison of Vector Quantization Techniques in Transform And Subband Coding of Imagery*. Signal Processing: Image Communication 3, Elsevier 1991, pp. 91–105
3. P. D' Alessandro, P. Formenti, R. Lancini: *Pyramid Vector Quantization Applied To A Video Coding Scheme For Contribution Quality*. Proc. of IEEE ICASSP'93, New York, USA, 1993, pp. 249–252
4. T.R. Fisher: *A Pyramid Vector Quantization*. IEEE Trans. on Information Theory, vol. IT-32, no 4, July 1986
5. A. Gersho, R.M. Gray: *Vector Quantization and Signal Compression*. Kluwer Academic 1992
6. B.M. Gordon, E. Tsern, T.H. Meng: *Design of Low Power Video Decompression Chip Set for Portable Applications*. Journal of VLSI Signal Processing Systems vol. 13, 1997, pp. 125–142
7. B. Jung, W.P. Burlison: *VLSI Array Architectures for Pyramid Vector Quantization*. The Journal of VLSI Signal Processing vol. 18, 1998, pp. 141–154
8. E.K. Tsern, T.H. Meng: *A Low Power Video-Rate Pyramid VQ Decoder*. IEEE Journal of Solid State Circuits, vol. 31, no. 11, Nov. 1996
9. E.K. Tsern, T.H.Y. Meng: *Image Coding Using Pyramid Vector Quantization Of Subband Coefficients*. Proc. of IEEE ICASSP94, New York, USA, 1994, vol. 5, pp. 601–604
10. J. Villasenor, C. Jones, B. Schoner: *Video Communications using Rapidly Reconfigurable Hardware*. IEEE Trans. on Circuits and Systems for Video Technology, vol. 5, Dec. 1995, pp. 565–567
11. N. Won, T.H. Meng: *A Low-Power Encoder For Pyramid vector Quantization of Subband Coefficients*. Journal of VLSI Signal Processing Systems, vol. 16, 1997, pp. 9–23
12. Xilinx: *The Programmable Logic Data Book*. San Jose, USA, 2001

P. RUSSEK, K. WIATR

RECONFIGURABLE VECTOR QUANTIZER FOR IMAGE COMPRESSION IN REAL-TIME SYSTEMS

Summary

This paper concerns implementation of Pyramid Vector Quantization (PVQ) in FPGA. The PVQ encoding algorithm is modified for easier hardware solution. The quantizer architecture is proposed and its hardware resources are evaluated. High data throughput characterises presented architecture. Based on evaluated hardware resources PVQ parameters are inherited for targeting several FPGA chips. The Xilinx's Virtex FPGA family is used for real implementation. The reconfigurable system is proposed where hardware reconfiguration is used for quantizer control in video compression.

Keywords: image processing, image compression, real-time systems, programmable logic device

Zagadnienia ustalania częstotliwości drgań generatorów z rezonatorami kwarcowymi pobudzanymi do drgań polem równoległym

KRZYSZTOF WEISS, BARBARA GNIEWIŃSKA, ALEKSANDER LISOWIEC

*Instytut Tele- i Radiotechniczny
03-450 Warszawa ul. Ratuszowa 11
kweiss@itr.org.pl*

*Otrzymano 2002.07.03
Autoryzowano 2002.10.22*

Artykuł stanowi kontynuację opisu wyników prac badawczych prowadzonych w Instytucie Tele- i Radiotechnicznym nad rezonatorami pobudzanymi do drgań modów anharmonicznych za pomocą pola elektrycznego skierowanego równolegle do powierzchni płytki kwarcowej, których wyniki zostały przedstawione w artykule prezentowanym w Kwartalniku Elektroniki i Telekomunikacji nr 3/2000. W artykule przedstawiono problematykę dostrajania do częstotliwości znamionowej generatorów z wykorzystaniem rezonatorów pobudzanych polem równoległym. Dla oceny zagadnienia przedstawiono w zarysie wpływ procesu technologicznego na dokładność kalibracji częstotliwości rezonatorów oraz możliwości przestrajania częstotliwości rezonatora w generatorze. Przedstawiono również porównanie parametrów i związków z nimi zachowań w procesie ustalania częstotliwości rezonatorów pobudzanych polem równoległym do drgań modu głównego oraz drgań anharmonicznych z rezonatorami klasycznymi pobudzanymi polem prostopadłym do drgań modu głównego.

Słowa kluczowe: rezonator kwarcowy, pole równoległe, generator wysokostabilny

WSTĘP

W generatorach kwarcowych o najwyższej stałości częstotliwości są stosowane rezonatory o specjalnej konstrukcji umożliwiającej osiągnięcie bardzo dobrych parametrów generatora, a w szczególności długo- i krótkoterminowej stałości częstotliwości i małej wrażliwości na wpływ temperatury. Są to zazwyczaj rezonatory cięcia SC o konstrukcji bezelektrodowej [1], lub rezonatory pobudzane polem równoległym [2]. Uzyskiwane przy zastosowaniu tych rezonatorów parametry generatorów są następujące:

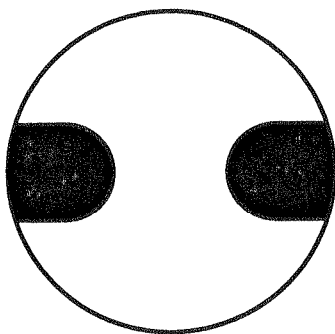
– niestaość długoterminowa częstotliwości — $< 5 \times 10^{-11}/24 \text{ h}$

- niestalość krótkoterminowa częstotliwości — $<1 \times 10^{-12}/s$
- temperaturowy współczynnik zmian częstotliwości — $<2 \times 10^{-12}/K$

Osiągnięcie takich parametrów wymaga bardzo dobrego termostatowania generatora i maksymalnego ograniczenia wpływu współpracujących z rezonatorem układów elektronicznych, a w szczególności układu wzbudzającego i układu regulacji częstotliwości. Cechą wspólną obydwu typów stosowanych rezonatorów jest trudność uzyskania z dużą dokładnością zadanej częstotliwości rezonatora w fazie produkcji oraz zmniejszona podatność na przestrajanie częstotliwości przez zewnętrzne układy elektroniczne. Szczególnie małą podatność na przestrajanie wykazują rezonatory pobudzone polem równoległym, charakteryzujące się bardzo małą pojemnością dynamiczną. Cecha ta bardzo utrudnia konstrukcję układu regulacji częstotliwości, którego zadaniem jest w fazie produkcji umożliwienie dostrojenia generatora do częstotliwości znamionowej, a podczas eksploatacji generatora w okresie ponad 10 lat zagwarantowanie możliwości korekcji zmian częstotliwości wynikających z procesu starzenia.

2. PARAMETRY REZONATORÓW POBUDZANYCH POLEM RÓWNOLEGLYM

Klasyczne rezonatory pobudzone są do drgań za pomocą dwóch elektrod umieszczonych po przeciwnych stronach płytki kwarcowej. Elektrody takie wytwarzają pole elektryczne prostopadłe do powierzchni płytki. Możliwe jest również pobudzanie płytki kwarcowej do drgań za pomocą elektrod wytwarzających pole elektryczne równoległe do jej powierzchni. Taką przykładową konfigurację elektrod umożliwiającą wzbudzenie drgań modu głównego przedstawiono na rysunku 1.

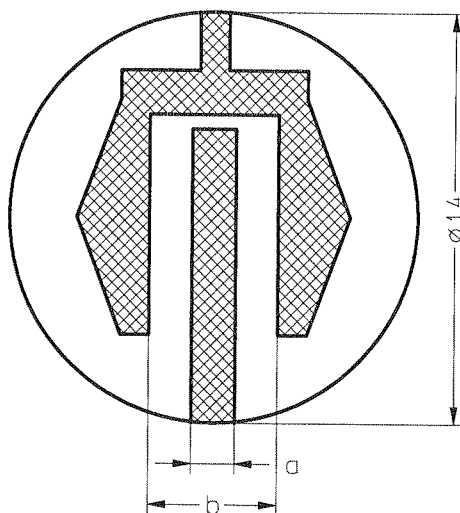


Rys.1. Konfiguracja elektrod wytwarzających pole równoległe

Fig. 1. The electrodes configuration for lateral electric field generation

Główną zaletą pobudzania polem równoległym jest usunięcie elektrod z obszaru drgań płytki. Zgodnie z jedną z teorii opisujących czynniki wpływające na długoterminową

stałość częstotliwości rezonatora głównymi czynnikami są zmiany zachodzące w czasie w elektrodzie (rekrytalizacja i dyfuzja) [3]. Tak więc usunięcie elektrod z obszaru drgań płytki powinno dać podwyższenie stałości długoterminowej rezonatora (analogicznie jak w rezonatorach BVA). Rezonatory te skrótowo oznacza się literami LFE (Lateral Field Excited). Alternatywnym rozwiązaniem w stosunku do rezonatorów pobudzanych do drgań modu głównego polem równoległym mogą być rezonatory pobudzane również polem równoległym ale do drgań anharmonicznych. Rezonatory takie zostały omówione szczegółowo w publikacji [4]. Wzbudzenie drgań anharmonicznych wymaga odpowiednio dobranego kształtu elektrod (rys.2.)

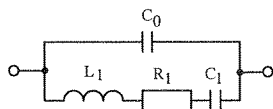


Rys.2. Konfiguracja elektrod wykorzystywana do wzbudzania polem równoległym drgań anharmonicznych modu n01

Fig. 2. The electrodes configuration for lateral field anharmonic n01 mode excitation

Rezonatory takie ze względu na obecność w nich kilku obszarów drgań o tej samej częstotliwości (dla modu n01 dwa a dla modu n11 cztery) zachowują się tak jak kilka rezonatorów o tej samej częstotliwości pracujących synchronicznie, czyli wykazują znacznie niższą rezystancję dynamiczną przy większej pojemności dynamicznej. W Instytucie Tele- i Radiotechnicznym wykonano cały szereg modeli rezonatorów pobudzanych polem równoległym do drgań modu głównego oraz modów anharmonicznych[5].

Przyjmując układ zastępczy rezonatora jak na rysunku 3 w tabeli 1 podano zestawienie porównawcze parametrów tych rezonatorów o częstotliwości 10 MHz z parametrami rezonatorów pobudzanych polem prostopadłym.



Rys.3. Elektryczny układ zastępczy rezonatora,

Fig. 3. Equivalent electric circuit of resonator

Zgodnie z układem zastępczym częstotliwości drgań rezonatora określa zależność (1).

$$f_s = \frac{1}{2\pi\sqrt{L_1 C_1}} \quad (1)$$

Tabela 1

Parametry rezonatorów o częstotliwości 10 MHz pobudzanych do drgań polem prostopadłym oraz polem równoległym

The parameters of 10 MHz resonators excited for vibrations by perpendicular field and by lateral field

Rodzaj drgań	$R_1[\Omega]$	$C_1[\text{aF}]$	$C_0[\text{pF}]$	$Q \times 10^6$
Pole równoległe drgania główne (mod 300)	500÷1000	10,8÷11,8	0,13÷0,14	1,30÷1,36
Pole równoległe drgania anharmoniczne (mod 301)	340÷410	36,5÷40,5	0,3	1,03÷1,16
Pole równoległe drgania anharmoniczne (mod 311)	300÷320	39÷53	0,48	1,0÷1,15
Pole prostopadłe drgania główne	40÷60	330÷360	2,2÷2,4	1,0÷1,2

Przeprowadzone badania zachowania w czasie generatorów z rezonatorami pobudzanymi polem równoległym wykazały bardzo dobre ich własności. Z całej populacji 15 sztuk generatorów wyposażonych w te rezonatory 80% wykazało dobrą stałość częstotliwości lepszą niż 5×10^{-11} . Przykładową zależność częstotliwości generatora od czasu pokazano na rys.4. Po skorygowaniu wpływu na częstotliwość zmian ciśnienia atmosferycznego[6] dobowa stałość częstotliwości wynosi 2×10^{-11} .

Type = "OCXO-2001" F_{nom} = "10.0" Number = 301

Measurement_period = "13.12.00 08.04.02"

Calculation_period = "05.11.01 30.11.01"

Calculation results**A. Not corrected****B. Corrected**

Average twenty-four hours frequency change /E-11/ -->

 $\delta f = 0.98$ $\delta f_{sk} = 1.13$

Calculation accuracy /E-11/ -->

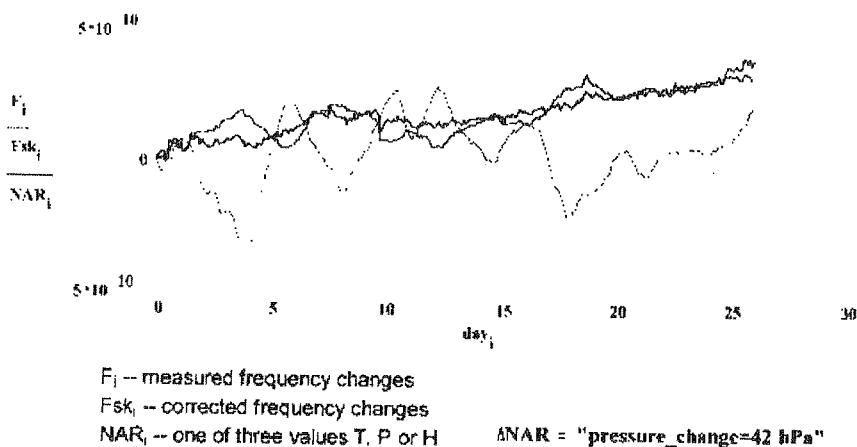
 $\varepsilon = 0.11$ $\varepsilon_{sk} = 0.07$

Standard deviation /E-10/ -->

 $\sigma = 0.51$ $\sigma_{sk} = 0.29$

ność (1).

(1)

AMBIENT_FACTOR "P"

Rys.4. Zależność częstotliwości od czasu generatora z rezonatorem pobudzonym polem równoległym do drgań modu 301

Fig. 4. The aging of oscillator with anharmonic LFE 301 mode resonator

3. UKŁAD REGULACJI CZĘSTOTLIWOŚCI

Układ regulacji częstotliwości stosowany w wysokostabilnych generatorach kwarcowych stanowi włączony w szereg z rezonatorem dwójnik złożony z diody pojemnościowej sterowanej napięciem i połączonych z nim szeregowo i równolegle reaktancji stałych.

Stosowany układ regulacji musi zapewniać:

- dostrojenie generatora do częstotliwości nominalnej przy określonej przez możliwości wykonawcze tolerancji wykonania rezonatora
- zakres zmian częstotliwości przy określonych zmianach napięcia sterującego diodą zapewniający możliwość korekty częstotliwości do wartości nominalnej w całym okresie eksploatacji generatora (co najmniej 10 lat)

Jednocześnie układ nie może obniżać możliwych do osiągnięcia parametrów generatora. Negatywne oddziaływanie układu regulacji na parametry generatora następuje przez:

- obniżenie skuteczności stabilizacji rezonatora przez włączenie w szereg elementów obniżających dobroć rezonatora
- zwiększenie rezystancji definiującej warunki wzbudzenia generatora
- dodatkowe zmiany częstotliwości w funkcji temperatury wynikające z wrażliwości termicznej czwórnika.

Uzyskiwane w wyniku włączenia reaktancji przestrojenie generatora wynika z zależności:

$$\frac{\Delta f}{f} = \frac{\frac{-x}{|x_0|}}{1 - \frac{x}{|x_0|}} \cdot \frac{C_1}{2 \cdot C_0} \quad (2)$$

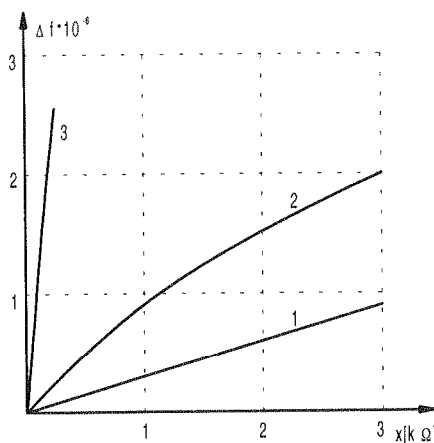
gdzie: — x — reaktancja dwójnika

C_1 — pojemność szeregową rezonatora

C_0 — pojemność równoległa rezonatora

$$|x_0| = \frac{1}{\omega C_0}$$

Zależność przestrojenia generatora od wartości włączonej w szereg reaktancji dla różnych typów rezonatorów przedstawiono na rysunku 5.



Rys.5. Przestrojenie generatora w funkcji reaktancji szeregowej

1. — rezonator z polem równoległym drganie główne
2. — rezonator z polem równoległym drganie anharmoniczne
3. — rezonator z polem prostopadłym drganie główne

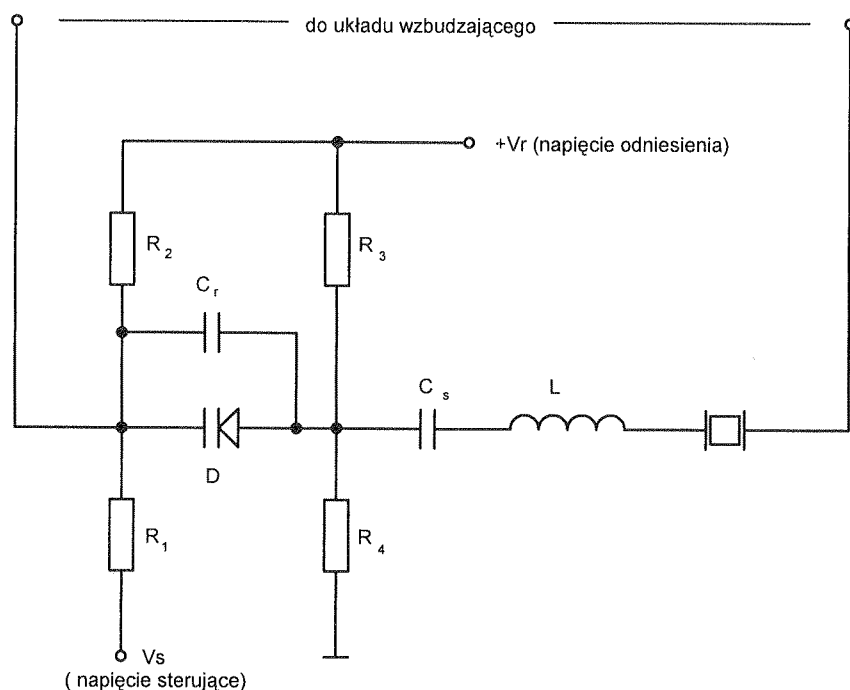
Fig. 5. Oscillator frequency tuning dependence on series reactance

1. — LFE main mode resonator
2. — LFE anharmonic mode resonator
3. — Main mode classical resonator

Dla rezonatorów pobudzanych polem prostopadłym przyjmuje się tolerancję wykonania $\pm 2 \times 10^{-6}$, a niezbędny zakres regulacji dla korekcji częstotliwości w czasie eksploatacji $\pm 5 \div 10 \times 10^{-7}$. Duża w porównaniu z rezonatorami pobudzanymi polem równoległym przestrajalność rezonatorów pozwala stosować układy, w których indukcyjność L jest zbędna a udział diody pojemnościowej w całej reaktancji jest ograniczony przez włączenie równolegle i szeregowo do diody dodatkowych pojemności dla uzyskania wymaganego zakresu korekcji. W czasie eksploatacji można dobrać odpowiedni przedział zmian napięcia sterującego przez dobór odpowiednich wartości rezystancji rezystorów $R_1 + R_4$ (rys.6).

(2)

ancji dla



Rys.6. Układ regulacji częstotliwości

Fig. 6. Frequency trimming circuit

Dzięki temu przy odpowiednim doborze parametrów układu można stosunkowo łatwo spełnić wymagania wynikające z parametrów rezonatora. Włączenie dodatkowych pojemności stałych pozwala na skuteczne ograniczenie wpływów termicznych wnoszonych głównie przez diodę pojemnościową, charakteryzującą się dużym i zmiennym w funkcji napięcia przestrająjącym współczynnikiem termicznym.

Dla rezonatorów pobudzanych polem równoległym w związku z poprawą stałości długoterminowej zakres zmian częstotliwości potrzebnych do korekcji w czasie eksploatacji można ograniczyć do wartości $\pm(2 \div 3) \times 10^{-7}$. Niestety nawet przy takim

ograniczeniu dla rezonatorów pracujących na drganiu głównym zmienność reaktancji diody pojemnościowej w zakresie praktycznie występujących zmian napięcia sterującego $1 \div 10V$ wystarcza na pokrycie zaledwie 80% tego zakresu, a dla rezonatorów anharmonicznych, o korzystniejszych parametrach wymaga wykorzystania ponad 50% tego zakresu. Pozostający w tych warunkach margines pokrycia zakresu przestrojenia wynikającego z dokładności wykonania rezonatora zawiera się w granicach $(0 \div 1) \times 10^{-6}$. Rozszerzenie zakresu regulacji osiąga się przez włączenie w szereg z diodą indukcyjności. Niestety dla uzyskania niezbędnych zmian częstotliwości wartości indukcyjności są duże. Ich przykładowe wartości w zależności od zakresu przestrajania i rodzaju rezonatora przedstawiono w tabeli 2.

Tabela 2

Przykładowe wartości indukcyjności stosowanych przy przestrajaniu generatorów z rezonatorami pobudzonymi polem równoległym.

Example of inductance values used for frequency trimming of oscillators with LFE resonator

Przestrojenie [10^{-6}]	Indukcyjność [μH]	
	Rezonator na drganiach głównych	Rezonator na drganiach anharmonicznych
1	80	20
2	-	40

Cewki o takich indukcyjnościach i konstrukcji przystosowanej do montażu powierzchniowego charakteryzują się dużym współczynnikiem termicznym, a praktycznie dopuszczalne do stosowania wartości indukcyjności ze względu na zbliżanie się rezonansu równoległego cewki nie powinny przekraczać $40 \mu H$. Oznacza to, że nawet przy korzystniejszym z punktu widzenia dostrojenia wariacie rezonatorów anharmonicznych dopuszczalna ze względów technicznych tolerancja wykonania rezonatora nie może być większa niż $1,5 \times 10^{-6}$. Przy takim założeniu współczynnik termiczny dwójnika przestrajającego może wynosić nawet $4 \times 10^{-4}/K$, a odpowiadające temu zmiany częstotliwości generatora $(0,5 \div 1) \times 10^{-9}/K$. Oznacza to, że do uzyskania wymaganej dla generatora wartości współczynnika temperaturowego $\leq 2 \times 10^{-12}/$ konieczne jest bardzo dobre termostatowanie układu regulacji częstotliwości. Skomplikowany jest również proces strojenia. Praktycznie sprowadza się do indywidualnego doboru elementów układu.

4. OGRANICZENIA DOKŁADNOŚCI WYKONANIA REZONATORA

Złożoność problemu wykonania rezonatora z bardzo dużą dokładnością wynika z faktu, że rezonator kwarcowy jest układem mechanicznym, którego dokładność wykonania

i pomiaru jest o wiele rzędów niższa od precyzji zdefiniowania częstotliwości w wyrobie końcowym, którym jest generator. Schematyczny proces wytwarzania rezonatora przedstawiono w tabeli 3.

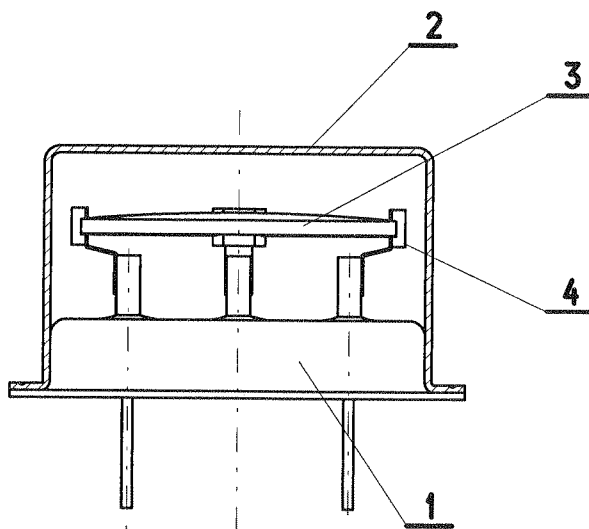
Tabela 3

Uproszczony proces wytwarzania rezonatora kwarcowego

The simplified process of resonator fabrication

L.p.	Operacja
1	Teoretyczne wyliczenie wymiarów płytki
2	Wykonanie mechaniczne płytki i pomiar jej grubości
3	Polerowanie powierzchni i wstępna korekta częstotliwości
4	Nanoszenie elektrod i montaż w podstawach
5	Wstępny pomiar częstotliwości i temperatury ekstremum
6	Kalibracja
7	Hermetyzacja
8	Pomiary końcowe

Uzyskanie wymaganej dokładności wyrobu potwierdzonej przez pomiary końcowe jest wynikiem kolejnego świadomego zawężania tolerancji. Konstrukcja rezonatora została przedstawiona na rysunku 7.



Rys.7. Rezonator wysokostabilny w obudowie TO-3. 1 — podstawa, 2 — osłona, 3 — płytka wibratora, 4 — element wsporczy

Fig. 7. Precision resonator in TO-3 enclosure. 1 — base, 2 — cover, 3 — vibrator, 4 — support.

4.1. WYZNACZANIE WYMIARÓW PŁYTEK

Wibrator kwarcowy rezonatora wysokostabilnego ma zazwyczaj kształt okrągłej płytki płasko- wypukłej. Proces doboru wymiarów płytki polega na rozwiązaniu metodą iteracyjną [7] zespołu przedstawionych poniżej równań.

$$f = \frac{N}{h} \quad (3)$$

gdzie: N' — stała częstotliwości

h — grubość płytki [mm]

f — częstotliwość [kHz]

Dla rezonatora cięcia SC o orientacji $\phi = 22^\circ 10'$ i $\theta = 34^\circ 05'$ ustalono zależność częstotliwości od promienia krzywizny R i grubości h wyrażoną wzorem:

$$N' = N + 280 \frac{d_1}{R} \quad (4)$$

gdzie:

$$d_1 = 2\sqrt{2R - h^2} \quad (5)$$

N — stała częstotliwości płytki nieskończenie rozciąglej (dla przedstawionej orientacji wynosi ona 1800)

Dokładność wyznaczenia parametrów płytki może być określona w programie iteracyjnym. Wystarczające praktycznie jest ustalenie promienia krzywizny z dokładnością 1 mm a grubości płytki z dokładnością 0,1 μm .

4.2. WYKONANIE MECHANICZNE

Obróbka mechaniczna płytek polega na stopniowym szlifowaniu płasko-równoległym, a w końcowym etapie na szlifowaniu sfery z dokładnością rzędu 0,5 μm . Określenie tego parametru jest umowne gdyż w rzeczywistości nie mierzy się rzeczywistej grubości płytek a jedynie częstotliwość jej drgań pomiędzy elektrodami wytwarzającymi pole prostopadłe, lub w niektórych procesach pole równoległe do powierzchni płytek. Częstotliwość drgań płytki mierzy się z dokładnością rzędu kilkuset herców a ustala z dokładnością rzędu 5÷25 kHz zależnie od częstotliwości wykonywanych wibratorów i rzędu drgań.

4.3. POLEROWANIE POWIERZCHNI

Dla rezonatorów cięcia SC stosuje się tak zwane polerowanie chemiczne polegające na odtrawianiu w odpowiednich roztworach trawiących warstwy kwarcu zdefektowanej

okrągłej
metoda

(3)

ależność

(4)

(5)

orientacji

ogramie
adnością

ównoleg-

kreślenie

grubości

ymi pole

i płytek.

a ustala

bratorów

olegające

ktowanej

podczas obróbki mechanicznej. Zwykle do trawienia stosuje się nasycone roztwory kwaśnego fluorku amonu i kwasu fluorowodorowego z dodatkiem środków powierzchniowo czynnych. Proces prowadzi się w podwyższonej temperaturze rzędu $50 \div 80^\circ\text{C}$. Poza wygładzeniem powierzchni kwarcu ma on również na celu zmniejszenie rozrzutów częstotliwości płytek. Po szlifowaniu mierzy się częstotliwość płytek i na podstawie znajomości szybkości trawienia ustala się czas niezbędny do odtrawienia warstwy nieznacznie większej niż głębokość defektów powierzchni spowodowanych szlifowaniem. Zwykle odtrawia się warstwę grubości co najmniej $3\mu\text{m}$ na stronę. Czas trawienia ustala się tak aby dla najcieńszej płytki ze wsadu uzyskać po tym procesie częstotliwość wymagana na danym etapie procesu technologicznego. Wsad płytek wyjmuje się z kąpeli trawiącej myje, suszy i mierzy ich częstotliwość. Płytki, które osiągnęły częstotliwość nominalną w zadanych tolerancjach odłącza się od wsadu, a pozostałe trawi się dalej dobierając czas trawienia tak aby następne płytki osiągnęły częstotliwość nominalną. Proces ten powtarza się do osiągnięcia zadanej częstotliwości przez wszystkie płytki. Można w ten sposób osiągnąć tolerancję częstotliwości płytek poniżej $\pm 1 \times 10^{-4}$ częstotliwości nominalnej.

4.4. NANOSZENIE ELEKTROD I MONTAŻ W PODSTAWKACH

Elektrody są nakładane na powierzchnię kwarcu przez próżniowe naparowanie warstwy metalu. W rezonatorach wysokostabilnych elektrody wykonuje się ze złota na podkładzie chromu lub tytanu. Grubość elektrod zależy od rzędu drgań rezonatora. Dla drgań podstawowych stosuje się grubość elektrod w zakresie $100 \div 150\text{ nm}$. Dla trzeciego drgania nadpodstawowego $70 \div 120\text{ nm}$, a dla piątego drgania nadpodstawowego $50 \div 80\text{ nm}$. Warstwa metalu jest naparowywana ze źródła przez szablony decydujące o kształcie i wymiarach elektrod. Ponieważ elektrody połączone są z wyprowadzeniami elektrycznymi, do których w następnym etapie montowane są sprężynki mocujące płytkę w podstawie, bardzo ważne jest takie ustawienie płytki kwarcowej w szablonie, aby w końcowym efekcie uzyskać możliwie najmniejszy wpływ mocowania na parametry rezonatora. W rezonatorach klasycznych pobudzanych do drgań polem prostopadłym nanoszona warstwa metalizacji obniża częstotliwość drgań płytki kwarcowej zgodnie z zależnością:

$$\frac{\Delta f}{f} = - \frac{2\rho_{\text{met}} \cdot h_{\text{met}}}{\rho_k \cdot h_k} \quad (6)$$

gdzie: ρ_{met} — gęstość metalu

h_{met} — grubość warstwy

ρ_k — gęstość kwarcu

h_k — grubość płytki

Grubość naparowanej warstwy jest ustalana za pomocą miernika grubości warstwy i automatycznego układu wyłączającego proces parowania po osiągnięciu zadanej grubości warstwy. Niepowtarzalność grubości warstwy w różnych partiach rezonatorów o tej samej

częstotliwości znamionowej nie przekracza ± 50 Hz. Praktycznie można przyjąć, że po naparowaniu elektrod rozrzut częstotliwości wibratorów jest taki sam jak po trawieniu.

W przypadku rezonatorów pobudzanych polem równoległym elektrody znajdują się praktycznie poza obszarem drgań płytki. W związku z tym ich wpływ na częstotliwość drgań płytki kwarcowej jest znacznie mniejszy (od kilku do kilkudziesięciu razy). Również w tym przypadku można przyjąć, że rozrzut częstotliwości po naniesieniu elektrod jest taki sam jak po trawieniu. Po naparowaniu elektrod płytki montuje się w podstawach za pomocą kleju przewodzącego tak aby elektrody zostały trwale połączone z wyprowadzeniami elektrycznymi podstawki, a następnie bardzo starannie myje. Kierunek mocowania płytek musi być zachowany z bardzo dużą dokładnością (pojedynczych stopni). Parametry tak zmontowanego wibratora kwarcowego można mierzyć z dużą precyzją.

4.5. POMIAR CZĘSTOTLIWOŚCI I TEMPERATURY EKSTREMUM

Wykorzystanie odbiciowej metody pomiaru częstotliwości rezonatora [8] umożliwia dokładny pomiar częstotliwości umieszczonego w komorze o regulowanej temperaturze rezonatora dołączonego do układu pomiarowego za pomocą kabla koncentrycznego o impedancji 50Ω . Mierzając zmiany częstotliwości rezonatora w funkcji zmian temperatury uzyskuje się informację o częstotliwości wibratora w temperaturze ekstremum, czyli w temperaturze w jakiej docelowo ma pracować rezonator i dla której określana jest częstotliwość znamionowa rezonatora. Częstotliwość wibratora rezonatorów klasycznych, których kalibracja częstotliwości odbywa się przez dopalenie warstwy metalu elektrod na tym etapie procesu technologicznego jest zwykle wyższa od częstotliwości znamionowej rezonatora. W rezonatorach pobudzanych polem równoległym kalibracja częstotliwości jest realizowana przez odtrawianie warstwy kwarcu, a więc częstotliwość musi być niższa od znamionowej. Dokładny jej pomiar pozwala określić wielkość poprawki częstotliwości, o którą należy ją zmienić w procesie kalibracji. Na błąd wyznaczenia częstotliwości f_s rezonatora składa się błąd wyznaczenia częstotliwości odpowiadającej minimalnej wartości współczynnika odbicia dla f_r oraz występującego przesunięcia między częstotliwością f_s a f_r . Błąd ten można wyznaczyć korzystając z zależności:

$$f_s = f_r \left[1 - \frac{\frac{C_1}{C_0}}{2 \left[1 + \frac{1}{(z_0 2\pi f_r C_0)^2} \right]} \right] \quad (7)$$

gdzie: C_0, C_1 — parametry rezonatora zgodnie z rys.2.

z_0 — impedancja charakterystyczna układu pomiarowego.

Dla parametrów rezonatorów wysokostabilnych, przy starannym skalibrowaniu układu i typowych dokładnościach wyznaczania parametru f_r oraz uwzględnieniu niepewności wyznaczenia wartości C_0 i C_1 dokładność wyznaczenia f_s jest na poziomie kilku $\times 10^{-8}$. Dokładność wyznaczenia temperatury ekstremum typowo jest nie gorsza niż $\pm 0,2$ K. Tak więc praktycznie błędy popełniane przy tej operacji są pomijalne.

4.6. KALIBRACJA CZĘSTOTLIWOŚCI

Proces kalibracji rezonatorów klasycznych polega na naparowaniu na powierzchnię jednej lub obydwu elektrod rezonatora niewielkiej ilości złota. Pogrubienie elektrody zgodnie z zależnością (6) powoduje obniżenie częstotliwości drgań rezonatora. Proces naparowania prowadzi się przy równoczesnym pomiarze częstotliwości wibratora. Automatyczny układ sterowania wyłącza proces parowania po osiągnięciu przez rezonator zadanej częstotliwości. Dokładność ustalenia częstotliwości rezonatora jest wysoka. Praktycznie można osiągnąć powtarzalność ustalania częstotliwości rzędu 1 ppm.

Inaczej przedstawia się sprawa kalibracji częstotliwości w przypadku rezonatorów pobudzanych polem równoległym. Analogicznie jak przy naparowaniu elektrod ich pogrubianie nie wpływa a istotny sposób na częstotliwość rezonatora, czyli nie można stosować do ich kalibracji tej metody. W Instytucie Tele- i Radiotechnicznym do kalibracji częstotliwości tych rezonatorów zastosowano trawienie chemiczne. Metoda polega na tym, że elektrody naparowuje się na płaskiej stronie wibratora i montuje się go w podstawę stroną wypukłą pozbawioną elektrod do góry. Tak zmontowany rezonator umieszcza się w kąpeli trawiącej tak aby zanurzona była tylko wypukła powierzchnia wibratora. W tych warunkach nie może być mowy o dokładnym pomiarze częstotliwości podczas kalibracji, dlatego trawienie kalibracyjne wykonuje się na czas. Po upływie określonego czasu wibrator wyjmuje się z kąpeli, starannie myje, suszy i mierzy jego częstotliwość. Jeśli częstotliwość nie osiągnęła wymaganej wartości proces powtarza się. Ponieważ szybkość trawienia zależy od wielu czynników jest to proces niezbyt precyzyjny. Aby osiągnąć tolerancję wymaganą dla dostrojenia rezonatora w generatorze (patrz rozdział 3) proces kalibracji trzeba powtarzać kilkakrotnie. Dla osiągnięcia dużej precyzji ustalenia częstotliwości stosuje się roztwory trawiące o bardzo małej szybkości trawienia (rzędu 1 Hz/s). Czasy trawienia dobiera się tak aby nie przekroczyć zadanej tolerancji częstotliwości. W ostatnich etapach czasy poszczególnych odtrawień są rzędu pojedynczych sekund. Pomiar częstotliwości prowadzi się przy zastosowaniu poprzednio omówionej metody z takim samym błędem. Głównym źródłem błędów jest zmiana temperatury podczas pomiaru. Proces trawienia i pomiary prowadzone są w temperaturze pokojowej. Przy tej temperaturze zależność częstotliwości rezonatora SC od temperatury wynosi około $1 \times 10^{-6}/K$, czyli przykładowo przy zmianie temperatury otoczenia o $0,1^\circ C$ dla rezonatora 10 MHz następuje zmiana częstotliwości o 1 Hz. Uwzględniając błąd wynikający z racjonalnej dyskretyzacji procesu oraz błąd wynikający ze stałości warunków termicznych maksymalną osiągalną dokładność kalibracji można ocenić jako $\pm 5 \times 10^{-7}$.

4.7. HERMETYZACJA

Proces hermetyzacji polega na zamknięciu rezonatora w próżni metodą zimnego zgniotu. Podczas tego procesu następuje zmiana częstotliwości rezonatora na skutek zmiany obciążenia powierzchni wibratora gazem, oraz na skutek wprowadzenia dodatkowych pojemności pomiędzy elektrodami a obudową rezonatora. Obniżenie ciśnienia

gazu wewnątrz obudowy rezonatora powoduje zmniejszenie tłumienia drgań wibratora, a przez to wzrost jego częstotliwości, natomiast wprowadzenie dodatkowych pojemności zwiększa C_0 , a przez to obniża częstotliwość. Praktyczne zmiany częstotliwości rezonatorów klasycznych w zależności od częstotliwości podano w tabeli 4.

Tabela 4

Zmiana częstotliwości rezonatorów pobudzanych polem
prostopadłym podczas hermetyzacji

The classical resonator frequency change during
hermetization process

Częstotliwość [MHz]	Zmiana częstotliwości [Hz]
5	3÷5
10	10÷15
20	35÷45

W przypadku rezonatorów klasycznych względna zmiana C_0 jest niewielka, natomiast w przypadku rezonatorów z polem równoległym, które posiadają wielokrotnie niższe wartości C_0 (patrz tabela 1) czynnik ten nabiera znaczenia. Dodatkowo zwiększa ten wpływ fakt, że w rezonatorach pobudzanych polem równoległym elektrody tworzą układ otwarty (pole elektryczne nie zamyka się pomiędzy elektrodami a wychodzi poza płytkę kwarcową) i zewnętrzna metalowa osłona znajdująca się w bezpośrednim ich sąsiedztwie oddziałuje znacznie silniej na parametry wibratora niż w rezonatorze klasycznym. Wprawdzie parametry rezonatorów pobudzanych polem równoległym mierzy się przed hermetyzacją z nałożoną na wibrator osłonką, ale podczas hermetyzacji zmienia się jakość kontaktu elektrycznego pomiędzy osłonką a podstawą. Dodatkowo osłonka się obniża o około 0,1 mm i powoduje zmianę pojemności statycznej rezonatora, a przez to zmianę jego częstotliwości. Sumaryczna zmiana częstotliwości w rezonatorach pobudzanych polem równoległym jest jednak średnio o połowę mniejsza niż w rezonatorach klasycznych. Zjawisko to nie byłoby niebezpieczne gdyby było powtarzalne, jednak wysokość mocowania płytki kwarcowej w podstawie zmienia się zależnie od wykrepowania uchwytów, a przez to w jakimś stopniu jest niepowtarzalna. Przez to i zmiany częstotliwości podczas hermetyzacji są niepowtarzalne. Niepowtarzalność zmian częstotliwości w rezonatorach o częstotliwości 10 MHz na drganiach anharmonicznych dochodzi do pięciu herców a dla rezonatorów pracujących na drganiu głównym rzędu 2 Hz. Błąd ustalenia częstotliwości wynikający z hermetyzacji jest niestety błędem nienaprawialnym.

4.8. POMIARY KOŃCOWE

Pomiary końcowe parametrów rezonatorów prowadzone są analogicznie jak pomiary przed hermetyzacją. Jednak dzięki temu, że obudowa rezonatora jest zamknięta a wewnątrz

atora,
ności
onato-

panuje bardzo niskie ciśnienie rezonatory takie wykazują wyższą dobroć i nie są podatne na zaburzenia pomiaru przez czynniki zewnętrzne takie jak wilgotność powietrza, ewentualne zanieczyszczenia powierzchni elektrod czy niepowtarzalność pozycji pokrywki obudowy. Dzięki temu możliwe jest osiągnięcie jeszcze wyższej dokładności i powtarzalności pomiaru. Rezonatory, których częstotliwość wykracza poza dopuszczalne tolerancje stanowią braki.

5. PODSUMOWANIE

Uwzględniając rozrzut częstotliwości podczas kalibracji i zmiany częstotliwości podczas hermetyzacji można określić sumaryczny rozrzut częstotliwości rezonatorów. Porównanie tych wartości z możliwym do uzyskania zakresem przestrajania częstotliwości generatora daje ocenę możliwości uzyskania generatorów o zadanej częstotliwości.

Tabela 5

Porównanie parametrów przestrajania generatorów z możliwościami wykonawczymi rezonatorów

The comparison of oscillator tuning parameters to resonator fabrication possibilities

L.p.	Typ rezonatora	Zakres przestrajania [10 ⁻⁷]	Możliwości wykonawcze [10 ⁻⁷]
1	Pole prostopadłe mod 300	>50	±10
2	Pole równoległe mod 300	±3	±6
3	Pole równoległe mod 301	±5	±6,5
4	Pole równoległe mod 311	±7	±7

omiast
niższe
wpływu
otwarty
rcową)
działuje
awdzie
tyzacja
ontaktu
oło 0,1
ę jego
polem
cznych.
ysokość
owania
zmiany
częstot-
ochodzi
z. Błąd
alnym.

omiary
ewnątrz

Z porównania wymaganych dla uzyskania i utrzymania przez określony okres czasu częstotliwości znamionowej generatora tolerancji częstotliwości rezonatora i możliwych do uzyskania w praktyce wynika jednoznacznie, że jest to dla rezonatorów pobudzanych polem równoległym bardzo poważny problem. Rezonatory pobudzone polem równoległym do drgań modu głównego nie dają się skalibrować tak, aby wszystkie osiągnęły w generatorze pełny zakres przestrajalności. Nieco lepiej przedstawia się sytuacja rezonatorów na drganiach anharmonicznych. Wprawdzie występują większe zmiany częstotliwości przy hermetyzacji obudów, ale szerszy zakres przestrajania umożliwia w większości przypadków dostrojenie generatora do wymaganej częstotliwości. Praktycznie przy wykonywaniu generatorów o częstotliwości 10MHz udało się tak skalibrować częstotliwość rezonatorów aby uzyskać dostrojenie do częstotliwości znamionowej dla 40% rezonatorów wykorzystujących drgania modu głównego, 60% na drganiach modu 301 i 100% rezonatorów na drganiach modu 311. Dla rezonatorów na drganiach modu

głównego osiągnięcie lepszych wyników jest mało prawdopodobne, natomiast dla rezonatorów na drganiach anharmonicznych można osiągnąć wyższy uzysk jeżeli uda się zminimalizować niepowtarzalność zmian częstotliwości w procesie hermetyzacji obudów. Ponieważ rezonatory pobudzone polem równoległym w porównaniu z rezonatorami klasycznymi wykazują zdecydowanie lepsze parametry użytkowe wdrożyć je do dalszej pracy dla poprawy uzysku. Ale nawet przy obecnym uzysku produkcja tych rezonatorów powinna być opłacalna, gdyż ich parametry (głównie starzenie i poziom szumów fazowych) nieznacznie tylko ustępują parametrom opracowanych we Francji a produkowanych jedynie w Szwajcarii rezonatorów BVA, których pracochłonność, a przez to i cena wielokrotnie przewyższa pracochłonność rezonatorów pobudzanych polem równoległym.

6. BIBLIOGRAFIA

1. R.J. Besson: *A new „electrodeless” resonator design*. Proc. 31st AFCS 1977.
2. A.W. Warner: *Use of parallel field excitation in design of quartz crystal units*. Proc. 17th AFCS. 1963
3. J.R. Vig, T.R. Meeke r: *The aging of bulk acoustic wave resonators, filters and oscillators*. Proc. 45th ASFC 1991.
4. K. Weiss, W. Szulc, M. Świdorski: *Rezonatory cięcia SC pobudzone polem równoległym do drgań anharmonicznych*. Kwartalnik Elektroniki i Telekomunikacji nr 3/2000, ss.431-452
5. K. Weiss, W. Szulc, B. Gniewińska, A. Masiukiewicz, B. Kalinowska: *Parameters comparison of different resonators excited by lateral field*. Proc. Joint Meeting EFTF-FCS. 1999.
6. B. Kalinowska, B. Gniewińska, M. Wójcicki: *Analiza długookresowego zachowania generatorów i opracowanie metod predykcji starzenia*. Sprawozdanie wewnętrzne ITR z pracy statutowej B3-S-97. Zadanie C. 1997.
7. K. Weiss: *A simple method of SC resonators design*. IEEE Transactions on Ultrasonics, Ferroelectrics and Frequency Control. Vol.48, No.6, Nov. 2001.
8. M. Wójcicki, A. Lisowiec: *A method for determination of basic electrical parameters of piezoelectric resonators in wide frequency range*. Electron Technology, 27, 3/4, pp.143-154, Institute of Electron Technology. Warszawa 1994.

K. WEISS, B. GNIEWIŃSKA, A. LISOWIEC

The problems of resonant frequency adjusting of the oscillators utilizing lateral field excited quartz resonators

Summary

The paper presents further results of research work done at ITR on lateral field excited anharmonic mode resonators, whose first results were presented in the paper published in Electronic and Telecommunications Quarterly no.3/2000. In the paper the problem of oscillator with LFE resonator tuning has been described. To estimate the problem, the influence of technological process on resonator calibration accuracy has been outlined. Comparison of parameters and their behavior during frequency settling of LFE normal and anharmonic mode resonators with classical perpendicular field excited main mode resonators has also been presented

Keywords: quartz resonator, lateral field, high stability oscillator

Modelowanie cewki z rdzeniem ferrytowym w programie SPICE z uwzględnieniem samonagrzewania

KRZYSZTOF GÓRECKI

*Katedra Radioelektroniki Morskiej, Akademia Morska w Gdyni
ul. Morska 83, 81-225 Gdynia
e-mail: gorecki@am.gdynia.pl*

*Otrzymano 2002.11.07
Autoryzowano 2003.01.18*

W pracy zaproponowano nowy elektrotermiczny model cewki z rdzeniem ferromagnetycznym, bazujący na izotermicznym modelu Jilesa-Athertona. Opracowany model uwzględnia takie właściwości rdzenia ferromagnetycznego jak pętla histerezy, nasycenie rdzenia, temperatura Curie oraz samonagrzewanie, spowodowane stratami energii zarówno w rdzeniu, jak i w uzwojeniu. Nowy model zaimplementowano w programie SPICE w postaci podukładu i porównano wyniki obliczeń z danymi katalogowymi. Niezbędne wartości parametrów elektrycznych i termicznych wyznaczono na podstawie odpowiednich pomiarów oraz danych katalogowych materiałów ferromagnetycznych.

Słowa kluczowe: rdzenie ferromagnetyczne, SPICE, analiza elektrotermiczna

1. WPROWADZENIE

Cewki są elementami składowymi wielu układów elektronicznych, np. obwodów rezonansowych i filtrów, jak również układów energoelektronicznych, np. stabilizatorów impulsowych, w których pełnią one rolę elementów przetwarzających energię elektryczną w energię pola magnetycznego i na odwrót. Duże wartości indukcyjności cewek przy zachowaniu niewielkich rozmiarów geometrycznych uzyskuje się dzięki wykorzystaniu rdzeni ferromagnetycznych.

Zastosowanie rdzenia ferromagnetycznego, dzięki jego dużej przenikalności magnetycznej, umożliwia zmniejszenie wymiarów geometrycznych cewki, z drugiej jednak strony prowadzi do nieliniowych charakterystyk cewki oraz zależności jej indukcyjności m. in. od wartości płynącego przez nią prądu [1]. Zależność ta jest istotna zwłaszcza

w układach dławikowych przetwornic dc-dc, w których wartość indukcyjności dławika determinuje zakres dopuszczalnych zmian prądu obciążenia. Dodatkowo, w układach mocy poważny problem stanowi odprowadzanie ciepła wydzielanego w rdzeniu [2] w czasie pracy cewki oraz silny wpływ temperatury na właściwości i parametry materiałów ferromagnetycznych [3, 4].

Z drugiej strony, przy projektowaniu i analizie układów elektronicznych coraz powszechniej wykorzystywane są komputerowe narzędzia symulacyjne i modele elementów elektronicznych. Od dokładności tych modeli zależy poprawność i dokładność uzyskanych wyników obliczeń. Obecnie jednym z najpopularniejszych programów do analizy układów elektronicznych jest program SPICE, posiadający wbudowane modele podstawowych elementów elektronicznych oraz bogate biblioteki zawierające wartości parametrów tych modeli, w tym parametry modeli wybranych rdzeni ferromagnetycznych.

Wykorzystywany w programie SPICE model indukcyjności nieliniowej (z rdzeniem ferromagnetycznym), opisany w pracach [5, 6], uwzględnia zjawisko samoindukcji oraz podstawowe właściwości rdzenia ferromagnetycznego, np. pętlę histerezy charakterystyki $B(H)$, natomiast nie uwzględnia strat energii w uzwojeniu ani w rdzeniu, a także nie uwzględnia wpływu temperatury na właściwości rdzenia. Pominięcie tych zjawisk powoduje niepoprawne modelowanie charakterystyk oraz parametrów cewek w zakresie wysokich temperatur.

Z praktyki inżynierskiej oraz z badań przeprowadzonych przez autora wynika, że na skutek wydzielania w cewce energii elektrycznej i zamiany jej na ciepło (efekt samonagrzewania), temperatura wnętrza cewki może różnić się znacznie od temperatury otoczenia, przekraczając nawet niekiedy temperaturę Curie. Z drugiej strony, jak wynika z prac [4, 7, 8], temperatura wpływa w istotny sposób na właściwości rdzenia ferromagnetycznego, np. na przenikalność magnetyczną, a zatem efekt samonagrzewania powinien być uwzględniony w modelu cewki.

W pracy zaproponowano nowy elektrotermiczny model cewki z rdzeniem ferromagnetycznym (EMCR), uwzględniający samonagrzewanie, wykorzystując przy jego tworzeniu strategię postępowania, stosowaną przy formułowaniu elektrotermicznych modeli elementów półprzewodnikowych [9]. Zgodnie z tą strategią, stworzenie modelu elektrotermicznego wymaga sformułowania zależności, opisujących charakterystyki izotermiczne cewki, słuszne dla całego dopuszczalnego zakresu zmian temperatury, następnie modelu termicznego, stanowiącego zależność temperatury wnętrza cewki od wydzielanej w niej mocy oraz modelu mocy czynnej, opisującego zależność tej wielkości od napięć i prądów zaciskowych cewki oraz od indukcji pola magnetycznego w rdzeniu tej cewki. Przy formułowaniu modelu izotermicznego cewki posłużono się izotermicznym modelem rdzenia ferromagnetycznego z pracy [5].

W rozdziale drugim przedstawiono równania opisujące model cewki z rdzeniem ferromagnetycznym, w rozdziale trzecim zaprezentowano sposób implementacji tego modelu w programie SPICE, w rozdziale czwartym omówiono sposób estymacji wartości parametrów modelu, natomiast rozdział piąty zawiera wyniki pomiarów i obliczeń charakterystyk cewek uzyskanych za pomocą EMCR oraz modelu wbudowanego w programie SPICE.

2. POSTAĆ MODELU

Jak wynika z fizycznych podstaw magnetyzmu [10, 11], prąd i płynący przez cewkę o długości l , posiadającą z zwojów, wytwarza wewnątrz cewki pole magnetyczne o natężeniu H danym wzorem

$$H = z \cdot i / l \quad (1)$$

Pole to wywołuje indukcję pola magnetycznego B opisaną zależnością

$$B = \begin{cases} \mu_0 \cdot (H + M) & \text{dla } T_r < T_c \\ \mu_0 \cdot H & \text{dla } T_r \geq T_c \end{cases} \quad (2)$$

gdzie μ_0 to przenikalność magnetyczna próżni, M jest magnetyzacją, charakteryzującą stopień uporządkowania domen magnetycznych wewnątrz rdzenia, T_r oznacza temperaturę rdzenia, natomiast T_c — temperaturę Curie. Stopień uporządkowania tych domen zależy od natężenia pola magnetycznego H . Zależność $M(H)$ odpowiadająca pętli histerezy przedstawiana jest w literaturze za pomocą różnych wzorów [5, 12, 13, 14]. Niektóre spośród tych wzorów, np. zawierające funkcje wymierne lub logarytmiczne, są słuszne jedynie w zakresie dodatnich wartości natężenia H [12], natomiast inne, zawierające w swym opisie funkcje hiperboliczne lub wielomiany, spełniają kryterium nieparzystości funkcji $M(H)$.

W programie SPICE stosowany jest model rdzenia ferromagnetycznego, stanowiący zmodyfikowaną wersję modelu Jiles'a-Atherton'a [5], w którym zależność pomiędzy magnetyzacją a natężeniem pola magnetycznego opisana jest równaniem różniczkowym o postaci [6]

$$\frac{dM}{dH} = \frac{M_a - M}{(1 + C) \cdot \delta \cdot K} + \frac{C}{1 + C} \cdot \frac{dM_a}{dH} \quad (3)$$

gdzie M_a oznacza magnetyzację wyznaczoną na podstawie przebiegu krzywej pierwotnego magnesowania, C — stałą elastycznych odkształceń ścian domen, K — stałą nieelastycznych odkształceń ścian domen, natomiast δ jest równa 1, gdy natężenie pola magnetycznego rośnie oraz -1, gdy natężenie to maleje.

Magnetyzacja dla rdzenia nienamagnesowanego opisana jest zależnością [5, 6]

$$M_a = M_s \cdot F\left(\frac{H + \alpha \cdot M_a}{a}\right) \quad (4)$$

gdzie M_s oznacza magnetyzację w nasyceniu, α — średni parametr pola, a - parametr kształtu, natomiast funkcja $F(\cdot)$ zdefiniowana jest następującym wzorem

$$F(x) = \operatorname{ctgh}(x) - x^{-1} \quad (5)$$

Jak wynika z badań autora [15], stosowanie zależności (5) w modelu cewki przy małych wartościach natężenia pola magnetycznego H oraz dokładności obliczeń stosowanej standardowo w programie SPICE, powoduje wystąpienie oscylacji numerycznych. Z tego względu w opracowanym przez autora modelu cewki wykorzystano opis zależności $M_d(H)$ za pomocą wzoru (4), przy czym funkcję $F(\cdot)$ wyrażono za pomocą sumy funkcji eksponencjalnych [15]

$$F(x) = \operatorname{sgn}(x) \cdot \left[1 - 0,9 \cdot \exp\left(-\frac{|x|}{2,5}\right) - 0,1 \cdot \exp\left(-\frac{|x|}{25}\right) \right] \quad (6)$$

gdzie funkcja $\operatorname{sgn}(x)$ jest równa 1 dla $x > 0$ oraz -1 dla $x < 0$.

Zaproponowana postać funkcji $F(\cdot)$, wyrażona wzorem (6) umożliwia uzyskanie wyników obliczeń bez oscylacji, a różnice między wartościami funkcji $F(x)$ uzyskanymi z wzorów (5) i (6) w całym zakresie zmian argumentu x nie przekraczają 2%.

Występujące we wzorach (3, 4) parametry M_s , C , K , a , α zależne są od przebiegu pętli histerezy oraz parametrów katalogowych rdzenia. Szczegółowe wzory, opisujące te zależności, mają następującą postać [6]

$$C = \frac{\mu_i - 1}{\frac{B_r}{\mu_0 \cdot H_c} - \mu_i} \quad (7)$$

$$K = H_c \quad (8)$$

$$M_s = \frac{B_s}{\mu_0} - H_s \quad (9)$$

$$\alpha = \frac{3 \cdot a}{M_s} - \frac{C}{(\mu_i - 1) \cdot (1 + C)} \quad (10)$$

$$\alpha = \frac{H_x - C \cdot \frac{M_x}{(\mu_i - 1) \cdot (1 + C)}}{\frac{1}{1 - \frac{M_x}{M_s}} - 3 \cdot \frac{M_x}{M_s}} \quad (11)$$

gdzie μ_i oznacza przenikalność początkową rdzenia, B_r – indukcję szczątkową, H_c – pole powściągające, B_s – indukcję nasycenia przy polu H_s , natomiast M_x oraz H_x oznaczają współrzędne punktu leżącego na krzywej pierwotnego magnesowania.

Występujące w powyższych zależnościach współrzędne punktów, leżących na pętli histerezy są zależne od temperatury T [4, 7, 8]. W niniejszej pracy przyjęto, że zależności te mają charakter liniowy i opisane są następującymi wzorami

$$B_r(T) = B_{r0} \cdot [1 + \alpha_r \cdot (T - T_0)] \quad (12)$$

$$B_s(T) = B_{s0} \cdot [1 + \alpha_s \cdot (T - T_0)] \quad (13)$$

$$H_c(T) = H_{c0} \cdot [1 + \alpha_c \cdot (T - T_0)] \quad (14)$$

gdzie B_{r0} , B_{s0} , H_{c0} oznaczają odpowiednio indukcję szczątkową, indukcję nasycenia oraz pole powściągające w temperaturze odniesienia T_0 , natomiast α_r , α_s oraz α_c to temperaturowe współczynniki zmian tych wielkości.

Zależność przenikalności początkowej μ_i od temperatury ma różny przebieg dla różnych rdzeni [4]. Na podstawie analizy przebiegu krzywych $\mu_i(T)$, zamieszczonych w danych katalogowych wielu typów rdzeni ferromagnetycznych [8], przyjęto odcinkami liniowy model rozważanej zależności o następującej postaci

$$\mu_i(T) = \begin{cases} \mu_{i0} \cdot [1 + \alpha_v \cdot (T - T_0)] & \text{dla } T < T_c \\ \mu_{i0} \cdot [1 + \alpha_v \cdot (T_c - T_0)] \cdot [1 - 0,06 \cdot (T - T_c)] & \text{dla } T_c < T < T_c + 17 \\ 1 & \text{dla } T > T_c + 17 \end{cases} \quad (15)$$

gdzie μ_{i0} oznacza początkową przenikalność magnetyczną rdzenia w temperaturze odniesienia T_0 , natomiast α_v oznacza temperaturowy współczynnik zmian przenikalności początkowej.

Zaindukowana w cewce siła elektromotoryczna e jest proporcjonalna do szybkości zmian strumienia pola magnetycznego, który z kolei jest proporcjonalny do wektora indukcji pola magnetycznego B , co ilustruje następująca zależność

$$e = -z \cdot S \cdot dB/dt \quad (16)$$

gdzie S oznacza pole przekroju czynnego rdzenia.

Powyższe rozważania są słuszne dla cewki bezstratnej. Jednak w rzeczywistej cewce występują straty energii elektrycznej spowodowane nieidealnością materiałów, z których została ona wykonana. Można wyróżnić trzy składniki strat energii w cewce: straty w uzwojeniu, straty w karkasie oraz straty w rdzeniu. Straty w karkasie mają charakter strat dielektrycznych i są zwykle pomijalnie małe. Straty w uzwojeniu wynikają ze skończonej rezystancji przewodu oraz ze zjawiska naskórkowości, którego znaczenie rośnie wraz ze wzrostem częstotliwości f prądu płynącego przez cewkę. Rezystancja R reprezentująca całkowite straty w uzwojeniu jest wyrażona wzorem [15]

$$R = \rho \cdot (1 + \alpha \cdot (T - T_0)) \cdot \frac{l_p}{S_p} \cdot \left(1 + \frac{\sqrt{S_p \cdot f \cdot \mu}}{2} \right) \quad (17)$$

gdzie S_p oznacza pole przekroju poprzecznego przewodu, l_p – długość tego przewodu, ρ – rezystywność przewodu, μ – przenikalność magnetyczną przewodu, α – temperaturowy współczynnik zmian rezystywności.

Z kolei straty w rdzeniu wynikają z istnienia pętli histerezy oraz z prądów wirowych [7, 13, 16], przy czym w rdzeniach ferrytowych, ze względu na dużą rezystywność, straty na prądy wirowe są pomijalnie małe [7], natomiast straty energii W_h , wynikające z istnienia pętli histerezy są proporcjonalne do powierzchni pętli histerezy oraz do objętości czynnej ferromagnetyka V_e .

W katalogach [7, 8] definiuje się stratność w rdzeniu $tg\delta$, która może być wykorzystywana do oceny strat w rdzeniu przy małych wartościach pola magnetycznego, natomiast w zakresie dużych pól operuje się pojęciem gęstości mocy strat w rdzeniu p_v wyrażonej w W/m^3 . Z przedstawionych w katalogach wykresów zależności $p_v(f, B, T)$ widać, że p_v rośnie ze wzrostem częstotliwości f oraz wraz ze wzrostem amplitudy składowej zmiennej B_m indukcji pola magnetycznego, natomiast zależność od temperatury może być przybliżona funkcją kwadratową. Do opisu zależności $p_v(f, B, T)$ zastosowano zależność empiryczną o postaci

$$p_v = A \cdot f^a \cdot B_m^b \cdot [1 - D \cdot (T - T_m)^2] \quad (18)$$

gdzie A , D , a , b , T_m są parametrami modelu. Typowo wartość parametru $a \approx 1$, natomiast $b \approx 2$ [4].

Energia elektryczna wydzielana w uzwojeniu i w rdzeniu na skutek samonagrzewania zamieniana jest na ciepło i powoduje wzrost temperatury cewki, tzn. zarówno uzwojenia, jak i rdzenia. Ponieważ rdzeń i uzwojenie mają strukturę jednorodną, a zjawiska decydujące o wydzielaniu mocy zachodzą z taką samą intensywnością w całej ich objętości, uzasadnione jest przyjęcie skupionego modelu termicznego cewki, w którym występują dwa elementy termicznie czynne — rdzeń i uzwojenie. Wydzielana w każdym z tych elementów moc powoduje wzrost temperatury zarówno samego elementu, w którym jest ona wydzielana (na skutek samonagrzewania), jak również drugiego elementu (na skutek wzajemnych sprzężeń termicznych). Ponieważ moce wydzielane w obu elementach są różne, różne będą również temperatury rdzenia T_{jr} oraz uzwojenia T_{ju} . Przyjęto postać skupionego modelu termicznego cewki z rdzeniem identyczną, jak dla elementów półprzewodnikowych umieszczonych na wspólnym podłożu. Model ten opisany jest za pomocą następującego układu równań zawierającego całki spłotu [9]

$$T_{jr} = T_a + \int_0^t Z_{rr}'(x) \cdot p_r(t-x) \cdot dx + \int_0^t Z_{ru}'(x) \cdot p_u(t-x) \cdot dx \quad (19)$$

$$T_{ju} = T_a + \int_0^t Z_{uu}'(x) \cdot p_u(t-x) \cdot dx + \int_0^t Z_{ur}'(x) \cdot p_r(t-x) \cdot dx \quad (20)$$

gdzie p_r oraz p_u oznaczają odpowiednio moce wydzielane w rdzeniu oraz w uzwojeniu, $Z_{rr}'(t)$ oraz $Z_{uu}'(t)$ to czasowe pochodne przejściowej impedancji termicznej rdzenia oraz uzwojenia, natomiast $Z_{ru}'(t)$ oraz $Z_{ur}'(t)$ to pochodne wzajemnych przejściowych impedancji termicznych między rdzeniem a uzwojeniem. Impedancje te mogą być opisane wzorem [9]

$$Z(t) = R_{th} \cdot \left[1 - \sum_{i=1}^N a_i \cdot \exp \left(-\frac{t}{\tau_{thi}} \right) \right] \quad (21)$$

gdzie R_{th} oznacza rezystancję termiczną, τ_{thi} – i-tą termiczną stałą czasową związaną ze współczynnikiem wagowym a_i , natomiast N oznacza ilość termicznych stałych czasowych występujących w modelu.

Tak więc, sformułowanie elektrotermicznego modelu cewki z rdzeniem ferromagnetycznym (EMCR) wymaga zastąpienia temperatury T , występującej we wzorach (12-15) oraz (18), temperaturą rdzenia T_{jr} , natomiast we wzorze (17) — temperaturą uzwojenia T_{ju} .

3. IMPLEMENTACJA MODELU W PROGRAMIE SPICE

Model EMCR opisują zależności, które ze względów formalnych nie mogą być w sposób bezpośredni zapisane w programie SPICE, np. pochodne dM/dH , czy też rezystancje zależne od częstotliwości, ponieważ w programie tym pochodne mogą być wyznaczane tylko względem czasu, natomiast rezystory są elementami liniowymi. A zatem, implementacja przedstawionego modelu w programie SPICE wymaga realizacji obwodowej niektórych funkcji występujących w tym modelu.

Przykładowo, wartości pochodnych wyznaczane są w układzie, stanowiącym równoległe połączenie pojemności oraz źródła napięciowego o wydajności odpowiadającej wielkości różniczkowanej. Prąd płynący w tym układzie jest wprost proporcjonalny do pochodnej napięcia panującego na źródle napięciowym.

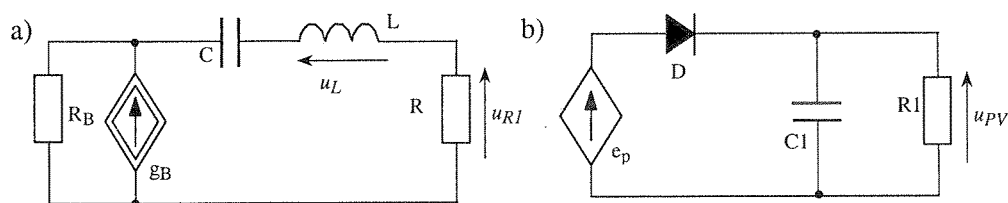
Z kolei, wartość magnetyzacji, opisanej wzorem (3) wyliczana jest w układzie stanowiącym połączenie równoległe kondensatora o jednostkowej pojemności oraz źródła prądowego o wydajności odpowiadającej iloczynowi prawej strony wzoru (3) i czasowej pochodnej natężenia pola magnetycznego H . Napięcie panujące na zaciskach źródła prądowego odpowiada magnetyzacji M . Ze względu na ograniczenia programu SPICE, który nie dopuszcza nieskończonej rezystancji między węzłami układu a punktem odniesienia, równoległe do źródła prądowego włączono rezystor o rezystancji $1 \text{ M}\Omega$.

Z uwagi na wymienione na początku niniejszego rozdziału ograniczenia programu SPICE, wprowadzenie do zaproponowanego modelu rezystancji R (wzór (17)), reprezentującej straty w uzwojeniu, wymagało wyznaczenia odwrotnej transformaty Laplace'a napięcia panującego na tej rezystancji. W wyniku otrzymano następującą zależność, opisującą wydajność źródła napięciowego u_R modelującego spadek napięcia na rezystancji R

$$u_R = \rho \cdot (1 + \alpha \cdot (T_j - T_0)) \cdot \frac{l_p}{S_p} \cdot \left(i + \frac{\sqrt{S_p \cdot u \cdot p}}{4 \cdot \pi} \cdot \frac{di}{dt} * \frac{1}{\sqrt{\pi \cdot t}} \right) \quad (22)$$

przy czym wartość spłotu $\frac{di}{dt} * \frac{1}{\sqrt{\pi \cdot t}}$ wyliczana jest w obwodzie pomocniczym

Ponieważ w zależności (18), opisującej straty energii w rdzeniu, występuje częstotliwość i amplituda indukcji, które nie mogą być wyznaczone bezpośrednio w analizie stanów przejściowych, w celu wyznaczenia strat w rdzeniu p_v konieczne było wprowadzenie obwodu pomocniczego o strukturze przedstawionej na rys.1.



Rys. 1. Realizacja układowa obwodu do wyznaczania stratności rdzenia

Fig. 1. The realisation of the circuit for calculations of the core losses

Na rys. 1a, wydajność prądowa źródła g_B równa jest indukcji pola magnetycznego B . Kondensator C umożliwia odseparowanie składowej stałej indukcji, a zatem przez cewkę L oraz rezystor R płynie prąd równy co do wartości składowej zmiennej indukcji. Jeżeli prąd ten ma przebieg okresowy, to można go opisać za pomocą sumy składników harmoniczných. Dla każdej częstotliwości harmoniczných f_i , napięcie u_L na cewce L można opisać wzorem

$$u_L(t) = 2 \cdot \pi \cdot f_i \cdot L \cdot B_m \cdot \sin(2 \cdot \pi \cdot f_i \cdot t) \quad (23)$$

natomiast prąd i_L tej cewki wynosi

$$i_L(t) = \frac{B_m \cdot \cos(2 \cdot \pi \cdot f_i \cdot t)}{R} \quad (24)$$

Wielkość pomocnicza p , stanowiąca iloczyn napięcia u_L oraz prądu i_L dana jest zależnością

$$p(t) = u_L \cdot i_L = \frac{B_m^2}{R} \cdot \pi \cdot f_i \cdot L \cdot \sin(4 \cdot \pi \cdot f_i \cdot t) \quad (25)$$

A zatem amplituda sygnału p odpowiada gęstości mocy strat w rdzeniu p_v danej wzorem (18), przy czym $a = 1$, $b = 2$, $A = \pi \cdot L/R$. Zastosowanie klasycznego detektora wartości szczytowej (rys. 1b), złożonego ze źródła napięcia e_p o wydajności

$$e_p = 2 \cdot u_{R1} \cdot u_L \cdot (1 + D \cdot (T - T_m)), \quad (26)$$

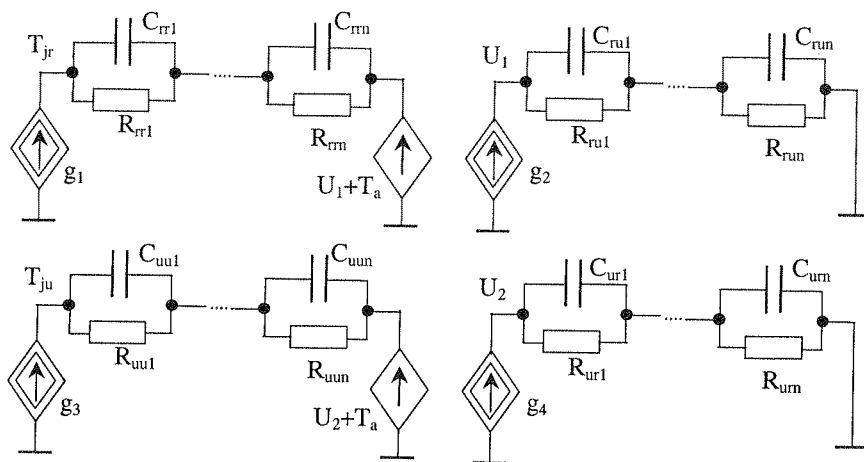
diody D , kondensatora $C1$ oraz rezystora $R1$, umożliwia uzyskanie pożądanego przebiegu gęstości mocy strat w rdzeniu równej co do wartości napięciu u_{PV} na rezystorze $R1$.

Straty energii, występujące w rdzeniu wpływają na dobroć cewki i są uwzględniane w jej schemacie zastępczym jako rezystancja szeregową R_h , na której spadek napięcia u_{Rh} opisany jest wzorem

$$u_{Rh} = \frac{p_V \cdot V_e}{i} \quad (27)$$

gdzie V_e jest objętością ekwiwalentną rdzenia, natomiast i oznacza prąd modelowanej cewki.

Model termiczny zrealizowano stosując analog elektryczno-termiczny. W reprezentacji obwodowej modelu termicznego, przedstawionej na rys.2, występują cztery generatory prądowe oraz cztery sieci RC. Wydajności dwóch spośród tych generatorów (g_1 oraz g_4) odpowiadają mocy wydzielanej w rdzeniu (iloczyn napięcia u_{PV} oraz objętości czynnej rdzenia V_e), natomiast wydajności dwóch pozostałych generatorów (g_2 oraz g_3) odpowiadają mocy wydzielanej w uzwojeniu (iloczyn prądu cewki i napięcia u_R). Obciążenie każdego źródła prądowego stanowi sieć reprezentująca przejściową impedancję termiczną własną ($Z_{rr}(t)$ dla źródła g_1 oraz $Z_{uu}(t)$ dla źródła g_3) lub wzajemną ($Z_{ru}(t)$ dla źródła g_2 oraz $Z_{ur}(t)$ dla źródła g_4). Napięcie na zaciskach źródła prądowego g_1 odpowiada temperaturze rdzenia T_{jr} , natomiast napięcie na zaciskach źródła prądowego g_3 odpowiada temperaturze uzwojenia T_{ju} .



Rys. 2. Reprezentacja obwodowa modelu termicznego cewki z rdzeniem ferromagnetycznym

Fig. 2. The circuital representation of the thermal model of the coil with a ferromagnetic core

Reprezentacja obwodowa modelu EMCR składa się z dwóch zasadniczych części. Obwód główny, włączony między zaciski zewnętrzne cewki, stanowi szeregowe połączenie trzech sterowanych źródeł napięciowych o wydajnościach odpowiadających odpowiednio sile elektromotorycznej samoindukcji e oraz spadkowi napięcia wynikającego ze strat w uzwojeniu u_R oraz strat w rdzeniu u_h . Z kolei obwód pomocniczy stanowią opisane wcześniej układy, umożliwiające wyznaczenie wartości odpowiednich pochodnych, całek oraz strat energii w rdzeniu i uzwojeniu, a także temperatur rdzenia i uzwojenia.

4. ESTYMACJA WARTOŚCI PARAMETRÓW

Parametry występujące w modelu rdzenia można wyznaczyć na podstawie danych katalogowych. W szczególności, z petli histerezy lub danych tabelarycznych dla temperatury pokojowej (typowo 25°C) odczytuje się wartości B_{so} , H_{co} , B_r . Na podstawie wartości parametrów B_s , H_c , B_r odpowiadających dwom temperaturom (typowo 25°C i 100°C) wyznacza się temperaturowe współczynniki zmian wymienionych wielkości α_s , α_c , α_r .

Z kolei z części rosnącej wykresu $\mu_i(T)$ wyznaczane jest nachylenie tej charakterystyki α_u , natomiast maksimum charakterystyki występuje przy temperaturze Curie T_c .

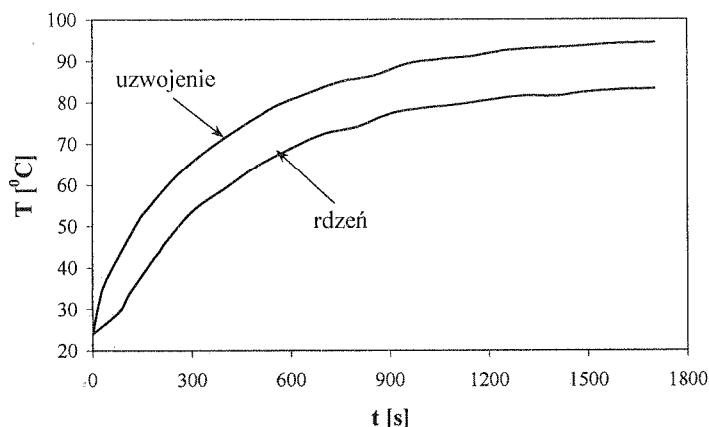
Na podstawie wykresów $p_v(T)$ wyznacza się wartość temperatury T_m , w której funkcja osiąga minimum. Na podstawie dwóch punktów leżących na charakterystykach $p_v(T)$ przy $B_m = \text{const}$ wyliczane są parametry A oraz D . Wartości pozostałych parametrów elektrycznych modelu odczytać można wprost z katalogu lub z tablic fizycznych.

Z kolei parametry modelu termicznego wyznaczono na drodze pomiarowej, pobudzając cewkę prądem stałym i rejestrując czasowe zmiany temperatury uzwojenia $T_{ju}(t)$ i rdzenia $T_{jr}(t)$ za pomocą kamery termograficznej. Zastosowano kamerę V-20 oraz urządzenie do ultra pomiarów temperatury firmy Vigo System [17, 18]. Poszczególne impedancje termiczne wyznaczono z zależności

$$Z_{uu}(t) = \frac{T_{ju}(t) - T_a}{P} \quad \text{oraz} \quad Z_{ur}(t) = \frac{T_{ru}(t) - T_a}{P} \quad (28)$$

Przyjęto, że własne impedancje termiczne rdzenia i uzwojenia są identyczne $Z_{rr} = Z_{uu}$, jak również identyczne są wzajemne przejściowe impedancje termiczne $Z_{ru} = Z_{ur}$. Parametry modelu poszczególnych impedancji termicznych wyznaczono metodą, opisaną w pracach [9, 19].

Z pomiarów rozkładów temperatury cewki umieszczonej w rdzeniu kubkowym [20] wynika, że w stanie ustalonym całe uzwojenie ma praktycznie jednakową temperaturę, wyższą od temperatury rdzenia, której rozkład przestrzenny także jest równomierny.



Rys. 3. Przebieg temperatury uzwojenia i rdzenia cewki w czasie nagrzewania mocą $P = 2,5 \text{ W}$

Fig. 3. The temperatures of the coil and the core while heating with power $P = 2.5 \text{ W}$

Czasowe przebiegi temperatury rdzenia i uzwojenia, uzyskane przy pobudzeniu uskokiem mocy o amplitudzie 2,5 W przedstawiono na rys.3.

Jak widać, w rozważanym przykładzie, stan termicznie ustalony występuje po upływie około 30 minut od chwili włączenia mocy. Widać też, że w początkowym okresie nagrzewania się cewki (do około 100 s), temperatura rdzenia rośnie wolniej niż temperatura uzwojenia.

5. WERYFIKACJA POPRAWNOŚCI MODELU

W celu zbadania praktycznej przydatności zaprezentowanego modelu cewki z rdzeniem (EMCR) przeprowadzono szereg analiz małosygnałowych oraz wielkosygnałowych prostego układu testowego, stanowiącego szeregowo połączenie prądowego źródła zasilania, badanej cewki oraz rezystora. Poprawność części obliczeń została zweryfikowana przez porównanie wyników obliczeń z charakterystykami katalogowymi wybranych rdzeni ferrytowych. Do badań wybrano dwa kubkowe rdzenie ferrytowe, wykonane z materiałów N26 oraz N27 [8], na których nawinięto uzwojenie złożone z 8 zwojów drutu miedzianego w emalii o średnicy 0,23 mm. Na podstawie danych katalogowych rdzeni wyznaczono następujące wartości parametrów rdzenia:

dla materiału N26: $T_C = 420$ K, $T_0 = 300$ K, $B_{r0} = 140$ mT, $H_{C0} = 23$ A/m, $\alpha_r = -1,9e-3$ K⁻¹, $\alpha_C = -1e-3$ K⁻¹, $\alpha_U = 4,33e-3$ K⁻¹, $B_{S0} = 380$ mT, $\alpha_S = -4,21e-3$ K⁻¹, $H_S = 1200$ A/m, $B_X = 345$ mT, $H_X = 190$ A/m, $\mu_i = 1950$,

dla materiału N27: $T_C = 490$ K, $T_0 = 300$ K, $D = 0,2e-3$ K⁻¹, $T_m = 353$ K, $B_{r0} = 250$ mT, $H_{C0} = 25$ A/m, $\alpha_r = -1e-3$ K⁻¹, $\alpha_C = -2,32e-3$ K⁻¹, $\alpha_U = 11,5e-3$ K⁻¹, $B_{S0} = 480$ mT, $\alpha_S = -1,67e-3$ K⁻¹, $H_S = 1200$ A/m, $B_X = 420$ mT, $H_X = 120$ A/m, $\mu_i = 2200$.

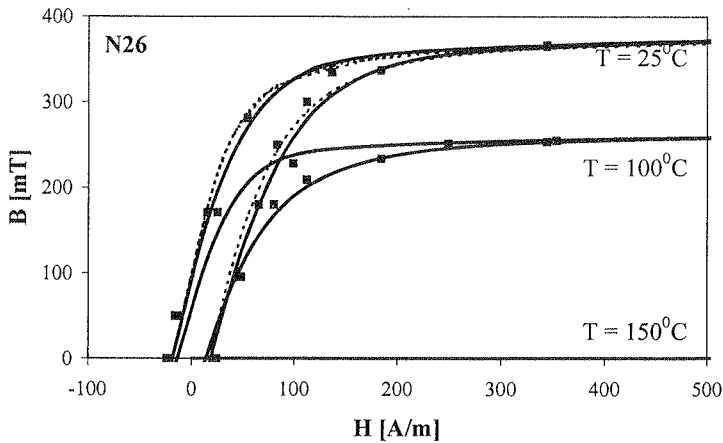
Dla obu cewek uzwojenie ma identyczne parametry o następujących wartościach: $l_p = 0,3$ m, $S_p = 0,04e-6$ m², $\rho_0 = 1,8e-8$ Ω m, $\alpha = 3e-4$ K⁻¹, $z = 8$. Również rozmiary geometryczne obu rdzeni są identyczne i opisane następującymi wartościami parametrów: $V_e = 2e-6$ m³, $S = 2,5e-4$ m², $l = 0,8e-2$ m.

Wyznaczone na podstawie pomiarów termograficznych wartości własnych rezystancji termicznych rdzenia i uzwojenia są równe $R_{thrr} = R_{thuu} = 28,2$ K/W, natomiast wzajemne rezystancje termiczne między rdzeniem a uzwojeniem są równe $R_{thru} = R_{thur} = 23,7$ K/W.

Na rys.4 przedstawiono izotermiczne (dla $R_{th} = 0$) charakterystyki $B(H)$ rdzenia z materiału N26, odpowiadające trzem wartościom temperatury otoczenia, równym 25°C, 100°C oraz 150°C, uzyskane za pomocą obliczeń przy wykorzystaniu modelu EMCR (linie ciągłe), modelu wbudowanego w programie SPICE [6] (linia kreskowa) oraz wyniki pomiarów katalogowych [8] (punkty). Z uwagi na symetrię pętli histerezy względem środka układu współrzędnych pokazano jej przebieg tylko w pierwszej i drugiej ćwiartce tego układu.

Jak widać, uzyskano zadawalającą zgodność wyników pomiarów oraz obliczeń, otrzymanych za pomocą obu modeli w temperaturze $T = 25^\circ\text{C}$. Model wbudowany w programie SPICE nie uwzględnia wpływu temperatury na charakterystyki cewki (krzywe wyznaczone dla trzech temperatur są nierozróżnialne), natomiast model EMCR prawidłowo odzwierciedla wyniki pomiarów. Dla temperatury $T = 150^\circ\text{C} > T_C$ rdzeń traci

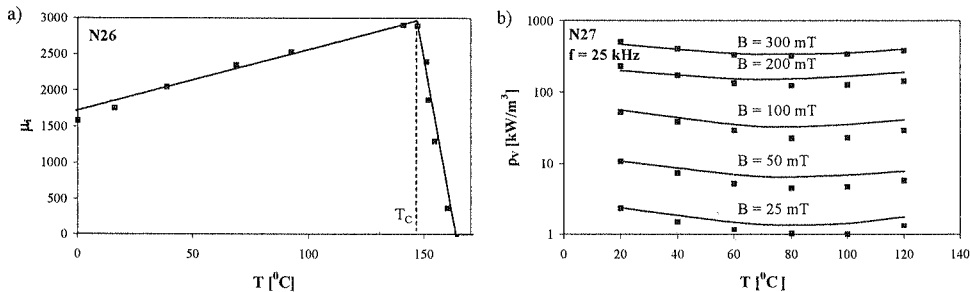
właściwości ferromagnetyczne, tzn. wartość indukcji B jest bliska zera i znika pętla histerezy (wykres pokrywa się z osią poziomą). Cewka z rdzeniem zachowuje się wówczas identycznie jak cewka powietrzna.



Rys. 4. Obliczone i pomierzone izotermiczne pętla histerezy dla rdzenia z materiału N26

Fig. 4. The calculated and measured isothermal hysteresis loops for the core from the material N26

Z kolei na rys.5 przedstawiono izotermiczne zależności wybranych parametrów rdzenia od temperatury otoczenia, uzyskane na podstawie pomiarów (punkty) oraz obliczeń przy użyciu EMCR (linie ciągłe). Rys. 5a dotyczy przenikalności początkowej, natomiast rys.5b - gęstości strat mocy w rdzeniu.



Rys. 5. Izotermiczne zależności temperaturowe przenikalności początkowej (a) oraz gęstości strat w rdzeniu (b)

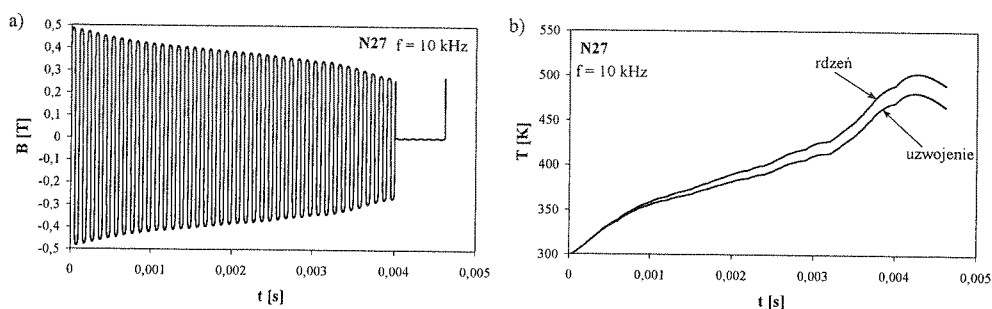
Fig. 5. The isothermal temperature dependences of the initial permeability (a) and the thickness of core losses (b)

Jak widać w obu przypadkach uzyskano zadawalającą zgodność wyników pomiarów katalogowych i obliczeń wykonanych przy wykorzystaniu EMCR. Przenikalność początkowa rośnie ze wzrostem temperatury aż do temperatury Curie (T_C), której

wartość, przyjęta w obliczeniach, wynosiła około 150°C , a następnie gwałtownie maleje do wartości odpowiadającej przenikalności próżni. Z kolei zależność $p_v(T)$ ma charakter paraboliczny, a krzywe odpowiadające różnym wartościom amplitudy indukcji pola magnetycznego B posiadają płaskie minimum przy temperaturze równej około 80°C .

Na kolejnych rysunkach przedstawiono wyniki analiz elektrotermicznych, zrealizowanych przy wykorzystaniu tego modelu. Na rys. 6 przedstawiono czasowy przebieg indukcji w rdzeniu przy pobudzeniu cewki z rdzeniem N27 prądem o przebiegu sinusoidalnym (rys. 6a) oraz przebiegi temperatury rdzenia i uzwojenia (rys. 6b). W analizach tych, w celu skrócenia czasu trwania obliczeń niezbędnych do uzyskania stanu ustalonego, przyjęto w modelu termicznym jedną termiczną stałą czasową o wartości $\tau_{th} = 1,5$ ms.

Jak widać na rys. 6b, w początkowym okresie, po włączeniu zasilania, na skutek samonagrzewania rosną temperatury zarówno rdzenia, jak i uzwojenia, powodując spadek wartości międzyszczytowej indukcji (rys. 6a). Przekroczenie przez temperaturę rdzenia wartości temperatury Curie, powoduje gwałtowny spadek indukcji do wartości bliskiej zera. W ślad za tym maleje moc tracona w rdzeniu, co powoduje spadek temperatury rdzenia. Gdy jej wartość spadnie poniżej temperatury Curie, rdzeń odzyska właściwości ferromagnetyczne co objawia się wzrostem wartości indukcji. Ponieważ w rozważanym przypadku dominują straty w rdzeniu, jego temperatura jest wyższa od temperatury uzwojenia.



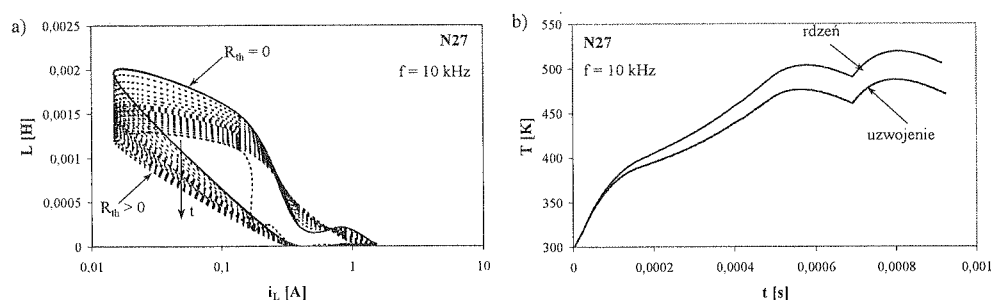
Rys. 6. Nieizotermiczne przebiegi indukcji (a) oraz temperatur rdzenia i uzwojenia (b)

Fig. 6. Nonisothermal traces of induction (a) and temperatures in the core and in the coil (b)

Warto przy tym zaznaczyć, że przy identycznym pobudzeniu w warunkach izotermicznych amplituda przebiegu $B(t)$ jest stała. A zatem, powyższy przykład dowodzi, że przy pracy cewki w temperaturze pokojowej, na skutek samonagrzewania, w istotny sposób mogą ulec zmianie jej charakterystyki.

Na skutek temperaturowych zmian w przebiegu pętli histerezy rdzenia, występują także zmiany indukcyjności cewki nawiniętej na tym rdzeniu, co zilustrowano na rys. 7a, gdzie linią ciągłą oznaczono izotermiczną zależność $L(i)$, natomiast linią kreskową — zależność nieizotermiczną. Jak widać, w obu przypadkach uzyskano zależność niejednoznaczną (występuje pętla na charakterystyce). Dodatkowo widać przesuwanie się nieizotermicznych charakterystyk $L(i)$ ku mniejszym wartościom indukcyjności wraz ze wzrostem

temperatury. Na rys. 7b przedstawiono obliczone przebiegi temperatur rdzenia i uzwojenia cewki, odpowiadające charakterystykom z rys. 7a. Zmiany indukcyjności cewki dla wybranego prądu mogą zmieniać się na skutek samonagrzewania nawet dwukrotnie.



Rys. 7. Izotermiczna i nieizotermiczna zależność indukcyjności cewki od prądu (a) oraz czasowe przebiegi temperatur (b)

Fig. 7. The isothermal and nonisothermal dependence of inductance from current (a) and time traces of temperatures (b)

6. UWAGI KOŃCOWE

W pracy zaproponowano nowy elektrotermiczny model cewki z rdzeniem ferromagnetycznym dla programu SPICE, bazujący na literaturowym izotermicznym modelu Jiles'a-Atherton'a [5]. Oprócz podstawowych zjawisk elektrycznych i magnetycznych, których opis ujęto w modelu z pracy [5], opracowany przez autora model EMCR uwzględnia dodatkowo wpływ na charakterystyki i parametry robocze cewki takich czynników jak: temperatura otoczenia, straty energii w rdzeniu oraz w uzwojeniu, a także samonagrzewanie.

Przedstawione na rys. 4 i 5 wyniki obliczeń wskazują, że opracowany przez autora model pozwala na otrzymanie izotermicznych charakterystyk cewki zgodnych z pomiarami i modelem literaturowym w pewnej temperaturze odniesienia oraz, w przeciwieństwie do modelu z pracy [5], poprawnie modeluje zmiany charakterystyk pod wpływem zmian temperatury otoczenia i samonagrzewania. Estymacja wartości parametrów nowego modelu jest prosta, gdyż większość parametrów modelu stanowią dane katalogowe rdzenia lub stałe materiałowe i rozmiary geometryczne drutu nawojowego oraz rdzenia. Przedstawione w pracy wyniki obliczeń nieizotermicznych wskazują na silny wpływ samonagrzewania na właściwości cewek. Wpływ ten może powodować zmiany przebiegu charakterystyk i wartości parametrów roboczych układów, zawierających cewki z rdzeniem ferromagnetycznym, np. dławikowych przetwornic dc-dc.

7. PODZIĘKOWANIA

Autor serdecznie dziękuje panom profesorom Januszowi Zarębskiemu i Witoldowi J. Stepowiczowi za cenne uwagi merytoryczne i redakcyjne, które w istotny sposób wpłynęły na ostateczną wersję niniejszej pracy.

8. BIBLIOGRAFIA

1. B. Grzesik, M. Stępień: *New Construction of High Frequency Transformer for Power Electronics Appliances*. V Krajowa Konferencja Naukowa Sterowanie w Energoelektronice i Napędzie Elektrycznym, Łódź, 2001, p. 167.
2. A. Van den Bossche, V. Valchev, J. Melkebeek: *Improved Thermal Modelling of Magnetic Components for Power Electronics*. EPE Journal, Vol. 12, No. 2, 2002, p. 7.
3. R.H. Coit: *Designing with Magnetic Cores at High Temperatures*. PCIM 2001 Europe Official Proc. of the 43th International Power Electronics Conference, Nurnberg, 2001, p. 607.
4. K. Górecki, J. Zarębski, J. Krupa: *Wpływ temperatury na właściwości cewki z rdzeniem ferromagnetycznym*. Elektronizacja, 2002. Nr 11, s. 21.
5. D.C. Jiles, D.L. Atherton: *Theory of ferromagnetic hysteresis*. Journal of Magnetism and Magnetic Materials, Vol. 61, 1986, p. 48.
6. J. Izydorczyk: *Pspice. Komputerowa symulacja układów elektronicznych*. Gliwice, Wydawnictwo Helion, 1993.
7. W. Dąbrowski, J. Król, M. Sobóń: *Materiały i rdzenie ferrytowe, ferryty magnetycznie miękkie ferroxyd*. Warszawa, Wydawnictwa Przemysłu maszynowego „Wema”, 1976.
8. SIFERRIT Materials. Materiały firmowe EPCOS AG, Munchen, 2000.
9. J. Zarębski: *Modelowanie, symulacja i pomiary przebiegów elektrotermicznych w elementach półprzewodnikowych i układach elektronicznych*. Prace Naukowe WSM w Gdyni, Gdynia 1996.
10. M. Herman, A. Kalestyński, L. Widomski: *Podstawy fizyki dla kandydatów na wyższe uczelnie*. Warszawa, PWN, 1984.
11. N. Mohan, T.M. Undeland, W.P. Robbins: *Power Electronics: Converters, Applications, and Design*. New York, John Wiley & Sons, 1995.
12. R. Kurdziel: *Obwody magnetyczne*. Warszawa, PWN, 1962.
13. J. Kuryłowicz: *Badania materiałów magnetycznych*. Warszawa, WNT, 1962.
14. J. Fuzi, E. Helerea: *Experimental Construction of Preisach Models for Ferromagnetic Cores*. International Conference ZM Communications GmbH (PCIM 1998), Power Conversion, Nurnberg, 1998, p. 661.
15. K. Górecki, J. Zarębski: *Elektrotermiczny model cewki z rdzeniem ferromagnetycznym dla programu SPICE*. XXV Międzynarodowa Konferencja z Podstaw Elektrotechniki i Teorii Obwodów IC-SPETO'2002, Ustroń, 2002, Vol. 2, s. 245.
16. Z. Celiński: *Materiałoznawstwo elektrotechniczne*. Warszawa, Oficyna Wydawnicza Politechniki Warszawskiej, 1998.
17. Instrukcja obsługi kamery termograficznej serii V-20, Vigo-System, Warszawa 2001.
18. Instrukcja obsługi urządzenia do ultraszybkich pomiarów temperatury, Vigo-System, Warszawa 2001.
19. J. Zarębski, K. Górecki: *Thermal Transients in Bipolar Transistors; Measurements and Simulations*. Int. Conf. on Information Systems, Methods Applied to Engineering Problems, Malta 1993, V. 2, p. 111.
20. T. Suchenia, M. Szczerkowski: *Badanie wpływu mocowania wybranych elementów półprzewodnikowych na ich rezystancję termiczną*. Praca dyplomowa inżynierska, Wydział Elektryczny, Akademia Morska w Gdyni, 2002.

K. GÓRECKI

SPICE-AIDED MODELLING OF COILS WITH THE FERRITE CORE
WITH SELFHEATING TAKEN INTO ACCOUNT

Summary

Coils are electronic devices which are often used in electronic small-power circuits, for example in filters, and also in high-power circuits, for example in switched dc-dc converters. The use of a ferromagnetic cores makes it possible to diminish the dimensions of a the coil, but it causes nonlinearity of their characteristics simultaneously. Due to the temperature effects, the strong temperature dependences of the coil characteristics are observed. In considerations presented in the paper the inner temperature of the coil, understood as the sum of the ambient temperature and the component resulting from selfheating, is taken into account. The aim of this paper is to formulate the model of the coil with the ferromagnetic core for SPICE.

Unfortunately, the isothermal model (i.e. without selfheating taken into account) of the coil with the ferromagnetic core, used in the popular software SPICE for the analysis of electronic circuits, omits the influence of temperature on the characteristics and the parameters of the core. In the paper the new electrothermal model of the coil with the ferromagnetic core (EMCR) designed for this software is proposed. The strategy of the formulation of EMCR model is the same as for the electrothermal models of semiconductor devices. According to this strategy, the coil isothermal model with the parameters dependent on temperature, based on the literature Jiles-Atherton model was elaborated at first. Next, the lumped thermal model which takes into account the power losses, both in the coil and in the core, resulting from selfheating in both coil components and additionally, the mutual thermal interactions between them was formulated.

EMCR includes the most important properties of the ferromagnetic core, as the hysteresis of $B(H)$ characteristic, the saturation of the core, Curie temperature and selfheating caused by energy losses, both in the core and in the coil, respectively.

The new EMCR model was implemented into SPICE software using a proper constructed subcircuit. Because of the limitations of SPICE software, some dependencies, existing in EMCR, describing for example the power losses in the core are realised by the special subcircuits. The lumped thermal model is represented by its electrical analog, consisting of the current sources representing power dissipated in the core and in the coil, and RC networks representing own transient thermal impedances of the coil and of the core, and the mutual transient thermal impedances between them.

At first, some isothermal calculations of the coil with the core characteristics using EMCR model and SPICE built-in model were performed with the ambient temperature as a parameter, without selfheating taken into account. The correctness of these calculations was compared with the catalogue data. The parameter values of the model were estimated on the basis of the catalogue data and the proper measurements. As it results from the investigations performed at the room temperature, the calculation results obtained with the use of both models fit well to the data, given in the catalogue. In the wide range of the ambient temperature only the results obtained with the use of EMCR fit well to the data, because SPICE built-in model of the ferromagnetic core does not include the influence of temperature on its characteristics and parameters.

Additionally, the electrothermal analyses of the coil with the core characteristics using EMCR model were performed. The results of these analyses show the strong influence of selfheating on the characteristics of the coil with the ferromagnetic core. This influence is observed as the decrease of the value of the coil inductance or the loss of ferromagnetic properties of the core as the result of exceeding Curie temperature. Of course, the change of the characteristics and parameter values of the coil as the result of selfheating can influence the characteristics of switched dc-dc converters.

Keywords: ferromagnetic core, SPICE, electrothermal analysis

System translacji automatycznej oparty na uogólnionych przykładach translacyjnych

MIROSLAW GAJER

*Katedra Automatyki AGH
al. Mickiewicza 30, 30-059 Kraków
mgajer@ia.agh.edu.pl*

*Otrzymano 2002.03.18
Autoryzowano 2002.08.27*

Translacja automatyczna o wysokiej jakości pomiędzy językami naturalnymi jest od wielu lat niedościgłym marzeniem dla wielu badaczy związanych z tą niezwykle pasjonującą, interdyscyplinarną dziedziną zastosowań informatyki. Translacja automatyczna była jedną z pierwszych aplikacji, jakie zostały zaproponowane dla komputerów, a pierwsze próby czynione w tym względzie pochodzą aż z lat pięćdziesiątych dwudziestego wieku. Niestety pomimo pierwszych sukcesów odniesionych w tej dziedzinie szybko okazało się, że zautomatyzowanie translacji jest zadaniem o wiele trudniejszym, ale zarazem i o wiele ciekawszym z naukowego punktu widzenia, niż pierwotnie sądzono. Powstała całkiem niedawno technika translacji automatycznej oparta na tzw. przykładach translacyjnych wydaje się być poważną alternatywą dla istniejących już metod automatyzacji translacji pomiędzy językami naturalnymi. W artykule zaproponowano wykorzystanie techniki translacji automatycznej opartej na przykładach w systemie, który byłby zdolny tłumaczyć dowolne teksty należące do języka ogólnego. Autor zaproponował własną, oryginalną modyfikację techniki translacji automatycznej opartej na przykładach, która umożliwiałaby uwzględnienie osobliwości gramatycznych języka polskiego. Rezultaty wstępne, uzyskane podczas przeprowadzonych eksperymentów wyglądają niezwykle obiecująco, a zaproponowana metoda wydaje się być krokiem postawionym we właściwym kierunku, na drodze do budowy w pełni automatycznego systemu translacji, który byłby zdolny do dokonywania przekładów o wysokiej jakości, przy czym przekłady te dokonywane byłyby dla dowolnych tekstów, a nie tylko dla tekstów ograniczonych do pewnej wybranej dziedziny działalności ludzkiej.

Słowa kluczowe: przetwarzanie języka naturalnego, lingwistyka komputerowa, translacja automatyczna, translacja oparta na przykładach

1. WPROWADZENIE

Pomysł zastosowania komputerów do dokonywania przekładów pomiędzy językami naturalnymi (np. pomiędzy szwedzkim i hiszpańskim) jest prawie równie stary, jak sam wynalazek komputera. Jako pierwszy z pomysłem tym wystąpił Amerykanin Warren Weaver już w 1947 roku [3]. W roku 1949 uczony ten wystosował do amerykańskiej fundacji wspierającej naukę (Rockefeller Foundation) memorandum, w którym postulował podjęcie badań w tym kierunku. Pierwszy zespół badawczy ukierunkowany specjalnie na automatyzację translacji został powołany już w 1951 roku w Austin w Teksasie. Z kolei w 1954 roku odbył się pierwszy publiczny pokaz systemu, który tłumaczył proste zdania z języka rosyjskiego na angielski. Niestety możliwości tego systemu były bardzo ograniczone (korzystał zaledwie z 16 reguł gramatycznych i bilingwicznego słownika o pojemności 250 haseł), podobnie zresztą, jak ograniczone były możliwości ówczesnych komputerów. W miarę postępów badawczych szybko okazało się jednak, że zadanie zautomatyzowania translacji jest o wiele trudniejsze, ale zarazem o wiele ciekawsze niż początkowo sądzono. Kłopoty na jakie natrafiono podczas prób zautomatyzowania translacji dały impuls do różnych krytycznych wystąpień. Miedzy innymi w 1959 roku pomysł zastosowania komputerów w translacji został skrytykowany przez filozofa Bar Hillela, który uważał, że zadanie to jest nierealizowalne nie tylko z powodów natury czysto technicznej (być może kiedyś do przezwyciężenia), ale w ogóle z zasady (za pomocą maszyny Turinga nie można w pełni symulować procesów zachodzących w umyśle człowieka). Krytyka ta zaowocowała ostatecznie wydaniem przez Amerykańską Akademię Nauk w 1964 roku specjalnego raportu (tzw. raport ALPAC), w którym instytucja ta postulowała wstrzymanie finansowania jakichkolwiek prac związanych z automatyzacją translacji. Renesans nastąpił dopiero pod koniec lat siedemdziesiątych. Pierwszym systemem translacji automatycznej, który odniósł sukces rynkowy był opracowany w Kanadzie system METEO, którego zadaniem było tłumaczenie raportów meteorologicznych i prognoz pogody z języka angielskiego na francuski. Począwszy od tego momentu prace nad translacją automatyczną trwają nieprzerwanie, a państwa takie jak: Japonia, Stany Zjednoczone oraz Unia Europejska przeznaczają na ten cel znaczne środki. Pomimo tego całkowicie automatyczna translacja dla języka ogólnego, która byłaby równie dobra jak wynik pracy przysięgłego tłumacza, wciąż pozostaje niedoścignionym ideałem.

Istnieje bardzo wiele różnorodnych przyczyn, które przesadzają o tym, że zautomatyzowanie translacji jest tak niezwykle trudnym zadaniem. Do przyczyn tych można zaliczyć w pierwszym rzędzie różnice pomiędzy systemami gramatycznymi różnych języków. Konstrukcje gramatyczne istniejące w jednym języku często nie mają swych bezpośrednich odpowiedników w innym języku. Również słownictwo każdego języka naturalnego stanowi autonomiczny i niezwykle złożony system, przy czym często jednostki leksykalne występujące w jednym języku nie mają swych bezpośrednich odpowiedników w innym języku — są to tak zwane dziury leksykalne (ang. lexical holes), które mogą zostać przetłumaczone jedynie w sposób opisowy. Jednakże prawdziwym utrapieniem badaczy zajmujących się translacją automatyczną jest wieloznaczność wypowiedzi, która jest immanentną cechą każdego języka naturalnego. Wieloznaczność wypowiedzi występuje

na poziomie składniowym w sytuacji, gdy rozbioru gramatycznego zdania można dokonać na co najmniej dwa alternatywne sposoby. Można mówić również o wieloznaczności semantycznej, gdy całą wypowiedź można rozumieć różnie, w zależności od jej kontekstu. Jednakże najpoważniejszym rodzajem wieloznaczności w przypadku systemów translacji automatycznej jest wieloznaczność występująca na poziomie leksykalnym, ponieważ jest ona najbardziej rozpowszechniona. W istocie, otwierając jakikolwiek duży dwujęzyczny słownik trudno jest do prawdy znaleźć wyrazy, które nie posiadałyby więcej niż jedno znaczenie. Niestety na obecnym etapie rozwoju komputerowych technik przetwarzania informacji nie opracowano jeszcze algorytmów, które byłyby w stanie rozwiązywać w sposób automatyczny wieloznaczności wypowiedzi. Bardziej szczegółowe informacje dotyczące problemów na jakie napotyka się nieuchronnie podczas wszelkich prób zautomatyzowania translacji, zainteresowany czytelnik może uzyskać studiując inne prace autora, np. [1] i [2].

2. PRZEGLĄD STOSOWANYCH ROZWIĄZAŃ

Do chwili obecnej opracowano wiele różnorodnych podejść do zagadnienia automatyzacji translacji. Systemy translacji automatycznej, które powstały najwcześniej charakterem swej pracy przypominały elektroniczne słowniki. W systemach takich zdania były tłumaczone wyraz po wyrazie, poprzez proste wyszukiwanie znaczeń wyrazów w dwujęzycznym słowniku i następnie podstawianie ich w miejsce wyrazów języka źródłowego. Oczywiście jakość uzyskanego w ten sposób przekładu była bardzo niska, ponieważ nie były uwzględniane żadne reguły gramatyki języka źródłowego, ani języka docelowego [3].

Kolejnym krokiem naprzód było opracowanie systemów translacji automatycznej opartych na tak zwanym transferze struktur syntaktycznych. Systemy takie oznaczane są skrótem **TBMT** (ang. Transfer-Based Machine Translation). Metoda transferu polega na dokonaniu w pierwszej kolejności analizy gramatycznej tłumaczonego zdania. Celem tej analizy jest dokonanie rozbioru gramatycznego zdania i ustalenie, który z wyrazów tłumaczonego zdania jest jego podmiotem, orzeczeniem, dopełnieniem bliższym, dopełnieniem dalszym, przydawką podmiotową, partykułą, przydawką dopełnieniową, okolicznikiem miejsca, czasu, sposobu, celu itp. Następnie po zakończeniu analizy gramatycznej tłumaczonego zdania dokonywany jest jego transfer do odpowiadającej mu struktury gramatycznej języka docelowego [4]. Proces ten zostanie zilustrowany na następującym przykładzie. Dane jest zdanie w języku angielskim:

If my father had had money my brother would have bought that car.

Po dokonaniu analizy gramatycznej tego zdania w systemie translacji automatycznej opartym na metodzie transferu, okazuje się, że zdanie to jest typowym zdaniem reprezentującym tzw. trzeci tryb warunkowy, które posiada następującą budowę:

<partykuła „if”> <podmiot 1> <orzeczenie 1> <dopełnienie 1> <podmiot 2>
<orzeczenie 2> <dopełnienie 2>.

Polski odpowiednik takiego zdania ma bardzo podobną budowę gramatyczną: <partykuła „gdyby”> <podmiot 1> <orzeczenie 1> <dopełnienie 1> <podmiot 2> <orzeczenie 2> <dopełnienie 2>.

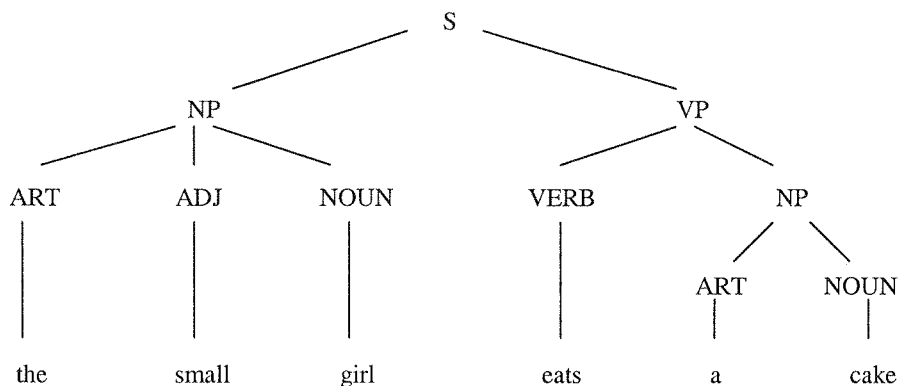
Pozostaje zatem dokonać tłumaczenia podmiotu, orzeczenia i dopełnienia zdania podrzędnego (<podmiot 1> <orzeczenie 1> <dopełnienie 1>) oraz zdania nadrzędnego (<podmiot 2> <orzeczenie 2> <dopełnienie 2>), aby uzyskać poprawny polski przekład tego zdania, który ma postać następującą:

Gdyby mój ojciec miał pieniądze mój brat kupiłby tamten samochód.

Z teoretycznego punktu widzenia wszystko wygląda jak najbardziej w porządku. Niestety w praktyce metoda transferu posiada dwie poważne wady. Pierwsza z nich polega na tym, że u podstaw metody transferu leży milczące założenie, zgodnie z którym tłumaczone zdanie posiada poprawną i w miarę przejrzystą budowę gramatyczną, co niestety w praktyce nader często okazuje się być nieprawdą [4]. Istniejące obecnie parsery (programy do analizy syntaktycznej) potrafią dokonać poprawnej analizy gramatycznej dla jedynie stosunkowo prostych zdań. Na przykład korzystając z reguł gramatyki bezkontekstowej angielskie zdanie:

The small girl eats a cake.

może zostać przedstawione w postaci następującej reprezentacji:



Niestety w przypadku zdań wielokrotnie złożonych podrzędnie i współrzędnie o skomplikowanej i zawitej budowie gramatycznej, nawet najlepsze parsery zawodzą i nie są w stanie dokonać prawidłowego rozbioru gramatycznego takich zdań. Na przykład jest rzeczą wysoce wątpliwą, czy kiedykolwiek zostanie stworzony dla języka polskiego parser, który potrafiłby bezbłędnie określić funkcje gramatyczne spełniane wszystkie wyrazy następującego zdania:

„Dla prowincji było rzeczą mało ważną, w czym imieniu są gnębione czy rządzone; bodźcem, który kierował ich postępowaniem, była władza w danej chwili istniejąca, więc gdy tylko ta władza uległa jakiejś większej sile, spieszyły błagać o litość zwycięzcę, który, mając olbrzymi dług do spłacenia, musiał poświęcić najbardziej wierne kraje na rzecz chciwości swoich żołnierzy”.

Powyższe zdanie zostało zaczerpnięte z książki Edwarda Gibbona pt. „Zmierzch Cesarstwa Rzymskiego”. Niestety, w praktyce językowej tego typu długie i zawite zdania wcale nie należą do rzadkości, co rzutuje w sposób oczywisty na jakość przekładów uzyskiwanych metodą TBMT.

Drużga z wad rozważanej metody translacji automatycznej polega na tym, że metoda ta nie posiada żadnych środków, które pozwałyby się jej zmierzyć z wieloznacznością występującą na poziomie leksykalnym [3]. Bowiern w sytuacji, gdy nawet zostanie dokonana w sposób prawidłowy gramatyczna analiza tłumaczonego zdania i właściwie zostanie wybrana odpowiadająca jej struktura gramatyczna języka docelowego, podczas dokonywania przekładów poszczególnych wyrazów pojawia się nieuchronnie problem, który z ich licznych ekwiwalentów, występujących w języku docelowym, należy wybrać. Na przykład niemieckie słowo „Kiefer” posiada dwa polskie ekwiwalenty znaczeniowe: „sosna” i „szczeka”. Oczywiście, nie jest rzeczą obojętną, który z nich zostanie podstawiony w miejsce niemieckiego oryginału. Dokonanie błędnego wyboru skutkuje najprawdopodobniej całkowitym niezrozumieniem przez odbiorcę dokonanego na drodze automatycznego przekładu. Jest to naprawdę poważny problem, bowiem wieloznaczność występująca na poziomie leksykalnym jest immanentną cechą każdego języka naturalnego.

Wymienione wady systemów translacji automatycznej oparte na metodzie transferu doprowadziły wielu badaczy do wniosku, iż nie można prawidłowo tłumaczyć tekstów, bez ich zrozumienia. Sformułowanie takich wniosków doprowadziło bezpośrednio do opracowania nowej klasy systemów translacji automatycznej, które zostały nazwane systemami opartymi na wiedzy. W literaturze systemy takie oznaczane są skrótem **KBMT** (ang. Knowledge-Based Machine Translation). Celem twórców systemów KBMT było wbudowanie do programów komputerowych dokonujących translacji wiedzy dotyczącej języka źródłowego i docelowego, jak również wiedzy dotyczącej dziedziny, z którą tłumaczone teksty są związane. Niestety w przypadku dokonywania przekładów tekstów dotyczących tematyki ogólnej, w komputerowy program automatycznej translacji musiałaby również zostać wbudowana ogólna wiedza o świecie, co zapewne jest sprawą jeszcze bardzo odległej przyszłości. Dlatego też, jeszcze do chwili obecnej nie powstał żaden system typu KBMT, który byłby w stanie tłumaczyć teksty należące do języka ogólnego. Natomiast systemy takie są z powodzeniem stosowane dla potrzeb translacji tekstów należących do pewnej specjalistycznej i bardzo wąskiej dziedziny wiedzy, której model może zostać z powodzeniem zaimplementowany w programie komputerowym. Na przykład typowym systemem KBMT jest system KANT opracowany na Carnegie-Mellon University, którego zadaniem jest tłumaczenie podręczników obsługi sprzętu ciężkiego, np. koparek, dźwigów, podnośników, spychaczy, wywrotek itp. [5].

Reasumując, istota systemów translacji automatycznej opartych na wiedzy polega na dokonaniu analizy tekstu w języku źródłowym i jego przekształceniu do pewnej abstrakcyjnej reprezentacji. Następnie w oparciu o tą reprezentację dokonywana jest synteza wypowiedzi w języku docelowym. Jeżeli rozważana reprezentacja posiada formę podobną do języka naturalnego, wówczas reprezentację taką określa się mianem *interlingua*, a cały system nazywa się systemem translacji automatycznej opartym na koncepcji języka *interlingua* i oznacza w literaturze skrótem **IBMT** (ang. Interlingua-

Based Machine Translation). Podstawowa zaleta koncepcji języka *interlingua* ujawnia się dopiero w wielojęzycznych systemach translacji automatycznej zawierających n języków, w sytuacji gdy trzeba dokonywać przekładów w dowolnym kierunku pomiędzy arbitralnie wybraną parą języków.

W podejściu klasycznym należałoby opracować tutaj $n(n-1)$ programów translacji automatycznej podczas, gdy stosując podejście oparte na języku *interlingua* wystarczy dla każdego z języków występujących w systemie opracować tylko po jednym programie syntezy i analizy, co daje łącznie zaledwie $2n$ programów. Niestety nikt jeszcze nie opracował takiego uniwersalnego języka, który byłby w stanie uwzględnić wszelkie zjawiska występujące w dowolnym z języków naturalnych świata [3]. Niektórzy badacze uważają, że taki uniwersalny język powinien naśladować naturalny tok rozumowania człowieka – wspólny dla wszystkich ludzi – ale droga do opracowania takiego języka jest jeszcze zapewne daleka. Podejmowane były natomiast próby zastosowania jako języka *interlingua* języka *esperanto*.

Koncepcja IBMT wydaje się być z teoretycznego punktu widzenia słusznym podejściem, ale jej praktyczne zrealizowanie dla języka ogólnego to na razie całkowite *science-fiction*. Dlatego też wielu badaczy zajmujących się translacją automatyczną postanowiło pójść w całkowicie innym kierunku i zerwać raz na zawsze ze stosowaniem w takich systemach jakiegokolwiek wiedzy lingwistycznej bądź dziedzicznej. Takie postawienie sprawy zaowocowało wypracowaniem kilku niezwykle interesujących koncepcji.

Pierwszą z rozważanych koncepcji jest tzw. tłumaczenie oparte na analizie przypadków. Systemy automatycznej translacji, w których stosowane jest takie podejście oznaczane są skrótem **CBMT** (ang. Case-Based Machine Translation). Idea leżąca u podstaw systemów CBMT jest bardzo prosta. Załóżmy, że mamy pewien bilingwiczny tekst, ot chociażby następujący:

„*During the whole of a dull, dark, and soundless day in the autumn of the year, when the clouds hung oppressively low in the heavens, I had been passing alone, on horseback, through a singularly dreary tract of country, and at length found myself as the shades of the evening drew on, within view of the melancholy House of Usher*”.

„*Przez cały mroczny, głuchy i smętny dzień jesienny, pod ciężką posową obłoków, co wlokły się nisko po niebie, nie opuszczałem siodła przebiegając samotnie na koniu dziwnie zamartłą, połąć kraju i dopiero z nastaniem wieczornych zmierzchów ujrzałem przed sobą posepny Dom Usherów*”.

Górna część zamieszczonego powyżej tekstu to fragment opowiadania Edgara Allana Poe’a zatytułowanego „The Fall of the House of Usher”, natomiast jego dolna część to polski przekład tego opowiadania noszący tytuł „Zagłada Domu Usherów”, dokonany przez wybitnego tłumacza Stanisława Wyrzykowskiego.

Jeżeli teraz z angielskojęzycznej części rozważanego bilingwicznego tekstu wybierze się pewną frazę, na przykład taką, jaka została w nim podkreślona (*as the shades of the evening drew on*), a następnie w polskojęzycznej części odszuka się jej odpowiednik (*dopiero z nastaniem wieczornych zmierzchów*), to istnieje duże prawdopodobieństwo, ale nie całkowita pewność, że jeżeli taka angielska fraza pojawi się jeszcze raz w tłumaczonym

angielskim tekście, to podstawiając w jej miejsce rozważany polski ekwiwalent, otrzyma się ponownie jej prawidłowy przekład.

Oczywiście, ktoś może zadać pytanie, dlaczego nie ma stuprocentowej pewności, że tak dokonany przekład będzie prawidłowy? Powód jest bardzo prosty, proszę przypomnieć sobie jeszcze uwagi na temat trudności, na jakie napotyka się podczas zautomatyzowania translacji – zwłaszcza różnice systemowe w gramatyce rozważanych języków, różnice w ich systemach słownictwa oraz wieloznaczność wypowiedzi występująca na poziomie leksykalnym, syntaktycznym i semantycznym. Zagadnienia te zostaną zilustrowane teraz na kilku przykładach.

Między innymi, węgierską frazę:

Ö alacsony.

można przełożyć na język polski na dwa alternatywne, całkowicie równoprawne sposoby:

On jest niski.

lub

Ona jest niska.

Powodem takiego stanu rzeczy jest fakt, polegający na tym, że język węgierski w ogóle nie zna pojęcia rodzaju gramatycznego. Powstaje zatem problem, który z tych dwóch przykładów translacyjnych należy umieścić w bazie danych? Decydując się na jeden z nich, całkowicie ignoruje się drugi, co skutkuje prawdopodobnie tym, że połowa dokonanych przekładów będzie nieprawidłowa pod względem gramatycznym.

Duży problem sprawiać mogą tutaj wieloznaczności wypowiedzi występujące na poziomie leksykalnym. Na przykład angielska fraza:

It was found in the pen.

może zostać przełożona na język polski na co najmniej trzy alternatywne sposoby:

To zostało znalezione w piórze.

To zostało znalezione w zagrodzie.

To zostało znalezione w łabędzicy.

w zależności od tego, jaki polski ekwiwalent znaczeniowy angielskiego słowa „pen” zostanie wybrany. Należy zauważyć, że wszystkie wymienione propozycje przekładów są dopuszczalne, bowiem wszystko zależy od tego, co zostało znalezione. Nawet ostatni przekład, mówiący o tym, że coś zostało znalezione w łabędzicy nie jest całkowicie pozbawiony sensu, bowiem rzecz może dotyczyć na przykład jakiegoś drobnego przedmiotu, który łabędzica mogła połknąć. Zatem, który z powyższych wzorców translacyjnych należało by zamieścić w bazie danych systemu automatycznej translacji?

Aby czytelnik nie wyrobił sobie fałszywego poglądu, że trudności takie występują tylko w przypadku tłumaczenia z języków obcych na polski, poniżej zostanie zamieszczonych kilka przykładów tłumaczenia w przeciwnym kierunku. Na przykład, chcąc przetłumaczyć na język hebrajski polską frazę:

Kim ty jesteś?

można to uczynić na dwa alternatywne sposoby:

מיהו אתה

(mi at)

albo

מִי אָתָּה

(mi ata)

przy czym pierwszą z nich stosuje się, w przypadku kierowania wypowiedzi do kobiety, a drugą do mężczyzny. Jest to bezpośrednią konsekwencją faktu rozróżniania w języku hebrajskim rodzaju męskiego i żeńskiego w drugiej osobie gramatycznej.

Podobnie tłumacząc na język arabski zdanie:

Oni są w Egipcie.

ponownie do wyboru są dwie równoprawne możliwości:

هُمَا فِي مِصْرَ

(huma fi misru)

albo

هُمْ فِي مِصْرَ

(hum fi misru)

Pierwszy przypadek odnosi się do sytuacji, gdy mowa jest o dokładnie dwóch osobnikach, a drugi dotyczy przypadku, gdy tych osobników jest więcej. Jest to konsekwencja istnienia w gramatyce języka arabskiego kategorii liczby podwójnej.

Również chcąc przetłumaczyć na język węgierski polską frazę:

To jest brat.

do wyboru są dwie możliwości:

Az öcs.

oraz

Az bátya.

Pierwszy przekład odnosi się do przypadku, w którym mowa jest o bracie młodszym, a drugi, gdy mowa jest o bracie starszym. Bowiem w języku węgierskim nie istnieje jedna jednostka leksykalna, łącząca obie wymienione kategorie znaczeniowe. Zatem, który z wymienionych przekładów należało by zamieścić w bazie danych z przykładami translacyjnymi.

Analogiczne przykłady zaczerpnięte z różnych języków świata można by wymieniać chyba w nieskończoność. Wszelkie tego typu przykłady pokazują, jak niezwykle trudnym jest zadanie translacji pomiędzy językami naturalnymi i to nawet dla człowieka, a co dopiero w przypadku powierzenia tego zadania bezrozumnym komputerom.

Pomimo braku gwarancji odnośnie poprawności przekładu dokonanego metodą CBMT, rozważana metoda translacji coraz częściej okazuje się być bardzo poważną alternatywą wobec stosowanych dotychczas technik automatycznej translacji. Dzieje się tak zwłaszcza w sytuacji, gdy języki uwikłane w proces przekładu są blisko ze sobą spokrewnione. Spektakularny przykład w tym zakresie pochodzi z Hiszpanii, gdzie od 1999 roku wydawana jest w języku katalońskim gazeta „El Periodico de Cataluña”, która jest tłumaczona z języka hiszpańskiego na drodze całkowicie automatycznej, właśnie dzięki zastosowaniu techniki CBMT [8].

Jest rzeczą oczywistą, że metoda CBMT wymaga opracowania bazy danych o pokazujących rozmiarach, która jest w stanie pomieścić dużą liczbę przekładów fraz należących do języka źródłowego. Oczywiście, teoretycznie liczba takich fraz jest nieskończenie wielka,

ale w praktyce okazuje się, że aby dokonywać przekładów w miarę o dobrej jakości, wystarczy, że system automatycznej translacji będzie uwzględniał od kilku do kilkunastu milionów najczęściej występujących w danym języku fraz.

Swoistą odmianę techniki CBMT stanowi technika translacji oparta na przykładach translacyjnych, która w literaturze oznaczana jest skrótem **EBMT** (ang. Example-Based Machine Translation). W metodzie EBMT nie jest specjalnie tworzona przez lingwistów baza danych zawierająca tłumaczenia fraz, natomiast system automatycznej translacji dysponuje znacznym zbiorem dwujęzycznych tekstów, przy czym zarówno frazy z tekstu w języku źródłowym, jak i ich domniemane przekłady są wybierane przez system w sposób automatyczny.

Pewną wadą metody EBMT jest konieczność dostarczenia zbioru bilingwicznych tekstów o bardzo dużych rozmiarach, przez co jego proces przeszukiwania celem odnalezienia w języku źródłowym danej frazy i w języku docelowym jego tłumaczenia, może okazać się bardzo czasochłonny. Pewnym rozwiązaniem jest tutaj zmodyfikowana wersja metody EBMT zwana translacją opartą na uogólnionych przekładach, która w literaturze oznaczana jest skrótem **GBMT** (ang. Generalized-example-Based Machine Translation). Istota tej metody polega na ponownym wbudowaniu do systemu translacji opartego na przykładach pewnej dawki wiedzy lingwistycznej. Zostanie to wyjaśnione na następującym przykładzie. Załóżmy, że celem systemu automatycznej translacji jest dokonanie przekładu na język polski następujących angielskich zdań:

I have bought a dog.

I have bought a cat.

I have bought a horse.

Stosując metodę EBMT należałoby zamieścić, aż trzy wzorce tłumaczeń tych zdań:

Kupiłem psa.

Kupiłem kota.

Kupiłem konia.

Należy zwrócić uwagę, że wzorce te nie mogą być krótsze, ponieważ wówczas tłumaczenie nie byłoby poprawne (np. *Kupiłem pies, kot, koń* itp.), ponieważ nie mogłyby być uwzględniony fakt, że w języku polskim biernik rzeczowników żywotnych jest równy dopełniaczowi, a nie mianownikowi, jak ma to miejsce w przypadku rzeczowników nieżywotnych (np. *Kupiłem dom, stół, dywan* itp.). Natomiast dodatkowe wbudowanie do systemu prostej reguły gramatycznej całkowicie rozwiązuje ten problem. Mianowicie system musi uwzględniać fakt, że tłumaczeniem angielskiej frazy:

I have bought <a noun>.

jest jej polski odpowiednik postaci:

Kupiłem <rzeczownik w bierniku>.

Widać zatem, że w metodzie GBMT bardzo wiele wzorców tłumaczeń (lista nazw obiektów ożywionych jest zapewne bardzo długa) zostało zastąpionych tylko jednym uogólnionym wzorcem translacji.

Główną zaletą wszelkich podejść opartych na wzorcach translacyjnych jest fakt, że w systemach takich w naturalny sposób minimalizowane jest ryzyko pojawienia się wieloznaczności na poziomie leksykalnym, ponieważ podczas translacji poszczególne

wyrazy nie są tłumaczone oddzielnie, ale jednocześnie tłumaczona jest cała ich grupa, dzięki czemu system korzysta z dodatkowych informacji kontekstowych. Ponadto przekłady dokonane metodą opartą na wzorcach translacyjnych brzmią dla człowieka w sposób bardziej naturalny, podobny do efektu pracy zawodowego tłumacza. Na przykład system translacji automatycznej oparty na transferze przetłumaczyłby angielską frazę *natural gas* zapewne jako *naturalny gaz*, podczas gdy prawidłowe tłumaczenie to *gaz ziemny* – zatem, jak widać, tłumaczenie całych fraz może wydatnie zwiększyć jakość automatycznych przekładów.

Innym podejściem do translacji automatycznej, w którym dokonano całkowitego zerwania z jakąkolwiek wiedzą lingwistyczną, jest technika translacji oparta na metodach statystycznych, oznaczana w literaturze skrótem **SBMT** (ang. Statistical-Based Machine Translation). Ze statystycznym podejściem do translacji automatycznej wiążą się takie pojęcia jak: statystyczne modelowanie języka (ang. language modelling) oraz statystyczne modelowanie translacji (ang. translation modelling).

Statystyczne modelowanie języka udziela odpowiedzi na pytanie o wielkość prawdopodobieństwa pojawienia się w tekście danego ciągu wyrazów. Prawdopodobieństwo to oznaczane jest jako $P(S)$, gdzie S jest pewnym ciągiem wyrazów.

$$S = S_1, S_2, S_3, \dots, S_{n-1}, S_n$$

Prawdopodobieństwo $P(S)$ może zostać policzone jako:

$$P(S) = P(S_1) \cdot P(S_2 | S_1) \cdot P(S_3 | S_1, S_2) \cdot P(S_4 | S_1, S_2, S_3) \cdot \dots$$

Prawdopodobieństwo $P(S_n | S_1, S_2, S_3, \dots, S_{n-1})$ nazywane jest N -gramem. W praktyce wartości N -gramów przybliżane są tzw. bigramami $P(S_n | S_{n-1})$ lub trigramami $P(S_n | S_{n-2}, S_{n-1})$. Bardzo rzadko stosuje się w praktyce N -gramy wyższych rzędów. Powodem tego są wielkie trudności z oszacowaniem wartości prawdopodobieństw takich N -gramów na podstawie dostępnych tekstów zapisanych w modelowanym języku. Bowiem już nawet w przypadku pewnych trigramów, które pojawiają się w danym języku bardzo rzadko, oszacowanie prawdopodobieństwa ich pojawienia się nastręcza duże trudności, a uzyskiwane na podstawie dostępnych tekstów wartości i tak są obarczone bardzo dużym błędem.

Z kolei statystyczne modelowanie translacji dostarcza wartości prawdopodobieństwa $P(T|S)$ wystąpienia ciągu wyrazów T w języku docelowym translacji, pod warunkiem, że zadany jest ciąg wyrazów w języku źródłowym S .

Ostatecznie prawdopodobieństwo wystąpienia paralelnych ciągów wyrazów w językach źródłowym i docelowym wyznaczane jest ze wzoru:

$$P(S, T) = P(T | S) \cdot P(S)$$

Następnie korzystając ze wzoru Bayesa otrzymuje się:

$$P(S | T) = \frac{P(T | S) \cdot P(S)}{P(T)}$$

Reasumując, statystyczne podejście do translacji polega na znalezieniu sekwencji wyrazów S , która posiada najwyższe prawdopodobieństwo wystąpienia, pod warunkiem, że zadana jest sekwencja wyrazów T .

Niestety praktyczne zastosowania statystycznych metod translacji automatycznej ograniczają się jedynie do prób tłumaczenia wypowiedzi sformułowanych w tak zwanych

językach kontrolowanych, stanowiących bardzo niewielki i ubogi podzbiór języków naturalnych, z których się pierwotnie wywodzą. Natomiast zastosowanie metod statystycznych do translacji wypowiedzi należących do języka ogólnego to jeszcze sprawa bardzo odległej i bliżej nieokreślonej przyszłości.

Niemniej jednak tego typu próby są wciąż podejmowane. Dość charakterystyczny przykład w tym zakresie pochodzi z uniwersytetu w Aachen, gdzie opracowany został system automatycznej translacji z języka niemieckiego na język angielski noszący nazwę *Verbmobile*, którego zasadę działania oparto właśnie na metodach statystycznych. Pole działania rozważanego systemu jest jednak dość ograniczone, gdyż służy on wyłącznie do tłumaczenia prostych tekstów reprezentujących dialog pomiędzy dwoma osobami, którego celem jest zaaranżowanie spotkania, ustalenie jego miejsca, terminu itp. [6] Niestety uzyskane rezultaty nie są zbyt zachęcające, co można zaobserwować analizując poniższy przykład, w którym zamieszczono próbkę tekstu niemieckiego, a poniżej jego angielski przekład uzyskany za pomocą systemu *Verbmobile*.

„Also, ich habe Unterlagen über Flüge da. Flug dauert siebzig Minuten. Ich denke, man kann nach Hannover auch mit dem Zug fahren, das dauert etwa viereinhalb Stunden mit dem Zug, und wenn Sie bedenken, bis wir in München am Flughafen sind, und in Hannover dasselbe noch mal. Ich würde fast vorschlagen, nachdem wir eigentlich Zeit haben, ist es etwas gemütlicher, im Zug zu fahren. Können wir etwas arbeiten”.

„Well, I have got papers flights over then. Flight takes seventy minutes. I think, we can also go by train to Hanover, that takes about four and half hours by train, and if you, we are in Munich to the airport, and in Hanover the same the same again. I would suggest since we actually have time, it is a bit more comfortable, in the train to go. We can bit work”.

Należy zauważyć, że zamieszczony powyżej niemiecki tekst został mimo wszystko w pewnych miejscach przetłumaczony poprawnie, czego przykładem może być np. zdanie: *Flight takes seventy minutes*. Niestety w większości przypadków znaczenia angielskiego przekładu można się jedynie domyślać, jak np. w przypadku zdania: *We can bit work*. Natomiast w niektórych przypadkach odgadnięcie sensu angielskiego przekładu jest wręcz niemożliwe. Dzieje się tak w przypadku następującej frazy: *and if you, we are in Munich to the airport, and in Hanover the same the same again*.

3. ZAPROPONOWANA PRZEZ AUTORA MODYFIKACJA TRANSLACJI OPARTEJ NA PRZYKŁADACH

Śledząc historię rozwoju systemów translacji automatycznej widać potwierdzenie się pewnej obiegowej opinii, według której „historia kołem się toczy”. Na początku pierwsze systemy translacji automatycznej przypominały swym kształtem elektroniczne słowniki i nie wykorzystywały żadnej dodatkowej wiedzy o języku źródłowym czy docelowym. Dość mierne rezultaty, jakie uzyskiwano w przypadku praktycznych zastosowań takich systemów skłoniły wielu badaczy do powzięcia próby wbudowania do takich systemów wiedzy lingwistycznej. W ten sposób powstały systemy oparte na zasadzie transferu TBMT. Zwiększając dawkę wiedzy lingwistycznej, w jaką wyposażano systemy

translacji automatycznej osiągnięto skrajność, która zmanifestowała się pojawieniem się koncepcji systemów opartych w całości na wiedzy KBMT. Jednakże w praktyce idea ta okazała się niemożliwa do zrealizowania. Z tych powodów część badaczy ponownie popadła w przeciwną skrajność, czego wyrazem było opracowanie systemów opartych na analizie przypadków CBMT i przykładach translacyjnych EBMT oraz systemów opartych na metodach statystycznych SBMT. We wszystkich tych systemach postanowiono zerwać z jakąkolwiek wiedzą o języku, a tłumaczony tekst traktować, jako zbiór symboli pojawiających się w sposób całkowicie arbitralny. W przeciwieństwie do systemów SBMT systemy CBMT i EBMT coraz częściej okazują się być poważną alternatywą dla istniejących wcześniej systemów translacji automatycznej. Jednakże prowadzone badania, mające na celu podniesienie jakości działania rozważanych systemów doprowadziły badaczy do wniosku, że aby osiągnąć zamierzony cel należy ponownie wbudować wiedzę lingwistyczną do systemów opartych na przykładach translacyjnych. W efekcie zaczęto podejmować próby dokonywania pewnych uogólnień przykładów translacyjnych, aby zmniejszyć ich liczbę i jednocześnie uczynić je bardziej elastycznymi. Wysiłki badawcze zaowocowały opracowaniem systemów GBMT, łączących technikę translacji opartą na przykładach z pewną dozą wiedzy lingwistycznej.

Wniosek wypływający z powyższych rozważań nasuwa się wręcz sam, a mówi on, że przyszłość translacji automatycznej należy bez wątpienia do systemów hybrydowych łączących w sobie najlepsze cechy i idee pochodzące z różnego typu rozwiązań [7].

Mając na uwadze powyższą przesłankę, autor artykułu zaproponował własne podejście do translacji automatycznej, które łączy ideę translacji opartej na przykładach z wiedzą o gramatyce docelowego języka przekładu.

Zaproponowany system opiera się na przesłance, w myśl której, jeżeli w pewnym zdaniu języka źródłowego wystąpiła fraza, która została przez tłumacza profesjonalistę przełożona w pewien określony sposób, to jeżeli identyczna fraza pojawi się jeszcze raz w innym tłumaczonym tekście i w jej miejsce podstawia się jej uprzedni przekład, to istnieje wysokie prawdopodobieństwo (ale nie całkowita pewność!), że postąpi się w ten sposób właściwie, czyli dokonany na tej drodze przekład będzie poprawny.

Wprowadzenie przez autora wiedzy lingwistycznej do zaproponowanego systemu polega na wyróżnieniu trzech klas przykładów translacyjnych oraz wzbogaceniu bazy danych, w której są zgromadzone przykłady translacyjne o cztery atrybuty: osobę gramatyczną Osoba, liczbę Liczba, rodzaj Rodzaj i przypadek Przypadek. W rozważanym systemie można wyróżnić przykłady translacyjne podlegające deklinacji NP (ang. Noun Phrase), przykłady translacyjne podlegające koniugacji VP (ang. Verb Phrase) oraz zwykłe przykłady translacyjne P (ang. Phrase).

Najprostszą postać mają zwykłe przykłady translacyjne P, które stanowią odpowiadającą sobie parę wyrażen w języku źródłowym i docelowym, co pokazano na poniższym przykładzie (translacja z języka angielskiego):

<i>since many years</i>
<i>od wielu lat</i>

Z kolei przykłady translacyjne podlegające deklinacji NP mają nieco bardziej złożoną budowę, gdyż zamiast jednego ekwiwalentu znaczeniowego w języku docelowym ich liczba wynosi aż sześć, w zależności od wartości przyjmowanej przez atrybut <Przypadek>, co pokazano na poniższym przykładzie (translacja z języka francuskiego).

<Przypadek>	<i>la présente édition</i>
1	<i>niniejsze wydanie</i>
2	<i>niniejszego wydania</i>
3	<i>niniejszemu wydaniu</i>
4	<i>niniejsze wydanie</i>
5	<i>niniejszym wydaniem</i>
6	<i>niniejszym wydaniu</i>

Postać przykładów translacyjnych podlegających koniugacji jest jeszcze bardziej złożona, bowiem zależy od wartości aż trzech atrybutów takich, jak: <Osoba>, <Liczba> i <Rodzaj>, co zobrazowano na następującym przykładzie (translacja z języka szwedzkiego):

<Osoba>	<Liczba>	<Rodzaj>	<i>har varit</i>
1	1	1	<i>byłem</i>
1	1	2	<i>byłam</i>
1	2	1	<i>byliśmy</i>
1	2	2	<i>byłyśmy</i>
2	1	1	<i>byłeś</i>
2	1	2	<i>byłaś</i>
2	2	1	<i>byliście</i>
2	2	2	<i>byłyście</i>
3	1	1	<i>był</i>
3	1	2	<i>była</i>
3	1	3	<i>było</i>
3	2	1	<i>byli</i>
3	2	2	<i>były</i>
3	2	3	<i>były</i>

Jak można zauważyć, atrybut <Przypadek> może przyjmować sześć wartości: 1, 2, 3, 4, 5 i 6, co odpowiada: mianownikowi, dopełniaczowi, celownikowi, biernikowi, narzędnikowi i miejscownikowi (wołacz tłumaczony jest za pośrednictwem zwykłych przykładów translacyjnych). Z kolei atrybut <Osoba> przyjmować może wartości: 1, 2 i 3, co

dopowiada pierwszej, drugiej i trzeciej osobie gramatycznej. Atrybut <Liczba> może przyjmować tylko wartości 1 lub 2, co odpowiada liczbie pojedynczej i mnogiej. Natomiast atrybut <Rodzaj> może przyjmować wartości: 1, 2 i 3, co odpowiada rodzajowi męskiemu, żeńskiemu i nijakiemu.

W jaki sposób podczas procesu translacji ustalane są wartości atrybutów? Na wstępie należy zauważyć, że prawdopodobnie w każdym języku świata występują takie elementy zdania jak: podmiot S (ang. subject), orzeczenie V (ang. verb) i dopełnienie O (ang. object). W zdecydowanej większości języków świata naturalnym szykiem rozważanych elementów zdania jest struktura typu SVO, czyli najpierw występuje podmiot, później orzeczenie, a na samym końcu dopełnienie zdania. W języku polskim (oraz w wielu innych językach) pomiędzy podmiotem zdania a jego orzeczeniem występuje tzw. związek zgody, tzn. orzeczenie zdania zgadza się z jego podmiotem pod względem osoby, liczby i rodzaju gramatycznego. Z kolei pomiędzy orzeczeniem zdania a dopełnieniem występuje tzw. związek rządu, polegający na tym, że orzeczenie wymusza użycie dopełnienia w odpowiednim przypadku gramatycznym.

Zatem sposób postępowania podczas dokonywania translacji z języka o szyku elementów zdania typu SVO jest następujący. Na początku, jeszcze przed rozpoczęciem translacji nowego zdania należy przypisać atrybutowi <Przypadek> wartość 1, ponieważ podmiot zdania zawsze występuje w mianowniku. Następnie rekord bazy danych zawierający przykład tłumaczenia frazy podmiotowej, musi również zawierać informacje o wartościach atrybutów <Osoba>, <Liczba> i <Rodzaj>. Wartości tych atrybutów odpowiadają osobie, liczbie i rodzajowi frazy, która stanowi przekład. Następnie tłumaczona jest fraza orzeczeniowa. Ponieważ wartości atrybutów <Osoba>, <Liczba> i <Rodzaj> zostały już ustalone system nie ma najmniejszych problemów z wyborem spośród wielu wariantów tego, który jest właściwy, tzn. spełnia związek zgody pomiędzy podmiotem a orzeczeniem zdania. Z kolei rekord bazy danych zawierający przekład frazy stanowiącej orzeczenie zdania, musi również zawierać informację o wartości, jaką należy przypisać atrybutowi <Przypadek>. Wartość atrybutu <Przypadek> niesie informację o tym, w jakim przypadku gramatycznym powinna wystąpić fraza stanowiąca dopełnienie zdania. Oprócz fraz stanowiących podmiot, orzeczenie i dopełnienie zdania, w typowym zdaniu występują również frazy stanowiące okoliczniki zdania, partykuły, spójniki itp. Jednakże frazy te nie podlegają ani koniugacji, ani deklinacji, w związku z czym tłumaczone są za pośrednictwem zwykłych przykładów translacyjnych.

Jak widać sposób postępowania w przypadku translacji zdań o budowie SVO jest stosunkowo prosty. Szyk SVO jest typowy dla większości języków indoeuropejskich, niestety występują również języki, w których stosowany jest inny szyk elementów zdania. Przykładem może być tutaj klasyczny język arabski, w którym (w przeciwieństwie do licznych jego współczesnych dialektów o szyku SVO) stosowany jest szyk VSO. Szyk taki posiada na przykład poniższe arabskie zdanie:

(szariba radżulun qahuatan)

które w dosłownym przekładzie (zachowując oryginalny szyk VSO języka arabskiego) znaczy:

شرب رجل قهوة

Wypił mężczyzna kawę.

W takim przypadku, jako pierwsza tłumaczona musi być fraza podmiotowa (występująca w arabskim zdaniu jako druga w kolejności) tak, aby przed translacją frazy orzeczeniowej mogły zostać ustalone wartości atrybutów Osoba, Liczba i Rodzaj. W rozważanym zdaniu fraza orzeczeniowa tłumaczona jest jako druga z kolei, a fraza dopełnieniowa jako trzecia.

Jako inny przykład posłużyć może język perski należący do irańskiej grupy indo-europejskiej rodziny językowej, w którym typowym szykiem elementów zdania jest SOV [26]. Jako przykład może posłużyć następujące perskie zdanie:

ایران زر دارد
(iran zar darad)

które w dosłownym przekładzie (zachowując oryginalny szyk SOV języka perskiego) znaczy:

Iran złoto ma.

W rozważanym przypadku języka perskiego najpierw należy przetłumaczyć (podobnie jak w przypadku szyku SVO) frazę podmiotową, a następnie frazę orzeczeniową, która jednakże zajmuje w zdaniu ostatnią, trzecią pozycję. Zajmująca w perskim zdaniu środkową pozycję fraza dopełnieniowa tłumaczona jest dopiero na końcu.

Sposób funkcjonowania zaproponowanego przez autora systemu translacji automatycznej zostanie zilustrowany na następującym przykładzie. Niech zadaniem pewnego hipotetycznego systemu translacji automatycznej, zbudowanego według metody zaproponowanej przez autora, będzie dokonanie przekładu następującego francuskiego zdania:

La liste des noms géographiques et le tableau des verbes irréguliers forment un supplément au présent dictionnaire.

Ponieważ zdania języka francuskiego posiadają szyk SVO, na samym początku ustalana jest wartość atrybutu <Przypadek> tak, aby podmiot polskiego przekładu był w mianowniku.

<Przypadek> = 1;

Następnie tłumaczona jest fraza podmiotowa, która należy do klasy fraz podlegających deklinacji NF:

<Przypadek>	<i>la liste des noms géographiques</i>
1	<i>lista nazw geograficznych</i>
2	<i>listy nazw geograficznych</i>
3	<i>liście nazw geograficznych</i>
4	<i>listę nazw geograficznych</i>
5	<i>listą nazw geograficznych</i>
6	<i>liście nazw geograficznych</i>

Jednocześnie ustalane są wartości atrybutów <Osoba>, <Liczba> i <Rodzaj>.

<Osoba> = 3;

<Liczba> = 1;

<Rodzaj> = 2;

Jako kolejna tłumaczona jest fraza zwykła F:

<i>et</i>
<i>i</i>

Następnie tłumaczona jest kolejna fraza podlegająca deklinacji NF:

<Przypadek>	<i>le tableau des verbes irréguliers</i>
1	<i>tablica czasowników nieregularnych</i>
2	<i>tablicy czasowników nieregularnych</i>
3	<i>tablicy czasowników nieregularnych</i>
4	<i>tablicę czasowników nieregularnych</i>
5	<i>tablicą czasowników nieregularnych</i>
6	<i>tablicy czasowników nieregularnych</i>

Jednocześnie ustalane są wartości atrybutów <Osoba>, <Liczba> i <Rodzaj>.

<Osoba> = 3;

<Liczba> = 1;

<Rodzaj> = 2;

Następnie tłumaczona jest fraza stanowiąca orzeczenie francuskiego zdania, która należy do klasy fraz podlegających koniugacji VF:

<Osoba>	<Liczba>	<Rodzaj>	<i>forment</i>
1	1	1	—
1	1	2	—
1	2	1	—
1	2	2	—
2	1	1	—
2	1	2	—
2	2	1	—
2	2	2	—
3	1	1	<i>stanowi</i>
3	1	2	<i>stanowi</i>
3	1	3	<i>stanowi</i>
3	2	1	<i>stanowią</i>
3	2	2	<i>stanowią</i>
3	2	3	<i>stanowią</i>

Jednocześnie z bazy danych odczytywana jest wartość atrybutu <Przypadek>, aby pomiędzy orzeczeniem a dopełnieniem zdania był spełniony związek rzędu. W rozważanym przypadku:

<Przypadek> = 4;

ponieważ dopełnienie w polskim przekładzie tego zdania musi być w bierniku.

W tym miejscu konieczna jest pewna dygresja. Jak widać w powyższym zestawieniu polskich form fleksyjnych francuskiego czasownika *forment* pozycje odpowiadające pierwszej i drugiej osobie nie zostały wypełnione. Powodem tego jest fleksyjność czasowników francuskich, które posiadają odrębne formy dla różnych osób gramatycznych. Rozważana tutaj forma *fleksyjna* *forment* odpowiada trzeciej osobie liczby mnogiej, dlatego zatem w zestawieniu jej polskich przekładów uwzględniono również formy odpowiadające liczbie pojedynczej? Otóż niekiedy tak się zdarza, że odpowiedniki pewnych wyrażen występujące w danym języku w liczbie pojedynczej, w innym języku występują w liczbie mnogiej i *vice versa*. Jako przykład może posłużyć hiszpański rzeczownik *la puerta*, który występuje w liczbie pojedynczej, natomiast jego polski odpowiednik, czyli „drzwi” — to rzeczownik w liczbie mnogiej. Zatem, jak widać w rozważanym zestawieniu polskich form fleksyjnych przekładu wypełnienie wszystkich pozycji, odpowiadających trzeciej osobie gramatycznej, jest jak najbardziej konieczne.

Jako kolejna tłumaczona jest fraza stanowiąca dopełnienie zdania i należąca do klasy fraz podlegających deklinacji NF:

<Przypadek>	<i>un supplémnt au présent dictionnaire</i>
1	<i>dodatek do niniejszego słownika</i>
2	<i>dodatku do niniejszego słownika</i>
3	<i>dodatkowi do niniejszego słownika</i>
4	<i>dodatek do niniejszego słownika</i>
5	<i>dodatkiem do niniejszego słownika</i>
6	<i>dodatku do niniejszego słownika</i>

Ponadto ustalane są wartości atrybutów <Osoba>, <Liczba> i <Rodzaj>.

<Osoba> = 3;

<Liczba> = 1;

<Rodzaj> = 1;

W rezultacie otrzymuje się następujący polski przekład francuskiego zdania:
Lista nazw geograficznych i tablica czasowników nieregularnych stanowi dodatek do niniejszego słownika.

Należy zauważyć, że otrzymany polski przekład jest zarówno poprawny pod względem budowy gramatycznej, jak również jest to przekład wierny, brzmiący ponadto naturalnie i przypominający efekt pracy człowieka tłumacza.

Szerszego omówienia wymaga jeszcze kwestia dotycząca sposobu, w jaki system translacji automatycznej odszukuje w bazie danych przykłady translacyjne.

Na początku całe tłumaczone zdanie traktowane jest jako jeden, integralny przykład translacyjny o dużych rozmiarach. Może się bowiem tak zdarzyć, że tłumaczone zdanie niesie ze sobą pewne treści o charakterze idiomatycznym, to znaczy takie, których nie można bezpośrednio odczytać analizując jedynie znaczenia poszczególnych elementów składowych zdania. Jako przykład może posłużyć następujące niemieckie zdanie:

Er traf den Nagel auf den Kopf.

które w całości stanowi następujący przykład translacyjny:

Utrafił w sedno problemu.

Gdy system translacji automatycznej nie odnajdzie mimo wszystko w bazie danych tłumaczenia dla całego zdania, odrzuca jego prawy skrajny element i ponawia próbę poszukiwań. Po odnalezieniu przykładu translacyjnego w bazie danych system zapamiętuje ten przykład (wraz z wartościami odpowiednich atrybutów), po czym skraca tłumaczone zdanie o fragment odnalezionego przykładu i powtarza na nowo całą procedurę. Powracając do rozważanego francuskiego zdania, poszukiwanie przez system jego przykładów translacyjnych wyglądałoby w sposób następujący:

<<La liste des noms géographiques et le tableau des verbes irréguliers forment un supplément au présent dictionnaire>> — **porażka** (nie odnaleziono w bazie danych)

<<La liste des noms géographiques et le tableau des verbes irréguliers forment un supplément au présent>> — **porażka** (nie odnaleziono w bazie danych)

<<La liste des noms géographiques et le tableau des verbes irréguliers forment un supplément au>> — **porażka** (nie odnaleziono w bazie danych)

...

<<La liste des noms géographiques et>> — **porażka** (nie odnaleziono w bazie danych)

La liste des noms géographiques>> — sukces (odnaleziono w bazie danych)

<<et le tableau des verbes irréguliers forment un supplément au présent dictionnaire>> — **porażka** (nie odnaleziono w bazie danych)

<<et le tableau des verbes irréguliers forment un supplément au présent>> — **porażka** (nie odnaleziono w bazie danych)

...

<<et le>> — **porażka** (nie odnaleziono w bazie danych)

<<**et**>> — **sukces** (odnaleziono w bazie danych)

<<le tableau des verbes irréguliers forment un supplément au présent dictionnaire>> — **porażka** (nie odnaleziono w bazie danych)

<<le tableau des verbes irréguliers forment un supplément au présent>> — **porażka** (nie odnaleziono w bazie danych)

...

<<le tableau des verbes irréguliers forment>> — **porażka** (nie odnaleziono w bazie danych)

<<**le tableau des verbes irréguliers**>> — **sukces** (odnaleziono w bazie danych)

<<forment un supplément au présent dictionnaire>> — **porażka** (nie odnaleziono w bazie danych)

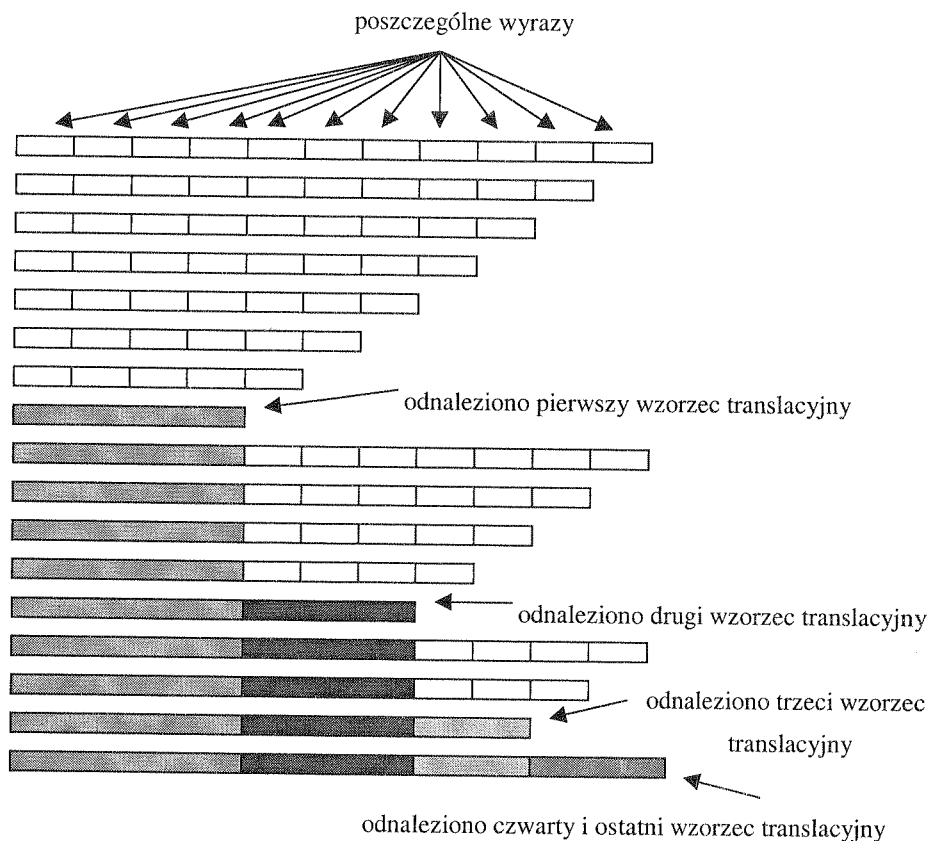
<<forment un supplément au présent>> — **porażka** (nie odnaleziono w bazie danych)

<<forment un supplément au>> — **porażka** (nie odnaleziono w bazie danych)

...

<<formet un>> — **porażka** (nie odnaleziono w bazie danych)<<formet>> — **sukces** (odnaleziono w bazie danych)<<un supplément au présent dictionnaire>> — **sukces** (odnaleziono w bazie danych)

Sposób działania zaproponowanego systemu translacji automatycznej został zilustrowany na rysunku 1.



Rys. 1. Sposób poszukiwania wzorców translacyjnych w zaproponowanym systemie automatycznej translacji

Fig. 1. The illustration of the way in which the machine translation system finds the translation patterns

Na rys. 1 każdy biały prostokąt symbolizuje jeden wyraz tłumaczonego zdania należącego do języka źródłowego przekładu. System translacji automatycznej odrzucając kolejne wyrazy i poruszając się z prawa na lewo poszukuje pierwszego wzorca translacyjnego. Po jego znalezieniu cała procedura powtarzana jest na nowo, aż do znalezienia drugiego wzorca translacyjnego. Operacje te są powtarzane do momentu wypełnienia całego zdania przykładami translacyjnymi pochodzącymi z bazy danych, w wyniku czego otrzymuje się przekład zdania na język docelowy.

Jest sprawą oczywistą, że o jakości otrzymywanych w ten sposób przekładów decyduje w pierwszym rzędzie rozmiar bazy danych, w której zgromadzono przykłady translacyjne. W najgorszym przypadku, gdy rozważana baza danych jest uboga i nie zawiera więcej haseł niż typowy dwujęzyczny słownik, przekład zostanie dokonany wyraz po wyrazie — podobnie jak ma to miejsce w przypadku systemów TBMT. Dalsze uzupełnianie bazy danych o kolejne przykłady translacyjne prowadzić będzie do stopniowego zwiększania się jakości translacji, ponieważ w takim przypadku nie będą tłumaczone tylko pojedyncze wyrazy, ale całe frazy, stanowiące zlepki kilku wyrazów, które zawierają w sobie dużą dawkę informacji kontekstowej, dzięki czemu minimalizowane jest prawdopodobieństwo dokonania ich błędnego przekładu na przykład w skutek pojawienia się wieloznaczności leksykalnej. Zatem kluczem do pełnego sukcesu jest baza danych zawierająca wzorce translacyjne o dość dużych rozmiarach. Opracowanie takiej bazy danych przekracza możliwości pojedynczego człowieka. Zadanie takie może natomiast zostać zrealizowane przez duży interdyscyplinarny zespół badaczy złożony zarówno z informatyków, jak i lingwistów.

Blizszego wyjaśnienia wymaga również sytuacja, w której system translacji automatycznej poszukując wzorca translacyjnego nie tylko nie odnajdzie takiego wzorca w postaci zlepku co najmniej dwóch wyrazów, ale nie odnajdzie w swojej bazie danych nawet pojedynczego wyrazu. W takiej sytuacji jedynym wyjściem będzie pozostawienie takiego wyrazu nie przetłumaczonego i ewentualnie jego zaznaczenie w tekście przekładu odpowiednim kolorem. Będzie to jednocześnie znak dla tłumacza dokonującego korekty przełożonego w sposób automatyczny tekstu, że wyraz taki nie został przetłumaczony i w związku z tym interwencja człowieka jest w tym miejscu konieczna. Interwencja taka polegać będzie najczęściej nie tylko na przetłumaczeniu nieznanego wyrazu, ale na opracowaniu nowego wzorca translacyjnego obejmującego rozważany wyraz i jego najbliższe sąsiedztwo. Wzorec taki byłby automatycznie wprowadzany do bazy danych zawierającej przykłady translacyjne, która w ten oto sposób powiększałaby systematycznie swoje rozmiary, podnosząc jednocześnie jakość działania rozważanego systemu translacji automatycznej.

Jak już wspomniano, pewną wadą metody translacji automatycznej opartej na przykładach translacyjnych jest fakt, że nie daje ona całkowitej pewności, że podstawiając w miejsce odnalezionego w bazie danych wzorca jego przykład translacji, postąpi się właściwie. W szczególności zjawisko wieloznaczności wypowiedzi występujące na poziomie leksykalnym, składniowym i semantycznym może spowodować, że dokonany w ten sposób przekład nie będzie poprawny.

Zapewne ciekawą koncepcją byłoby wyposażenie rozważanego systemu translacji automatycznej w pewnego rodzaju miernik prawdopodobieństwa dokonania poprawnego przekładu zdania. Mając zadaną wejściową sekwencję wyrazów (zdanie, które ma zostać przetłumaczone) S oraz jej przekład uzyskany na drodze automatycznej T , rozważane prawdopodobieństwo można wyliczyć ze wzoru:

$$P(T|S) = P(S|T) \cdot P(T)$$

Jednakże, aby tego dokonać należy dysponować statystycznym modelem języka docelowego oraz modelem translacji z języka docelowego na źródłowy. Niestety opracowanie tego typu modeli nie jest w praktyce sprawą łatwą i w rezultacie tego w chwili obecnej modele takie istnieją tylko dla nielicznych języków naturalnych.

Rozważany system translacji automatycznej po wyliczeniu prawdopodobieństwa dokonania poprawnej translacji, w zależności od uzyskanej jego wartości, mógłby zaznaczać przełożone zdania różnymi kolorami. Na przykład wyświetlenie przekładu w kolorze zielonym świadczyłoby o dużym stopniu pewności dokonania prawidłowego przekładu. Jeżeli przekład byłby nieco wątpliwy, zdanie to wyświetlane byłoby w kolorze żółtym. Natomiast w przypadku uzyskania zdecydowanie niskiej wartości rozważanego prawdopodobieństwa — taki podejrzany przekład wyświetlany byłby w kolorze czerwonym. Dysponując tego rodzaju informacją użytkownik systemu wiedziałby, które z partii uzyskanego na drodze automatycznej przekładu wymagają ewentualnego skonsultowania z tłumaczem profesjonalistą.

4. ZAKOŃCZENIE

Z zamieszczonych powyżej rozważań widać, że pomysł zautomatyzowania translacji pomiędzy językami naturalnymi wciąż pozostaje niedoścignionym ideałem, można by wręcz powiedzieć — św. Gralem poszukiwań badawczych. Warto być chyba w tej materii, mimo wszystko, optymistą. Bowiem, nawet jeśli wymarzony ideał nigdy nie zostanie osiągnięty, to i tak jest rzeczą wysoce prawdopodobną, że dzięki nieustannie wzrastającej mocy obliczeniowej komputerów i dzięki intensywnym badaniom prowadzonym przez interdyscyplinarne zespoły złożone z informatyków i lingwistów, jakość pracy systemów automatycznej translacji zbliży się, przynajmniej w stopniu zadawalającym, do pożądanego ideału, jakim jest system całkowicie automatycznej translacji dostarczający przekładów o wysokiej jakości dla tekstów należących do języka ogólnego. Należy ponadto zauważyć, że w przypadku języków o bliskim stopniu wzajemnego pokrewieństwa ideał ten już w zasadzie osiągnięto, czego przykładem może być automatyczne tłumaczenie gazety „Periodico de Catalua” z języka hiszpańskiego na kataloński [8]. Zatem w przyszłości komputery będą zapewne tłumaczyć teksty pomiędzy różnymi językami równie biegle, jak już dzisiaj komputery potrafią np. grać w szachy. Niestety na obecnym etapie rozwoju nauki i techniki nie może natomiast być żadnej nadziei na to, że komputery będą cokolwiek z tłumaczonych przez siebie tekstów rozumieć.

5. BIBLIOGRAFIA

1. M. G a j e r: *System translacji automatycznej oparty na dialogu z użytkownikiem i przykładach translacyjnych*. Pro Dialog 14(2002), Poznań, Wydawnictwo NAKOM, ss. 77-115
2. M. G a j e r: *Fakty i mity na temat systemów translacji automatycznej*. Elektronizacja nr 1-2/2002, Warszawa, Wydawnictwo SIGMA-NOT, ss. 33-36

3. D. Arnold, L. Balkan, S. Meijer, R. L. Humphreys, L. Sadler: *Machine Translation: An Introductory Guide*. London, NCC Blackwell, 1994
4. P. Whitelock, K. Kilby: *Linguistic and Computational Techniques in Machine Translation System Design*. London, UCL Press, 1995
5. E. Nyberg, T. Mitamura, J. Carbonell: *The KANT Machine Translation System: From R&D to Initial Deployment*. Pittsburgh, USA, Center for Machine Translation, Carnegie Mellon University, 1996
6. H. Ney, S. Nicen, F. J. Och, H. Sawaf, C. Tillmann, S. Vogel: *Algorithms for Statistical Translation of Spoken Language*. IEEE Transactions on Speech and Audio Processing, vol. 8, no. 1, January 2000, pp. 24-36
7. J. G. Carbonell, T. Mitamura, E. H. Nyberg: *The KANT Perspective: A Critique of Pure Transfer (and Pure Interlingua, Pure Statistics, ...)*, Pittsburgh, USA, Center for Machine Translation, Carnegie Mellon University, 1998
8. C. Rico: *From Novelty to Ubiquity: Computers and Translation at the Close of Industrial Age*. <http://www.accurapid.com/journal/15mt2.htm>

M. GAJER

THE EXAMPLE BASED MACHINE TRANSLATION SYSTEM

Summary

High-quality machine translation between human languages has for a long time been an unattainable dream for many computer scientists involved in this fascinating and interdisciplinary field of the application of computers.

Machine translation is a science about writing computer programs that are able to translate in a fully automatic way between natural languages, e.g. between Spanish and Korean. Machine translation was invented by an American Warren Weaver, who on the 15th of July 1949 sent his famous memorandum to Rockefeller Foundation with which he started the scientific research in the field. The first research group was constituted on the MIT in 1952. The first public demo of an operating machine translation system took place in New York on the 7th January 1954. During this demo the machine translation system translated 49 simple sentences from Russian into English.

But, after over 50 year of intensive scientific research machine translation is still far away from its satisfactory solution!

There are many reasons why machine translation is so extremely difficult. One of them is the ambiguity of natural language. The ambiguity can manifest on the word level (words can have many different equivalents in the target language), syntactic level (the syntactic analysis of a given sentence can be performed in many different ways) and semantic level (the whole sentence can be understood in many different ways).

The other reason of fundamental difficulties with translation process are the differences between the grammar systems of source and target languages. Also very important are differences in the vocabularies of source and target languages, where it is often difficult to find the exact equivalents of source words. Last but not least, is that many elements of the source text must be translated in an idiomatic way.

The developed quite recently example-based machine translation technique seems to be a serious alternative to the existing automatic translation techniques. In the paper the usage of the example based machine translation technique for the development of a system, which would be able to translate an unrestricted text is proposed. The new approach to the example-based machine translation technique that takes into account the peculiarity of the Polish grammar is developed.

The Polish language, which belongs to the group of Slavonic languages has a quite complex grammar. The main feature of Polish is its flexion nature, which makes learning it extremely difficult for foreigner. In the Polish grammar there can be found such phenomena as declination of nouns (there are 6 cases) and conjugation of verbs. Also there are singular and plural forms of words and 3 genders (masculine, feminine and neutral).

These phenomena make the direct usage of the Example-Base Machine Translation for Polish language difficult and non-effective. In order to implement this technique for the Polish language it must be modified thoroughly. This author proposed such modification, which takes into account the peculiarities of the Polish grammar and that allows for its effective implementation for the purpose of translating into Polish from the most of West European languages.

The obtained primary results of the development of proposed system seem to be very promising and appear to be a step made in the right direction towards a fully-automatic high quality machine translation system for unrestricted text.

Keywords: natural language processing, computational linguistics, machine translation, example-based translation

Pracę zrealizowano w ramach grantu KBN nr 8T11C 01917 „Formalne metody i narzędzia wspomagania analizy oraz projektowania baz danych i baz wiedzy”.

E
s
n
s
n
A
h
re
E
C

V
A
w
w
pe
U
-
-
-
-
-

Ob

Str
wz
w p
- V

Spe
Tek
pod
(np
pon
strz

INFORMACJE DLA AUTORÓW

Redakcja przyjmuje do publikowania prace oryginalne, przeglądowe i monograficzne wchodzące w zakres szeroko pojętej elektroniki. Ponieważ KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJI jest czasopismem 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 powinny 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, w tym rysunki i tabele.

Wymagania podstawowe.

Artykuły należy nadsyłać na wyraźnym, jednostronnym, czarno-białym wydruku komputerowym. Wydruk w formacie A4 powinien mieć znormalizowaną liczbę wierszy i znaków w wierszu (30 wierszy po 60 znaków w wierszu), w dwóch egzemplarzach, w języku polskim lub angielskim wybranym przez autora. Do wydruku powinna być dołączona dyskietka z elektronicznym tekstem artykułu. Preferowane edytory to WORD 6 lub 8. Układ artykułu (w wersji podstawowej) musi być następujący:

- Tytuł.
- Autor (imię i nazwisko autora/ów).
- Miejsce pracy (nazwa instytucji, miejscowość, adres. + ew. adres elektroniczny (e-mail)).
- Zwięzłe streszczenie powinno być w języku takim, w jakim jest pisany artykuł (wraz ze słowami kluczowymi).
- Tekst podstawowy powinien mieć następujący układ:

1. WPROWADZENIE
2. np. TEORIA
3. np. WYNIKI NUMERYCZNE
- 3.1.
- 3.2.
4.
5.
6. PODSUMOWANIE
7. ew. PODZIĘKOWANIA
8. BIBLIOGRAFIA

- Układ streszczenia w dodatkowej wersji językowej powinien być następujący:

AUTOR (inicjał imienia i nazwisko).

TYTUŁ (w języku angielskim – o ile artykuł pisany jest w języku polskim i na odrwót).

Obszerne do 3600 znaków streszczenia (wraz z słowami kluczowymi) w języku:

- a. angielskim, gdy artykuł pisany jest w języku polskim.
- b. polskim, gdy artykuł pisany jest w języku angielskim.

Streszczenie to powinno pozwolić czytelnikowi na uzyskanie istotnych informacji zawartych w pracy. Z tego względu w streszczeniu tym mogą być cytowane numery istotnych wzorów, rysunków i tabel zawartych w podstawowej wersji językowej.

- Wszystkie strony muszą mieć numerację ciągłą.

Sposób pisania tekstu.

Tekst powinien być pisany bez używania wyróżnień, a w szczególności nie dopuszcza się spacjiowania, 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 za pomocą przyjętych znaków adjustacyjnych, np. podkreślenie linią przerywaną oznacza spaciowanie (rozstrzelanie), podkreślenie linią ciągłą – pogrubienie, podkreślenie wężykami — kursywa.

Tekst powinien być napisany z podwójnym odstępem między wierszami, tytuły i podtytuły małymi literami. Marginesy z każdej strony powinny mieć około 35 mm. Wielkość czcionki wydruku powinna być zbliżona co najmniej co wielkości czcionki maszyny do pisania (minimum 12 punktów). Przy podziale pracy na rozdziały i podrozdziały cyfrowe ich oznaczenia nie powinny być większe niż II stopnia (np. 4.1.1.).

Sposób pisania tabel.

Tabele powinny być pisane na oddzielnych stronach. Tytuły rubryk pionowych i poziomych powinny być napisane małymi literami z podwójnym odstępem między wierszami. Przypisy (notki) dotyczące tabel należy pisać bezpośrednio pod tabelami. Tabele należy numerować kolejno liczbami arabskimi, u góry każdej tabeli podać tytuł dwujęzyczny. W pierwszej kolejności w podstawowej wersji językowej, a później w dodatkowej wersji językowej. Tabele umieścić na końcu maszynopisu. Przyjmowane są tabele algorytmów i programy na wydrukach komputerowych. W tym przypadku zachowany jest ich oryginalny układ. Tabele powinny być cytowane w tekście.

Sposób pisania wzorów 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 Electrotechnical Commission) oraz ISO (International Organization of Standardization).

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 1. F. Valdoni: A new millimetre wave satellite. E.T.T. 1990, vol. 2, no 5, pp. 141–148
- nieperiodyczne 2. K. Andersen: A resource allocation framework. XVI International Symposium, Stockholm (Sweden), may 1991, paper A 2,4
- książki 3. Y.P. Tvidis: Operation and modeling of the MOS transistors. New York, McGraw-Hill, 1987, p. 553

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 (preferowany edytor Corel Draw). 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 oraz numer rysunku. Spis podpisów pod rysunki i fotografie powinien być umieszczony na oddzielnej stronie. Podpisy pod rysunkami (fotografiami) powinny być dwujęzyczne: w pierwszej kolejności w podstawowej wersji językowej, a później w dodatkowej wersji językowej. Rysunki powinny być cytowane w tekście.

Uwagi końcowe.

Na odrębnej stronie powinny być podane następujące informacje:

- adres do korespondencji z kodem pocztowym (domowy lub do miejsca pracy),
- telefon domowy i/lub do miejsca pracy,
- adres e-mailowy (jeśli autor posiada).

Autorowi przysługuje bezpłatnie 20 odbitek artykułu. Dodatkowe egzemplarze odbitek, lub cały zeszyt Autor może zamówić u wydawcy na własny koszt.

ni.
co
ly

Autora obowiązują korekta autorska, którą powinien wykonać w ciągu 3 dni od daty otrzymania tekstu z Redakcji oraz zwrócić osobiście lub listownie pod adres 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.

ć
ty
li
ej
na
ć

i.
ne
e,
y
y
ul

,
i.

-
l,

ż
b
e
ót
n
e:
i

r

INFORMATION FOR AUTHORS OF K.E.T.

The editorial staff will accept for publishing only original monographic and survey papers concerning widely understood electronics. Because of the fact that KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJI is a journal of the Committee for Electronics and Telecommunications of Polish Academy of Science, it presents scientific works concerning theoretical bases and applications from the field of electronics, telecommunications, microelectronics, optoelectronics, radioelectronics and medical electronics.

Articles should be characterised by original depiction of a problem, its own classification, critical opinion (concerning theories or methods), discussion of an actual state or a progress of a given branch of a technique and discussion of development perspectives.

An article published in other magazines can not be submitted for publishing in K.E.T. The size of an article can not exceed 30 pages, 1800 character each, including figures and tables.

Basic requirements

The article should be submitted to the editorial staff as a one side, clear, black and white computer printout in two copies. The article should be prepared in English or Polish. Floppy disc with an electronic version of the article should be enclosed. Preferred wordprocessors: WORD 6 or 8.

Layout of the article.

- Title.
- Author (first name and surname of author/authors).
- Workplace (institution, adress and e-mail).
- Concise summary in a language article is prepared in (with keywords).
- Main text with following layout:
 - Introduction
 - Theory (if applicable)
 - Numerical results (if applicable)
 - Paragraph 1
 - Paragraph 2
 -
 -
 - Conclusions
 - Acknowledgements (if applicable)
 - References
- Summary in additional language:
 - Author (firs name initials and surname)
 - Title (in Polish, if article was prepared in English and vice versa)
 - Extensive summary, hawever not exceeching 3600 characters (along with keywords) in Polish, if artide was prepared in English and vice versa). The summary should be prepared in a way allowing a reader to obtaoin essential information contained in the artide. For that reason in the summary author can place numbers of essential formulas, figures and tables from the article.

Pages should have continues numbering.

Main text

Main text can not contain formatting such as spacing, underlining, words written in capital letters (except words that are commonly written in capital letters). Author can mark suggested formatting with pencil on the margin of the article using commonly accepted adjusting marks.

Text should be written with double line spacing with 35 mro left and right margin. Titles and subtitles should be written with small letters. Titles and subtitles should be numberd using no mor than 3 levels (i.e. 4.1.1.).

Tables

Tables with their titles should be places on separate page at the end of the article. Titles of rows and columns should be written in small letters with double line spacing. Annotations concerning tables should be placed directly below

the table. Tables should be numbered with Arabic numbers on the top of each table. Table can consist algorithm and program listings. In such cases original layout of the table will be preserved. Table should be cited in the text.

Mathematical formulas

Characters, numbers, letters and spacing of the formula should be adequate to layout of main text. Indexes should be properly lowered or raised above the basic line and clearly written. Special characters such as lines, arrows, dots should be placed exactly over symbols which they are attributed to. Formulas should be numbered with Arabic numbers placed in brackets on the right side of the page. Units of measure, letter and graphic symbols should be printed according to requirements of IEE (International Electrotechnical Commission) and ISO (International Organisation of Standardisation).

References

References should be placed at the end of the main text with the subtitle „References”. References should be numbered (without brackets) adequately to references placed in the text. Examples of periodical [1], non-periodical [2] and book [3] references:

1. F. Valdoni: A new millimetre wave satellite. E.T.T. 1990, vol. 2, no 5, pp. 141–148
2. K. Anderson: A resource allocation framework. XVI International Symposium (Sweden). May 1991. paper A 2.4
3. Y.P. Tividis: Operation and modeling of the MOS transistors. New York. McGraw-Hill. 1987. p. 553

Figures

Figures should be clearly drawn on plain or millimetre paper in the format not smaller than 9×12 cm. Figures can be also printed (preferred editor – CorelDRAW). Photos or diapositives will be accepted in black and white format not greater than 10×15 cm. On the margin of each drawing and on the back side of each photo author's name and abbreviation of the title of article should be placed. Figure's captions should be given in two languages (first in the language the article is written in and then in additional language). Figure's captions should be also listed on separate page. Figures should be cited in the text.

Additional information

On the separate page following information should be placed:

- mailing address (home or office),
- phone (home or/and office),
- e-mail.

Author is entitled to free of charge 20 copies of article. Additional copies or the whole magazine can be ordered at publisher at the author's expense.

Author is obliged to perform the author's correction, which should be accomplished within 3 days starting from the date of receiving of the text from the editorial staff. Corrected text should be returned to the editorial staff personally or by mail. Correction marks should be placed on the margin of copies received from the editorial staff or if needed on separate pages. In the case when the correction is not returned in said time limit, correction will be performed by technical editorial staff of the publisher.

In case of changing of workplace or home address Authors are asked to inform the editorial staff.