POLSKA AKADEMIA NAUK
KOMITET ELEKTRONIKI I TELEKOMUNIKACJI

KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJI

ELECTRONICS AND TELECOMMUNICATIONS QUARTERLY

TOM 51 — ZESZYT 2

WARSZAWA 2005

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

Dyżury Redakcji: środa 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. 11,56 Ark. druk. 9,25	Podpisano do druku w maju 2005 r.
Papier offset. kl. III 80 g. B-1	Druk ukończono w maju 2005 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 Quartery jest kontynuatorem tradycji powstałego 51 lat temu kwartalnika p.t. „Rozprawy Elektrotechniczne".

Kwartalnik jest czasopismem Komitetu Elektroniki i Telekomunikacji Polskij Akademii Nauk. Wydawany jest przez Warszawską Drukarnię Naukową PAN. Kwartalnik jest czasopismem naukowym, na którego łamach są publikowane artykuły i komunikaty prezentujące wyniki oryginalnych prac teoretycznych i doświadczalnych, a także przeglądowych. Związane są one z szeroko rozumianymi dziedzinami współczesnej elektroniki, telekomunikacji, mikroelektroniki, 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 wyników prac naukowych zrealizowanych w ramach „GRANTów" Komitetu Badań Naukowych spełnia więc jeden z wymogów stawianych tym pracom.

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

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

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

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

Redakcja

SPIS TREŚCI

CONTENTS

KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJI, 2005, 51, z. 2, ss. 223–238

MODELOWANIE KROPEK KWANTOWYCH

MARIUSZ MĄCZKA

Katedra Podstaw Elektroniki, Politechnika Rzeszowska
35-959 Rzeszów, W.Pola 2
e-mail: mmaczka@prz.rzeszow.pl

Otrzymano 2004.03.04
Autoryzowano 2004.05.30

Niniejszy artykuł dotyczy kropek kwantowych formowanych elektrostatycznie wewnątrz tranzystora jednoelektronowego. Przyrząd ten wytwarzany jest w oparciu o technologię bramek powierzchniowych z wykorzystaniem struktury ISIS (ang. Inverted Semiconductor Insulator Semiconductor). Praca jest kontynuacją badań nad doskonaleniem numerycznego modelu omawianego tu przyrządu i dotyczy uproszczenia polegającego na traktowaniu elektrod polaryzujących jako nieskończenie cienkie powierzchnie metaliczne. W niniejszej pracy przedstawiono wpływ grubości elektrod na rozkład potencjału w obszarze formowania kropki kwantowej w oparciu o nowy model przyrządu.

Słowa kluczowe: Tranzystory jednoelektronowe, kropki kwantowe, dwuwymiarowy gaz elektronowy, równanie Poissone'a

1. WPROWADZENIE

W połowie lat siedemdziesiątych John R. Schrieffer [1] wysunął sugestię, że zamknięcie elektronów w warstwie łączącej dwa różne materiały doprowadzi do zjawisk, w których ujawni się kwantowa natura elektronu. Nie można było tego pomysłu zrealizować, dopóki nie rozwinięto techniki wytwarzania — tzw. epitaksji z wiązek molekularnych (ang. Molecular Beam Epitaxy — MBE) [2] W tej technice wzrost kryształu uzyskuje się przez nakładanie kolejnych pojedynczych warstw atomowych. Kontrolując starannie rodzaj nakładanych atomów, można wytworzyć wiele „sztucznych" struktur krystalicznych, w których przerwy energetyczne dopasowane są do potrzeb, a nie wybierane spośród tych, które dostarcza przyroda. Przy pomocy tej technologii można wytworzyć bardzo małe półprzewodnikowe przyrządy kwantowe. Jeden z nich — kropki kwantowe formowane elektrostatycznie są bezpośrednio związane z tematem niniejszej

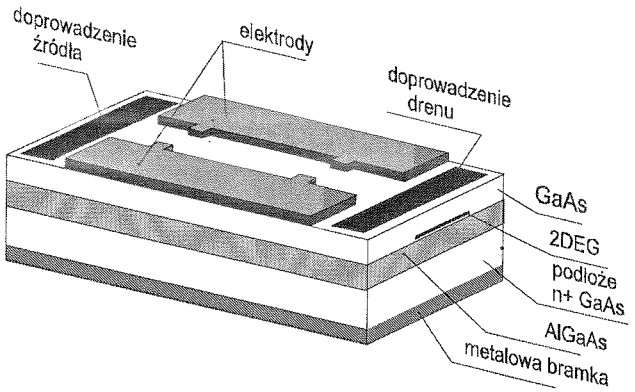
pracy. Są one wytwarzane w oparciu o technologię bramek powierzchniowych z wykorzystaniem struktury ISIS (ang. Inverted Semiconductor Insulator Semiconductor)[3]. W przyrządach tych dwuwymiarowy gaz elektronowy (ang. 2DEG) ograniczony przez potencjał do bardzo małych rozmiarów stanowi o ich podstawowych właściwościach.

Obszar kropki kwantowej jest tak mały, że prawa fizyki kwantowej ingerują w ruch elektronów we wszystkich kierunkach. Kropkę kwantową możemy traktować jak duży „sztuczny" atom [4], w którym elektrony utrzymywane są w małej objętości na ściśle określonych poziomach energetycznych. W przeciwieństwie jednak do prawdziwego atomu, gdzie siły ograniczające ruch elektronu są określone z góry przez strukturę jądra i prawie się nie zmieniają, siły ograniczające w sztucznym atomie można dokładnie kontrolować za pomocą geometrii układu i czynników zewnętrznych. Można tego dokonywać na dwa sposoby. Pierwszy to właściwa polaryzacja bramek sterujących a drugi to odpowiednia ich konfiguracja przestrzenna. Przykładając ujemne napięcie do bramek powierzchniowych powodujemy zubożenie 2DEG pod nimi na tym większym obszarze im większa jest wartość przyłożonego napięcia. Odpowiedni kształt i rozmiar bramek pozwala uzyskać większe lub mniejsze kropki kwantowe. W przypadku badań prowadzonych przez M.A. Kastnera [5] osiągały one rozmiary 500-1000 nm. Kropka kwantowa jest obszarem odizolowanym od reszty 2DEG, dlatego jedynym sposobem aby dodać do niej nowe elektrony jest tunelowanie kwantowe. Zjawisko to polega w przybliżeniu na przepuszczeniu przez kropkę prądu składającego się z pojedynczych elektronów.

Przyrząd, który umożliwia dodawanie pojedynczych elektronów do obszaru kropki kwantowej jest nazywany tranzystorem jednoelektronowym (ang. Single Electron Transistor – SET). Przykładową strukturę takiego przyrządu przedstawiono na Rys. 1. Kropka kwantowa jest integralną częścią tranzystora jednoelektronowego na tyle, ważną że bez niej przyrząd ten po prostu nie działa. Zapewne dlatego nazwy tranzystor jednoelektronowy i kropka kwantowa używane są często w literaturze w odniesieniu do tego samego przyrządu.

Podobnie jak kropka kwantowa jest kresem tendencji do miniaturyzacji układów, tak tranzystor jednoelektronowy stanowi kres ewolucji urządzeń kontrolujących przepływ prądu. Ma on za zadanie kontrolować prąd elektryczny porcjami, po jednym elektronie. Sugeruje to jego ewentualne zastosowania w cyfrowych obwodach logicznych, gdzie obecność lub nieobecność elektronu symbolizowałaby odpowiednio 0 lub 1.

Tranzystor jednoelektronowy jest przedmiotem badań numerycznych w naszym zespole od 1994 r. Badania te poświęcone są analizie pewnych własności przyrządu na drodze analizy modeli numerycznych [6],[7],[8]. Prezentowana praca jest kontynuacją rozważań nad doskonaleniem modelu tranzystora i dotyczy stosowanego do tej pory uproszczenia polegającego na traktowaniu elektrod polaryzujących jako nieskończenie cienkie powierzchnie metaliczne. W niniejszym artykule przedstawiono wpływ grubości elektrod na rozkład potencjału w obszarze formowania kropki kwantowej w oparciu o nowy model przyrządu. Istotnym wydaje się tu fakt, że następuje zmiana w podejściu do dolnej elektrody. Do tej pory w modelu tranzystora była to płaska powierzchnia me-



Rys. 1. Struktura tranzystora jednoelektronowego

Fig. 1. The structure of the single- electron transistor

taliczna przesunięta do obszaru złącza n+ GaAs–GaAs. Podstawą tego uproszczenia było założenie, że w całym obszarze n+ GaAs oraz na elektrodzie bramki dolnej panuje ten sam potencjał wymuszony przez źródło zasilania. W nowo opracowanym modelu uwzględniono nie tylko samą płaską elektrodę, ale cały obszar gdzie występuje ten sam potencjał, co na elektrodzie bramki dolnej.

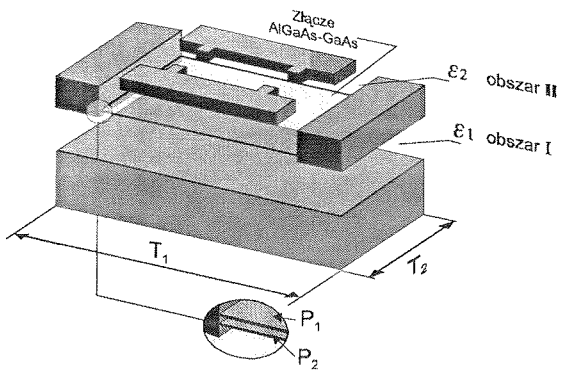
2. OPIS MODELU

Symulacje zostały przeprowadzone w oparciu o dynamiczny model elektrodowy tranzystora jednoelektronowego, iteracyjnie korygowany w trakcie kolejnych etapów obliczeniowych. Model wyjściowy (Rys. 2) składa się z pięciu obszarów przewodzących o ustalonych potencjałach odpowiadających metalicznym elektrodom tranzystora tj. bramki G, źródła S, drenu D oraz dwóch specyficznie ukształtowanych górnych elektrod E formujących poprzeczne bariery potencjału.

Polaryzację elektrod badanego modelu przedstawia Rys. 3. Na powierzchniach tych elektrod określa się funkcje gęstości potencjału warstwy pojedynczej, fizycznie odpowiadające powierzchniowej gęstości ładunku. Warstwę AlGaAs reprezentuje obszar dielektryczny I o przenikalności ε_1, zaś warstwę GaAs — obszar dielektryczny II o przenikalności ε_2 (półprzewodniki te nie są domieszkowane i do celów obliczeń pola elektrycznego mogą być z dobrym przybliżeniem traktowane jako dielektryki).

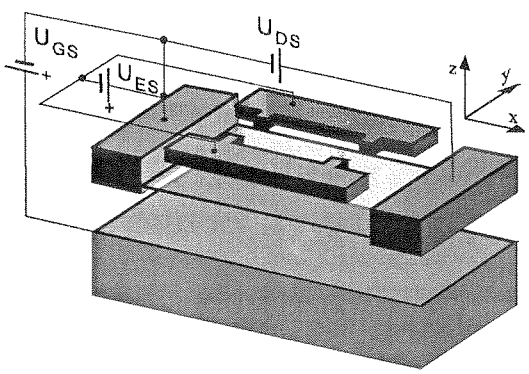
Na powierzchni rozgraniczającej obszary I i II wprowadzono dodatkowo dwie bardzo bliskie siebie powierzchnie P_1 i P_2, na których również wprowadza się funkcje gęstości potencjału warstwy pojedynczej. Ich celem jest uwzględnienie nieciągłości normalnej składowej natężenia pola elektrycznego na granicy obu obszarów.

Odpowiednia polaryzacja elektrod sprawia, że w obszarze GaAs powstaje dwuwy-



Rys. 2. Model wyjściowy tranzystora jednoelektronowego

Fig. 2. The model of the single-electron transistor



Rys. 3. Polaryzacja elektrod badanego przyrządu

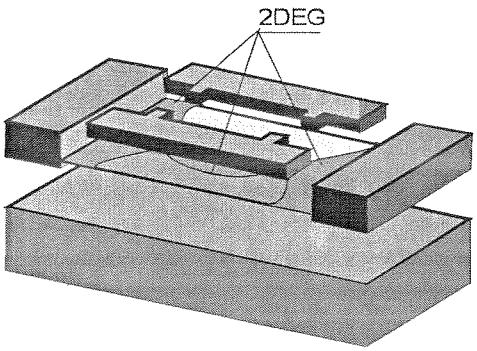
Fig. 3. Polarization of electrodes in the studied device

miarowy gaz elektronowy (2DEG), który stanowi dobrze przewodzącą cienką warstwę pomiędzy źródłem i drenem rozdzieloną dwiema barierami potencjału.

Zakładając, że potencjały źródła i drenu są sobie bliskie [9] ($U_{DS} \ll U_{GS}$, $U_{DS} \ll U_{ES}$) można przyjąć, że potencjał warstwy 2DEG jest stały i równy potencjałom elektrod S i D. 2DEG stanowi, więc niejako przedłużenie elektrod źródła-S i drenu-D.

W pierwszym etapie obliczeń wyznacza się kształt obszaru, w którym powstaje 2DEG i na tej podstawie koryguje się model tak, aby uwzględniał obecność gazu elektronowego jak to przedstawia Rys. 4.

2DEG jest tu reprezentowany przez cienką powierzchnię o potencjale źródła (w przeprowadzonych symulacjach przyjęto dla uproszczenia $U_{DS} = 0$). W kolejnych etapach następuje dalsza modyfikacja modelu, której sposób zostanie zawarty w algorytmie symulacji przedstawionym w dalszej części pracy.



Rys. 4. Skorygowany model tranzystora jednoelektronowego

Fig. 4. Corrected model of the single-electron transistor

3. MATEMATYCZNE SFORMUŁOWANIE PROBLEMU

Zagadnienie polega na poszukiwaniu rozwiązania równania Poissone'a:

$$\Delta\varphi = -\frac{\rho}{\varepsilon}, \tag{1}$$

gdzie ρ jest gęstością objętościową ładunku elektrycznego (w rozpatrywanym przypadku niezrównoważony ładunek występuje w obszarze 2DEG), z warunkami brzegowymi dotyczącymi stałości potencjału φ na powierzchniach elektrod i obszaru 2DEG zaś na powierzchni rozgraniczającej obszary *I* i *II* — warunków ciągłości potencjału i normalnej składowej indukcji elektrycznej:

$$\begin{aligned} &\varphi_i|_{S_k} = V_k \qquad i = 1,2 \\ &\varphi_1|_{P_1} = \varphi_2|_{P_2} \\ &\varepsilon_1 \frac{\partial\varphi_1}{\partial z}\bigg|_{P_1} = \varepsilon_2 \frac{\partial\varphi_2}{\partial z}\bigg|_{P_2} \end{aligned} \tag{2}$$

gdzie indeksy 1, 2 dotyczą potencjałów w obszarach *I* i *II* odpowiednio, a V_k oznacza potencjał k-tej elektrody. Ponieważ funkcja ρ nie jest znana rozwiązując numerycznie postawione zagadnienie należy doprowadzić do iteracyjnego samouzgodnienia pól ρ i φ. W pierwszym kroku przyjęto więc $\rho = 0$, czyli rozwiązuje się równanie Laplace'a z ww. warunkami brzegowymi. Funkcji φ poszukuje się w postaci potencjałów warstwy pojedynczej:

$$\varphi_i(p_i) = \frac{1}{4\pi\varepsilon_i} \int_{S_i} \frac{\sigma_{S_i}}{r} ds_i, \qquad i = 1, 2 \tag{3}$$

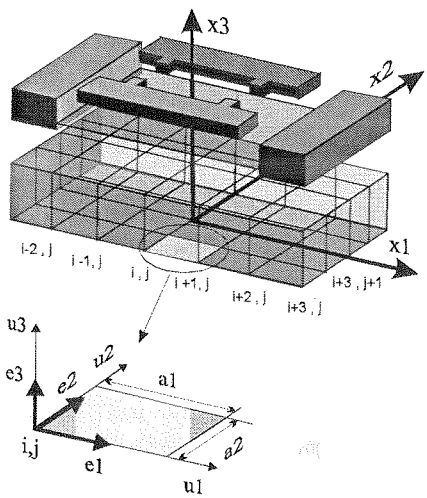
gdzie σ_{S_i} są gęstościami potencjałów warstwy pojedynczej na powierzchniach S_i stanowiących brzeg i-tego obszaru, zaś r jest odległością dowolnego punktu p_i tego obszaru od dowolnego punktu na jego brzegu.

Podstawiając (3) do warunków brzegowych otrzymuje się układ równań całkowych ze względu na funkcje σ_{S_i}. Po jego numerycznym rozwiązaniu poszukiwaną funkcję potencjału otrzymuje się na podstawie (3).

4. METODA OBLICZEŃ

W celu rozwiązania postawionego wyżej zagadnienia korzystano z Metody Elementów Brzegowych. Na elektrodach oraz na powierzchniach P_1, P_2 określona została prostokątna siatka dyskretyzacyjna, której elementy mają różną wielkość.

Umożliwiło to uzyskanie kompromisu między dokładnością obliczeń a wielkością modelu numerycznego (ograniczonego pojemnością pamięci operacyjnej komputera). Sposób dyskretyzacji jednej z elektrod tranzystora oraz oznaczenia stosowane w lokalnym i globalnym układzie współrzędnych przedstawia Rys. 5.



Rys. 5. Oznaczenia w globalnym i lokalnym układzie współrzędnych

Fig. 5. Sign in global and local system

Na szczególną uwagę zasługuje fakt, że z każdym elementem dyskretyzującym związany jest lokalny układ współrzędnych, który w zależności od swego umiejscowienia może mieć różną orientację w stosunku do globalnego układu współrzędnych, służącego do określenia struktury i wymiarów przyrządu. W celu określenia wzajemnych zależności pomiędzy globalnym a lokalnym układem współrzędnych konieczne było wyznaczenie wektorów bazowych: $e1$, $e2$, $e3$. Dzięki temu współrzędne punktów w każdym z lokalnych układów współrzędnych można było określić według zależności:

$$u_v\big|_{v=1,2,3} = \sum_{k=1}^{3} \left(x_k - x_k^{i,j}\right) \vec{e}\,_{v,k}^{\,i,j} \tag{4}$$

$$\vec{e}_{1,k}^{\,i,j} = \frac{x_k^{i+1,j} - x_k^{i,j}}{a_1^{i,j}} \tag{5}$$

$$\vec{e}_{2,k}^{\,i,j} = \frac{x_k^{i,j+1} - x_k^{i,j}}{a_2^{i,j}} \tag{6}$$

$$\vec{e}_{3,k}^{\,i,j} = \vec{e}_{1,k}^{\,i,j} \times \vec{e}_{2,k}^{\,i,j} \tag{7}$$

gdzie x_1, x_2, x_3, to współrzędne w globalnym układzie współrzędnych.

Określenie współrzędnych w lokalnym układzie współrzędnych jest niezbędne w procesie obliczania całek po powierzchniach elementów dyskretyzujących (patrz (8)).

Poszukiwane funkcje gęstości potencjałów są aproksymowane na elementach siatki funkcjami schodkowymi w rezultacie, czego wyrażenie (3) przybiera postać:

$$\varphi_P = \frac{1}{4\pi\varepsilon_i} \sum_{n=1}^{N_i} \sigma_{S\,in} \iint\limits_{S_{in}} \frac{ds_i}{r} \quad \text{dla} \quad P \in S_i \tag{8}$$

Gdzie S_{in} oznacza powierzchnię n-tego elementu siatki na powierzchni granicznej obszaru S_i. Po podstawianiu (8) do warunków brzegowych (2) otrzymuje się układ równań algebraicznych:

$$\begin{cases} \dfrac{1}{4\pi\varepsilon_1} \displaystyle\sum_{n=1}^{N_I} \sigma_{S_{1,n}} \iint\limits_{S_{1,n}} \frac{ds_1}{r} = V_{S_1} \\ \dfrac{1}{4\pi\varepsilon_2} \displaystyle\sum_{n=1}^{N_{II}} \sigma_{S_{2,n}} \iint\limits_{S_{2,n}} \frac{ds_2}{r} = V_{S_2} \\ \dfrac{1}{\varepsilon_1} \displaystyle\sum_{n=1}^{N_I} \sigma_{S_{1,n}} \iint\limits_{S_{1,n}} \frac{ds_1}{r} = \frac{1}{\varepsilon_2} \sum_{n=1}^{N_{II}} \sigma_{S_{2,n}} \iint\limits_{S_{2,n}} \frac{ds_2}{r} \\ \varepsilon_1 \displaystyle\sum_{n=1}^{N_I} \sigma_{S_{1,n}} \iint\limits_{S_{1,n}} \frac{\partial}{\partial z} \frac{1}{r} ds_1 = \varepsilon_2 \sum_{n=1}^{N_{II}} \sigma_{S_{2,n}} \iint\limits_{S_{2,n}} \frac{\partial}{\partial z} \frac{1}{r} ds_2 \end{cases} \tag{9}$$

gdzie: N_I ilość elementów dyskretyzujących brzeg obszaru I, N_{II} ilość elementów dyskretyzujących brzeg obszaru II

W układzie tym niewiadomymi są wartości $\sigma_{S_{i,n}}$. Całki w postaci (8) zostały obliczone analitycznie i wyrażają się dość złożonymi zależnościami [10].

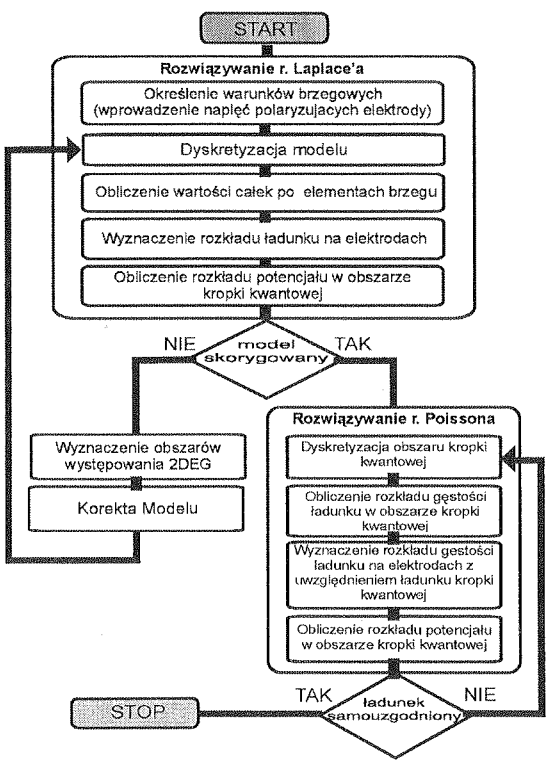
Wyprowadzony powyżej układ równań był rozwiązywany metodą eliminacji Gaussa. Stanowiło to wstępny etap obliczeniowy, po zakończeniu, którego możliwe było wyznaczenie, za pomocą wyrażenia (8), wartości potencjału w dowolnym punkcie P leżącym wewnątrz przyrządu.

Badania rozkładów energii potencjalnej skupiały się w pobliżu złącza AlGaAs-GaAs, gdzie formowała się kropka kwantowa (pomiędzy barierami potencjału).

Dalszy ciąg obliczeń polegał na samouzgadnianiu się ładunku w kropce kwantowej i elektrodach przyrządu. Wielkość ładunku w kropce kwantowej przy znanym po pierwszym etapie obliczeniowym potencjale ograniczającym obliczano na podstawie wyrażenia [11] określającego koncentrację dwuwymiarowego gazu elektronowego:

$$n_{2DEG} = \frac{m}{\pi\hbar^2} \int_{Ec}^{\infty} f(E, E_F) dE \tag{10}$$

gdzie f jest funkcją rozkładu Fermiego-Diraca, E_F — energią Fermiego, E_C — energią dna pasma przewodnictwa, m — masą efektywną elektronu. Powyższy wzór można stosować w przypadku wystarczająco dużych kropek kwantowych, tj. zawierających statystycznie znaczącą liczbę elektronów. Szczegółowy przebieg symulacji został przedstawiony w postaci algorytmu na Rys. 6.



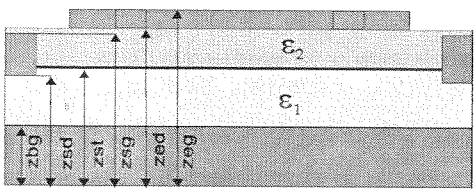
Rys. 6. Algorytm obliczeń rozkładów potencjału w kropce kwantowej

Fig. 6. The Algorithm of the calculations of the potential distribution in the quantum dot

Na podstawie powyższego algorytmu powstał program obliczeniowy napisany w języku C++, który został uruchomiony, przetestowany a następnie użyty do obliczeń na komputerze klasy PC.

5. WYNIKI SYMULACJI

Obliczenia rozkładów potencjału koncentrowały się wokół obszarów złącza AlGaAs – GaAs, gdyż właśnie w tym obszarze tworzy się odizolowany obszar kropki kwantowej.



Rys. 7. Określenie parametrów symulacji dla nowego modelu

Fig. 7. Qualification of the parameters of simulation for a new model

Na Rys.7 oraz w Tabeli.1 określone zostały parametry symulacji oraz ich wartości.

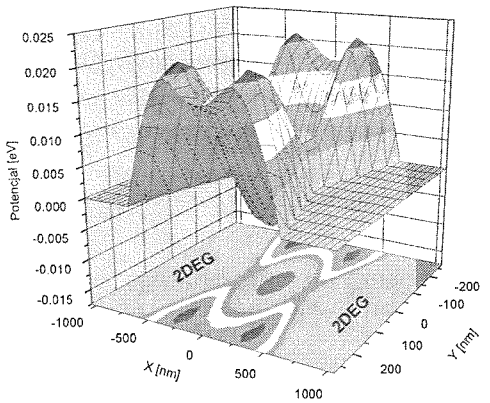
Tabela 1

Wartości parametrów symulacji dla modelu C

The parameters for a model C

Napięcia polaryzujące		
U_GS = 0.4V	UES=-0.22 V	UDS=0 V
Wymiary struktury		
T_x = 2400nm	zbg=300 nm	zsg=620 nm
Ty=600 nm	zst=600 nm	zeg=750 nm
zsd=580 nm		zed=700 nm
Współczynniki przenikalności dielektrycznej		
ε_2 = 13.2	ε_1 = 12.8	

Przykładowy rozkład potencjału w omawianym obszarze dla parametrów symulacji określonych na Rys.7 i w Tabeli 1 przedstawia Rys. 8



Rys. 8. Rozkład energii potencjalnej w pobliżu złącza AlGaAs-GaAs dla parametrów symulacji podanych w Tabeli 1

Fig. 8. The Potential distribution near around AlGaAs-GaAs junction, for simulation parameters of the given in Table 1

Cechą charakterystyczną przedstawionego powyżej rozkładu potencjału jest istnienie barier energetycznych [12] za pomocą których można w obszarze tranzystora wytworzyć kropkę kwantową w postaci odizolowanego obszaru 2DEG. Bardzo podobne wyniki symulacji można znaleźć w pracy opublikowanej przez grupę naukowców której przewodniczył U. Meirav [13].

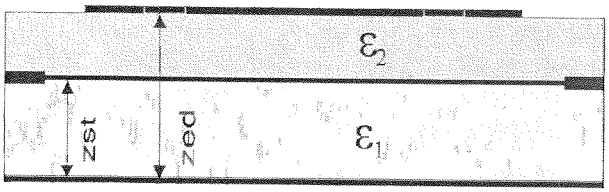
Głównym celem badań była odpowiedź na pytanie, w jakim stopniu uwzględnienie grubości elektrod wpływa na kształtowanie potencjału. W tym celu na początku dokonano porównania rozkładów potencjału dla poprzedniego i nowego modelu. W celu identyfikacji model tranzystora omawiany w ramach tej pracy nazwano modelem C zaś model dotychczasowy modelem B. Wyniki porównania dla parametrów symulacji podanych w Tabelach: 1,2 ilustruje Rys.10.

Table 2

Wartości parametrów symulacji dla modelu B

The parameters for a model B

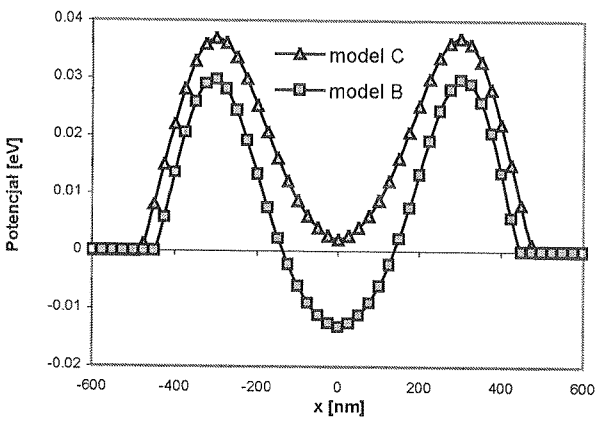
Napięcia polaryzujące		
$U_G S = 0.4V$	UES=-0.22 V	UDS=0 V
Wymiary struktury		
$T_x = 2400nm$	$T_y = 600nm$	
zed=400 nm	zst=300 nm	
Współczynniki przenikalności dielektrycznej		
$\varepsilon_2 = 13.2$	$\varepsilon_1 = 12.8$	



Rys. 9. Określenie parametrów symulacji dla modelu B

Fig. 9. Qualification of the parameters of simulation for model B

Wynika z niego fakt, że jakkolwiek charakter zmian rozkładów potencjału jest podobny, to jednak wartości napięć formujących odizolowany obszar kropki kwantowej są nieco inne dla obu modeli.

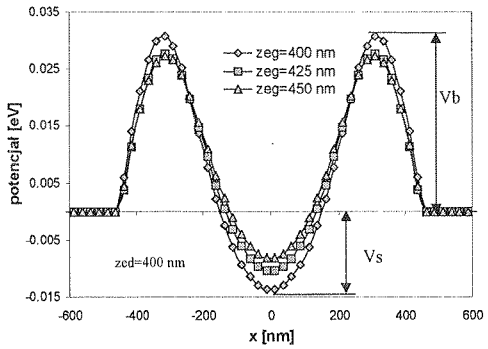


Rys. 10. Porównanie rozkładów energii potencjalnej wzdłuż środka kanału tranzystora dla rozpatrywanych modeli

Fig. 10. The comparison of the potential distributions along centre of the channel of the transistor for examination models

Widać, że ta sama konfiguracja napięć (U_{GS} = 0.4V, U_{DS} = 0V, U_{ES} = −0.22V) pozwala na wytworzenie kropki kwantowej w przypadku użycia poprzedniego modelu (model B) natomiast jest nieodpowiednia, gdy używany jest nowy model przyrządu (model C) uwzględniający grubość elektrod polaryzujących. Aby utworzyć odizolowany obszar 2DEG w przypadku nowego modelu konieczne jest zwiększenie napięcia U_{GS} lub zmniejszenie napięcia U_{ES}.

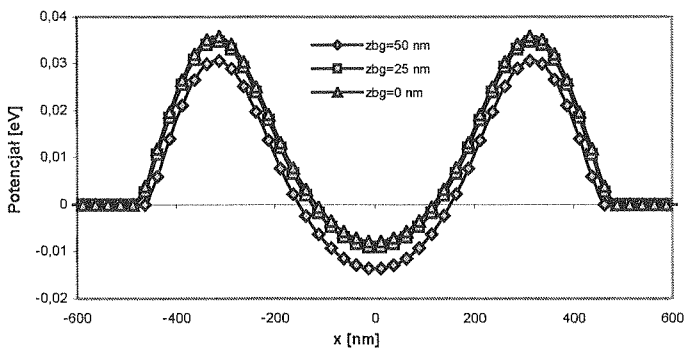
W tranzystorach jednoelektronowych kluczową rolę w ich działaniu odgrywa położenie i konfiguracja elektrod sterujących, dlatego następnym zadaniem było zbadanie wrażliwości potencjału na różne grubości elektrod polaryzujących wyniki tych badań przedstawiają wykresy na Rys. 11, 12, 13, Parametry dla symulacji znajdują się w Tabeli 1.



Rys. 11. Ilustracja wpływu grubości górnych elektrod na rozkład potencjału. dla parametrów symulacji zawartych w Tabeli 1

Fig. 11. The influence of the thickness of the upper electrodes on to potential distributions for simulation parameters of given in Table 1

Jak pokazuje Rys. 11. wartości parametrów potencjału formującego: głębokość studni (Vs) i wysokość barier (Vb) zmniejszają się nieliniowo w miarę wzrostu grubości elektrod górnych. Obserwując natomiast wykresy na Rys. 12 można zauważyć, że zwiększanie grubości elektrody dolnej powoduje zmniejszenie głębokości studni potencjału Vs, i jednocześnie jest przyczyną wzrostu wysokości bariery Vb.

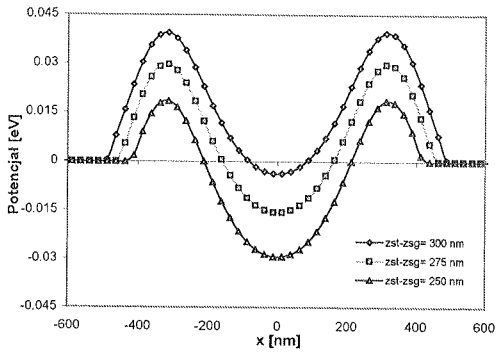


Rys. 12. Ilustracja wpływu grubości dolnej elektrody na rozkład potencjału. dla parametrów symulacji zawartych w Tabeli 1

Fig. 12. The influence of the thickness of the bottom electrode on to potential distribution for simulation parameters of given in Table 1

Wynika z tego wniosek, że im grubsza dolna elektroda tym bardziej potencjał formujący przesuwa się w kierunku wartości dodatnich. Jest się to nieco zaskakujące, gdyż wydawać by się mogło, że zwiększenie rozmiarów ujemnie naładowanej dolnej elektrody powinno raczej przyczyniać się do zwiększania głębokości studni potencjału (parametr Vs). W rzeczywistości jednak, to nie zmiany grubości elektrody ale jej

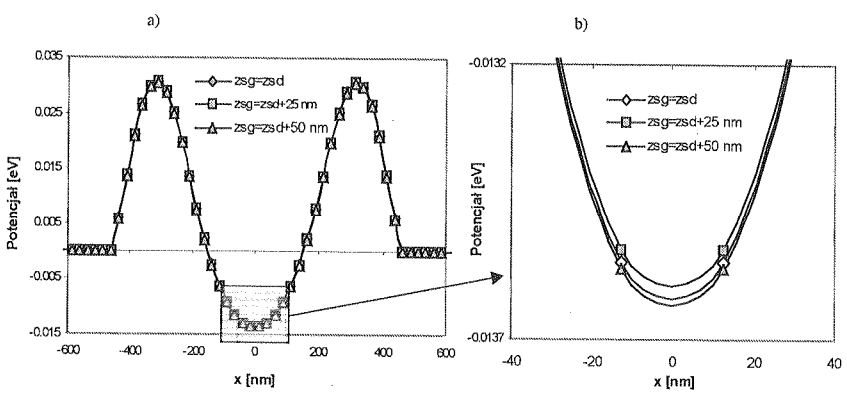
odległości od obszaru złącza AlGaAs-GaAs, gdzie tworzy się gaz elektronowy powodują znaczne zmiany parametrów Vs i Vb o czym przekonują nas kolejne wyniki symulacji przedstawione na Rys. 13. Widać tu, że odsunięcie dolnej elektrody o 50 nm od AlGaAs-GaAs powoduje prawie całkowity zanik studni potencjału (Vs≈0).



Rys. 13. Ilustracja wpływu odległości dolnej elektrody od obszaru formowania 2DEG na rozkład potencjału dla parametrów symulacji zawartych w Tabeli 1

Fig. 13. The influence of the distance of the bottom electrode from 2DEG area on to potential distributions for simulation parameters of given in Table 1

W trakcie numerycznych badań potencjału okazało się również, że zmiany grubości elektrod bocznych tranzystora (źródła-S i drenu-D) nie mają w zasadzie żadnego wpływu na rozkład potencjału w obszarze tworzenia się kropki kwantowej, o czym możemy się przekonać analizując wyniki przedstawione na Rys 14 a).



Rys. 14. a) Ilustracja wpływu grubości elektrod bocznych (źródła i drenu) na rozkład potencjału dla parametrów symulacji zawartych w Tabeli 1, b) powiększenie fragmentu wykresu a)

Fig. 14. a) The influence of the thickness of the side electrodes (the source and drain) on to potential distributions for simulation parameters of given in Table 1, b) the increase the fragment of graph a)

Zmiany w rozkładzie potencjału dla tego przypadku możemy zaobserwować dopiero kiedy wykres z Rys. 14 a) znacznie powiększymy. Powiększony fragment wykresu z Rys. 14 a) prezentuje Rys. 14 b). Widać na nim wyraźnie jak niewielkie zmiany w rozkładzie potencjału wywołuje zmiana grubości bocznych elektrod sterujących tranzystora. Oczywiście rozważamy to w kontekście pozostałych elektrod sterujących tranzystora.

6. PODSUMOWANIE

Badania numeryczne rozkładów potencjału w tranzystorze jednoelektronowym z użyciem nowego modelu pokazały, że uwzględnienie różnych grubości elektrod polaryzujących tranzystor zmienia w pewnym stopniu kształt potencjału formującego kropkę kwantową. Ma to swoje następstwa w postaci zmniejszenia rozmiarów kropki kwantowej w stosunku do tych, które były przy założeniu cienkich elektrod polaryzujących Charakter zmian potencjału formującego zależy od aktualnie rozpatrywanej elektrody. Ogólnie mówiąc potencjał formujący jest najbardziej czuły na zmiany grubości elektrod górnych a najmniej na zmiany grubości elektrod źródła i drenu. Można też zauważyć, że zastosowanie opisanego tu modelu powoduje nieznaczne zmniejszenie zakresu napięć polaryzujących, dla których w badanym przyrządzie można wytworzyć odizolowany obszar studni potencjału (wiąże się to ze zmniejszeniem sumy Vs+Vb)

Otrzymane wyniki potwierdzają wnioski znane z wcześniejszych symulacji. Pierwszy z nich mówi o tym, że kształt potencjału formującego kropkę kwantową zależy przede wszystkim od konfiguracji przestrzennej elektrod. Zastosowanie nowego dokładniejszego modelu pozwoliło jedynie zweryfikować jego ilościowe zmiany. Drugi wniosek jaki się nasuwa dotyczy rozmiarów kropki kwantowej, która zarówno dla wszystkich badanych dotychczas modeli jest płaskim dyskiem o grubości kilkunastu i średnicy kilkuset nanometrów. W świetle tych wniosków można stwierdzić, że przyjęte założenia o jednorodności struktury dla wcześniejszych symulacji nie powodowały zasadniczych błędów jakościowych i ilościowych. Uwzględnienie grubości elektrod pozwala jednak uzyskać dokładniejsze rezultaty.

Praca naukawa finansowana ze środków budżetowych na naukę w latach 2005/2007 jako projekt badawczy nr 3T11B69328

7. BIBLIOGRAFIA

1. J. R. Schrieffer, P. Soven: *Theory of the electronic structure.* Physics Today, vol. 28, no. 4, April 1975, USA, pp.24-30.
2. P. M Petroff, S. P. Denbaars: *MBE and MOCDV growth and properties of self-assembling quantum dot arrays in III-IV semiconductors structures.* Superlattices and Microstruktures, vol 15:1, 15, 1994.

3. H. Tomozawa, K. Jinushi, H. Okada, T. Hashizume H. Hasegawa: *Design and fabrication of GaAs/AlGaAs single electron transistors based on in-plane Schottky gate control of 2DEG.* Physica B, vol.227, no.1- 4, Sept. 1996, Netherlands, pp.112-15.
4. M.A. Kastner: *Artificial atoms.* Physics Today, vol. 46, 1993, pp. 24-31.
5. M.A. Kastner: *The single-elektron transistor.* Reviews of Modern Physics, vol. 64, no. 3, July 1992, pp 849-858.
6. S. Pawłowski, A Kusy, R. Sikora, M. Mączka, E. Machowska- Podsiadło: *Numerical studies of quantum dot electrical transport propertes in Metal/Non-Metal Microsystems.* Physics, Technology and Aplications, B. W. Licznerski, A. Dziedzic, Editors, Proc. SPIE 2780, 1995, str. 202-207.
7. S. Pawłowski, A. Kusy, M. Mączka, E. Machowska-Podsiadło: *Numerical studies of quantum dot: effect of negative dynamic conductance.* Proceedings of the 6th Conference on Electron Technology, ELTE'97, Krynica, May 1997, vol. 2, str. 554-557.
8. M. Mączka, S. Pawłowski, A. Kusy, E. Machowska-Podsiadło: *Badania numeryczne rozkładu potencjału w kropce kwantowej ze sterowaną barierą* Elektronika, 4'98, str 16 - 19.
9. D. B. Chklovskii, B. I. Shklovskii, L. I. Glazman: *Electrostatics of edge channels.* Physical Rev. B, 1992 r., vol. 46, no. 7, 15 Aug.
10. M. Mączka: *Kropki kwantowe formowane elektrostatycznie: badanie rozkładów potencjału.* Rozprawa doktorska, Politechnika Rzeszowska 2001
11. J.H. Davies: *The Physics of Low-Dimentional Semiconduktors.* Cambrigde University Press, 1996.
12. U. Meirav, M. A. Kastner, S. J. Wind: *Single Electron Charging and Periodic Conductance Resonances in GaAs Nanostructures.* Phys. Rev. Lett., 1990, vol.65, pp.771-774.
13. U. Meirav, P. L. Mc Euen, M. A. Kastner, E. B. Foxman, A. Kumar, S. J. Wind: *Conductance Oscillations and transport spectroscopy of a quantum dot.* Phys. B-Condensed Matter , 1991, vol. 85, pp. 357-366.

M. MĄCZKA

MODELING OF THE QUANTUM DOTS

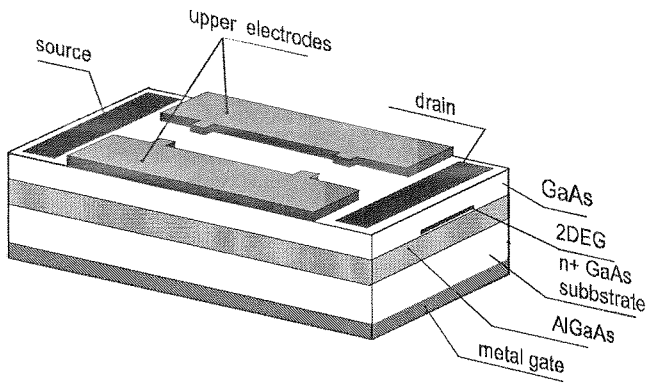
Summary

The work concerns quantum dots formed electrostatically inside the single-electron transistors. These devices are produced with superficial gate technology in inverted semiconductor-insulator- semiconductor structure. M.A. Kastner [5] proposed the first such a device in 1991.

The structure of the single-electron transistor In this article quantum dots are modeled with the use of the confined potential formed by electrostatic field. The numerical method of dissolving Poisson equation in these devices is described. The Method uses Boundary Elements. Potential distributions are calculated with the use potential density functions of the single layer. These functions physically are related to density of charge.

In previous works in the calculations an infinitely thin polarizing electrodes were assumed. In the present paper the influence of electrodes thickness onto potential distributions is examined. This require the previous model to be renewed. In the work the comparison of potential distributions in old and new numerical model is presented.

The numerical investigations of potential distributions, with the use of new model show, that the finite thickness of polarizing electrodes, to some extent changes the shape of the potential forming a quantum



The structure of the single-electron transistor

dot. This change depend on electrode, which is considered. Generally speaking confining potential is more sensitive to the thickness of upper electrodes than to thickness of electrodes of sources or drain.

Keywords: quantum dot, single electron transistor, potential distributions, Poisson equation

KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJI, 2005, 51, z. 2, ss. 239–253

Numerical modeling of passively Q-switched Nd:YAG lasers

JACEK ŚWIDERSKI

Institute of Optoelectronics, Military University of Technology
2 Kaliskiego Str., 00-908 Warsaw, Poland
tel.: +48 22 683 98 42; e-mail: jswiders@poczta.onet.pl

Otrzymano 2003.12.12
Autoryzowano 2004.06.09

Mathematical description of a passively Q-switched laser and the way it works are presented. Both theoretical and experimental model of this laser were worked out. Moreover, numerical analysis of saturable absorber influence on laser efficiency was made. The thing taken into account here is the saturable absorber characterized by excited state absorption (ESA). The optimization procedure points to the optimal circumstances of the laser system considered. The results obtained numerically are in very good agreement with those achieved experimentally.

Keywords: Q-switched solid state laser, saturable absorber, numerical simulation

1. INTRODUCTION

Q-switching of Nd:YAG lasers is widely applied in scientific research and for practical applications using active or passive devices. Active Q-switching methods (mechanical, electro-optic and acusto-optic) are usually used in Nd:YAG systems to obtain pulses with high repetition rate and high average output power [1]. The technique of active Q-switching is rather complicated but the advantage is that the repetition rate is independent of pumping power. The alternative technique of laser losses switching is an application of a saturable absorber which transmission changes versus internal laser flux intensity. In this technique, a material with high absorption at the laser wavelength is placed inside the laser resonator and prevents laser oscillation until the population inversion reaches a value exceeding the total optical losses (dissipative losses, saturable absorber losses, transmission losses) inside the laser cavity. As the photon density builds up following achievement of a net positive inversion, passive absorber rapidly bleaches into a high transmission state, thereby Q-switching the laser.

There are some advantages and disadvantages of both passive and active Q- switches. By using an active Q-switch, the repetition rate can be controlled by an external driver, whereas for a passive Q-switch the repetition rate is given by the internal parameters of the laser. The passive Q-switch is initiated by the laser intensity inside the resonator itself. Therefore, this Q-switch is simple — there is no need for drivers but, at the same time, there is reduced flexibility in choosing laser parameters. That means it is not possible to change the repetition rate for a given input pumping power and absorber transmission. At high depth of modulation the residual losses are also high and therefore the efficiency is low. Other disadvantages of a passive Q-switch are the large build-up of time fluctuations and the intensity instability from pulse to pulse. However, compared with active Q-switching methods, passive techniques that use saturable absorbers can significantly simplify the operation and the alignment, improve the reliability and the compactness, and reduce the costs of laser sources [2,3].

2. RATE EQUATIONS FOR PASSIVELY Q-SWITCHED ND:YAG LASERS

Passive Q-switching is currently the most attractive method of generation of ns and sub-ns laser pulses. The main characteristics of a Q-switch are as follows: concentration of absorption centers, ground state absorption (GSA) and excited state absorption (ESA) cross sections, lifetimes of the upper levels. To design the passively Q-switched laser properly it is essential to know all these parameters as well as the potential gain in the lasing medium (LM) and its scatter losses. There are some works [4,5] where

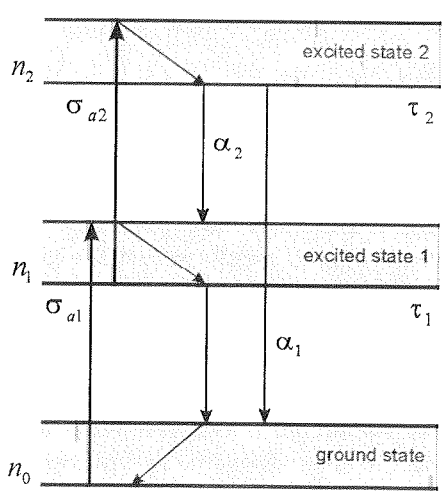


Fig. 1. Energy level diagram of a saturable absorber characterized by ESA

Rys. 1. Schemat poziomów energetycznych nieliniowego absorbera wykazującego zjawisko ESA

the problem of design and optimization of passively Q-switched lasers with saturable absorbers (SA) exhibiting ESA was solved. However, in these works the solutions were limited to the case of the so-called "slow absorber". This paper presents developed a detailed analysis of passively Q-switched lasers based on the normalized dimensionless rate equations taking into account: ESA and lifetimes of SA levels as well as the magnification in the laser cavity.

The case when a saturable absorber is characterized by only one excited state seldom occurs in nature. In reality, a saturable absorber is usually characterized by more than one excited state to which molecules from ground state are transported. The molecules, which are located in excited state, can be transported even to higher excited levels (on condition that the energy gap of adjacent excited levels equals energy of absorption quantum). This situation is depicted in Fig. 1. The parameters marked in this figure are as follows:

n_0, n_1, n_2 — population density of the ground state, the first and the second excited state, respectively, σ_{a1} — absorption cross section from the ground state to the first excited state of SA, σ_{a2} — absorption cross section from the first to the second excited state of SA, τ_1, τ_2 — lifetime of the first and the second excited state, respectively, α_1, α_2 — deactivation coefficient of SA centers from the second to the first excited state and to the ground state (the sum of α_1 and α_2 equals one, $\alpha_1 \ll \alpha_2$). According to literature reports, there are many saturable absorbers, in which excited state absorption phenomenon occurs. The most common are dielectrics doped with different active ions, like chromium, vanadium, erbium, cobalt, thulium and many others [6-9].

A passive Q-switched laser system with saturable absorber characterized by ESA can be described by four rate equations relating to average power density in a laser cavity (J), amplification in a laser medium (k), population density of the first (n_1) and second (n_2) excited state:

$$\frac{dJ(t)}{dt} = V_R \left\{ k(t) - \rho_s - \frac{l_{SA}}{l_{LM}} \sigma_{a1} \left[(N_0 - n_1(t) - n_2(t)) + \frac{\sigma_{a2}}{\sigma_{a1}} n_1(t) \right] \right\} J(t) \tag{1}$$

$$\frac{dk(t)}{dt} = \varpi_P [\chi - k(t)] - \frac{k(t)}{\tau_{LM}} - \frac{k(t)J(t)}{Es} \tag{2}$$

$$\frac{dn_1(t)}{dt} = M^2 \frac{J(t)}{Es_{a1}} \left\{ [N_0 - n_1(t) - n_2(t)] - \frac{\sigma_{a2}}{\sigma_{a1}} n_1(t) \right\} - \frac{n_1(t)}{\tau_1} + \frac{\alpha_2 n_2(t)}{\tau_2} \tag{3}$$

$$\frac{dn_2(t)}{dt} = M^2 \frac{J(t)}{Es_{a2}} n_1(t) - \frac{n_2(t)}{\tau_2} \tag{4}$$

with initial conditions:

$$k(t = 0) = k_0 \tag{5}$$

$$J(t = 0) = J_0 \tag{6}$$

where:
$Es_{a1} = h\nu_s/\sigma_{a1}$ — saturation energy of the nonlinear absorber to the first excited state;
$Es_{a2} = h\nu_s/\sigma_{a2}$ — saturation energy of the nonlinear absorber to the second excited state;
$Es = h\nu_s/\sigma_e$ — saturation energy of the laser medium (the energy that can be stored in LM);
$N_0 = n_0 + n_1 + n_2$ — total population density of SA absorbing centers;
$h\nu_s$ — photon energy; ω_p — pumping speed of the active medium; M — enlargement of an internal beam in a laser cavity (the ratio of cross-sectional area of a beam in SA to cross-sectional area of a beam in LM); l_{LM} and l_{SA} are the lengths of LM and SA, respectively; χ — theoretical maximum gain which can be obtained in LM ($\chi = n_0 \sigma_e$); σ_e — emission cross-section of LM; $V_R = c(l_{LM}/L_{opt})$ — the speed of light in the resonator; τ_{LM} — fluorescence lifetime of LM; ρ_s — static losses coefficient, which include dissipative (ρ_d) and transmission (ρ_t) losses; L_{opt} — optical length of the resonator.

To describe the Q-switched laser the point model was used. This model is good enough for passive Q-switching. The analysis and solution of the rate equations presented above allow to present the dynamics of laser generation in the system discussed.

3. NORMALIZED RATE EQUATIONS FOR PASSIVELY Q-SWITCHED ND:YAG LASERS

The mathematical description of a passively Q-switched solid-state laser, given in a prior section, includes functions and parameters whose values are dimensional and absolute. It makes more detailed physical interpretation of these equations difficult and permits to solve them only for specific cases. Obtaining the quantitative information, typical for analytical solution of the problem discussed, is impossible in this case. Therefore, the rate equations (1) — (4) should be transformed into a form in which their functions and parameters will have relative values, linked with certain material constants or laser system features. It guarantees to formulate more general conclusions related to a laser with saturable absorber as a passive switch and allow to its optimization.

In the analysis proposed the laser energy balance was taken into account. Depending on $M^2\delta_1$ parameter, being normalized pumping speed of a saturable absorber, the changes of energy generated in a laser with its division between energy emitted on static and dynamic losses were examined. The initial gain and different relations of static and dynamic losses in a laser cavity were changeable.

In connection with foregoing, the following normalized and dimensionless quantities are introduced (Tab. 1):

Table 1

Definitions of normalized constants and variables

Opis zastosowanych wielkości unormowanych

Constant	Symbol	Definition
normalized one-trip transit time of photon flux in the resonator of optical length L_{opt}	T_{LM}	$\frac{c \cdot t}{L_{opt}}$
normalized power density in the laser cavity	I	$\frac{J \cdot L_{opt}}{Es \cdot c}$
normalized gain of LM	K	$\frac{k}{\chi}$
normalized population of the first excited state of SA	N_1	$\frac{n_1}{N_0}$
normalized population of the second excited state of SA	N_2	$\frac{n_2}{N_0}$
normalized fluorescence lifetime of LM	T_{LM}	$\frac{\tau_{LM} \cdot c}{L_{opt}}$
normalized lifetime of the first excited state of SA	T_1	$\frac{\tau_1 \cdot c}{L_{opt}}$
normalized lifetime of the second excited state of SA	T_2	$\frac{\tau_2 \cdot c}{L_{opt}}$
normalized pumping speed of LM	W_p	$\frac{\varpi_p \cdot L_{opt}}{c}$
normalized static losses of the laser cavity	R_S	$\frac{\rho_s}{\chi}$
normalized initial dynamic losses of the laser cavity related to SA	Γ_{SA}	$\frac{\gamma_{SA}}{\chi}$
normalized transmission losses of the laser cavity	R_T	$\frac{\rho_t}{\chi}$
normalized dissipative losses of the laser cavity	R_D	$\frac{\rho_d}{\chi}$

Introduction of constants defined in Tab. 1 modifies the equations (1) – (4) to the following form:

$$\frac{dI}{dT} = l_{LM} \cdot \chi \cdot \{K - R_S - \Gamma_{SA} \cdot [(1 - N_1 - N_2) + \delta \cdot N_1]\} \cdot I \qquad (7)$$

$$\frac{dK}{dT} = W_p \cdot (1 - K) - \frac{K}{T_{LM}} - K \cdot I \tag{8}$$

$$\frac{dN_1}{dT} = M^2 \cdot I \cdot \delta_1 \cdot [(1 - N_1 - N_2) - \delta \cdot N_1] - \frac{N_1}{T_1} + \frac{\alpha \cdot N_2}{T_2} \tag{9}$$

$$\frac{dN_2}{dT} = M^2 \cdot I \cdot \delta_1 \cdot \delta \cdot N_1 - \frac{N_2}{T_2} \tag{10}$$

with initial conditions:

$$I(0) = \frac{\chi \cdot l_{LM}}{T_{LM}} (R_S + \Gamma_{SA}) \frac{\Omega}{4\pi} \tag{11}$$

$$\mathrm{K_{OA}}(0) = \mathrm{R_S} + \Gamma_{NA} \tag{12}$$

$$\mathrm{N_1}(0) = 0 \tag{13}$$

$$\mathrm{N_2}(0) = 0 \tag{14}$$

where:

$$\delta = \frac{\sigma_{a2}}{\sigma_{a1}} \tag{15}$$

$$\delta_1 = \frac{\sigma_{a1}}{\sigma_e} \tag{16}$$

$$\gamma_{SA} = \frac{l_{SA} \cdot N_0 \cdot \sigma_{a1}}{l_{LM}} \tag{17}$$

Ω — solid angle of a laser generation

Using the rate equations (7) — (10) it is possible to find expression describing the values of energy in a laser with a saturable absorber. It can be output energy, energy scattered on static, dynamic losses of a laser cavity or energy stored in a laser medium.

As it results from the principal of operation of the laser considered, during its pumping process, the energy is stored in it as long as the gain reaches the static losses and initial dynamic losses delivered by saturable absorber. After the generation threshold overflow, the gain in LM increases due to still working pump, however this increase is usually insignificant. Therefore, it can be assumed that energy stored in a laser medium corresponds to the state where the gain equals the initial cavity losses. Hence, the energy stored in 1 cm^2 of LM cross section surface can be given by:

$$E_{stored} = k \cdot E \cdot l_{LM} \tag{18}$$

Linking this expression with saturation energy of LM we receive the normalized stored energy:

$$E^N_{stored} = \frac{E_{stored}}{E_S} = \chi \cdot l_{LM} \cdot (R_S + \Gamma_{SA}) \tag{19}$$

After crossing the generation threshold in LM (T = 0) the lasing begins and continues until the laser medium saturates (T = T_k). The power generated is distributed on static and dynamic losses and can be described by:

$$E_{STAT} = \rho_S \cdot l_{LM} \cdot \int_0^{t_k} J dt \tag{20}$$

$$E_{DYN} = l_{LM} \int_0^{t_k} J(t) \cdot \rho_D(t)\, dt \tag{21}$$

Integrating and linking these expressions with saturation energy of an active medium we receive the normalized energy, emitting on static and dynamic losses of a laser cavity:

$$E^N_{STAT} = \frac{E_{STAT}}{E_S} = R_S \cdot \chi \cdot l_{LM} \cdot \int_0^{T_k} I dT \tag{22}$$

$$E^N_{DYN} = \frac{E_{DYN}}{E_S} = \Gamma \cdot \chi_{SA} \cdot l_{LM} \int_0^{T_k} I \cdot (1 - N_1 + \delta N_1 - N_2)\, dT \tag{23}$$

The normalized output energy is given by:

$$E^N_{out} = \frac{E_{out}}{E_S} = R_T \cdot \chi \cdot l_{LM} \cdot \int_0^{T_k} I dT \tag{24}$$

Normalized equations (7)-(10) are the nonlinear system of equations and they don't have trivial analytical solutions. Therefore they were solved by means of numerical methods.

4. NUMERICAL ANALYSIS RESULTS

The whole analysis was conducted in two aspects. In $M^2\delta_1$ parameter domain the situation where energy necessary to switch dynamic losses would be possibly small while energy emitted on static losses would be the highest was looked for. However, in dynamic to static losses ratio domain it was necessary to find the situation when the output energy of the laser would be maximum. To this end, the equations (7)-(10) were integrated and then suitable integrals occurring in equations (22)-(24) were calculated.

The whole calculations were conducted for the following date: $\chi \cdot l_{LM} = 300$, $W_p = 10^{-7}$, $T_{LM} = 10^4$, $R_s + \Gamma_{SA} = 0.01$. The participation of dynamic losses Γ_{SA} in

the laser cavity was variable and equaled 90%, 80% and 70% of the whole value of initial LM losses. The value of $M^2\delta_1$ parameter was variable in the range from 0 to 100. The normalized lifetimes of SA excited states equaled: $T_1 = 1000$ or 1 and $T_2 = 1$.

Fig. 2 and Fig. 4 depict dependence of energy emitted on static (dissipative and transmission) losses as a function of $M^2\delta_1$ for high initial gain and for two different values of T_1 time. E^N_{STAT} always rises along with the increase of $M^2\delta_1$ value. The most explicit changes occur in the range of intermediate values of $M^2\delta_1$, where the highest changes of E^N_{STAT} occur. The high value of $M^2\delta_1$ accompanies the insignificant increment of E^N_{STAT} . These relations depend on Γ/R_s ratio only to a minimum extent.

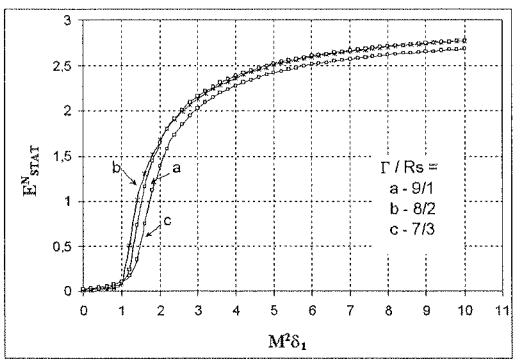


Fig. 2. Normalized energy emitted on static losses of the laser cavity vs. $M^2\delta_1$ parameter for SA without ESA. $T_1 = 1000$

Rys. 2. Unormowana energia wydzielona na stratach statycznych ośrodka aktywnego w zależności od $M^2\delta_1$ dla nieliniowego absorbera, który nie wykazuje ESA. $T_1 = 1000$

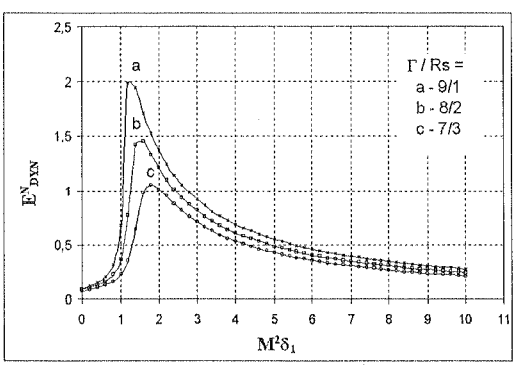


Fig. 3. Normalized energy emitted on dynamic losses of the laser cavity vs. $M^2\delta_1$ parameter for SA without ESA. $T_1 = 1000$

Rys. 3. Unormowana energia wydzielona na stratach dynamicznych ośrodka aktywnego w zależności od $M^2\delta_1$ dla nieliniowego absorbera, który nie wykazuje ESA. $T_1 = 1000$

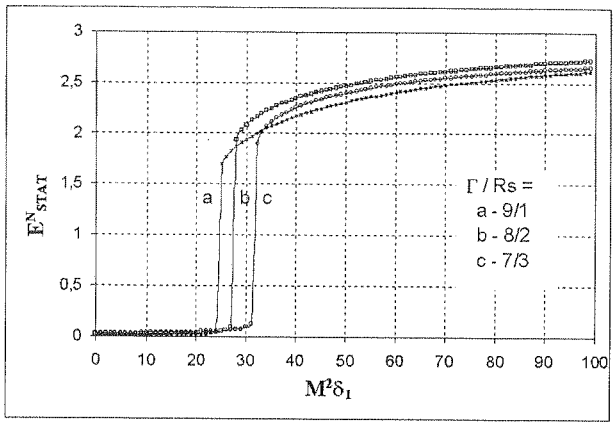


Fig. 4. Normalized energy emitted on static losses of the laser cavity vs. $M^2\delta_1$ parameter for SA without ESA. $T_1 = 1$

Rys. 4. Unormowana energia wydzielona na stratach statycznych ośrodka aktywnego w zależności od $M^2\delta_1$ dla nieliniowego absorbera, który nie wykazuje ESA. $T_1 = 1$

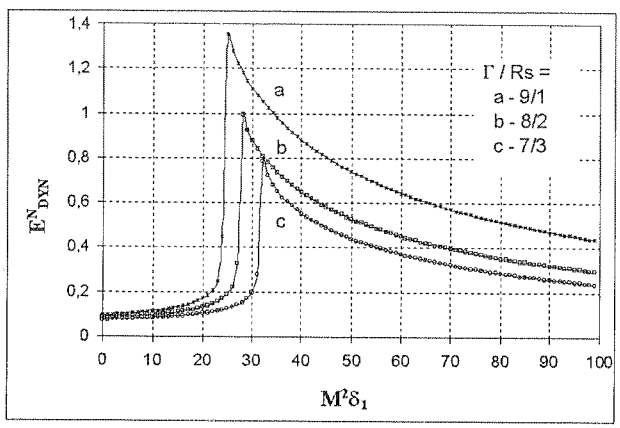


Fig. 5. Normalized energy emitted on dynamic losses of the laser cavity vs. $M^2\delta_1$ parameter for SA without ESA. $T_1 = 1$

Rys. 5. Unormowana energia wydzielona na stratach dynamicznych ośrodka aktywnego w zależności od $M^2\delta_1$ dla nieliniowego absorbera, który nie wykazuje ESA. $T_1 = 1$

In Fig. 3 and Fig. 5 the dependence of energy emitted on dynamic losses versus $M^2\delta_1$ parameter was presented. What can be quantitatively determined from this diagram is the energy which is necessary to saturate the SA. There is a clear maximum of the energy absorbed by SA — as far as effective laser losses switching is concerned, this range is the most relevant. Along with the increase of Γ/R_s ratio the maximum of E^N_{DYN} shifts towards the lowest values of $M^2\delta_1$ and the value of this maximum

decreases. It determines the necessity of using saturable absorbers characterized by higher absorption cross section or it forces us to make an artificial modification of $M^2\delta_1$ parameter by using higher enlargements M of a generated laser beam.

The behavior of a Q-switched laser with SA characterized by ESA was presented in Fig. 6 and Fig. 7.

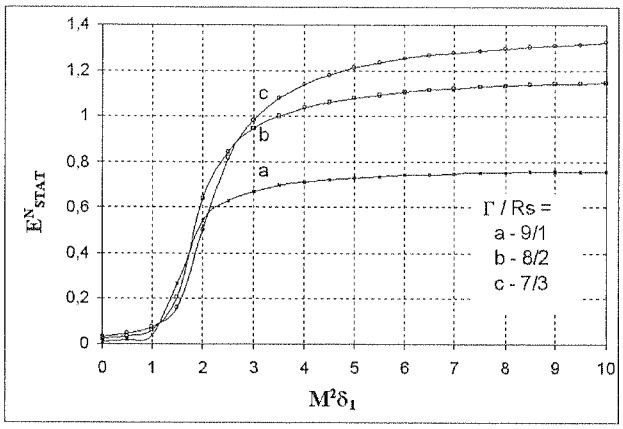


Fig. 6. Normalized energy emitted on static losses of the laser cavity vs. $M^2\delta_1$ parameter for SA with ESA. $T_1 = 1000$, $T_2 = 1$, $\delta = 0.2$

Rys. 6. Unormowana energia wydzielana na stratach statycznych ośrodka aktywnego w zależności od $M^2\delta_1$ dla nieliniowego absorbera wykazującego ESA. $T_1 = 1000$, $T_2 = 1$, $\delta = 0.2$

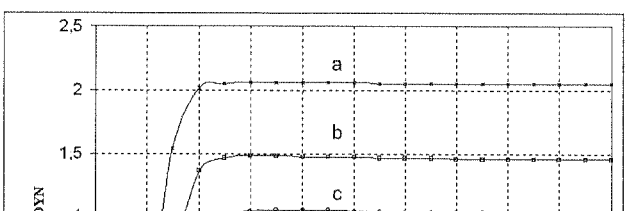


Fig. 7. Normalized energy emitted on dynamic losses of the laser cavity vs. $M^2\delta_1$ parameter for SA with ESA. $T_1 = 1000$, $T_2 = 1$, $\delta = 0.2$

Rys. 7. Unormowana energia wydzielana na stratach dynamicznych ośrodka aktywnego w zależności od $M^2\delta_1$ dla nieliniowego absorbera wykazującego ESA. $T_1 = 1000$, $T_2 = 1$, $\delta = 0.2$

It is easy to notice, that excited state absorption phenomenon in SA influences negatively laser generation efficiency causing the decrease of energy emitted on static losses and simultaneously increasing the energy which saturates the SA. The differences between E^N_{STAT} and E^N_{DYN} versus Γ/R_s ratio are higher than for SA without ESA.

From the diagrams presented above (depending on $M^2\delta_1$ parameter value) three different working regimes of a passive Q-switched laser can be specified:

1. The case when $M^2\delta_1 \gg 1$. Then the passively Q-switched laser is well designed. The laser medium is strongly saturated. In case of SA without ESA the small amount of energy stored in the laser cavity is destined for dynamic losses switching. Most of the energy stored is emitted on transmission and dissipative losses. However, as far as a laser with SA characterized by ESA is concerned, the energy saturating SA is constant versus $M^2\delta_1$ parameter value.
2. The case when $M^2\delta_1 < 1$. The gain in a LM is weakly saturated and the energy generated is very small. In extreme case, when $M^2\delta_1 \to 0$ (it can be interpreted it as using an absorber with absorption cross section close to zero and characterized by some finite transmission — resulting from absorbing centers concentration) the laser works in free running regime. The losses introduced by SA aren't changeable during the generation and such an absorber can't be called nonlinear.
3. The case when $M^2\delta_1 > 1$ (where energy changes are the highest). In this range along with the increase of $M^2\delta_1$ value the gain saturation of LM builds up rapidly, especially for high values of Γ/R_s ratio value. It is caused by absorption saturation of SA. There is an explicit maximum of energy switching the SA, whose location depends on Γ/R_s ratio. This range can be called ineffective Q-modulation.

From the analysis done so far it is in not difficult to notice that for a laser with a SA exhibiting ESA as well as a laser with a SA characterized by the lack of ESA phenomenon the relations of energy emitted on static losses RS change as a function of the ratio of the whole laser losses components. It points to the necessity of optimization of laser transmission losses R_T depending on dissipative losses R_D and $M^2\delta_1$ parameter.

Fig. 8 — Fig. 11 present dependences of normalized output laser energy as a function of Γ/R_s ratio. The dissipative losses were assumed at 1%, 5% and 10% of initial gain. On the basis of the characteristic curves presented above it is easy to point the optimum Γ/R_s ratio for which the output energy is maximum. This maximum is strongly dependent on dissipative losses level, which is obvious ($R_S = R_T + R_D$). The higher value of dissipative losses corresponds to lower laser output energy (Fig 10 — 11) and the higher value of $M^2\delta_1$ corresponds to higher output energy. Along with the increase of $M^2\delta_1$ the maximum of output energy shifts towards lower values of Γ/R_s ratio, and the level of this maximum increases. The optimal output energy is strongly limited especially from low values of Γ/R_s ratio direction. In this range the slight change of this ratio induces significant changes of output energy. This limitation occurs especially for higher values of dissipative losses (see Fig. 11).

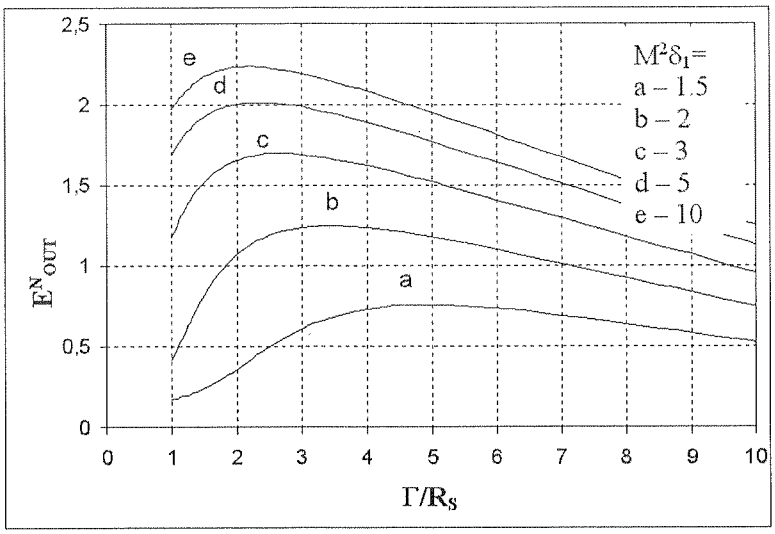


Fig. 8. Normalized output laser energy vs. Γ/Rs ratio for the laser with SA without ESA. Dissipative losses R_D = 5%. T_1 = 1000

Rys. 8. Unormowana energia wyjściowa w funkcji stosunku Γ/Rs dla lasera z nieliniowym absorberem, który nie wykazuje ESA. R_D = 5%. T_1 = 1000

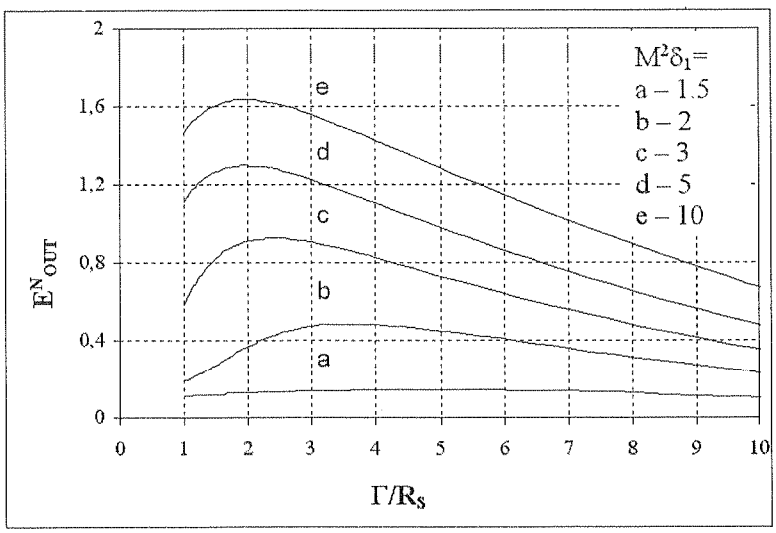


Fig. 9. Normalized output laser energy vs. Γ/Rs ratio for the laser with SA characterized by ESA. Dissipative losses R_D = 5%. T_1 = 1000, T_2 = 1, δ = 0.2

Rys. 9. Unormowana energia wyjściowa w funkcji stosunku Γ/Rs dla lasera z nieliniowym absorberem wykazującym ESA. R_D = 5%. T_1 = 1000, T_2 = 1, δ = 0.2

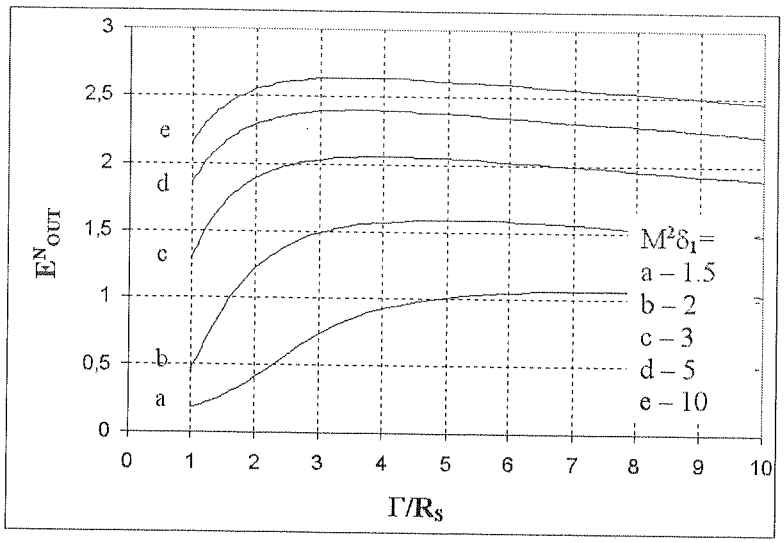


Fig. 10. Normalized output laser energy vs. Γ/Rs ratio for the laser with SA without ESA. Dissipative losses R_D = 1%. T_1 = 1000

Rys. 10. Unormowana energia wyjściowa w funkcji stosunku Γ/Rs dla lasera z nieliniowym absorberem, który nie wykazuje ESA. losses R_D = 1%. T_1 = 1000

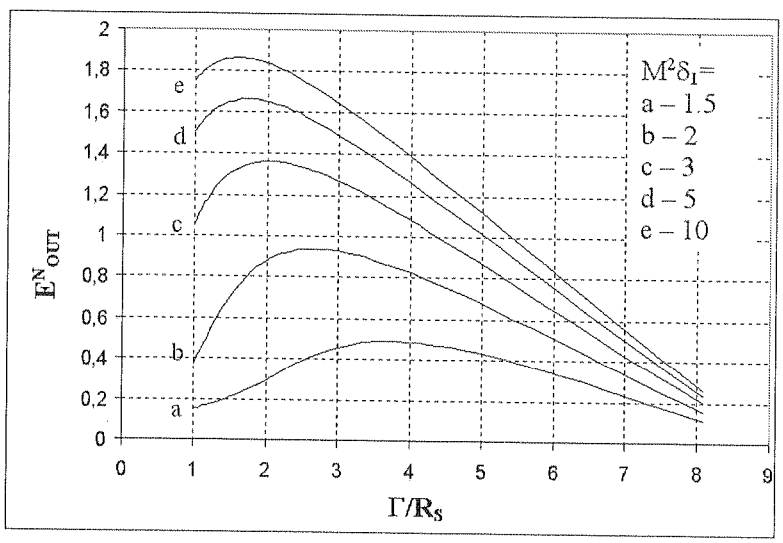


Fig. 11. Normalized output laser energy vs. Γ/Rs ratio for the laser with SA without ESA. Dissipative losses R_D = 10%. T_1 = 1000

Rys. 11. Unormowana energia wyjściowa w funkcji stosunku Γ/Rs dla lasera z nieliniowym absorberem, który nie wykazuje ESA. losses R_D = 10%. T_1 = 1000

5. CONCLUSIONS

In conclusion, the optimization of a passively Q-switched laser with saturable absorber characterized by ESA was considered. A detailed analysis of the passively Q-switched lasers based on the normalized dimensionless rate equations considering: ESA and lifetimes of the SA as well as the magnification in the laser cavity was developed. The following conclusions were drown from the analysis presented above:

- Efficient generation in a passively Q-switched laser takes place when a saturable absorber is characterized by the absence of ESA,
- In order to maximize the laser output energy the saturable absorber should be characterized by long lifetime of its first excited state as well as high value of the ratio of ground state absorption cross section of a SA to emission cross section of a LM. It guarantees high saturation of a laser medium,
- ESA phenomenon decreases the laser generation efficiency. It results from multiple absorption transitions between excited states of SA. The ESA phenomenon causes: the decrease of energy emitted on static (dissipative and transmission) losses of a laser cavity, the increase of energy level which is necessary to "switch" saturable absorber, and the fall of laser medium gain saturation,
- The level of energy emitted on the whole laser cavity losses depends on the ratio of its particular components. This relation occurs especially in case of SA characterized by ESA.

6. ACKNOWLEDGEMENTS

The author would like to thank Marek Skorczakowski and Andrzej Zajac from Laser Technology Laboratory (Institute of Optoelectronics) for their constructive comments.

7. REFERENCES

1. W. Koechner: *Solid State Laser Engineering.* 5th edition, Springer-Verlag, 2001.
2. L. Chen, S. Zhano, J. Zheng, Z. Cheng, H. Chen:*Characteristics of a passively Q-switched Nd^{3+}:$NaY(WO_4)_2$ laser with Cr^{4+}:YAG saturable absorber.* Optics & Laser Technology, vol. 34, 2002, pp. 347-350.
3. J. Liu, J. Yang, J. He: *High repetition rate passively Q- switched diode pumped Nd:YVO_4 laser.* Optics & Laser Technology, vol. 35, 2003, pp. 431-434.
4. J. Degnan: *Optimization of Passively Q–Switched Lasers.* IEEE Journal of Quantum Electronics, vol. 31, 1995, pp. 1890-1901.
5. X. Zhang, S. Zhao, Q. Wang, Q. Zhang, L. Sun, S. Zhang: *Optimization of Cr^{4+}-doped saturable — absorber Q-switched lasers.* IEEE Journal of Quantum Electronics, vol. 33, 1997, pp. 2286-2294.
6. M. B. Camargo, R. D. Stultz, M. Birnbaum: *Co^{+2}:YSGG saturable absorber Q-switched for infrared erbium lasers.* Optics Letters, vol. 20, 1995, pp. 339-341.

7. V. B. Sigacher, T. T. Basier, M. E. Doroshenko, V. V. Osiko, A. G. Popashvili: *1.3μm neodymium laser passively Q — switched with Nd^{+2}:SrF_2 and V^{+3}:YAG crystal and Raman shifting to eye — safe region.* OSA proceedings on Advanced Solid — State Lasers, vol. 24, 1995, pp. 460-463.
8. Y. K. Kuo, M. Birnbaum, W. Chen, K. Yue, M. F. Huang: *Passive Q — switching of the Tm, Cr:YAG 2μm laser with a HLF solid — state saturable absorber. OSA proceedings on Advanced Solid — State Lasers.* vol. 24, 1995, pp. 445-449.
9. U. Griebner, R. Koch: *Passively Q — switched Nd:glass fibre — bundle laser.* Electronics Letters, vol. 31, 1995, pp. 205-208.

J. ŚWIDERSKI

MODELOWANIE NUMERYCZNE DZIAŁANIA LASERÓW Nd:YAG Z PASYWNĄ MODULACJĄ DOBROCI REZONATORA

Streszczenie

W niniejszym opracowaniu została przedstawiona idea pracy oraz matematyczny opis dynamiki generacji lasera z pasywną modulacją dobroci. Analiza dotyczyła wpływu parametrów nieliniowego absorbera na sprawność generacji lasera. Pod uwagę został tutaj wzięty nieliniowy absorber wykazujący zjawisko absorpcji ze stanów wzbudzonych (ESA). Dokonana analiza wskazała na optymalne warunki pracy omawianego układu laserowego.

Z przeprowadzonej analizy wynikają następujące wnioski:

- Sprawna generacja w układzie laserowym z pasywną modulacją dobroci ma miejsce wówczas, gdy nieliniowy absorber nie wykazuje zjawiska ESA.
- Celem maksymalizacji energii wyjściowej lasera nieliniowy absorber powinien charakteryzować się długim czasem życia poziomu wzbudzonego oraz dużą wartością stosunku jego absorpcyjnego przekroju czynnego do emisyjnego przekroju czynnego ośrodka aktywnego. Gwarantuje to dobre wysycenie ośrodka aktywnego.
- Zjawisko ESA zmniejsza sprawność generacji rozpatrywanego układu. Wynika to z wielokrotnych przejść absorpcyjnych pomiędzy pierwszym i drugim poziomem wzbudzonym. Zjawisko ESA powoduje:
  - — malenie poziomu energii wydzielonej na stratach statycznych rezonatora,
  - — wzrost poziomu energii niezbędnej na przełączenie nieliniowego absorbera,
  - — zmniejszenie wysycenia ośrodka aktywnego.
- W dziedzinie stosunku strat statycznych i dynamicznych występuje wyraźne maksimum energii wyjściowej lasera. Przy wzroście parametru $M^2\delta_1$ maksimum to przesuwa się w kierunku mniejszych wartości wspomnianego stosunku, przy czym wartość tego maksimum wzrasta.

Optymalizacja konstrukcji lasera z przełączaniem jest zagadnieniem oczywistym i bardzo istotnym zarazem – zwłaszcza, że układy te znajdują szerokie zastosowanie w systemach automatyki przemysłowej, w urządzeniach optoelektronicznych dla celów medycznych, ochrony środowiska i metrologii, jak również w nadajnikach dalmierzy, układach naprowadzania rakiet i oświetlania celów.

Słowa kluczowe: lasery ciała stałego z przełączaniem strat rezonatora, nieliniowy absorber, analiza numeryczna

KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJI, 2005, 51, z. 2, ss. 255–275

Optymalizacja z wykorzystaniem zmodyfikowanej sieci Hopfielda

ZBIGNIEW NAGÓRNY, ANDRZEJ KOS

Zbigniew Nagórny
Studium Generale Sandomiriense
Wyższa Szkoła Humanistyczno-Przyrodnicza w Sandomierzu
ul. Krakowska 26
27-600 Sandomierz
e-mail: nagorny@interia.pl

Andrzej Kos
Akademia Górniczo-Hutnicza w Krakowie
Katedra Elektroniki
al. Mickiewicza 30
30-059 Kraków
e-mail: kos@agh.edu.pl

Otrzymano 2004.03.18
Autoryzowano 2004.06.16

W pracy przedstawiono oryginalną metodę rozwiązywania problemów optymalizacyjnych z wykorzystaniem zmodyfikowanej sieci neuronowej typu Hopfielda. W klasycznym rozwiązaniu wartości wag połączeń między neuronami w sieci Hopfielda są obliczane przed symulacją i nie ulegają zmianie. W niniejszej pracy zbadano wpływ modyfikacji wartości sygnałów wejściowych neuronów lub wag podczas symulacji sieci, na otrzymywane rozwiązania. Zauważono, że modyfikacja sygnałów wejściowych lub wag umożliwia uzyskanie lepszych rozwiązań. W pracy przedstawiono rezultaty otrzymane zmodyfikowaną siecią dla problemu komiwojażera. Dla problemów o wymiarze równym 10, sieć dla 100% prób generowała optymalne rozwiązanie. Dla większych problemów rozwiązania są lepsze od otrzymanych klasyczną siecią. Przedstawiono wnioski płynące z zastosowania opisanej metody.

Słowa kluczowe: optymalizacja, sieć neuronowa, sieć Hopfielda, funkcja energii, problem komiwojażera

1. WPROWADZENIE

Od wielu lat prowadzone są badania naukowe zmierzające do znalezienia nowych metod optymalizacji, ze względu na niedogodność stosowania metod tradycyjnych. Rozwiązanie problemu optymalizacyjnego polega na znalezieniu minimum globalnego funkcji kosztu, która w sposób sformalizowany opisuje ten problem. Optymalne rozwiązanie zapewnia więc osiągnięcie celu, przy minimalnych kosztach. Podstawową wadą tradycyjnych metod optymalizacji jest zbyt długi czas potrzebny na znalezienie rozwiązania. Istnieje cały szereg metod rozwiązywania problemów NP-zupełnych: heurystyczne [1, 5, 8, 14, 21], wykorzystujące sieci neuronowe [1-4, 6-17, 21-25], algorytmy genetyczne [14, 26], symulowane wyżarzanie [6-11, 14, 18-20, 23], metaheurystyka tabu [14], Ant Colony Optimization [14].
W szczególności sieci neuronowe cieszą się dużym zainteresowaniem oraz uznaniem ze względu na zdolność współbieżnego przetwarzania danych przez wiele pracujących równocześnie elementów sieci, co umożliwia uzyskanie rozwiązania w krótkim czasie. W dodatku wzrost wymiaru problemu nie pociąga za sobą istotnego wzrostu czasu oczekiwania na wynik w przypadku realizacji sprzętowej sieci [7]. Do rozwiązywania problemów optymalizacyjnych stosowany jest m. in. perceptron wielowarstwowy [1], sieci samoorganizujące się [2, 4, 9, 15, 16], sieć Hopfielda [1, 3, 6-14, 17, 21-25].
W niniejszej pracy do rozwiązania problemu NP-zupełnego zastosowano sieć neuronową Hopfielda. Oryginalny wkład autorów polega na modyfikacji wartości sygnałów wejściowych neuronów lub wartości wag. Metoda opiera się na zaproponowanej przez autorów regule zastosowanej przy projektowaniu optymalnej topografii układów VLSI [27], metoda ta może być stosowana do rozwiązywania innych problemów optymalizacyjnych. Wprowadzona modyfikacja umożliwia poprawę skuteczności sieci. Opis klasycznej sieci Hopfielda stosowanej w problemach optymalizacyjnych znajduje się w rozdziale 2. Modyfikacja tej sieci, umożliwiająca poprawę skuteczności działania jest przedstawiona w rozdziale 3. Rozdział 4 zawiera wyniki uzyskane przy pomocy zmodyfikowanej sieci dla problemu komiwojażera. W tej części odniesiono się również do znanych z literatury rozwiązań. Wnioski płynące z zastosowania przedstawionej metody zamieszczono w rozdziale 5.

2. ZASTOSOWANIE SIECI HOPFIELDA DO ROZWIĄZYWANIA PROBLEMÓW OPTYMALIZACYJNYCH

W niniejszym rozdziale przedstawione są podstawy teoretyczne zastosowania sieci Hopfielda do rozwiązywania problemów optymalizacyjnych. Opis sieci oraz postać funkcji energii znajduje się w podrozdziale 2.1. Analog elektryczny oraz sposób symulacji sieci jest przedstawiony w podrozdziale 2.2.

2.1. OPIS SIECI

W 1985r. J. J. Hopfield i D. W. Tank przedstawili sposób zastosowania sieci rekurencyjnej do rozwiązywania problemów optymalizacyjnych [6]. Do tego celu została zastosowana sieć rekurencyjna z ciągłym czasem oraz z ciągłą funkcją aktywacji neuronów, zamiast dwuwartościowej 0/1 lub ±1. Liczne zastosowania sieci można znaleźć w pracach [1, 3, 7-14, 21-25]. Stosowanie dwuwartościowych neuronów zwykle nie przynosi dobrych rezultatów w rozwiązywaniu problemów optymalizacyjnych [9], jakkolwiek zdarzają się wyjątki [8].

W modelu ciągłym najczęściej stosowana jest funkcja aktywacji neuronów w postaci funkcji sigmoidalnej lub tangens hiperboliczny, funkcje te określone są wzorami [1, 3, 6-14, 21-25]

$$V = f(U) = \frac{1}{1 + e^{-\alpha U}} \tag{1}$$

$$V = f(U) = 1/2(1 + \tanh(\alpha U)) \tag{2}$$

gdzie: U — sygnał wejściowy neuronu, V — sygnał wyjściowy, α — parametr.

Dla dużych wartości α funkcja aktywacji jest bardzo stroma i zbliża się do funkcji skoku jednostkowego.

Dla tak określonej sieci wprowadza się pojęcie energii sieci. Duże wartości α we wzorach 1 oraz 2, prowadzą do postaci funkcji energii bez wyrażeń całkowych [1, 6-8, 10-13]

$$E = -1/2 \sum_{i=1}^{N} \sum_{j=1}^{N} \omega_{ij} V_i V_j - \sum_{i=1}^{N} I_i V_i \tag{3}$$

gdzie: E — energia sieci, N — liczba neuronów, ω_{ij} — waga połączenia wyjścia neuronu j z wejściem neuronu i, I_i — zewnętrzny sygnał wejściowy neuronu i.

Dla sieci z symetryczną macierzą wag połączeń między neuronami funkcja energii jest funkcją nierosnącą w czasie [6-12]. Z powyższej własności można wyciągnąć wniosek, że sieć zainicjowana pewnymi wartościami początkowymi będzie zmierzać do stanu stabilnego, odpowiadającego minimum lokalnemu lub globalnemu. Jeżeli rozwiązywany problem optymalizacyjny da się sprowadzić do problemu minimalizacji funkcji kwadratowej to sieć może posłużyć do rozwiązania tego problemu. Sieć neuronowa minimalizując swoją energię może jednocześnie znaleźć minimum funkcji, która jest sformalizowanym zapisem problemu. Funkcja ta zwykle nazywa się funkcją kosztu lub celu i może być zapisana w postaci

$$E(V) = \frac{A}{2} h_1(V) + \frac{B}{2} h_2(V) + \frac{C}{2} h_3(V) + \frac{D}{2} h_4(V) \tag{4}$$

gdzie: E — funkcja kosztu, h_1, h_2, h_3, h_4 — funkcje stanowiące składniki funkcji kosztu, A, B, C, D — stałe, V — sygnał wyjściowy neuronu.

We wzorze 4 należy zastosować takie funkcje $h(V)$, które prawidłowo określą dany problem optymalizacyjny, razem ze wszystkimi ograniczeniami wymaganymi od

poprawnego rozwiązania. W tym celu należy określić liczbę niezbędnych neuronów tworzących sieć oraz sposób interpretowania sygnałów wyjściowych poszczególnych neuronów. Do rozwiązania danego problemu optymalizacyjnego angażowana jest cała sieć neuronowa. Następnie należy określić wartości wag połączeń ω_{ij} między neuronami oraz wartości zewnętrznych sygnałów wejściowych I_i — poprzez przyrównanie wzoru 3 określającego energię sieci do wzoru 4 wyrażającego funkcję kosztu problemu. Podczas rozwiązywania problemów optymalizacyjnych z użyciem sieci Hopfielda zwykle pojawia się problem, polegający na osiąganiu przez sieć minimów lokalnych funkcji energii. Zjawisko to powoduje pogorszenie skuteczności metody. Rozwiązaniem w ramach klasycznego modelu sieci jest wykonywanie wielu prób z inicjalizacją sieci losowymi wartościami. Wielokrotne próby uwzględniające różne punkty startowe zwiększają prawdopodobieństwo znalezienia optymalnego rozwiązania. Niestety takie rozwiązania dają stosunkowo małe prawdopodobieństwo osiągnięcia minimum globalnego. Klasyczna sieć Hopfielda często dostarcza znacznie gorszych wyników w porównaniu z innymi metodami [8, 14, 20, 21, 23]. Ta obserwacja stała się inspiracją dla usprawnienia tejże sieci w możliwie prosty sposób.

2.2. ANALOG ELEKTRYCZNY SIECI HOPFIELDA

Model sieci Hopfielda z ciągłą charakterystyką neuronów jest najczęściej stosowany do rozwiązywania problemów optymalizacyjnych. Rysunek 1 przedstawia schemat układu elektronicznego, realizującego model ciągły sieci.
Rolę neuronów pełnią wzmacniacze operacyjne o nieliniowej charakterystyce. Rezystor R_{ij} decyduje o wartości wagi połączenia wyjścia neuronu j z wejściem neuronu i. Rezystancja R_i oraz pojemność C_i odpowiadają rezystancji oraz pojemności wejściowej membrany rzeczywistego neuronu, w naszym przypadku i-tego neuronu [6, 9-11]. Napięcie na wejściu oraz na wyjściu wzmacniacza operacyjnego odpowiada wejściowemu oraz wyjściowemu sygnałowi neuronu. Wartość prądu dostarczanego przez źródło prądu I_i określa zewnętrzny sygnał wejściowy i-tego neuronu.

Dla układu z rysunku 1 można zapisać równanie Kirchhoffa dla i-tego węzła

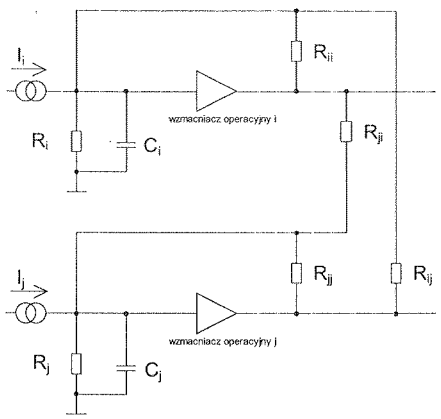
$$C_i\frac{dU_i}{dt} = \sum_{j=1}^{N}\frac{1}{R_{ij}}\left(V_j - U_i\right) + I_i - \frac{U_i}{R_i} = -U_i\left(\frac{1}{R_i} + \sum_{j=1}^{N}\frac{1}{R_{ij}}\right) + \sum_{j=1}^{N}\frac{1}{R_{ij}}V_j + I_i \qquad (5)$$

Wprowadza się rezystancję zastępczą r_i taką, że

$$\frac{1}{r_i} = \frac{1}{R_i} + \sum_{j=1}^{N}\frac{1}{R_{ij}} \qquad (6)$$

Równanie sieci otrzymuje postać

$$C_i\frac{dU_i}{dt} = \sum_{j=1}^{N}\frac{1}{R_{ij}}V_j + I_i - \frac{U_i}{r_i} = \sum_{j=1}^{N}\omega_{ij}V_j + I_i - \frac{U_i}{r_i} \qquad (7)$$



Rys. 1. Model elektryczny dwu-neuronowego wycinka sieci Hopfielda

Fig. 1. Two-neuron segment of the Hopfield net in electronic components

gdzie: $\omega_{ij} = 1/R_{ij}$ jest wagą połączenia wyjścia neuronu j z wejściem neuronu i. Dobierając pojemności kondensatorów oraz rezystancje

$$C_i = C_j = C \qquad \text{dla} \qquad \text{i, j} = 1, \ldots, \text{N} \tag{8}$$

$$r_i = r_j = R \qquad \text{dla} \qquad \text{i, j} = 1, \ldots, \text{N} \tag{9}$$

ω_{ij} zawiera arbitralnie dobrane stałe, a sygnał I_i może przyjmować dowolną wartość, po podzieleniu równania 7 przez C i po przedefiniowaniu $\omega_{ij} := \omega_{ij}/C$, $I_i := I_i/C$ równanie 7 przyjmuje postać

$$\frac{dU_i}{dt} = \sum_{j=1}^{N} \omega_{ij} V_j + I_i - \frac{U_i}{RC} \tag{10}$$

gdzie: $R\,C$ jest stałą czasową sieci, która oznaczana jest przez τ.
Otrzymano układ równań różniczkowych, który można rozwiązać numerycznie stosując metodę Eulera. Układ równań różniczkowych jest zastępowany wówczas układem równań różnicowych [8, 10, 24]

$$U_i(t + \Delta t) = U_i(t) + \Delta t \left(\sum_{j=1}^{N} \omega_{ij} V_j + I_i - \frac{U_i(t)}{\tau} \right) \tag{11}$$

gdzie: Δt jest krokiem czasowym metody.
W dalszej części zostanie przedstawiona modyfikacja klasycznej sieci Hopfielda.

3. MODYFIKACJA SIECI HOPFIELDA

W tradycyjnej metodzie optymalizacji z użyciem sieci Hopfielda wartości wag są wyznaczane na początku symulacji i nie ulegają zmianie [1, 3, 6-14, 21-25]. W

niniejszej pracy przedstawiono oryginalne rozwiązanie polegające na modyfikacji wartości sygnałów wejściowych neuronów lub wag sieci Hopfielda podczas symulacji. Obydwie metody modyfikacji są równoważne. W pracy zbadano wpływ modyfikacji sygnałów wejściowych lub wag na otrzymywane rozwiązania problemu optymalizacyjnego, na przykładzie problemu komiwojażera (ang. Travelling Salesman Problem, TSP). Do badań napisano specjalny program w języku C++ symulujący pracę sieci. Program dokonywał obliczeń na podstawie wzoru 11, a krok czasowy Δt w tym wzorze odpowiadał jednej iteracji.

Metoda polega na wykonaniu zadanej liczby prób, podczas których dokonywana jest modyfikacja sygnałów wejściowych lub wag. Przed rozpoczęciem symulacji ustala się wartości stałych A, B, C, D na podstawie symulacji wstępnych. Na początku każdej próby sieć jest inicjowana losowymi wartościami sygnałów wejściowych neuronów. Następnie w sposób losowy wybierana jest liczba iteracji *nnar*, podczas wykonywania tych iteracji następuje modyfikacja sygnałów wejściowych neuronów, polegająca na tłumieniu sygnałów dochodzących do poszczególnych neuronów przez sprzężenia zwrotne. Liczba iteracji *nnar* wybierana jest losowo z przedziału $2000 \div 5000$ dla $\Delta t = 0,01\tau$ tak, aby czas trwania tych iteracji był kilkadziesiąt razy większy od stałej czasowej sieci τ. Wartości zewnętrznych sygnałów wejściowych neuronów nie są poddawane żadnym zmianom. Należy przyjąć taki sposób modyfikacji sygnałów wejściowych, aby po wykonaniu *nnar* iteracji nie występowało już tłumienie tych sygnałów. Na końcu każdej próby jest wykonywana pewna liczba iteracji, ale już bez modyfikacji sygnałów wejściowych, w celu uzyskania stabilnego rozwiązania. Otrzymane rozwiązanie jest następnie zapisywane. Po wykonaniu zadanej liczby prób wybierane są najlepsze rozwiązania. Rysunek 2 przedstawia algorytm postępowania.

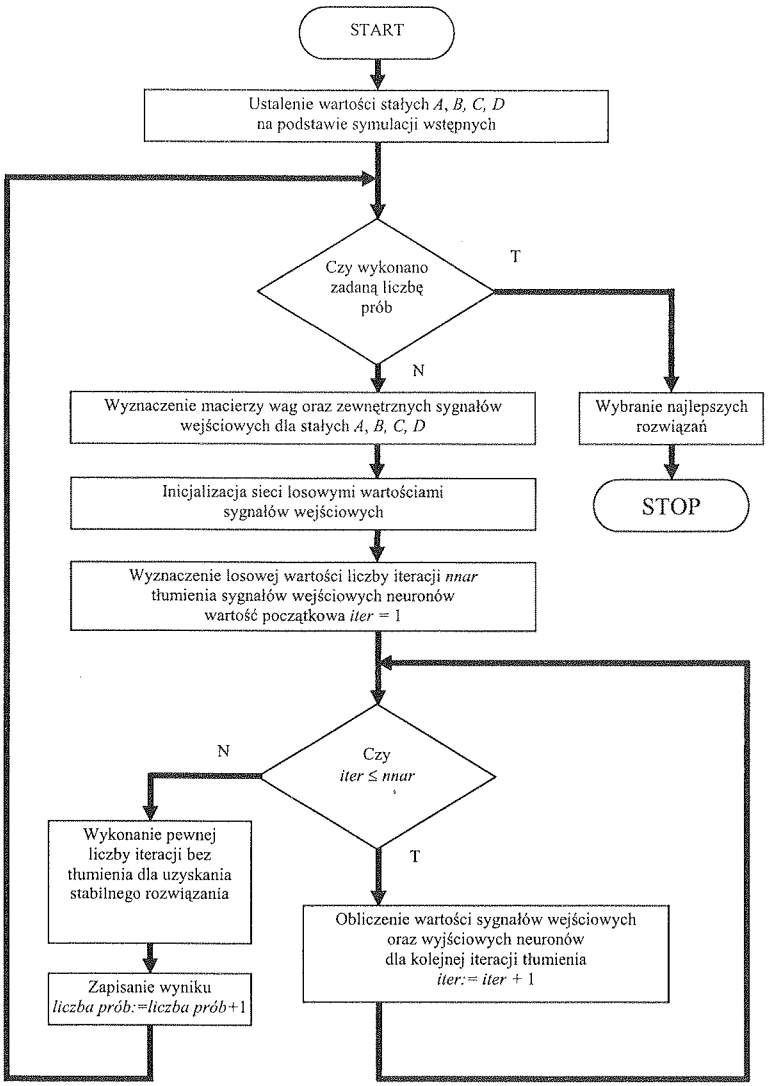
W niniejszej pracy zbadano jeden ze sposobów modyfikacji sygnałów wejściowych, taki sam dla wszystkich sygnałów, polegający na liniowej zmianie tłumienia sygnałów wejściowych w kolejnych iteracjach. Wzór 11 przyjmuje dla tego przypadku postać:

$$\begin{aligned} U_i(t+\Delta t) &= U_i(t) + \Delta t\left(\sum_{j=1}^{N} \omega_{ij}\frac{iter}{nnar}V_j + I_i - \frac{U_i(t)}{\tau}\right) = \\ &= U_i(t) + \Delta t\left(\frac{iter}{nnar}\sum_{j=1}^{N} \omega_{ij}V_j + I_i - \frac{U_i(t)}{\tau}\right) \end{aligned} \tag{12}$$

gdzie: *iter* jest numerem kolejnej iteracji, $iter = 1, \ldots, nnar$.

Drugi sposób realizacji nowej metody polega na zmianie wartości wag i jest równoważny metodzie z modyfikacją sygnałów wejściowych neuronów. Po zainicjowaniu sieci następuje wykonanie *nnar* iteracji, podczas których zmieniane są wartości wag. Należy przyjąć taki sposób zmiany wag, aby po wykonaniu *nnar* iteracji sieć posiadała takie same wartości wag jak sieć stosowana w tradycyjnej metodzie — w wyniku porównania wzorów 3 oraz 4. Sposób ze zmianą wag jest jednak znacznie trudniejszy w realizacji sprzętowej oraz znacznie mniej wydajny w realizacji programowej, ponieważ dla sieci złożonej z N^2 neuronów wymagana jest modyfikacja N^4 wag w każdej

iteracji zmiany wag. Nie można w tym przypadku zastosować uproszczenia użytego we wzorze 12 dla metody z modyfikacją sygnałów wejściowych, gdzie wykorzystano własność, że wszystkie sygnały docierające do danego neuronu przez sprzężenia zwrotne są tłumione w takim samym stopniu. Własność ta pozwala na pominięcie N^4 operacji, podczas których w ogólnym przypadku wyznaczane są wartości stłumionych sygnałów, docierających do neuronów przez poszczególne sprzężenia zwrotne.



Rys. 2. Algorytm metody z modyfikacją sygnałów wejściowych

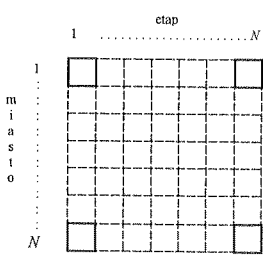
Fig. 2. The algorithm of the method with input signals modifying

4. WYNIKI UZYSKANE PRZY POMOCY ZMODYFIKOWANEJ SIECI HOPFIELDA

Niniejszy rozdział zawiera wyniki otrzymane przy pomocy zmodyfikowanej sieci Hopfielda dla problemu komiwojażera. W podrozdziale 4.1 znajduje się opis problemu. Postać funkcji kosztu stosowanej w problemie komiwojażera oraz wzory określające wartości wag i zewnętrznych sygnałów wejściowych neuronów przedstawione są w podrozdziale 4.2. Wyniki uzyskane dla sześciu przykładów problemu komiwojażera zamieszone są w podrozdziale 4.3.

4.1. PROBLEM KOMIWOJAŻERA

Problem komiwojażera jest jednym z klasycznych kombinatorycznych problemów optymalizacyjnych, jest to tzw. problem NP-zupełny [6-12, 14, 16-18]. Problem ten w najczęściej spotykanej postaci polega na znalezieniu trasy przejazdu komiwojażera łączącej N miast tak, aby koszt przejazdu był najmniejszy z możliwych, każde miasto było odwiedzone dokładnie jeden raz oraz podróż komiwojażera zakończyła się w mieście, z którego wyruszył. Ponadto zakłada się, że koszt przejazdu między dwoma dowolnymi miastami spośród wszystkich N miast nie zależy od kierunku przejazdu między tymi miastami. Dla tak postawionego problemu dla $N > 2$ miast istnieje $N!/2N$ różnych tras przejazdu, ponieważ każda trasa spośród $N!$ możliwych tras może rozpoczynać się w jednym spośród N miast oraz przebiegać w jednym z dwóch możliwych kierunków. W celu znalezienia optymalnej drogi dla problemu komiwojażera o wymiarze N buduje się sieć, w której liczba neuronów wynosi N^2. Sieć składa się z N wierszy, zawierających N neuronów każdy, zgodnie z rysunkiem 3. Każdy neuron w sieci jest więc indeksowany dwoma wskaźnikami. Pierwszy wskaźnik określa numer miasta, a drugi oznacza pozycję miasta na trasie. Jeżeli neuron o współrzędnych (x, i) po osiągnięciu przez sieć stanu stabilnego posiada wartość $V_{xi} = 1$, wówczas oznacza to, że miasto x powinno być odwiedzone podczas i-tego etapu podróży. Należy dodać, że podział sieci na wiersze ma charakter umowny i nie ma żadnego wpływu na topologię połączeń między neuronami.



Rys. 3. Podział sieci na wiersze i kolumny

Fig. 3. The division of the net for rows and columns

4.2. POSTAĆ FUNKCJI KOSZTU

Można teraz przystąpić do określenia funkcji kosztu, która będzie poddawana procesowi minimalizacji. Funkcja kosztu będzie zawierać składnik zawierający sumę kosztów poszczególnych etapów podróży, ale musi zawierać także ograniczenia gwarantujące poprawność otrzymanej trasy. Postać funkcji kosztu dla problemu komiwojażera, która została zaproponowana przez Hopfielda i Tanka [6] nie jest postacią optymalną z punktu widzenia skuteczności sieci [8, 20, 23, 24]. Sieć często dostarcza rozwiązań, które są syntaktycznie niepoprawne. Powodem tego jest globalne ograniczenie, określające liczbę neuronów w całej sieci, posiadających sygnał wyjściowy równy 1. Ograniczenie to powoduje, że wpływ sygnałów wejściowych poszczególnych neuronów na poprawność rozwiązania jest uśredniany, ponieważ wszystkie sygnały wyjściowe neuronów są traktowane w sposób równorzędny [8]. W celu wyeliminowania tego zjawiska została wprowadzona nowa postać energii, w której ograniczenie globalne zastąpiono ograniczeniami określającymi dokładną liczbę neuronów posiadających sygnał wyjściowy równy 1, w poszczególnych wierszach oraz kolumnach sieci. W rezultacie funkcja kosztu będzie składała się z następujących składników (ograniczeń) [8, 23]:

a) Ograniczenie 1

Każde miasto może być odwiedzone dokładnie jeden raz. Uwzględniając podział sieci na wiersze zgodnie z rysunkiem 3, ograniczenie to wymaga, aby w stanie stabilnym w każdym wierszu sieci dokładnie jeden neuron posiadał na wyjściu sygnał równy 1. Wówczas składnik ten będzie miał wartość 0. Ograniczenie wyrażone jest wzorem

$$
\begin{aligned}
E_1 &= A/2 \sum_{x=1}^{N} \left(\sum_{i=1}^{N} V_{xi} - 1 \right)^2 = \\
&= A/2 \sum_{x=1}^{N} \left(\left(\sum_{i=1}^{N} V_{xi} \sum_{i=1}^{N} V_{xi} \right) - 2 \sum_{i=1}^{N} V_{xi} + 1 \right) = \\
&= A/2 \sum_{x=1}^{N} \sum_{i=1}^{N} \sum_{i'=1}^{N} V_{xi} V_{xi'} - A \sum_{x=1}^{N} \sum_{i=1}^{N} V_{xi} + AN/2 = \\
&= A/2 \sum_{x=1}^{N} \sum_{i=1}^{N} \sum_{x'=1}^{N} \sum_{i'=1}^{N} \delta_{xx'} V_{xi} V_{x'i'} - A \sum_{x=1}^{N} \sum_{i=1}^{N} V_{xi} + AN/2
\end{aligned}
\tag{13}
$$

gdzie: A jest stałą, $\delta_{xx'} = \begin{cases} 1 & \text{gdy} \quad x = x' \\ 0 & \text{gdy} \quad x \neq x' \end{cases}$.

b) Ograniczenie 2

W każdym etapie podróży komiwojażer powinien odwiedzić dokładnie jedno miasto. Składnik ten jest równy 0, gdy w stanie stabilnym w każdej kolumnie sieci do-

kładnie jeden neuron będzie posiadał sygnał wyjściowy równy 1. Ograniczenie można zapisać w następujący sposób

$$\begin{aligned} E_2 &= B/2 \sum_{i=1}^{N} \left(\sum_{x=1}^{N} V_{xi} - 1 \right)^2 = \\ &= B/2 \sum_{i=1}^{N} \left(\left(\sum_{x=1}^{N} V_{xi} \sum_{x=1}^{N} V_{xi} \right) - 2 \sum_{x=1}^{N} V_{xi} + 1 \right) = \\ &= B/2 \sum_{i=1}^{N} \sum_{x=1}^{N} \sum_{x'=1}^{N} V_{xi} V_{x'i} - B \sum_{x=1}^{N} \sum_{i=1}^{N} V_{xi} + B\, N/2 = \\ &= B/2 \sum_{x=1}^{N} \sum_{i=1}^{N} \sum_{x'=1}^{N} \sum_{i'=1}^{N} \delta_{ii'} V_{xi} V_{x'i'} - B \sum_{x=1}^{N} \sum_{i=1}^{N} V_{xi} + B\, N/2 \end{aligned} \tag{14}$$

gdzie: B jest stałą.

c) Ograniczenie 3

Sygnały wyjściowe wszystkich neuronów w stanie stabilnym powinny wynosić 0 lub 1 lub być bliskie tym wartościom. Ograniczenie to zapisujemy następującym wzorem

$$\begin{aligned} E_3 &= C/2 \sum_{x=1}^{N} \sum_{i=1}^{N} V_{xi} (1 - V_{xi}) = C/2 \sum_{x=1}^{N} \sum_{i=1}^{N} V_{xi} - C/2 \sum_{x=1}^{N} \sum_{i=1}^{N} V_{xi} V_{xi} = \\ &= C/2 \sum_{x=1}^{N} \sum_{i=1}^{N} V_{xi} - C/2 \sum_{x=1}^{N} \sum_{i=1}^{N} \sum_{x'=1}^{N} \sum_{i'=1}^{N} \delta_{xx'} \delta_{ii'} V_{xi} V_{x'i'} \end{aligned} \tag{15}$$

gdzie: C jest stałą.

d) Ograniczenie 4

Składnik określający sumę kosztów poszczególnych etapów podróży komiwojażera wyrażony jest następującym wzorem

$$\begin{aligned} E_4 &= D/2 \sum_{x=1}^{N} \sum_{\substack{x'=1 \\ x' \neq x}}^{N} \sum_{i=1}^{N} d_{xx'} V_{xi} (V_{x',i+1} + V_{x',i-1}) = \\ &= D/2 \sum_{x=1}^{N} \sum_{i=1}^{N} \sum_{x'=1}^{N} \sum_{i'=1}^{N} d_{xx'} V_{xi} V_{x'i'} (\delta_{i',i+1} + \delta_{i',i-1}) \end{aligned} \tag{16}$$

gdzie: D jest stałą, $d_{xx'}$ jest kosztem drogi między miastami x oraz x', wskaźniki $i+1$ oraz $i-1$ obliczane są modulo N tzn.

$$i+1 = \begin{cases} i+1 & \text{dla} \quad i < N \\ 1 & \text{dla} \quad i = N \end{cases}, \qquad i-1 = \begin{cases} i-1 & \text{dla} \quad i > 1 \\ N & \text{dla} \quad i = 1 \end{cases}.$$

Czwarty składnik funkcji kosztu jest sumą kosztów przejazdów między danym miastem, a miastami odwiedzonymi w poprzednim oraz następnym etapie, suma jest obliczana dla wszystkich miast. Dla prawidłowego domknięcia trasy wskaźniki $i+1$ oraz $i-1$ we wzorze 16 są obliczane modulo N, ponieważ komiwojażer powinien wrócić do miasta, z którego wyruszył.

Funkcja kosztu, która będzie poddawana minimalizacji przez sieć będzie wyrażona więc następującym wzorem

$$\begin{aligned} E &= E_1 + E_2 + E_3 + E_4 = \\ &= 1/2\, N(A+B) - 1/2 \sum_{x=1}^{N}\sum_{i=1}^{N}\sum_{x'=1}^{N}\sum_{i'=1}^{N} [-A\delta_{xx'} - B\delta_{ii'} + C\delta_{xx'}\delta_{ii'} - \\ &\quad -Dd_{xx'}(\delta_{i',i+1} + \delta_{i',i-1})]\, V_{xi}V_{x'i'} - \sum_{x=1}^{N}\sum_{i=1}^{N}(A+B-C/2)\, V_{xi} \end{aligned} \tag{17}$$

Przyrównując wzór 17 do wzoru 3 można otrzymać wzory na wartości wag połączeń między neuronami oraz wartości zewnętrznych sygnałów wejściowych neuronów

$$\omega_{xi,x'i'} = -A\delta_{xx'} - B\delta_{ii'} + C\delta_{xx'}\delta_{ii'} - Dd_{xx'}(\delta_{i',i+1} + \delta_{i',i-1}) \tag{18}$$

$$I_{xi} = A + B - C/2 \tag{19}$$

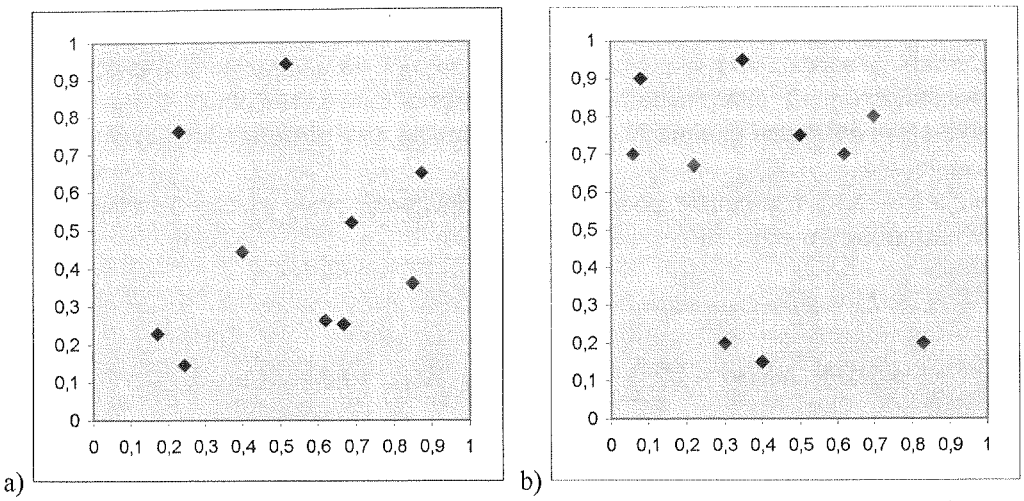
gdzie: $\omega_{xi,x'i'}$ — waga połączenia między neuronami o współrzędnych (x, i) oraz (x', i'), I_{xi} — zewnętrzny sygnał wejściowy neuronu (x, i).

Sieć neuronowa z wagami i zewnętrznymi sygnałami wejściowymi określonymi według wzorów 18 oraz 19 będzie minimalizować funkcję kosztu, czyli całkowity koszt trasy komiwojażera, z uwzględnieniem ograniczeń.

4.3. WYNIKI SYMULACJI

Celem przeprowadzonych symulacji było znalezienie optymalnej trasy komiwojażera dla sześciu przykładów. Pierwsze dwa przykłady są zbiorem 10 miast. Przykład 1 jest standardowym zbiorem porównawczym metod wykorzystujących sieci neuronowe. Pierwotnie został zastosowany w pracy [6]. Przykład 2 pochodzi z pracy [8]. Rysunek 4 przedstawia położenie 10 miast leżących w kwadracie jednostkowym dla przykładów 1 oraz 2.

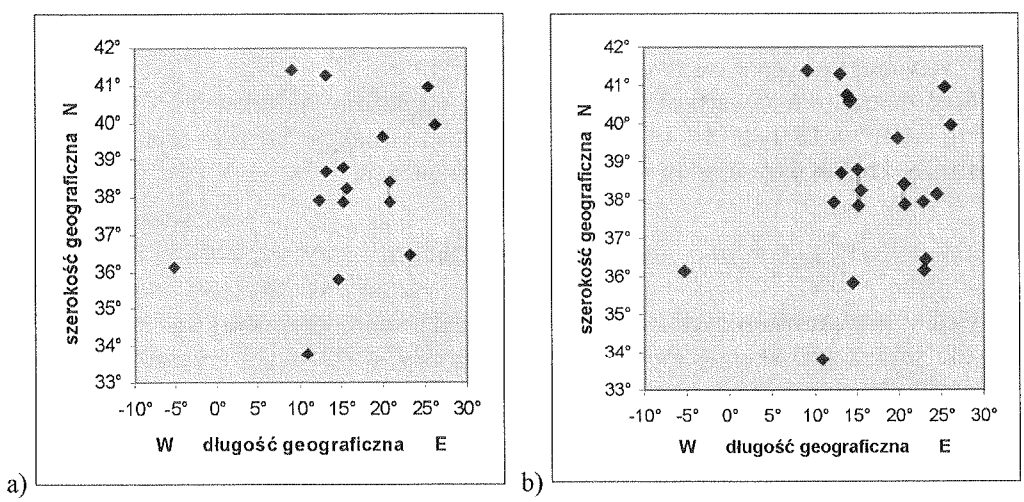
Przykłady 3, 4 oraz 6 zaczerpnięto z biblioteki problemu komiwojażera [28]. W bibliotece tej znajdują się również optymalne trasy dla tych przykładów razem z kosztem rozwiązań. Przykłady 3 oraz 4 (ulysses16.tsp oraz ulysses22.tsp) opisują podróż Odyseusza, która składa się odpowiednio z 16 oraz 22 etapów. Poszczególne etapy podróży wyznaczają współrzędne geograficzne punktów na Ziemi, ujemne wartości długości geograficznej odpowiadają długości geograficznej W. Dla obydwu przykładów podczas



Rys. 4. Położenie miast w kwadracie jednostkowym: a) dla przykładu 1, b) dla przykładu 2

Fig. 4. The position of the cities in the unit square: a) for 1st example, b) for 2nd example

symulacji zastosowano takie samo skalowanie jak w pracy [23]. W wyniku skalowania optymalna trasa dla przykładu 3 wynosi 2,36316 oraz 2,41279 dla przykładu 4. Rysunek 5 przedstawia współrzędne geograficzne punktów dla przykładów 3 oraz 4.



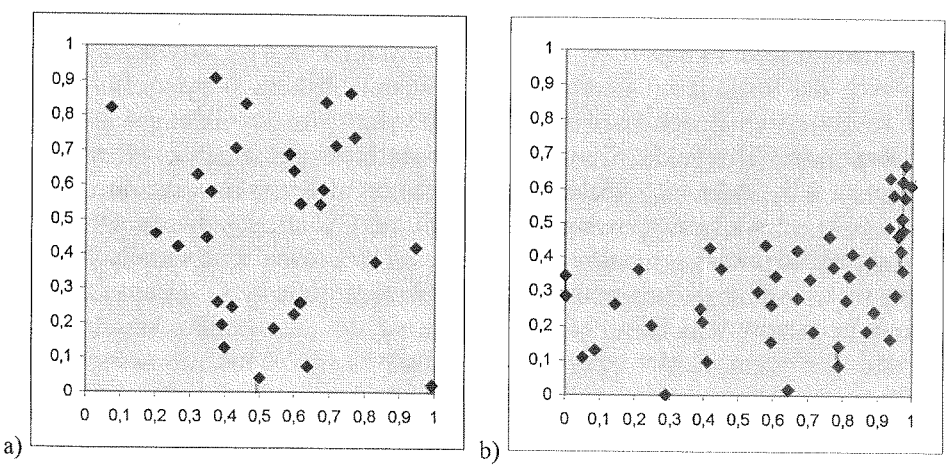
Rys. 5. Współrzędne geograficzne punktów na Ziemi: a) dla przykładu 3, b) dla przykładu 4

Fig. 5. The coordinates of the points on the earth: a) for 3rd example, b) for 4th example

Przykład 5 jest zbiorem 30 miast, pierwotnie został zastosowany w pracy [6]. Przykład 6 (att48.tsp) tworzą stolice 48 stanów w USA. W przykładzie 6 zastosowano takie same

skalowanie jak w pracy [23] tak, aby koszt optymalnego rozwiązania dla tego przykładu wynosił 4,32452. Rysunek 6 przedstawia położenie miast w kwadracie jednostkowym dla przykładów 5 oraz 6.

Każdy z przykładów został rozwiązany z użyciem klasycznej sieci Hopfielda oraz sieci z modyfikacją sygnałów wejściowych, a uzyskane wyniki zostały ze sobą porównane. Obydwa rodzaje sieci zostały zaimplementowane programowo. Symulacje przeprowadzono dla sieci o stałej czasowej $\tau = 1s$, z zastosowaniem metody Eulera według wzoru 11 dla klasycznej sieci oraz według wzoru 12 dla sieci z modyfikacją sygnałów wejściowych. Krok czasowy w metodzie Eulera wynosił $\Delta t = 0,01\tau$, natomiast funkcja aktywacji neuronów określona była wzorem 2, dla $\alpha = 10$. Liczba iteracji *nnar* w metodzie z modyfikacją sygnałów wejściowych była losowana z przedziału $2000 \div 5000$. Dla innych kroków czasowych Δt należy odpowiednio zmienić przedział, z którego losowana jest wartość *nnar*. Dla obydwu sieci zastosowano te same wartości stałych we wzorach na wartości wag połączeń oraz zewnętrznych sygnałów wejściowych neuronów, aby było możliwe porównanie wyników. Wartości stałych dobrano tak, aby prawdopodobieństwo uzyskania poprawnego rozwiązania było bliskie 100%. W programie symulującym obydwa rodzaje sieci zastosowano synchroniczną aktualizację, tzn. w danej chwili czasowej równocześnie obliczane są sygnały wejściowe oraz wyjściowe wszystkich neuronów. Dla obydwu metod zastosowano takie same kryterium zakończenia danej próby. Otrzymane rozwiązanie jest uznawane za stabilny stan sieci, jeżeli dla wszystkich neuronów bezwzględna wartość zmiany sygnału wyjściowego neuronu w danej iteracji jest mniejsza od 10^{-6}. Zastosowanie mniejszej wartości niż 10^{-6} nie ma wpływu na otrzymywane rozwiązania, natomiast powoduje wzrost liczby iteracji niezbędnych do zakończenia próby [24].



Rys. 6. Położenie miast w kwadracie jednostkowym: a) dla przykładu 5, b) dla przykładu 6

Fig. 6. The position of the cities in the unit square: a) for 5th example, b) for 6th example

Tabela 1 przedstawia wyniki otrzymane dla przykładów 1-6. Przeprowadzono 100 prób dla przykładów 1-4, 50 prób dla przykładu 5 oraz 25 prób dla przykładu 6. W tabeli 1 zebrane są następujące wyniki: koszt optymalnej trasy, średni koszt otrzymanych rozwiązań, stosunek średniego kosztu do kosztu optymalnego rozwiązania, koszt najlepszego rozwiązania otrzymanego daną metodą oraz prawdopodobieństwo otrzymania prawidłowego rozwiązania w jednej próbie. Dla obydwu metod zbadano również dwa sposoby inicjalizacji neuronów różnymi, losowymi wartościami początkowymi sygnałów wejściowych z przedziału

$$\text{a)} \qquad -0,3 \le U_{xi} \le 0,3 \qquad \text{dla} \qquad x,i = 1,\ldots,N \tag{20}$$

$$\text{b)} \qquad -\frac{1}{10\alpha} \le U_{xi} \le \frac{1}{10\alpha} \qquad \text{dla} \qquad x,i = 1,\ldots,N \tag{21}$$

W pierwszym sposobie wartości początkowe sygnałów wejściowych są losowane z równomiernym rozkładem prawdopodobieństwa z przedziału $-0,3$ do $0,3$ — ponieważ dla funkcji aktywacji określonej wzorem 2 i $\alpha = 10$, wartości sygnałów wyjściowych dla $-0,3$ oraz $0,3$ są bliskie -1 oraz 1. Drugi sposób polega na inicjalizacji neuronów wartościami sygnałów wejściowych bliskimi 0, ponieważ wartości sygnałów wyjściowych będą wówczas bliskie 0,5 i żaden neuron nie będzie uprzywilejowany. Taki sposób inicjalizacji uważany jest za najbardziej korzystny [8, 23]. Losowe zaburzenie z przedziału $\pm 1/10\alpha(\pm 0,01)$ przyjęto podobnie jak w pracach [6, 8, 24].

Analiza wyników zawartych w tabeli 1 pozwala wyciągnąć wniosek, że wyniki otrzymane metodą z modyfikacją sygnałów wejściowych są lepsze od wyników osiągniętych klasyczną metodą. Wyniki otrzymane obydwoma metodami dla przykładów 1 oraz 2 są bardzo zbliżone, a koszt rozwiązań jest równy lub bliski kosztowi optymalnego rozwiązania. Począwszy od przykładu 3 widoczna jest przewaga metody z modyfikacją sygnałów wejściowych, przewaga ta wzrasta wraz ze wzrostem wymiaru problemu. Dla przykładu 6 średni koszt rozwiązań otrzymanych metodą z modyfikacją sygnałów wejściowych dla inicjalizacji neuronów wartościami bliskimi 0 był o 19,64% większy od kosztu optymalnego rozwiązania, co jest lepszym wynikiem w porównaniu do klasycznej sieci Hopfielda. Średni koszt rozwiązań osiągniętych z wykorzystaniem klasycznej sieci był o 27,45% większy od kosztu optymalnego rozwiązania. Na uwagę zasługuje również większe prawdopodobieństwo otrzymania poprawnego rozwiązania dla metody z modyfikacją sygnałów wejściowych. Przeprowadzone symulacje potwierdzają również, że inicjalizacja neuronów sygnałami wejściowymi bliskimi 0 (sygnały wyjściowe neuronów w pobliżu 0,5) przyczynia się do wzrostu skuteczności obydwu metod pod względem jakości otrzymywanych rozwiązań. Wyniki uzyskane metodą z modyfikacją sygnałów wejściowych dla obydwu sposobów inicjalizacji neuronów są jednak lepsze od wyników otrzymanych klasyczną metodą, niezależnie od sposobu inicjalizacji neuronów. Dla przykładów 3-6 obydwoma metodami nie udało się osiągnąć optymalnego rozwiązania, jedynie dla przykładu 5 metodą z modyfikacją sygnałów wejściowych otrzymano rozwiązanie o koszcie o 0,03% większym od kosztu optymalnego rozwiązania. Koszt najlepszego rozwiązania uzyskanego klasyczną siecią dla

przykładów 3-6 był od 0,76% do 12,43% większy od kosztu optymalnego rozwiązania, dla inicjalizacji neuronów sygnałami wejściowymi bliskimi 0. Dla tego samego sposobu inicjalizacji metoda z modyfikacją sygnałów wejściowych pozwoliła dla przykładów 3-6 na osiągnięcie najlepszego rozwiązania o koszcie od 0,03% do 9,65% większym od kosztu optymalnego rozwiązania, co jest również lepszym wynikiem w porównaniu do klasycznej sieci. W trakcie symulacji sprawdzono również jakie wyniki można otrzymać metodą z modyfikacją sygnałów wejściowych dla większych wartości *nnar* i inicjalizacji neuronów wartościami bliskimi 0. W tabeli 2 zamieszczone są wyniki otrzymane dla przykładów 3-6 dla *nnar* losowanego z przedziału $4000 \div 8000$ oraz $8000 \div 12000$.

Analizując wyniki zamieszczone w tabeli 2 można zauważyć, że średni koszt rozwiązań jest mniejszy dla większych wartości *nnar*. Dla przykładu 5 średni koszt otrzymanych rozwiązań był większy od kosztu optymalnego rozwiązania jedynie o 5,03% dla *nnar* z przedziału $8000 \div 12000$, w porównaniu do 8,26% dla *nnar* z przedziału $2000 \div 5000$. Dla przykładu 6 i *nnar* z przedziału $8000 \div 12000$ średni koszt otrzymanych rozwiązań był o 12,59% większy od kosztu optymalnego rozwiązania. Dla tego samego przykładu i *nnar* z przedziału $2000 \div 5000$ średni koszt rozwiązań był o 19,64% większy od kosztu optymalnego rozwiązania. Koszt najlepszego rozwiązania dla przykładu 6 był większy od kosztu optymalnego rozwiązania o 5,29% dla *nnar* z przedziału $8000 \div 12000$, w porównaniu do 9,65% dla *nnar* z przedziału $2000 \div 5000$. Średni koszt rozwiązań dla przykładów 3 oraz 4 jest również mniejszy dla większych wartości *nnar*. Koszt najlepszych rozwiązań dla tych przykładów i *nnar* z przedziału $4000 \div 8000$ oraz $8000 \div 12000$ jest jednak większy niż dla *nnar* z przedziału $2000 \div 5000$. Dla większych wartości *nnar* koszt otrzymanych rozwiązań dla przykładów 3 oraz 4 charakteryzuje się mniejszym odchyleniem od wartości średniej.

Wyniki otrzymane podczas symulacji dla przykładów 1-6 są lepsze od wyników przedstawionych w innych pracach [6, 8, 22-24]:

a) przykład 1

Sieć neuronowa z postacią funkcji energii zaproponowana po raz pierwszy w pracy [6] umożliwiła otrzymanie prawidłowego rozwiązania dla przykładu 1 z prawdopodobieństwem równym 15% [22], 3,75% [24], 36% [8]. Sieć znalazła optymalne rozwiązanie dla tego przykładu dla 0,5% prób [24], nie znalazła tego rozwiązania [22]. Średni koszt rozwiązań wynosił 3,35, a optymalne rozwiązanie zostało otrzymane [8]. Dla postaci energii określonej wzorem 17 średni koszt rozwiązań wynosił ok. 3,10, a prawdopodobieństwo otrzymania optymalnego rozwiązania ok. 7% [23]. W przypadku zastosowania sieci z mnożnikami Lagrange'a wszystkie rozwiązania były optymalnymi rozwiązaniami o koszcie równym 2,691 [23]. Średni koszt rozwiązań dla modelu stochastycznego z białym szumem wynosił 2,85, a prawdopodobieństwo otrzymania poprawnego rozwiązania 100% [8]. Model stochastyczny z pulsującym szumem umożliwił uzyskanie prawidłowych rozwiązań dla 100% prób, o średnim koszcie 2,95 [8]. Modele stochastyczne z białym szumem oraz pulsującym szumem umożliwiły otrzy-

manie optymalnego rozwiązania. W przypadku zastosowania Maszyny Gaussowskiej ponad 70% prób zakończyło się uzyskaniem optymalnego rozwiązania, średni koszt rozwiązań wynosił ok. 2,70, a ponad 80% rozwiązań było poprawnych [8]. Chaotyczna sieć neuronowa w realizacji asynchronicznej umożliwiła osiągnięcie optymalnego rozwiązania dla 98,9%-100% prób, jedynie 0%-0,5% prób zakończyło się otrzymaniem niepoprawnego rozwiązania [8].

b) przykład 2

Klasyczna sieć neuronowa typu Hopfielda z postacią funkcji energii zaproponowaną w pracy [6] umożliwiła otrzymanie dla przykładu 2 średniego kosztu rozwiązań równego 3,24, a 45% prób zakończyło się otrzymaniem poprawnego rozwiązania, wśród których było optymalne rozwiązanie [8]. Średni koszt dla modelu stochastycznego z białym szumem wynosił 2,89, a prawdopodobieństwo otrzymania poprawnego rozwiązania 100% [8]. Model stochastyczny z pulsującym szumem umożliwił uzyskanie prawidłowych rozwiązań dla 100% prób, o średnim koszcie 2,96 [8]. Modele stochastyczne z białym szumem oraz pulsującym szumem umożliwiły osiągnięcie optymalnego rozwiązania. W przypadku zastosowania chaotycznej sieci neuronowej w realizacji synchronicznej 0,3% prób zakończyło się otrzymaniem optymalnego rozwiązania, dla 96,7% prób otrzymano drugie w kolejności rozwiązanie, dla 2% prób niepoprawne rozwiązanie, a średni koszt rozwiązań stanowił 100,3% kosztu optymalnego rozwiązania [8].

c) przykład 3, 4 oraz 6

Klasyczna sieć Hopfielda z postacią energii określoną wzorem 17 umożliwiła w przypadku przykładów 3, 4 oraz 6 osiągnięcie średniego kosztu rozwiązań równego odpowiednio ok. 3,00, 3,30 oraz 7,30 [23]. Rezultaty otrzymane z użyciem sieci z mnożnikami Lagrange'a były lepsze, średni koszt rozwiązań wynosił odpowiednio ok. 2,65, 2,65 oraz 6,00 [23]. Średni koszt rozwiązań dla przykładów 3, 4 oraz 6 z wykorzystaniem sieci z mnożnikami Lagrange'a był odpowiednio ok. 10%, 10% oraz 40% większy od kosztu optymalnego rozwiązania, a 100% prób zakończyło się otrzymaniem poprawnego rozwiązania [8]. Dla przykładów 3, 4 oraz 6 żadną z metod nie udało się otrzymać optymalnego rozwiązania [8, 23]. Koszt najlepszych rozwiązań otrzymanych siecią z mnożnikami Lagrange'a wynosił odpowiednio ok. 2,45, 2,55 oraz 5,25 [23].

d) przykład 5

Sieć Hopfielda zaproponowana w pracy [6] umożliwiła otrzymanie najlepszego rozwiązania o koszcie o 19,01% większym od kosztu optymalnego rozwiązania. Koszt najlepszego rozwiązania uzyskanego zmodyfikowaną siecią Hopfielda był o 9,72% większy od kosztu optymalnego rozwiązania, a jedynie 2,2% prób zakończyło się otrzymaniem poprawnego rozwiązania [24]. Optymalne rozwiązanie udało się otrzymać metodą zaproponowaną w pracy [16].

Wyniki otrzymane w niniejszej pracy zmodyfikowaną siecią Hopfielda są dla wszystkich przykładów lepsze od przedstawionych wyników znanych z literatury. Również

wyniki otrzymane podczas symulacji klasyczną siecią Hopfielda są lepsze od wyników uzyskanych tą siecią w innych pracach.

Tabela 1

Wyniki otrzymane dla poszczególnych przykładów

Achieved results for the individual examples

Numer przykładu/ wartości stałych	Wynik	Klasyczna metoda		Metoda z modyfikacją sygnałów wejściowych	
		inicjalizacja $-0,3 \div 0,3$	inicjalizacja $-0,01 \div 0,01$	inicjalizacja $-0,3 \div 0,3$	inicjalizacja $-0,01 \div 0,01$
przykład 1	koszt optymalnej trasy	2,6907	2,6907	2,6907	2,6907
$A = 5$	średni koszt	2,7899	2,6915	2,6907	2,6907
$B = 5$	śr. koszt / koszt opt. rozw.	1,0369	1,0003	1	1
$C = 0,5$	koszt najlepszego rozw.	2,6907	2,6907	2,6907	2,6907
$D = 2$	prawd. prawidł. rozw.	97%	100%	100%	100%
przykład 2	koszt optymalnej trasy	2,7818	2,7818	2,7818	2,7818
$A = 5$	średni koszt	2,8093	2,7818	2,7818	2,7818
$B = 5$	śr. koszt / koszt opt. rozw.	1,0099	1	1	1
$C = 0,5$	koszt najlepszego rozw.	2,7818	2,7818	2,7818	2,7818
$D = 2,5$	prawd. prawidł. rozw.	99%	100%	100%	100%
przykład 3	koszt optymalnej trasy	2,3632	2,3632	2,3632	2,3632
$A = 5$	średni koszt	2,9469	2,5108	2,4939	2,4996
$B = 5$	śr. koszt / koszt opt. rozw.	1,247	1,0625	1,0553	1,0577
$C = 0,5$	koszt najlepszego rozw.	2,4921	2,3811	2,3951	2,3942
$D = 0,9$	prawd. prawidł. rozw.	85%	99%	100%	100%
przykład 4	koszt optymalnej trasy	2,4128	2,4128	2,4128	2,4128
$A = 5$	średni koszt	3,2328	2,6498	2,6212	2,5859
$B = 5$	śr. koszt / koszt opt. rozw.	1,3399	1,0982	1,0864	1,0717
$C = 0,5$	koszt najlepszego rozw.	2,7179	2,4522	2,4649	2,4522
$D = 1,1$	prawd. prawidł. rozw.	72%	80%	96%	94%
przykład 5	koszt optymalnej trasy	4,2774	4,2774	4,2774	4,2774
$A = 5$	średni koszt	5,8368	4,9449	4,7139	4,6307
$B = 5$	śr. koszt / koszt opt. rozw.	1,3646	1,1561	1,1020	1,0826
$C = 0,5$	koszt najlepszego rozw.	5,0423	4,3781	4,2875	4,2786
$D = 1,6$	prawd. prawidł. rozw.	82%	86%	92%	94%
przykład 6	koszt optymalnej trasy	4,3245	4,3245	4,3245	4,3245
$A = 5$	średni koszt	6,8483	5,5116	5,2935	5,1737
$B = 5$	śr. koszt / koszt opt. rozw.	1,5836	1,2745	1,2241	1,1964
$C = 0,5$	koszt najlepszego rozw.	5,5322	4,862	4,6706	4,742
$D = 1,4$	prawd. prawidł. rozw.	100%	92%	92%	96%

Tabela 2

Wyniki otrzymane dla przykładów 3-6 metodą z modyfikacją sygnałów wejściowych dla *nnar* z przedziału 4000 ÷ 8000 oraz 8000 ÷ 12000

Achieved results by using the method with input signals modifying for 3-6 examples and *nnar* in the interval 4000 ÷ 8000 and 8000 ÷ 12000

Numer przykładu / wartości stałych	Wynik	Metoda z modyfikacją sygnałów wejściowych *nnar* 4000 ÷ 8000	Metoda z modyfikacją sygnałów wejściowych *nnar* 8000 ÷ 12000
		inicjalizacja $-0,01 \div 0,01$	inicjalizacja $-0,01 \div 0,01$
przykład 3	koszt optymalnej trasy	2,3632	2,3632
$A = 5$	średni koszt	2,4806	2,4472
$B = 5$	śr. koszt / koszt opt. rozw.	1,0497	1,0355
$C = 0,5$	koszt najlepszego rozw.	2,4463	2,4303
$D = 0,9$	prawd. prawidł. rozw.	100%	100%
przykład 4	koszt optymalnej trasy	2,4128	2,4128
A=5	średni koszt	2,5862	2,5796
B=5	śr. koszt / koszt opt. rozw.	1,0719	1,0691
C=0,5	koszt najlepszego rozw.	2,4898	2,5473
D=1,1	prawd. prawidł. rozw.	99%	100%
przykład 5	koszt optymalnej trasy	4,2774	4,2774
$A = 5$	średni koszt	4,5716	4,4925
$B = 5$	śr. koszt / koszt opt. rozw.	1,0688	1,0503
$C = 0,5$	koszt najlepszego rozw.	4,2875	4,2786
$D = 1,55$	prawd. prawidł. rozw.	100%	86%
przykład 6	koszt optymalnej trasy	4,3245	4,3245
$A = 5$	średni koszt	5,0438	4,8688
$B = 5$	śr. koszt / koszt opt. rozw.	1,1663	1,1259
$C = 0,5$	koszt najlepszego rozw.	4,5965	4,5534
$D = 1$	prawd. prawidł. rozw.	100%	92%

5. PODSUMOWANIE

Przeprowadzone symulacje wskazują na możliwość poprawy jakości rozwiązań osiąganych przez sieć Hopfielda w problemach optymalizacyjnych w wyniku modyfikacji wartości sygnałów wejściowych lub wag neuronów podczas pracy sieci. Najkorzystniejszym sposobem inicjalizacji sieci jest inicjalizacja neuronów wartościami

sygnałów wejściowych bliskimi 0. Średni koszt otrzymanych tą drogą rozwiązań dla problemu komiwojażera był dla wszystkich badanych przykładów mniejszy od średniego kosztu rozwiązań otrzymanych klasyczną siecią Hopfielda oraz innymi znanymi metodami. Dla problemów o wymiarze równym 10 zmodyfikowana sieć Hopfielda dla 100% prób generowała optymalne rozwiązanie. Rozwiązania otrzymane z użyciem tej sieci dla większych problemów były lepsze od rozwiązań uzyskanych klasyczną siecią oraz innymi metodami. Ponadto wraz ze wzrostem wartości średniej przedziału, z którego losowana była liczba iteracji *nnar*, następowało zmniejszenie średniego kosztu otrzymywanych rozwiązań. Lepsze rezultaty uzyskane w wyniku zastosowania metody z modyfikacją sygnałów wejściowych należy tłumaczyć tym, że podczas iteracji *nnar*, w wyniku malejącego tłumienia sygnałów wejściowych neuronów wzrasta energia sieci, określona wzorem 3. Losowe wartości liczby iteracji *nnar* określają również szybkość wzrostu energii sieci podczas tych iteracji. Pozwala to sieci na wyjście z minimów lokalnych funkcji energii dla pewnych wartości liczby *nnar*. Osiągnięcie minimum lokalnego przez klasyczną sieć Hopfielda uniemożliwia opuszczenie tego minimum. Dlatego rozwiązanie nie jest wtedy optymalne w sensie globalnym.
W symulacjach zastosowano modyfikację sygnałów wejściowych, ponieważ metoda ta jest bardziej wydajna w realizacji programowej i wymaga mniejszych nakładów w realizacji sprzętowej sieci.

6. BIBLIOGRAFIA

1. A. Kos: *Modelowanie hybrydowych układów mocy i optymalizacja ich konstrukcji ze względu na rozkład temperatury.* Kraków, Wydawnictwa AGH, 1994
2. P. Bratek, A. Kos: *Complex optimisation of topology of VLSI circuits with self-organising neural nets.* Proc. of the Mixed Design of Integrated Circuits and Systems, Łódź (Poland), June 1996, pp. 78-83
3. H. Mączka, P. Dziurdzia, A. Kos: *Neural algorithm for minimisation of total length of connections in VLSI circuits.* Proc. of the XIXth National Conference on Circuit Theory and Electronic Networks, Kraków-Krynica (Poland), October 1996, pp. II/319-324
4. P. Bratek, A. Kos: *Self-organising neural computations for optimisation of IC topography in thermal aspect.* Proc. of the XIXth National Conference on Circuit Theory and Electronic Networks, Kraków-Krynica (Poland), October 1996, pp. II/627-632
5. M. Gajęcki, A. Kos: *A problem of optimization of topography of VLSI circuit.* Proc. of the XIXth National Conference on Circuit Theory and Electronic Networks, Kraków-Krynica (Poland), October 1996, pp. II/307-312
6. J. J. Hopfield, D. W. Tank: *"Neural" computation of decisions in optimization problems.* Biological Cybernetics 1985, vol. 52, pp. 141-152
7. R. Tadeusiewicz: *Sieci neuronowe.* Warszawa, Akademicka Oficyna Wydawnicza, 1993
8. J. Mańdziuk: *Sieci neuronowe typu Hopfielda. Teoria i przykłady zastosowań.* Warszawa, Akademicka Oficyna Wydawnicza EXIT, 2000
9. J. Hertz, A. Krogh, R. G. Palmer: *Wstęp do teorii obliczeń neuronowych.* Warszawa, WNT, 1995
10. J. A. Freeman, D. M. Skapura: *Neural networks: algorithms, applications, and programming techniques.* Addison-Wesley Publishing Company, 1991
11. T. Khanna: *Foundations of neural networks.* Addison-Wesley Publishing Company, 1990

12. J. Żurada, M. Barski, W. Jędruch: *Sztuczne sieci neuronowe.* Warszawa, Wydawnictwo Naukowe PWN, 1996
13. G. A. Tagliarini, J. F. Christ, E. W. Page: *Optimization using neural networks.* IEEE Transactions on Computers 1991, vol. 40, no. 12, pp. 1347-1358
14. H. Kwaśnicka: *Efficiency of selected meta-heuristics applied to the TSP problem: a simulation study.* TASK Quarterly 2003, vol. 7, no. 1, pp. 73-91
15. A. Hemani, A. Postula: *Cell placement by self-organisation.* Neural Networks 1990, vol. 3, pp. 377-383
16. R. Durbin, D. Willshaw: *An analogue approach to the travelling salesman problem using an elastic net method.* Nature 1987, vol. 326, pp. 689-691
17. I. Tokuda, T. Nagashima, K. Aihara: *Global bifurcation structure of chaotic neural networks and its application to traveling salesman problems.* Neural Networks 1997, vol. 10, no. 9, pp. 1673-1690
18. S. Kirkpatrick, C. D. Gelatt, Jr., M. P. Vecchi: *Optimization by simulated annealing.* Science 1983, vol. 220, no. 4598, pp. 671-680
19. C. I. Park: *Predicting the annealing range by computing critical temperature in mean field annealing for the traveling salesman problem.* Artificial Neural Networks, Proceedings of the 1991 International Conference on Artificial Neural Networks, Amsterdam, North-Holland, 1991, pp. 1019-1024
20. D. E. Van den Bout, T. K. Miller III: *Improving the performance of the Hopfield-Tank neural network through normalization and annealing.* Biological Cybernetics 1989, vol. 62, pp. 129-139
21. X. Xu, W. T. Tsai: *Effective algorithms for the traveling salesman problem.* Neural Networks 1991, vol. 4, pp. 193-205
22. G. V. Wilson, G. S. Pawley: *On the stability of travelling salesman problem algorithm of Hopfield and Tank.* Biological Cybernetics 1988, vol. 58, pp. 63-70
23. S. Z. Li: *Improving convergence and solution quality of Hopfield-type neural networks with augmented Lagrange multipliers.* IEEE Transactions on Neural Networks 1996, vol. 7, no. 6, pp. 1507-1516
24. B. Kamgar-Parsi, B. Kamgar-Parsi: *On problem solving with Hopfield neural networks.* Biological Cybernetics 1990, vol. 62, no. 5, pp. 415-423
25. A. R. Bizzarri: *Convergence properties of a modified Hopfield-Tank model.* Biological Cybernetics 1991, vol. 64, no. 4, pp. 293-300
26. D. Rutkowska, M. Piliński, L. Rutkowski: *Sieci neuronowe, algorytmy genetyczne i systemy rozmyte.* Warszawa, Wydawnictwo Naukowe PWN, 1997
27. A. Kos, Z. Nagórny: *Minimalizacja długości połączeń w układach elektronicznych z wykorzystaniem sieci Hopfielda.* Kwartalnik Elektroniki i Telekomunikacji, praca przyjęta do druku 2005, z. 1
28. G. Reinelt: *Traveling salesman problem library.*
Internet: http://www.iwr.uni-heidelberg.de/groups/comopt/software/TSPLIB95/tsp/

Z. NAGÓRNY, A. KOS

OPTIMIZATION BY USING A MODIFIED HOPFIELD NETWORK

Summary

The paper presents a novel method for solving optimization problems by using a modified Hopfield network. In a conventional Hopfield network weight values of the net are calculated before simulation and are constant. Our method is a novel method, because the values of input signals or weights are

modified during simulation. The method makes use of the Hopfield net with continuous function of neurons according to Eq. (2). The model Hopfield net in electronic components is shown in Figure 1. An energy function of the neural net is described by Eq. (3). Comparing Eq. (3) and Eq. (4), which is a general form of an optimization problem cost function, we get weight and external input signal values. The Hopfield net is implemented in software in this work. A simulating program makes use of Eq. (11) to calculate the input of each neuron in the net. During the simulation the input signals are modified in accordance with Eq. (12). The duration of input signals modifying is defined by a random value *nnar*. Finally, a number of iterations to achieve a stable state of the net are done. A number of trials are performed for each optimization problem and the best results are chosen. Figure 2 shows the algorithm of the method. In this work, simulations were done for six examples of the travelling salesman problem. A cost function of the travelling salesman problem is described by Eq. (17). This function consists of four components: the total length of the salesman's tour, two terms, which ensure that the salesman's tour is valid and the term, which forces neurons to have the output signal equal zero or one. The method with input signal values modifying during simulation gives better results than the conventional one. Results are performed in Table 1 and 2. For ten-city problems the modified Hopfield net finds the optimal solution with 100% success rate. Some conclusions coming from using this method are presented.

Keywords: optimization, neural network, Hopfield net, energy function, travelling salesman problem

KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJI, 2005, 51, z. 2, ss. 277–290

Zastosowanie algorytmów ewolucyjnych w lingwistyce komputerowej celem oceny stopnia pokrewieństwa języków naturalnych

MIROSŁAW GAJER

Katedra Automatyki, Akademia Górniczo-Hutnicza
Al. Mickiewicza 30, 30-059 Kraków
e-mail: mgajer@ia.agh.edu.pl

Otrzymano 2003.11.25
Autoryzowano 2004.09.20

Artykuł został poświęcony zagadnieniom związanym z zastosowaniem algorytmów ewolucyjnych w zagadnieniach lingwistycznych. Celem opracowanego przez autora programu komputerowego bazującego na technice algorytmów ewolucyjnych jest dostarczenie liczbowej miary określającej stopień wzajemnego pokrewieństwa pomiędzy językami naturalnymi. Badanie genetycznego pokrewieństwa języków naturalnych i łączenie ich na tej podstawie w większe jednostki, jest jednym z centralnych zagadnień współczesnej lingwistyki. Pomimo wieloletnich badań wiele podstawowych kwestii dotyczących pochodzenia wybranych języków naturalnych nie zostało jeszcze ostatecznie wyjaśnionych. Lingwiści w toczonych pomiędzy sobą sporach odwołują się do argumentów różnorodnej natury, ale zwykle cała dyskusja jest prowadzona na poziomie charakterystycznym dla nauk humanistycznych, gdzie dojście do ostatecznej prawdy wydaje się być niezwykle trudne. Z drugiej jednak strony, spośród nauk humanistycznych, lingwistyka jest często uważana za naukę ścisłą. Konsekwencją tego jest coraz powszechniejsze zastosowanie komputerów w badaniach lingwistycznych, na przykład podczas prób rekonstrukcji wymarłych języków starożytnych, czy też podczas prób odczytania nieznanego systemu pisma. Autor artykułu kierując się oczywistymi analogiami, jakie zachodzą pomiędzy historycznym rozwojem języków naturalnych i powstaniem grup językowych, a ewolucją organizmów żywych i powstaniem gatunków, wykorzystał powstałą dzięki inspiracji procesami biologicznymi technikę algorytmów ewolucyjnych. Opracowany przez autora program komputerowy pozwala na uzyskanie precyzyjnej, wyrażonej w liczbach odpowiedzi na pytanie, jak blisko ze sobą mogą być spokrewnione dwa badane języki naturalne. Zdaniem autora opracowane przez niego metoda może okazać się bardzo przydatna dla lingwistów, ponieważ może dostarczyć w prowadzonych przez nich dyskusjach dodatkowych argumentów przemawiających za poparciem bądź odrzuceniem wysuwanej hipotezy dotyczącej pochodzenia badanego języka naturalnego.

Słowa kluczowe: lingwistyka komputerowa, algorytmy ewolucyjne, ewolucja języków naturalnych

1. WPROWADZENIE

Spośród nauk humanistycznych lingwistyka często bywa nazywana nauką ścisłą [1]. Nic zatem dziwnego, że wkrótce po pojawieniu się pierwszych komputerów cyfrowych, jako pierwsze z możliwych zastosowań komputerów zostały zaproponowane właśnie zastosowania ukierunkowane na analizę i przetwarzanie języka naturalnego [2]. Można wymienić w tym miejscu zastosowania związane z automatycznym wyszukiwaniem informacji w korpusie tekstów, automatycznym tworzeniem streszczeń tekstów, a także systemy przeznaczone do prowadzenia dialogu komputera z człowiekiem w języku naturalnym [3]. Odrębną gałęzią lingwistycznych zastosowań komputerów jest translacja automatyczna, czyli dziedzina zajmująca się opracowywaniem programów komputerowych zdolnych do dokonywania przekładów z jednego języka naturalnego na drugi (np. z polskiego na włoski) [4].

Z drugiej strony komputery okazały się również niezwykle przydatnym narzędziem dla lingwistów, którzy zaczęli je stosować podczas prac związanych z rekonstrukcją wymarłych języków oraz podczas prób odczytania nieznanych systemów pisma [5].

Z kolei pod koniec lat 60-tych badacze zainspirowani nowymi dokonaniami w dziedzinie biologii i nauk przyrodniczych, związanymi z odkryciem DNA i mechanizmu dziedziczenia informacji genetycznej organizmów żywych, podjęli udaną próbę przeniesienia mechanizmów ewolucji biologicznej w świat komputerów. W ten sposób narodziły się algorytmy ewolucyjne, które szybko okazały się być niezwykle uniwersalnym narzędziem optymalizacyjnym [6].

Obecnie zastosowania algorytmów ewolucyjnych są bardzo szerokie, począwszy od nauk technicznych, a na badaniach operacyjnych i naukach ekonomicznych kończąc. We wszystkich tych zastosowaniach algorytmy ewolucyjne wykorzystywane są do przeszukiwania przestrzeni możliwych rozwiązań danego problemu optymalizacyjnego celem znalezienia jak najlepszego rozwiązania [7]. W porównaniu z klasycznymi metodami optymalizacji opartymi głównie na badaniu pochodnych funkcji celu, algorytmy ewolucyjne wykazują znacznie większą odporność na występowaniu tzw. minimów lokalnych, w których przy użyciu metod klasycznych zwykle grzęźnie cały proces poszukiwania i w związku z tym znalezienie rozwiązania o odpowiednio dobrej jakości staje się niemożliwe.

Dzięki swojej prostocie i dużej uniwersalności, możliwych zastosowań algorytmów ewolucyjnych jest dosłownie bez liku. Niestety autorowi niniejszej publikacji jak dotychczas nie udało się natknąć w literaturze na zastosowania algorytmów ewolucyjnych związane z obszarem lingwistyki. Jest to tym bardziej dziwne, że zastosowania takie powinny wręcz wydawać się naturalnymi, bowiem historia rozwoju każdego języka naturalnego bardzo przypomina historię ewolucji gatunków organizmów żywych, ponieważ w obu przypadkach występują mechanizmy takie, jak dziedziczenie, zmienność

i selekcja naturalna. Podobnie jak organizmy żywe dziedziczą za pośrednictwem DNA cechy swoich przodków, tak również język naturalny przekazywany jest z pokolenia na pokolenie. Również podobnie, jak w przypadku organizmów żywych do ich materiału genetycznego zakodowanego w DNA wprowadzane są pewne drobne przypadkowe zmiany zwane mutacjami, tak samo każdy żywy język naturalny nieustannie się rozwija (pojawiają się nowe słowa, istniejące słowa otrzymują nowe znaczenia, zmienia się wymowa słów, ewoluują reguły gramatyczne, składnia zdania itp.). Wreszcie w obu rozważanych przypadkach wymienione losowe zmiany podlegają selekcji naturalnej, tzn. jeśli jakaś powstała w sposób przypadkowy nowa cecha organizmu żywego bądź języka okaże się być pod pewnym względem korzystna, wówczas istnieje duże prawdopodobieństwo jej utrwalenia się i przekazania kolejnym pokoleniom organizmów żywych bądź użytkownikom danego języka naturalnego.

2. PODSTAWOWE ZAGADNIENIA WSPÓŁCZESNEJ LINGWISTYKI

Jak wynika z przeprowadzonych uprzednio rozważań historia rozwoju języków naturalnych przypomina bardzo historię rozwoju gatunków organizmów żywych, a dzieje się tak dlatego, że w jednym i drugim przypadku procesem ewolucji rządzą bardzo podobne fundamentalne zasady oparte na zmienności i dziedziczeniu. Podobnie jak kiedyś z pierwotnej prakomórki wyewoluowały wszystkie znane obecnie formy organizmów żywych, podobnie istniejące współcześnie na Kuli Ziemskiej języki łączy się w większe jednostki zwane rodzinami językowymi, przy czym zakłada się, że wszystkie języki należące do danej rodziny językowej muszą pochodzić od wspólnego przodka — prajęzyka. Na przykład dla współczesnych języków romańskich wspólnym przodkiem była łacina. Językoznawcy podejmują również próby rekonstrukcji języka praindoeuropejskiego, którym posługiwano się około pięciu tysięcy lat temu, i który dał początek wszystkim używanym współcześnie językom indoeuropejskim.

Pomimo ponad stu lat badań językoznawstwo jest wciąż dyscypliną naukową, która może być uważana za stosunkowo młodą, i która wciąż skrywa przed badaczami wiele tajemnic. Fakt ten można w pełni uświadomić sobie, jeżeli weźmie się pod uwagę wielką różnorodność i ogrom materiału badawczego, który jest potencjalnie dostępny dla badaczy związanych zawodowo z lingwistyką [8]. Niestety na skutek postępującego wymierania licznych języków o niewielkiej liczbie użytkowników, w ostatnich czasach rozważany materiał badawczy zaczął się dość dramatycznie kurczyć, i wszystko wskazuje na to, iż jest to już proces nieodwracalny. Zatem lingwiści muszą się pospieszyć i zintensyfikować prowadzone przez siebie badania, bowiem w przeciwnym przypadku znaczna część ich materiału badawczego po prostu zniknie z powierzchni Ziemi [11].

Jak już wspomniano, jednym z zadań stawianych przed lingwistyką jest weryfikacja hipotez dotyczących pochodzenia języków od wspólnego przodka – prajęzyka. Kwestie pochodzenia języków mają znaczenie fundamentalne dla językoznawstwa, niestety na obecnym etapie badań wciąż w wielu punktach pozostaje znacznie więcej pytań niż odpowiedzi.

Na przykład jeszcze kilkanaście lat temu powszechnie uważano, że języki tybetańskie, chińskie, birmańskie i tajskie należą do jednej wielkiej chińsko-tybetańskiej rodziny językowej, czyli że wywodzą się wszystkie od wspólnego przodka. Na pokrewieństwo takie wskazywały przede wszystkim cechy typologiczne tych języków, a mianowicie to, iż wszystkie języki tej grupy są językami w znacznej mierze monosylabicznymi (wyrazy składają się z tylko jednej sylaby), tonalnymi (każdą sylabę można wypowiedzieć w jednym z kilku tonów, np.: wysokim, niskim, opadającym, wznoszącym, opadająco-wznoszącym itp., przy czym za każdym razem uzyskuje się zupełnie inne znaczenie wyrazu) oraz silnie analitycznymi tzn., że o funkcji pełnionej przez wyraz decyduje jego pozycja zajmowana w zdaniu (zmiana szyku zdania prowadzi do zmiany jego znaczenia lub też zdanie takie staje się dla odbiorcy zupełnie niezrozumiałe). Jednakże obecnie jedność chińsko-tybetańskiej rodziny językowej jest wśród badaczy zajmujących się tym problemem bardzo mocno kwestionowana. Wiele bowiem wskazuje na to, iż języki tybetańskie, chińskie, birmańskie i tajskie nie mają ze sobą nic wspólnego (w tym sensie, że nie wywodzą się od wspólnego praprzodka), a wymienione podobieństwa typologiczne powstały na skutek wzajemnych kontaktów użytkowników języków tybetańskich, birmańskich i tajskich z użytkownikami języków chińskich i po prostu pewne cechy języków chińskich przeniknęły do pozostałych języków czyniąc je wszystkie typologicznie podobnymi [9].

Z problemami podobnej natury borykają się zwolennicy bądź przeciwnicy hipotezy, w myśl której wszystkie języki należące do rodziny uralskiej (z językami ugrofińskimi takimi jak: węgierski, fiński, estoński, mordwiński, maryjski, udumurucki, komi i jeszcze z wieloma innymi) oraz do ałtajskiej rodziny językowej (z takimi językami jak: turecki, azerbejdżański, kazachski, turkmeński, uzbecki, kirgiski, ujgurski, tatarski, czuwaski, baszkirski, kałmucki, mongolski, madżurski i wieloma innymi) wywodzą się od wspólnego prajęzyka. Są również badacze, którzy idą jeszcze dalej i do tej wielkiej rodziny językowej dodają jeszcze języki koreański i japoński, które w myśl głoszonej przez nich hipotezy miałyby być najwcześniej oddzieloną grupą od rozważanej prarodziny językowej [1]. Jednakże nie wszyscy badacze są skłonni podzielać tego typu poglądy, co więcej istnieją wśród nich i tacy, którzy w ogóle kwestionują jedność ałtajskiej rodziny językowej [10]. Według tych uczonych rodzina ałtajska rozpada się na trzy mniejsze rodziny językowe: turecką, mongolską i mandżurską, które poza podobieństwami typologicznymi wykształtowanymi na skutek wielowiekowych wzajemnych kontaktów, nie mają ze sobą nic wspólnego, czyli nie pochodzą od jednego prajęzyka. Zatem sytuacja byłaby tutaj identyczna, jak w przypadku omówionej uprzednio chińsko-tybetańskiej rodziny językowej, gdzie liczne podobieństwa typologiczne wykształtowały się jedynie w skutek wzajemnych kontaktów użytkowników języków należących do różnych rodzin językowych.

Każda z wymienionych hipotez lingwistycznych posiada zarówno swoich gorących zwolenników, jak i zagorzałych przeciwników, którzy w toczonych przez siebie dyskusjach i zażartych sporach przytaczają różnorodne argumenty na poparcie głoszonych przez siebie poglądów. Niestety, jak dotychczas rozważane spory, dotyczące hipotez

lingwistycznych, prowadzone są na poziomie typowym dla nauk humanistycznych, gdzie prawda wydaje się być rzeczą subiektywną, uzależnioną całkowicie od odpowiednio sprytnego dobrania, podczas prowadzonej dyskusji, właściwych argumentów, charakteryzujących się dużą siłą przekonywania słuchaczy.

Pragnąc się przyczynić do zmiany istniejącego stanu rzeczy, autor podjął próbę opracowania ścisłej metody pozwalającej na uzyskanie liczbowej miary określającej stopień pokrewieństwa pomiędzy językami. Biorąc pod uwagę fakt istnienia dużych analogii pomiędzy ewolucją gatunków organizmów żywych a rozwojem języków naturalnych, autor zdecydował się na wybór algorytmów ewolucyjnych jako podstawowego narzędzia opracowywanej przez siebie metody.

3. IMPLEMENTACJA ALGORYTMÓW EWOLUCYJNYCH W LINGWISTYCE

Autor postanowił skoncentrować prowadzone przez siebie prace na analizie słownictwa należącego do różnych języków. Opracowana przez autora metoda pozwala na uzyskanie liczbowej miary, która mówi, jak dalece zbiór słownictwa jednego języka jest oddalony od zbioru analogicznego słownictwa drugiego języka. Oczywiście opracowana przez autora metoda nie ma mocy rozstrzygającej ewentualne spory lingwistów na temat pochodzenia języków, może jednakże dostarczyć dodatkowych (zdaniem autora silnych) argumentów przemawiających za poparciem bądź odrzuceniem rozważanej hipotezy.

Prowadzone prace zostały skoncentrowane na analizie fonetycznej badanego słownictwa. Postąpiono tak głównie z dwóch powodów. Po pierwsze, uwzględniono powszechnie uznawany w lingwistyce prymatu języka mówionego nad językiem pisanym, który jest jedynie formą wtórną, pewnym niedoskonałym odbiciem języka żywego, obecnego na co dzień w mowie posługującej się nim populacji [9]. Po drugie, wzięto pod uwagę fakt, polegający na tym, iż w wielu współczesnych językach pisownia wyrazów ma całkowicie historyczny charakter i w związku z tym ma z wymową tych wyrazów niewiele wspólnego. Spośród języków używanych w Europie wystarczy wymienić języki takie, jak chociażby angielski, francuski, duński czy irlandzki, gdzie w przypadku prawie każdego z wyrazów, opanowawszy jego wymowę należy jeszcze nauczyć się jego prawidłowego zapisu.

Zmiany fonetyczne zachodzące podczas ewolucji języków naturalnych polegają przede wszystkim na zjawiskach takich, jak utrata samogłosek (tworzenie tzw. form *alegro*), bądź wstawianie samogłosek w przypadku, gdy użytkownicy danego języka z pewnych bliżej nie określonych powodów przestają tolerować zbitki spółgłoskowe występujące w ich języku (proces ten nazywany jest tworzeniem tzw. form *lento*). Poza tym języki podczas swego historycznego rozwoju mogą również gubić spółgłoski (zwłaszcza na początku lub końcu wyrazów). Typowym zjawiskiem jest również występowanie tzw. przesunięć spółgłoskowych, podczas których np. spółgłoski dźwięczne zostają zastąpione ich bezdźwięcznymi odpowiednikami, oprócz tego głoski mogą ulegać nazalizacji bądź ją tracić itp. Ponadto trzeba pamiętać o tym, iż pozornie ta sama

głoska w każdym języku jest wymawiana nieco inaczej. Dobrym przykładem jest głoska „r", która wymawiana jest odmiennie w językach takich, jak: polski, niemiecki, niderlandzki, francuski, walijski, arabski.

Uwzględniając niemożność dokładnego oddania wymowy wyrazów za pomocą ich pisowni, autor przyjął maksymalnie uproszczony sposób zapisu wyrazów, który jednakże później okazał się całkowicie wystarczający do wyciągnięcia wniosków na temat pokrewieństwa języków. W zaproponowanym przez autora systemie do zapisu wymowy wyrazów zastosowano jedynie dwa znaki. Znak „c" oznaczający dowolną spółgłoskę (ang. consonat) oraz znak „v" oznaczający dowolną samogłoskę (ang. vowel).

Odrębnym, ale nie mniej ważnym zagadnieniem, jest wybór materiału porównawczego, w oparciu o który zostaną przeprowadzone eksperymenty i na podstawie ich wyników będą później wyciągane wnioski odnośnie pochodzenia języków oraz ich wzajemnego pokrewieństwa. Dobierając słownictwo jakiegoś języka, jako materiał badawczy trzeba być niezwykle ostrożnym. Na przykład, gdyby jako bazę porównawczą dla języka węgierskiego wybrać wyrazy takie, jak: „káposzta", „kolbásza", „drága", „tiszta", „szoba", „kulcs", „sonka", „uborka", „ebéd", „vácsora", „serda", „csütörtök" i „péntek", można byłoby wyciągnąć błędne wnioski prowadzące do konkluzji, iż węgierski jest genetycznie spokrewniony z językami słowiańskimi. Tak jednak nie jest, bowiem wszystkie wymienione wyrazy zostały przed wiekami jedynie zapożyczone od Słowian Panońskich przez madziarskich najeźdźców.

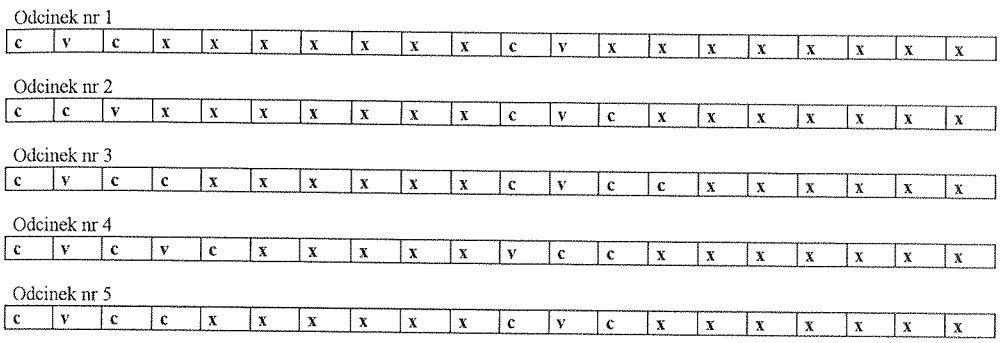
W przeprowadzonych za pomocą algorytmów ewolucyjnych badaniach, autor jako materiał porównawczy dla badanych języków wybrał liczebniki główne (od jeden do dziesięć). Autor kierował się przy tym przesłanką, w myśl której liczebniki główne należą do podstawowego zbioru słownictwa każdego języka i w związku z tym są dziedziczone bezpośrednio od prajęzyka, z którego dany język się wywodzi. Natomiast liczebniki prawie nigdy nie są zapożyczeniami z innych języków. Chociaż pamiętać trzeba o tym, że w językoznawstwie nie ma nigdy prawd absolutnych i zawsze od każdej reguły można gdzieś znaleźć jakieś wyjątki. Na przykład francuski liczebnik „huit" (osiem) pochodzi z języka Celtów. Podobnie w języku Swahili liczebniki „sita" (sześć) i „tisa" (dziewięć) wywodzą się z języka arabskiego, a kantoński liczebnik „go" (pięć) jest japońskim zapożyczeniem. Jednakże jak się w dalszej części wywodu okaże, wymienione fakty nie mają większego wpływu na uzyskane wyniki badań i wyciągnięte na ich podstawie wnioski.

Stosując uproszczoną formę zapisu fonetycznego dla liczebników języka angielskiego otrzymuje się następujące wyrazy, które następnie należy zakodować na materiale genetycznym chromosomu podlegającemu ewolucji:

1. one – cvc
2. two – cv
3. three – ccv
4. four – cvc
5. five – cvcc
6. six – cvcc

7. seven – cvcvc
8. eight – vcc
9. nine – cvcc
10. ten – cvc

Podczas badań przyjęto stałą długość chromosomu składającą się ze stu pozycji (łac. loci), przy czym każdej pozycji odpowiada dokładnie jeden gen, który może przybierać jedną z trzech możliwych wartości: „c", „v" lub „x". Ostatni z wymienionych symboli „x" oznacza brak jakiejkolwiek głoski na danej pozycji. Dla zapisu każdego z liczebników przeznaczono po dziesięć kolejnych pozycji. W przypadku, gdy dany wyraz zawierał mniejszą niż dziesięć liczbę głosek brakujące pozycje zostały wypełnione symbolami „x". W przypadku języka angielskiego uproszczony zapis fonetyczny liczebników, który został zamieszczony na chromosomie ma postać taką, jak na rys. 1. Ze względów technicznych następujące po sobie odcinki chromosomu zostały na rysunku zamieszczone jeden pod drugim.



Rys. 1. Postać chromosomu z zakodowaną wymową liczebników języka angielskiego

Fig. 1. The way of coding on a chromosome the pronunciation of English numerals

Zastosowana została tzw. dwuelementowa strategia ewolucyjna, w której występują jedynie dwa osobniki: osobnik rodzicielski oraz osobnik potomny. W przypadku dwuelementowej strategii ewolucyjnej jedyną stosowaną operacją genetyczną jest operacja mutacji. Osobnik potomny jest zawsze zmutowaną kopią osobnika rodzicielskiego. Zarówno dla osobnika rodzicielskiego, jak i potomnego obliczana jest wartość funkcji dopasowania. Jeżeli osobnik potomny posiada lepszą wartość funkcji dopasowania, wówczas osobnik rodzicielski zostaje usunięty z populacji, a osobnik potomny zajmuje jego miejsce, samemu stając się rodzicem dla kolejnego potomka będącego jego zmutowana kopią. W przypadku, gdy funkcja dopasowania potomka ma gorszą wartość niż funkcja dopasowania rodzica, potomek taki jest usuwany z populacji, a następnie poprzez zmutowanie materiału genetycznego rodzica tworzony jest kolejny potomek. Opisany proces powtarzany jest do skutku, tzn. do momentu powstania potomka o lepszej wartości funkcji dopasowania niż w przypadku osobnika rodzicielskiego. Wreszcie wygenerowanie osobnika spełniającego zadane kryterium poszukiwań kończy cały

proces, a osobnik taki wyprowadzany jest jako odnalezione rozwiązanie badanego zagadnienia.

W przypadku zastosowania dwuelementowej strategii ewolucyjnej do badania stopnia pokrewieństwa genetycznego pomiędzy językami, korpus słownictwa danego języka jest zawsze porównywany z pewnym wzorcem. Wzorcem tym jest zapisane za pomocą omówionej uprzednio metody słownictwo języka, co do którego istnieje podejrzenie, iż może mieć on powiązanie genetyczne z językiem, który badamy.

Zastosowany algorytm ewolucyjny działa w sposób następujący. Na początku tworzony jest osobnik potomny, który stanowi dokładną kopię osobnika rodzicielskiego. Następnie za pomocą generatora liczb losowych wybierany jest numer pozycji genu osobnika potomnego. Jak już wcześniej wspomniano, gen ten może mieć trzy różne wartości (tzw. allele): „c", „v" bądź „x". Następnie wartość tego genu porównywana jest z wartością genu wzorca. Jeżeli jest ona identyczna omówiony proces powtarzany jest od nowa. W przypadku przeciwnym wykonywana jest operacja mutacji, która polega na tym, że wartość genu zmieniana jest na jedną z dwóch różnych od niego wartości. Na przykład, jeżeli gen miał wartość „c" w procesie mutacji z równym prawdopodobieństwem może mu zostać przypisana wartość „v" lub „x". Zmutowany osobnik potomny podlega ocenie. Jako funkcję dopasowania (ang. fitness function) wybrano funkcję zliczającą liczbę genów, które są identyczne z genami wzorca. Jeżeli zmutowany osobnik potomny uzyska większą wartość funkcji dopasowania, wówczas osobnik rodzicielski jest usuwany, a osobnik potomny zajmuje jego miejsce. Po czym opisane operacje są powtarzane na nowo. Algorytm ewolucyjny kończy działanie w przypadku, gdy funkcja dopasowania dla zmutowanego potomka osiągnie wartość równą liczbie genów wzorca, co oznacza, że w wyniku ewolucji osobnik stał się tożsamy z wzorcem.

Jako miarę oddalenia badanego osobnika od wzorca przyjęto średnią liczbę pokoleń, jakie muszą kolejno po sobie nastąpić, aby osobnik ten upodobnił się całkowicie do wzorca. Jest rzeczą raczej oczywistą, że im bardziej wymowa liczebników zakodowanych na materiale genetycznym badanego osobnika różni się od wymowy liczebników zakodowanych na materiale genetycznym wzorca, tym więcej pokoleń musi upłynąć do momentu, w którym badany osobnik w wyniku następujących w kolejnych pokoleniach mutacji przekształci się we wzorzec. Przypuszczenia te zostały całkowicie potwierdzone podczas przeprowadzonych przez autora badań, których rezultaty zostały omówione w kolejnym punkcie.

4. WYNIKI BADAŃ

W przeprowadzonych przez autora badaniach, jako wzorzec przyjęto zapis wymowy liczebników głównych języka polskiego. Następnie badano zbieżność algorytmu ewolucyjnego operującego na materiale genetycznym stanowiącym zapis wymowy liczebników głównych wybranych języków świata. Uwzględniono zarówno języki indoeuropejskie, jak i języki nie należące do indoeuropejskiej rodziny językowej. Dla

każdego z badanych języków wykonano dziesięć eksperymentów, podczas których zapis wymowy liczebników badanego języka ewoluował aż do momentu, w którym stał się tożsamy z zapisem wymowy liczebników języka polskiego. Liczba pokoleń algorytmu ewolucyjnego, jakie nastąpiły po sobie podczas każdej z prób, została dla każdego z eksperymentów odnotowana w odpowiedniej tabeli. Ponadto w ostatniej kolumnie każdej tabeli zawarto wartość średnią liczby pokoleń algorytmu ewolucyjnego.

W pierwszej kolejność eksperymenty przeprowadzono dla języków słowiańskich. Ich rezultaty zostały zamieszczone w tab. 1.

Tabela 1

Wyniki eksperymentów uzyskane dla języków słowiańskich

The results of experiments that were conducted for Slavonic languages

język	1	2	3	4	5	6	7	8	9	10	średnia
czeski	35	26	30	27	26	31	28	25	36	34	29,8
słowacki	31	29	32	32	29	30	23	33	31	31	30,1
rosyjski	57	49	49	58	47	56	53	57	56	50	53,2
białoruski	42	22	41	33	41	35	35	38	28	40	35,5
ukraiński	56	53	53	50	54	64	55	48	51	54	53,8
słoweński	47	41	52	47	38	50	47	45	45	40	45,2
chorwacki	47	43	54	41	37	48	46	48	42	38	44,4
serbski	53	38	61	44	62	50	56	42	44	53	50,3
macedoński	58	55	52	53	55	61	59	50	56	59	55,8
bułgarski	55	52	58	54	54	50	61	58	46	60	54,8

Następnie eksperymenty przeprowadzono dla języków należących do grupy języków germańskich, a ich rezultaty zebrano w tab. 2.

Tabela 2

Wyniki eksperymentów uzyskane dla języków germańskich

The results of experiments that were conducted for Germanic languages

język	1	2	3	4	5	6	7	8	9	10	średnia
angielski	49	52	49	42	57	50	53	40	50	44	48,6
niemiecki	60	52	66	59	63	64	50	46	55	53	56,8
niderlandzki	56	57	61	50	60	59	49	47	48	56	54,3
afrikaans	59	57	54	53	69	63	53	48	45	55	55,6
szwedzki	60	53	51	49	62	55	48	47	43	44	51,2
norweski	70	67	63	57	62	60	52	50	51	60	59,2
duński	68	66	65	56	63	59	55	51	53	56	59,2
islandzki	56	49	51	48	61	50	44	51	55	56	52,1

Badania przeprowadzono również dla języków romańskich, a ich wyniki zamieszczono w tab. 3.

Tabela 3

Wyniki eksperymentów uzyskane dla języków romańskich

The results of experiments that were conducted for Romance languages

język	1	2	3	4	5	6	7	8	9	10	średnia
portugalski	65	51	58	55	63	62	55	54	54	59	57,6
hiszpański	55	39	54	49	52	45	47	45	47	44	47,7
kataloński	66	55	57	60	65	65	61	55	63	61	60,8
francuski	63	68	66	74	70	59	51	66	65	71	65,3
włoski	46	34	44	50	43	58	57	48	46	48	47,4
rumuński	57	48	60	60	61	57	55	48	52	66	56,4

Z kolei w tab. 4 zamieszczono wyniki eksperymentów uzyskane dla innych języków należących do indoeuropejskiej rodziny językowej, nie wchodzących jednakże w skład żadnej z wymienionej uprzednio grup języków słowiańskich, germańskich i romańskich. Uwzględniono tutaj między innymi języki celtyckie, indoirańskie, bałtyckie oraz języki takie jak grecki i albański.

Tabela 4

Wyniki eksperymentów uzyskane dla wybranych języków indoeuropejskich

The results of experiments that were conducted for the chosen Indo-European languages

język	1	2	3	4	5	6	7	8	9	10	średnia
litewski	58	70	63	69	77	60	58	83	65	53	65,6
łotewski	40	37	52	51	57	54	57	42	50	58	49,8
walijski	66	58	60	55	67	61	58	60	59	72	61,6
irlandzki	59	47	54	56	56	56	48	49	58	59	54,2
perski	70	64	61	63	60	54	49	57	57	57	59,2
urdu	76	63	70	68	67	65	58	60	62	68	65,7
albański	62	43	69	62	51	54	58	54	56	65	57,4
grecki	77	72	76	80	81	66	69	65	72	80	73,8

Natomiast tab. 5 prezentuje wyniki eksperymentów uzyskane dla wybranych języków nie należących do indoeuropejskiej rodziny językowej. Uwzględniono tutaj między innymi języki należące do rodzin afroazjatyckiej, ałtajskiej, uralskiej, bantu i chińskiej.

Tabela 5

Wyniki eksperymentów uzyskane dla wybranych języków nie należących do rodziny indoeuropejskiej

The results of experiments that were conducted for the chosen Non-Indo-European languages

język	1	2	3	4	5	6	7	8	9	10	średnia
węgierski	95	96	71	76	71	75	78	76	86	84	80,8
fiński	98	82	97	91	90	81	76	83	104	70	87,2
estoński	79	75	61	84	76	61	69	72	75	79	73,1
arabski	106	104	92	105	96	116	105	105	90	115	103,4
hebrajski	74	76	70	76	69	75	66	67	64	70	70,7
swahili	69	74	64	62	65	73	55	62	64	69	65,7
turecki	60	74	58	59	62	72	58	60	54	64	62,1
japoński	76	82	57	60	66	59	63	60	65	64	65,2
chiński mandaryński	73	85	58	56	65	62	66	53	42	59	61,9
chiński kantoński	77	88	71	68	71	69	60	55	66	81	70,6

5. WNIOSKI KOŃCOWE

Wyniki przeprowadzonych eksperymentów prezentują się bardzo ciekawie i to w sytuacji (co trzeba szczególnie podkreślić) wyboru bardzo skromnego pod względem objętości materiału porównawczego dla badanych języków (było to zaledwie dziesięć słów — liczebników głównych). Pomimo tak szczupłego materiału porównawczego uzyskane wyniki dają sporo do myślenia.

Na przykład dla języków najbliżej spokrewnionych z językiem polskim i należących do tej samej podgrupy języków zachodniosłowiańskich, czyli dla języków czeskiego i słowackiego uzyskano średnie wyniki wynoszące odpowiednio 29,8 i 30,1 pokoleń algorytmu ewolucyjnego. Niewiele więcej, bo zaledwie 35,5 pokoleń algorytmu ewolucyjnego musiało przeminąć zanim zapis wymowy liczebników języka białoruskiego przekształcił się w zapis tych liczebników charakterystyczny dla języka polskiego, który został umieszczony na wzorcu. Warto ostatni z wyników porównać z wynikiem uzyskanym dla języka rosyjskiego wynoszącym 53,2 pokoleń. Potwierdza to całkowicie głoszoną przez slawistów tezę, w myśl której język białoruski jest językiem przejściowym – łączącym ogniwem pomiędzy językami polskim i rosyjskim, a zatem jest znacznie bliższy polskiemu niż rosyjski. Interesująco prezentuje się również wynik uzyskany dla języka chorwackiego wynoszący 44,4 pokoleń, co zdaje się potwierdzać obiegową opinię o dużym podobieństwie tego języka do języka polskiego. Wyniki odpowiadające największej liczbie pokoleń algorytmu ewolucyjnego równe 54,8 i 55,8

uzyskano odpowiednio dla języków bułgarskiego i macedońskiego, które przez językoznawców są uważane za najbardziej odstające od pozostałych współczesnych języków słowiańskich. Jak widać uzyskane podczas eksperymentów wyniki zdają się tę tezę potwierdzać.

Ogółem dla języków należących do grupy słowiańskiej uzyskano wynik wynoszący 45,3 pokoleń algorytmu ewolucyjnego.

W przypadku ośmiu badanych języków należących do grupy germańskiej uzyskano wynik średni wynoszący 54,6 pokoleń algorytmu ewolucyjnego. Z kolei w przypadku sześciu badanych języków z grupy romańskiej wynik ten wynosi 55,9. W obu przypadkach uzyskane wyniki średnie są większe niż dla języków słowiańskich, co jest całkowicie zrozumiałe w świetle znacznie bliższego pokrewieństwa genetycznego języka polskiego z pozostałymi językami słowiańskimi, niż z językami germańskimi czy romańskimi.

Jeżeli chodzi o wyniki badań uzyskane dla innych wybranych języków należących do rodziny indoeuropejskiej, to warto zwrócić bliższą uwagę na język łotewski, dla którego uzyskano wynik równy 49,8 pokoleń algorytmu genetycznego oraz na język grecki, dla którego wynik ten wynosi 73,8. Otrzymane wyniki zdają się potwierdzać głoszoną już od dawna przez lingwistów hipotezę o wzajemnej bliskości języków słowiańskich i bałtyckich (języki te są łączone nawet we wspólną grupę bałtosłowiańską) i ich pochodzeniu od wspólnego prajęzyka bałtosłowiańskiego. Mimo wszystko językowi polskiemu jest znacznie bliżej do języka łotewskiego niż do języka greckiego, który oddzielił się prawdopodobnie najwcześniej od hipotetycznej prarodziny języków indoeuropejskich.

Ogółem dla badanych języków indoeuropejskich nie należących do grupy słowiańskiej, germańskiej i romańskiej otrzymano wynik średni wynoszący 60,9 pokoleń algorytmu ewolucyjnego. Zgodnie z oczekiwaniami średni wynik uzyskany dla języków nie należących do indoeuropejskiej rodziny językowej ma większą wartość wynosząca 74,1 pokoleń algorytmu ewolucyjnego. Jest to całkowicie zrozumiałe jeśli weźmie się pod uwagę bardzo dalekie pokrewieństwo genetyczne języka polskiego i rozważanych języków indoeuropejskich oraz zupełny brak takiego pokrewieństwa w przypadku języków należących do innych badanych rodzin językowych.

Na zakończenie autor chciałby jeszcze raz podkreślić, że uzyskane wyniki rokują duże nadzieje odnośnie możliwości praktycznego zastosowania opracowanej przez autora metody w badaniach lingwistycznych prowadzonych nad pochodzeniem języków naturalnych. Również autor w przyszłości zamierza zastosować opracowaną przez siebie metodę w badaniach nad pochodzeniem języka ormiańskiego [13], ukierunkowanych na weryfikacje hipotez wywodzenia się tego języka z grupy języków perskich, wymarłej już grupy języków anatolijskich [12] czy też z grupy, do której należał przodek współczesnego języka greckiego.

6. BIBLIOGRAFIA

1. Encyklopedia językoznawstwa ogólnego. Wydawnictwo Ossolineum, Wrocław, 1999
2. L. Bolc, M. Cichy, L. Różańska: *Przetwarzanie języka naturalneg.* Wydawnictwa Naukowo-Techniczne, Warszawa, 1982
3. Cz. Basztura: *Jak rozmawiać z komputerem ludzkim głosem?.* Wydawnictwo Politechniki Wrocławskiej, Wrocław, 1995
4. W. J. Hutchins: *Machine Translation — Past, Present, Future.* Ellis Horwood Series in Computers and Their Applications, London, 1986
5. A. Kondratow: *Zaginione cywilizacje.* Państwowy Instytut Wydawniczy, Warszawa, 1988
6. D. Goldberg: *Algorytmy genetyczne i ich zastosowania.* WNT, Warszawa, 1995
7. D. Rutkowska, M. Piliński, L. Rutkowski: *Sieci neuronowe, algorytmy genetyczne i systemy rozmyte.* Wydawnictwo naukowe PWN, Warszawa, 1997
8. J. A. Matisoff: *Zagrożona różnorodność: Języki i formy życia.* Świat Nauki, październik 2002, ss. 66-73
9. M. J. Künstler: *Języki chińskie.* Wydawnictwo Akademickie DIALOG, Warszawa, 2000
10. J. Tulisow: *Język mandżurski.* Wydawnictwo Akademickie DIALOG, Warszawa, 2000
11. A. F. Majewicz: *Języki świata i ich klasyfikowanie.* Państwowe Wydawnictwo Naukowe, Warszawa, 1989
12. M. Popko: *Ludy i języki starożytnej Anatolii.* Wydawnictwo Akademickie DIALOG, Warszawa, 1999
13. A. Pisowicz: *Gramatyka ormiańska (Grabar – Aszcharabar).* Wydawnictwo KSIĘGARNIA AKADEMICKA, Kraków, 2001

M. GAJER

THE IMPLEMENTATION OF EVOLUTIONARY ALGORITHMS IN COMPUTATIONAL LINGUISTICS FOR THE PURPOSE OF ESTIMATION OF THE RELATIONSHIP BETWEEN NATURAL LANGUAGES

Summary

The topic of the paper is aimed for the implementation of evolutionary algorithms in the field of linguistics. The aim of the research conducted by this author is the estimation of relationship between natural languages.

Among the arts linguistics is often considered a science. Considering this it is nothing strange that soon after the invention of the first digital computer there were proposed some non-numerical application that were aimed for natural language processing. These computer programs are mainly aimed on the spoken dialog between human and computer. Information processing techniques are used also by the linguists for the purpose of facilitation of their research.

For example computers were used by the linguists during the process of reconstruction of some ancient languages and for deciphering of an unknown alphabets. All this could of course be done also by a human but computers are much faster and accurate and though can save the linguists the great amount of boring work.

As far as this author knows the evolutionary algorithms have never been used in the field of linguistics. In this author's opinion that fact may seem strange because the mechanisms that rule the evolution of species and the evolution of natural languages are in fact based on the same principles. In both cases

slight mutations that affect the genetic material of a living organism or a language (its pronunciation or grammatical rules) can be propagated to the next generations if they are in some sense profitable or give some advantage over the rivals.

Linguistics is a discipline that is over one hundred years old, but despite that it is a still young discipline with many open problems that wait for their solution. These open problems are mainly affiliated with the questions about the origin of certain natural languages. For example the linguists are still quarrelling about the origin of Basque language and whether it is related with some other languages.

The method that was developed by this author cannot of course give a definite answer to such questions, but it can deliver some new and maybe strong arguments for supporting or rejection of given linguistic hypotheses.

As it was mentioned the method is based on the technique of evolutionary algorithms and its clue is the operation of mutation during which the pronunciation of words changes. The experiments were conducted on the pronunciation of cardinal numerals of several languages. The evolutionary algorithm operates as long as the pronunciation of mutated words converts into the pronunciation of the given pattern. The result is the average number of generation of evolutionary algorithm. The experiments conducted by this author revealed that such average numbers are able to say much about the relationship between natural languages. The results of experiments that were gathered in tables 1 — 5 show how the Polish language is closely related with other Slavonic languages and also with some Germanic and Romance languages and how far it is from other Non-Indo-European languages.

Keywords: computational linguistics, evolutionary algorithms, the evolution of natural languages

KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJI, 2005, 51, z. 2, ss. 291–303

Software Implemented Fault Detection and Fault Tolerance Mechanisms — Part I: Concepts and algorithms

PIOTR GAWKOWSKI, JANUSZ SOSNOWSKI

Institute of Computer Science,
Warsaw University of Technology,
ul. Nowowiejska 15/19 00-665 Warsaw
Email: jss@ii.pw.edu.pl

Otrzymano 2004.09.10
Autoryzowano 2004.12.30

The paper discusses the problem of eliminating hardware fault effects by means of software. We describe various error detection and fault handling software schemes, show their limitations and capabilities. On the basis of this analysis, we propose original fault handling procedures, which integrate hardware and software mechanisms. Special attention is paid to exception handling and error recovery procedures. The presented solutions have been verified for a wide spectrum of applications running on IBM PC environment.

Keywords: on-line testing, fault detection, fault tolerance, microprocessor systems

1. INTRODUCTION

Dependability is a key issue in many applications ranging from simple embedded to complex mainframe or distributed systems. System robustness (correct operation in the presence of faults) and fail-safe operation (avoiding generation of incorrect critical results or output signals) are becoming common requirements to modern civilisation. These properties specify acceptable (for the considered application) system behavior in the presence of various faults. In fact the acceptable behavior depends upon the application. For example in highly dependable banking systems it is important to assure low probability and short duration of system outages. Faults in traffic light controller should not result in collision situations, in the worst case the lights should be switched off (fail-stop state). Designing dependable systems we have to take into account the taxonomy of faults, capabilities of fault avoidance, fault masking and fault tolerance techniques. A good survey of these issues is given in [1,17,21].

Various fault detection (including on-line test techniques), diagnosis and error handling procedures have been developed (e.g. [4,15,21,22]). Software techniques are especially attractive when hardware design is fixed and cannot be modified e.g. COTS elements (commercial-off-the-shelf), IP (intellectual property) blocks in SoC (system-on-chip) design. They can be very effective for transient faults, which dominate in new technologies as well as in radiation environment [3,6,20,24].

Software techniques for decreasing system fault susceptibility are based on modification of the application code by introducing suitable procedures or instructions either at high-level source code or low level assembler code (e.g. [5,13-16,19]). Depending upon the algorithm these modifications are more or less complex and involve different levels of redundancy (data, code, time).

The techniques proposed in the literature concentrate mostly on software redundancy concepts (section 2). In fact, error detection should combine hardware and software mechanisms. Faults detected by hardware are signalled to software as exceptions. Our experiments show that they may cover 10-60% of faults. We have also identified other weak points related to software environment. In particular, typical neglected problems relate to synchronisation and error handling procedures in real system: taking over exceptions at the application level, localising and identifying faults, eliminating their effects etc. In section 3 we present an original approach, which integrates all these aspects. Some specific problems of error coverage are discussed in section 4. They relate to transient, intermittent and permanent fault discrimination. Moreover, we consider the impact of time and memory overhead for the hardened applications. The effectiveness of error handling procedures is analysed in part II of the paper [9].

2. IMPROVING SYSTEM DEPENDABILITY

Hardware faults can be detected using special on-line test mechanisms e.g. watchdog processors, control flow monitoring, error detection codes, invalid memory access or invalid instruction code detectors [21,22]. Some of these mechanisms are available in COTS chips and systems. Typically faults are signalled by interrupt mechanisms (exceptions). Error confinement capabilities of these mechanisms can be enhanced at the software level. Fault tolerance needs some kind of redundancy like massive redundancy in hardware (e.g. TMR systems), data redundancy (e.g. error detection and correction codes), algorithm, program and time redundancy. Hardware redundancy not only increases system cost but also limits the possibility of using COTS systems with matured production processes, which are the most reliable etc. Hence, recently there has been a lot of interest in software implemented fault detection and fault tolerance procedures.

Software implemented fault hardening mechanisms are targeted for error detection or correction. Error detection is assured by simple assertions and control flow monitoring ([2,14]) or by various program modifications: i) replicating instructions, program variables, ii) program modules, N-version programming or iii) using recovery blocks

[12] etc. Fault tolerance or masking can be achieved with retries, error recovery, voting etc. These methods are especially effective in case of transient hardware faults. A large class of permanent faults can also be tolerated with RESO (recomputing with shifted operands), redundancy in RAM and CPU (e.g. superscalar structure) etc. [10,16,21]. Moreover, many data structures or files comprise some natural redundancy useful in detecting or correcting faults [8].

Fault detection or fault tolerance techniques targeted for software faults (design faults) may not cover hardware permanent faults efficiently. In N-version programming different versions of a program are developed to avoid common design errors [12]. The result correctness is assured by voting whereas error detection can be obtained by comparison. Similarly, in recovery block techniques various versions of a program are used, but only one version is active at a time and the result is verified with acceptance tests. Another version is used if the previous one did not pass the acceptance test. Depending upon the acceptance test this technique may cover hardware faults. Hybrid techniques like consensus recovery, N self-checking have similar properties [12].

It is important to extend program or data diversity for hardware faults as well. The simplest technique is the use of AN codes in which each data word is multiplied by some constant A [10]. An error is detected by checking if the computation result is divisible by A. This requires division operation (expensive if A is not even number). AN code is only applicable to fixed point numbers. In ED^{4}I approach [15] data (and partially program) diversity is assured by transforming an original program according to some rules, performing computations with the original and transformed programs and comparison of results generated by two versions. In the transformed program, all variables and constants are multiplied by the diversity factor k. Depending on the value of k, the original and transformed programs may use different parts of hardware and propagate faults in different ways. Hence, it is highly probable that a fault will produce different results for both programs. Practically, this probability depends on the value of k and program features. In [15] it was shown that good results (highest error detection) can be achieved with some optimal values of k adjusted to specific logical blocks (e.g. k=-2, 4, -1 for the adder, multiplier/divider and shifter blocks). We have to select one k value for the whole program, depending on which blocks dominate calculations.

Program transformations are performed in basic blocks and branch statements. The basic block is a sequence of consecutive statements such that flow of control enters at the block beginning and leaves at its end (without branching within the block). Transformation of the basic block is limited to conversion of variables and constants to k-multiples. The branching condition transformation modifies expressions in the branch statement and keeps the control flow the same as in the original program. Detailed transformation algorithms are given in [15].

In hardware approach to fault detection an important issue is to assume not only high fault coverage but also to limit the circuit overhead (e.g. lower than duplication). An example of this is checking arithmetic results using residual codes or parity predic-

tion circuitry [10,21]. Transforming these ideas into software we can achieve similar code overhead as for classical techniques with replicated calculations.

The granularity of the recalculation or checking can be coarse-grained (final result verification as in the approaches presented above) or fine-grained. A fine-grained approach could lower the error detection latency (at the cost of program overhead) which depends on the granularity level: program modules, functions or even instructions. Some kind of fine- grained checking based on comparison was proposed in [19], where several rules of transforming a primary program into redundant structure by duplication were developed. These rules relate to variable duplication and control flow checking:

- Variable duplication rules (VDR):
  - #1: Every variable x must be duplicated: let $x1$ and $x2$ be the names of the two copies;
  - #2: Every write operation performed on x must be performed on $x1$ and $x2$;
  - #3: After each read operation on x, the two copies $x1$ and $x2$ must be checked for consistency and an error detection procedure should be activated if an inconsistency is detected.
- Control flow checking rules (CFCR):
  - #4: An integer value v_i, is associated with every basic i-th block (branch-free) in the code;
  - #5: A global execution check flag (*ecf*) variable is defined; a statement assigning to *ecf* the value of v_i is introduced at the very beginning of every basic i-th block; a test on the value of *ecf* is also introduced at the end of the basic block;
  - The rules #4 and #5 are extended for procedures etc. (more details in [19]).

To assure error correction, we can use the following error correction rules (ECR) defined in [23]:

— For one selected set of duplicated variables (primary or secondary) one checksum c is associated. The initial value of the checksum is equal to the EXOR of all of the already initialised variables in the set. Before every write on the variable from this set, the checksum is recomputed by EXORing its previous value with the last value of updated variable. Then, after writing the new value of the variable, it is updated by EXORing it with this value.

— If an error is detected, the error recovery procedure recomputes the EXOR on the set of variables associated with the checksum and compares it with the stored one. If both values are consistent, the associated variable set is copied to the secondary set; otherwise, the secondary set of variables is copied to the primary one.

The original transformation rules do not cover some important issues of high-level programming. In particular, the pointer variables cannot be treated the same way as the normal data, as they cannot be simply compared and operations cannot be doubled. To resolve this problem, we create two copies of the pointer (each one pointing to the same memory cell - primary variable) and the offset between them. The offset is used for comparison of the data copies: the primary one is pointed directly, whereas the

secondary one is pointed by the duplicated pointer with the offset. Since using only global variables as in [23] is too restrictive, we also have introduced extension of the above rules from [19,23] by independent treatment of local and global variables, which leads to separate checksums.

More advanced methods of detecting faults can be based on monitoring program and data flow during execution. As a program executes, it moves through transitions from one global state to another through normal execution or due to disturbance. Monitoring can verify the consistency between parts of the software module, the occupancy time the software spends in a particular state and the transition paths taken from one state to another. Two entities may have valid states individually, but the combination of their states may not be legal, so we have to check this consistency. The execution of consistency checks is based on software system requirements such as concurrency, synchronization, locking and security. We can verify the occupancy factors (probability to be in a given state, resource usage, synchronization delays) and the transition consistency (whether the path taken between two entities is consistent). The transition distribution looks at historical distribution of how often all the transition paths are taken compared with expected operational patterns.

While designing fault detection or fault tolerance mechanisms in software we can assume fine- or coarse-grained model. In the former, checking procedures are performed frequently, e.g. every execution of a duplicated instruction or a small piece of a program. In the letter, this process is initialized scarcely, e.g. at the end of calculations. Although fine granularity assures lower fault latency, it also introduces bigger memory and time overhead. In consequence, this may lead to higher sensitivity to faults, due to higher probability of disturbing fault-handling instructions (we illustrate this in part II [9]). Similar problems arise with error recovery procedures (essential for transient faults). Eliminating fault effects we face the problem of fault identification, which needs more complex fault handling procedures (section 3 and 4).

3. ERROR HANDLING PROBLEMS

In the literature authors restrict their considerations to general ideas of fault detection or correction. While implementing these ideas in a system environment in real applications, we face two neglected problems: i) handling exceptions, and ii) managing error detection and recovery procedures (in particular, those related to fault localization and discrimination). Built in hardware fault detectors (e.g. access violation, invalid opcode, memory misalignment errors) generate interrupts (exceptions), which are handled typically by the operating system. In a wide range of applications, we found that such events are generated in 10-60% depending upon fault localisation. Hence, even in the case of software procedures tolerating faults, quite a significant percentage of faults may not be corrected or masked due to generated exceptions. To overcome this problem, we have to include exception handling at the application level. Some operating systems as well as high-level languages deliver appropriate mechanisms to

facilitate resolving this problem. For instance, the *try* and *catch* construct in object oriented languages has the following form:

```
try {segment of the application code}
catch (exception filter1) {specification of exception
handling procedure}
.....
catch(exception filter n) {specification of exception
handling procedure}
```

It assures taking over exceptions (specified by the filter or all of them – *catch*(...)) generated during the execution of the code within the *try* brackets (*try* {*segment of the application code*}). For any specified exception (in the exception filter) we define an appropriate handling procedure. For example it may initiate reexecution of the program code starting from some specified checkpoint (previously established — backward recovery), suspending further execution of the thread etc. Error correction can be made, if the error is uniquely correlated with the disturbed program segment. In some cases this correlation may be impossible. For example, disturbing program stack may change return addresses of procedures, in consequence leading to access violation exception, which cannot be correctly correlated with code segment to recover (practically only full retry is possible — restart). Moreover in microcontroller platforms we have very poor hardware error detectors, so access violation exception is not available. To detect such errors, we can use a software control flow checking mechanism, but an efficient recovery cannot be done yet due to problems in fault localisation. In the case of fine-grained fault detection or correction, return addresses stored on the stack are not redundant and not visible at the program code (implicit use of pointer registers e.g. ESP, EIP and EBP registers in Intel processors).

To overcome the presented problem, we should assure redundancy of pointers stored on the stack and appropriate comparison or voting mechanisms (part II [9]), what can be done manually or with a post-processor. Another solution is executing the application in debugging mode with breakpoints set for stack reference instructions. By executing these instructions, we verify their correctness: valid addresses, comparison with redundant copy of stack or switching to safe addresses, if critical disturbances occur.

While handling exceptions or other error messages (e.g. generated by the application), we should notice that transient faults injected into registers and data memory may be efficiently recovered by retries or error recovery to the previous checkpoint. More problems relate to faults in the program code, as even transient faults might modify the program code. These modifications can be detected using various checksums and then recovered. In general, we can distinguish several levels of recovery. If the fault appears in the cache memory, then it can be recovered by flushing this memory (one or more levels of caches can be considered). A fault in the RAM memory may need reloading from a stable memory (e.g. a disc). To reduce recovery time, several copies

of the program segments can be stored in RAM and exchanged. In many applications we can assume that simple retry can eliminate most of transient faults, so only in the case of the second exception, we can decide to recover by checking segment code correctness and reload it. Various strategies are possible here and they can be adapted to the application. These relatively simple ideas are quite complex in implementation within any typical operational system environment.

For an illustration, we present the developed scheme of the coarse-grained fault detection and fault tolerance related to TMR system implemented in multithreaded environment (single or multiprocessor). The main thread manages calculations in the processing threads and determines the final result.

```
void main(void)
{InitializeInputData();
 CreateWorkingThreads(); // create processing threads
                  //they will be suspended waiting for data to compute
 while (Data_To_Process) // while there is new // data to process
 { PrepareInputData(); // prepare input data
  StartToCompute(); // start (resume) computations //of all processing threads
  WaitForResults(); // wait for minimal number of //results for the final decision
  SelectResults(); // determine reliable result
  PrepareNextIteration();} // prepare for the next-iteration
       // and if needed restart, terminate or suspend disturbed threads
}}
```

After some initialisation steps (initial input data preparation, creating processing threads), the main thread enters the main computation loop whose iteration begins with preparing input data for processing threads. Then, all processing threads receive the "green light" to begin computations. The main thread suspends itself waiting for results from the processing threads. It is waked-up when at least 2 processing threads deliver results or the time-out occurs. This is assured by the function *WaitForResults*. In Win32 such a function can be realised using a synchronisation mechanism *WaitForMultipleObjects*, which signals, if the specified wait criteria has been met: 1) the timeout interval elapsed, 2) any one or all of the specified objects are in the signalled state. In the sequel, we will also use simpler function *WaitForSingleObjects*, in which the above criteria 2 is reduced to a single object. In case of time-out situation, if any result is available, it could be considered as the correct one (if the processing thread was reliable in the past, some reliability specification can be also added to the result) or a fixed safe result value might be delivered. The optional solutions are retry or sending an error message to the user.

If the main thread receives results from 2 threads, it compares them. If they are consistent (according to the used voting algorithm e.g. [11]), the result of the given iteration is selected. In this case, there is no need to wait for the third thread computations (it can be cancelled and prepared for the next iteration). If two results are not consistent the main thread should wait for the third result (or for time-out), and

then the voting procedure should determine the final iteration result. In more advanced fault handling, some statistical information of the processing threads (related to the history of thread behaviour etc.) can also be used. It can be obtained from reliability monitoring in the main and processing threads.

The processing threads may detect a fault locally (e.g. by taking over exceptions) and perform some correction actions (e.g. an error recovery procedure). This process can be enhanced by some statistical information delivered from the main thread (e.g. consistency of the obtained result with a selected correct one, fault identification statistics in previous iterations etc.). If the processing thread activity is not successful (e.g. infinite loop), the main thread may terminate it, recreate it or reconfigure the system to the degraded structure (e.g. double modular redundancy - DMR). Here we give an outline of a processing thread.

```
Status WorkingThread(void)
{ try { CheckForConsistency(); } // testing the hardware and the thread
  catch(...) { return THREAD_INTERNAL_FAILURE; }
  try { while(!finalize) // the main computation loop
        {try { WaitForSingleObject(GreenLight); // suspend processing thread
                        // till the main thread synchronization
       MakeCheckpoint();// store initial state for error nrecovery
       CheckForConsistency();
       Compute(InputData); // data processing
       ValidateResults(); // local checking of results
       ReturnResults(Result, ResultStatus);// signal availability of results
                        // and its status to the main thread
       UpdateInternalStatistics();} //reliability statistics
  catch( InternalFailure )
  { return THREAD_INTERNAL_FAILURE;
  catch(ComputationFailure —— ...)
  { // local recovery from transient faults
    UpdateInternalStatistics();
    try{ RollbackRecoveryFromCheckpoint();
         CheckForConsistency();Compute(InputData);
         VaildateResults();ReturnResults(Result,ResultStatus);
         UpdateInternalStatisticks();}
    catch(InternalFailure)
     {return THREAD_INTERNAL_FAILURE;}
     catch(ComputationFailure —— ...) // unrecoverable
                  //computation fault but consistent thread state
    { ReturnResults(NOTHING,ResultStatus=INVALID);
      UpdateInternalStatistics();}} } }
  catch(...) { return THREAD_INTERNAL_FAILURE; }
  return OK;}
```

The main thread is synchronized with the processing threads via *GreenLight* (starting iteration processing) and *finalize* (end of calculations) flags. The processing thread generates synchronisation objects for the main thread in the function *ReturnResults*,

which are used by the function *WaitForMultipleResult* in the main thread. The presented processing thread implements internal reliability monitoring and retry in case of fault detection (rollback recovery using checkpointing). The consistency check is performed after thread creation as well as for each processing iteration. Such a frequent checking may be needed only in critical applications. This prevents the situation when some faults occur in the system during thread suspension. The processing thread implements local error containment. All these functions are optional and are specified in italic.

In general, different schemes of error handling can be used in the main as well as in the subordinate threads and optimised by taking into account application requirements, fault statistics etc. (compare [3,22]). In particular, we can move most of fault handling processing (in italic) to the main thread. Two types of exceptions are filtered in the processing thread: *InternalFailure* and *ComputationFailure*. The first one is the most critical and may relate to permanent faults in hardware, whereas the second one may be eliminated by error recovery. While performing recovery in response to exception, we have to be careful with the system state recovery. For example, we have observed that faults injected into FPU (floating-point unit) resulted in critical corruption, which had its impact during program re-execution. Hence, the FPU initialization is needed before the retry. If the recovery procedure is too simple, there is a risk that the recovery will generate a wrong result without exception. In consequence, we will loose more by handling the primary exception instead of signalling an error.

In simple systems with no multithreading capabilities the redundant processing code should be supported with an additional control code to integrate error detection and error handling procedures. All code segments are executed sequentially. We illustrate this for an application with DMR/TMR redundancy (coarse-grained).

```
try{ InputData.Checkpoint(); //PS1
   Process1(InputData, Result);
   AcceptanceTest(Result.FirstProposition);} // optional testing
catch(...){ConsistencyCheck();} //optional check and repair
try{ InputData.Rollback(); //PS2
   Process2(InputData, Result);
   AcceptanceTest(Result.SecondProposition);} // optional testing
catch(...){ConsistencyCheck();} //optional check and repair
try{ if (CompareTwoPropositions(Result))
        return OK;}
catch(...){ConsistencyCheck();} //optional check and repair
try{   InputData.Rollback(); PS3
       Process3(InputData, Result);
       return SelectResult(Result); }
catch(...){return FAILED;}
```

The function *Processi(InputData, Result)* performs calculations in the specified i-th processing segment *(InputData)* and delivers *Result*. The function *InputData.Checkpoint()* saves input data for recovery use. The function *InputData.Rollback*() recovers the original input data for the next processing segment. For each data pro-

cessing segment results are optionally checked with acceptance tests. After second processing, the result is compared with the first one and — in the case of consensus — further code is skipped; otherwise, the third processing segment is executed and the most probable result is selected or an error is signalled (fail). Any exception initiates the optional checking code consistency, error removing and execution of the next segment.

In the presented schemes, we assume that exceptions are handled at the application level. Should we encounter some problems, we can shift this to the system level by issuing *throw* command from the application. In general, error-handling process may be much more complex, if we add the capability of distinguishing transient, intermittent and permanent faults, which can be done using algorithms described in the next section. Notice that combining software and hardware redundancy leads to more complex error handling strategies (compare [22]).

To facilitate developing efficient fault hardened applications, some functions could be included in compilers or operating systems, e.g. procedures for application consistency checking periodically, cache invalidation, checkpointing and rollback. This is because many aspects of the execution environment cannot be managed at the application level. After introducing code redundancy into the application, we have observed that some compilers may detect and reduce it during code optimisation (part II [9]). So, it is reasonable to verify this effect and switch off the optimisation. In fact, some improvements of compilers towards dependable software are worth consideration.

4. ERROR COVERAGE PROBLEMS

There are two important issues in developing fault-hardening procedures. The first one relates to localization and identification of faults (in particular, transient, intermittent, and permanent). The other problem is the impact of the introduced space and time overhead on fault susceptibility.

Transient faults are caused by various external and internal disturbances (electrical, cosmic and ion radiation etc.). They appear randomly and their effects can be eliminated by recalculation etc. Permanent and intermittent faults relate to circuit defects and can be masked out by the circuit redundancy (massive or naturally present in some COTS elements). The crucial point is to distinguish these three classes of faults (the overwhelming majority of the occurring faults are transient). Algorithms dealing with this problem have been proposed in the literature [3,22]. In our methodology we rely on count and threshold scheme from [3]. In the single threshold count algorithm (STC), as time goes on, detected faults for a specified component (e.g. working thread) are counted with decreasing weighs as they get older, to decide the point in time when keeping a system component on-line is no longer beneficial. Here, we can use a simple filtering function:

$$\alpha^{L} = \alpha^{L-1} \cdot K \quad \text{if} \quad J^{L} = 0 \quad \text{else} \quad \alpha^{L-1} + 1, \quad 0 \le K \le 1$$

where: α is a score associated to each not-yet removed component to record information about failures experienced by that component, J^L denotes the L-th signal specifying detected fault (1) or the lack of detected fault (0) in the L-th time instant. The parameter K and the threshold α_T are tuned so as to find the minimum number of consecutive faults sufficient to consider a component as faulty (permanent, intermittent). Similar filtering function can be formulated as an additive expression [3]. In more complex double threshold count scheme (DTC) we have two thresholds α_L and α_H. Components with $\alpha < \alpha_L$ are considered as non-faulty (disturbed or not by transient faults), with $\alpha > \alpha_H$ as faulty (permanent or intermittent fault) and with $\alpha_L \leq \alpha \leq \alpha_H$ as suspected (further decision can be based on extensive testing).

The cost of fault hardening is time and space overhead of the application. To analyse this overhead, it is reasonable to distinguish static and dynamic as well as active and passive overhead. The static overhead shows the increase in static code, which relates to memory space requirements. The dynamic overhead relates to time and it shows the time increase in execution of the hardened application compared to the original one. The active overhead is limited to the code executed during application operation, if no error has occurred. The passive overhead includes the overhead related to error handling activity (if an error occurred). A large active (static and dynamic) overhead may result in proportionally higher fault rate (e.g. due to cosmic rays) as compared with the original application. So, an important issue is to achieve high error coverage to compensate this effect (neglected in the literature). The passive overhead influences system operation (if a fault has occurred). The code related to this overhead is also susceptible to faults, so there is a danger that it can be corrupted and result in bad error recovery (if fault occurs). As opposed to active overhead, the passive one is not checked during normal operation, so fault accumulation effect may appear during long periods of operations. Hence, an important issue is to check the integrity of this code periodically or before its usage (error scrubbing).

5. CONCLUSION

Error handling in COTS systems needs integration of hardware and software error detection mechanisms as well as error recovery procedures in the system environment. This results in quite complex fault handling scenarios, so various mechanisms supporting exception handling, process synchronisation etc. are needed. We have outlined some possibilities of improving these mechanisms in COTS systems. This paper also shows that beyond the basic redundancy concepts, there are many problems (neglected in the literature) related to the integration with the system environment.

Embedding software fault hardening procedures into the application leads to memory and time overhead. On one hand, it influences performance and cost (not critical in many applications). On the other hand, it makes the system more robust. However, the resulting overhead introduces new areas of fault disturbance what is especially critical in the case of fine-grained approaches (with low error latency). An important

issue is to check fault susceptibility of these mechanisms and harden them, which is analysed in part II [9].

6. ACKNOWLEDGEMENT

This work was supported by KBN grant no. 4T11C049 25.

7. REFERENCES

1. A. Avizienis, J. C. Laprie, B. Randel and C. Landwehr: *Basic concepts and taxonomy of dependable and secure computing.* IEEE Trans. on Dependable and Secure Computing, vol. 1, no. 2, pp. 11-33, Jan.- Mar. 2004.
2. A. Benso, S. Di Carlo, G. Di Nartale, P. Prinetto and L. Tagliaferri: *Control flow checking via regular expressions.* Proc. of the 10^{th} Asian Test Symposium, pp. 299-303, 2001.
3. A. Bondavalli, S. Chiaradonna, F. Di Giandomenico and F. Grandoni: *Threshold-based mechanisms to discriminate transient from intermittent faults.* IEEE Trans. on Computers, vol. 49, no. 3, pp. 230-245, March 2000.
4. D. C. Bossen, A. Kitamorn, K. F. Reick and S. Floyd: *Fault tolerant design of the IBM pSeries 690 system using Power 4 processor technology.* IBM J. Res.& Dev., vol. 46, no. 1, pp. 77-86, Jan. 2002.
5. P. Cheynet, et al.: *Experimentally evaluating an automatic approach for generating safety critical software with respect to transient errors.* IEEE Trans. on Nuclear Science, vol. 47, no. 6, pp. 231-236, Dec. 2000.
6. P. E. Dodd, L. W. Massengill: *Basic mechanisms and modelling of single-event upset in digital microelectronics.* IEEE Trans. on Nuclear Science, vol. 49, pp. 583-602, June 2003.
7. P. Gawkowski, J. Sosnowski: *Experimental evaluation of fault handling mechanisms.* Proc. of 20th Int. Conference SAFECOMP, Springer Verlag, LNCS 2187, pp. 109-118, 2001.
8. P. Gawkowski, J. Sosnowski: *Dependability evaluation with fault injection experiments.* IEICE Trans. Inf. & Syst., vol. E86D, no. 12, pp. 2642-2649, Dec. 2003.
9. P. Gawkowski, J. Sosnowski: *Software Implemented Fault Detection and Fault Tolerance Mechanisms, Part II, Experimental evaluation of error coverage.* Kwartalnik Elektroniki i Telekomunikacji, nr 3, 2005 (w druku).
10. B. W. Johnson: *Design and analysis of fault tolerant digital systems.* Addison Wesley, 1989.
11. G. Latif-Shabgahi, J. M. Bass and S. Bennett: *A taxonomy for software voting algorithms used in safety critical systems.* IEEE Trans. on Reliability, vol. 63, no. 3, pp. 319-328, Sept. 2004.
12. M. Lyu: *Software fault tolerance.* John Wiley & Sons, 1995.
13. B. Nicolescu, R. Velazco and M. S. Reorda: *Effectiveness and limitations of various software techniques for soft error detection, a comparative study.* Proc. of 7^{th} IEEE Int. On-line Testing Workshop, pp. 172-177, 2001.
14. N. Oh, P. P. Shirvani and E. J. McCluskey: *Control flow checking by software signature.*IEEE Trans. on Reliability, vol. 51, no. 1, pp. 111-122, March 2002.
15. N. Oh, S. Mitra and E. J. McCluskey: *ED^4I Error detection by diverse data and duplicated instructions.* IEEE Trans. on Computers, vol. 51, no. 2, pp. 180-199, Feb. 2002.
16. N. Oh, P. P. Shirvani and E. J. McCluskey: *Error detection by duplicated instructions in super scalar processors.* IEEE Trans. on Reliability, vol. 51, no. 1, pp. 63-75, March 2002.

17. F. Piedad M. Hawkins: *High Availability, design techniques and processes.* Prentice Hall Inc., 2001.
18. S.J. Piestrak: *Design of Self-Testing Checkers for Unidirectional Error Detecting Codes.* Scientific Papers of the Inst. of Tech. Cybern. of the Tech. Univ. of Wrocław. no. 92, Ser.: Monographs No. 24, Oficyna Wyd. Polit. Wrocł., Wrocław 1995.
19. M. Rebaudengo, M. S. Reorda and M. Violante: *A new software based technique for low cost fault tolerant application.* Proc. of Annual Reliability and Maintainability Symposium, pp. 25-28, 2003.
20. N. Seifert, Z. Xiaowei, L. W. Massengill: *Impact of scaling on soft-error rates in commercial microprocessors.* IEEE Trans. on Nuclear Science, vol. 49, pp. 3100-3106, Dec. 2002.
21. D. P. Siewiorek and R. S. Swarz: *Reliable Computer Systems: Design and Evaluation.* AK Peters, 1998.
22. J. Sosnowski: *Transient fault tolerance in digital systems.* IEEE Micro, pp. 24-35, February 1994.
23. F. Vargas, R. D. R. Fagundes, D. Barros and D. R. Brum: *Briefing a new approach to improve EMI immunity of DSP systems.* Proc. of the 12th IEEE Asian Test Symposium, pp. 468-471, 2003.
24. R. Velazco, S. Rezgui: *Assessing the soft error rate of digital architectures devoted to operate in radiation environment: a case study.* Journal of Electronic Testing, vol. 19, pp. 83-90, 2003.

P. GAWKOWSKI, J. SOSNOWSKI

PROGRAMOWE MECHANIZMY DETEKCJI I TOLEROWANIA BŁĘDÓW SPRZĘTU CZĘŚĆ I KONCEPCJE I ALGORYTMY

Streszczenie

W artykule przedstawiono programowe metody zwiększania odporności systemów mikroprocesorowych na błędy sprzętu. Omówiono różne techniki detekcji i tolerowania błędów, ich ograniczenia oraz możliwości. Na bazie przeprowadzonej analizy przedstawiono oryginalne procedury obsługi błędów, które integrują mechanizmy sprzętowe i programowe. Szczególną uwagę poświęcono problemowi obsługi wyjątków i mechanizmom odtwarzania. Zaproponowane rozwiązania zostały zweryfikowane dla szerokiego spektrum aplikacji.

Słowa kluczowe: testowanie równoległe, detekcja i tolerowanie błędów, systemy mikroprocesorowe

KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJI, 2005, 51, z. 2, s. 305–318

Wpływ tłumienia zależnego od polaryzacji na dyspersję polaryzacyjną w światłowodzie jednomodowym

ŁUKASZ MAKSYMIUK

Politechnika Warszawska, Instytut Telekomunikacji
Zakład Systemów Mikrofalowych i Optoelektronicznych
ul. Nowowiejska 15/19, 00-665 Warszawa
maksymiuk@tele.pw.edu.pl

Otrzymano 2004.06.03
Autoryzowano 2004.09.14

W niniejszym artykule autor prezentuje model kaskadowy światłowodu jako układu składającego się z kilkudziesięciu segmentów. Model ten jest odpowiedni do rozpatrzenia problemu oddziaływania zjawiska PDL na PMD. Autor na bazie owego modelu przedstawia wyniki symulacji numerycznych. Autor przedstawia proces zmian rozkładu gęstości prawdopodobieństwa różnicowego opóźnienia grupowego w zależności od PDL. Drugim aspektem rozważanym w niniejszym artykule jest zaburzenie ortogonalności modów polaryzacyjnych pod wpływem PDL. Na podstawie symulacji numerycznych zostaje przedstawiony stopień degradacji systemu transmisyjnego z PMD pod wpływem niezerowego PDL.

Słowa kluczowe: dyspersja polaryzacyjna, tłumienie zależne od polaryzacji, oddziaływanie PMD na PDL

1. WSTĘP — OPIS PROBLEMU

Obecnie pracuje się nad wdrożeniem systemów, których przepływności będą osiągać wartości 40 Gbit/s i więcej. Wiadomo, że jedną z poważnych trudności przy uruchamianiu takich systemów jest dyspersja polaryzacyjna (PMD ang. Polarization Mode Dispersion), która wywołuje poszerzenie impulsu mogące prowadzić do wystąpienia interferencji międzysymbolowych. Przyjmuje się, że średnie różnicowe opóźnienie grupowe (DGD ang. Differential Group Delay), które jest definiowane jako różnica czasów propagacji modów polaryzacyjnych LP_{01x} i LP_{01y}, nie powinno przekraczać 10% trwania impulsu (np. dla systemu 40 Gbit/s dopuszczalna wartość średniego DGD wynosi 2,5ps). Jeśli chodzi o tłumienie zależne od polaryzacji PDL (ang. Polarization De-

pendent Loss), nie istnieją normy określające całkowity dopuszczalny PDL systemu, określa się jedynie PDL dla niektórych wybranych komponentów [1]. Wiadomo jednak, że zbyt duży PDL negatywnie wpływa na jakość transmisji w systemie poprzez wprowadzanie niestabilności mocy optycznej sygnału, która może powodować błędne rozpoznanie wartości logicznej przez odbiornik. Przez długi czas oba zjawiska (PDL oraz PMD) były traktowane oddzielnie, wiele wskazuje na to, że było to błędne podejście. Należy sobie postawić pytanie: Czy tłumienie zależne od polaryzacji wpływa na wartość różnicowego opóźnienia grupowego? Jeśli tak faktycznie jest, to należy sobie postawić kolejne pytanie: Jaki jest charakter zależności pomiędzy PMD a PDL? W niniejszym artykule autor daje odpowiedzi na dwa powyższe pytania i pokazuje charakter zależności pomiędzy rozkładem prawdopodobieństwa DGD a PDL w światłowodzie jednomodowym wykorzystując do tego symulacje komputerowe bazujące na modelu matematycznym, który będzie opisany w dalszej części.

2. OPIS MATEMATYCZNY PDL I PMD

Istnieje kilka różnych matematycznych metod ujęcia PDL oraz PMD, które opisują oba zjawiska oddzielnie. W przypadku PMD najczęściej używaną jest metoda opisu za pomocą głównych stanów polaryzacji PSP (ang. Principal Polarization State) zaproponowana po raz pierwszy prze C.D. Poole'a (patrz [2]). Polega ona na wprowadzeniu dwóch ortogonalnych wektorów w formalizmie Stokesa $\vec{\Omega}_+$ i $\vec{\Omega}_-$ odpowiadających stanom polaryzacji, dla których światło pokonuje światłowód w najkrótszym i najdłuższym czasie. Jest to analogia do światłowodowych modów polaryzacyjnych LP_{01x} i LP_{01y}. Jeżeli chodzi o PDL, to najczęściej spotyka się opis za pomocą wektora $\vec{\Gamma}$ (patrz [3]), którego długość jest równa PDL (w skali liniowej), a jego kierunek i zwrot pokazują na sferze Poincaré'go stan polaryzacji najmniej stłumionej.

Aby zbadać oddziaływanie obu zjawisk potrzebny jest ich wspólny opis. W niniejszym artykule posłużono się zespoloną macierzą transmisji Jonesa [4], oznaczmy ją jako $T(\omega)$. Macierz ta opisuje pojedynczy element toru transmisyjnego (np. światłowód), bądź cały układ i zawiera informację zarówno o tłumieniu zależnym od polaryzacji, jak i dyspersji polaryzacyjnej [5]. Jest ona zależna od pulsacji ω. Jeżeli układ składa się z kilku komponentów, a nas interesuje jego charakterystyka wypadkowa, można ją wyznaczyć jako iloczyn poszczególnych macierzy Jonesa:

$$T_{wyp}(\omega) = T_n \cdot T_{n-1} \cdot \ldots \cdot T_2 \cdot T_1. \tag{1}$$

Macierz $T(\omega)$ ma tę własność, że można ją zdekomponować na dwie macierze: jedną odpowiedzialną za dyspersję polaryzacyjną, drugą za tłumienie zależne od polaryzacji [6]. Oznaczmy $A(\omega)$ jako macierz odpowiedzialną za PDL, a $U(\omega)$ jako macierz odpowiedzialną za PMD, możemy wtedy zapisać zależność [6]:

$$T(\omega) = A(\omega)U(\omega). \tag{2}$$

Ze wzoru (2) płynie konkluzja, że każdy element charakteryzujący się niezerowym tłumieniem zależnym od polaryzacji oraz dyspersją polaryzacyjną można przedstawić jako złożenie dwóch elementów: pierwszego odznaczającego się wyłącznie tłumieniem zależnym od polaryzacji i drugiego związanego wyłącznie z dyspersją polaryzacyjną.

Aby zdekomponować macierz $T(\omega)$ na macierze $A(\omega)$ i $U(\omega)$ należy posłużyć się następującymi zależnościami [6]:

$$A(\omega) = \sqrt{T(\omega)T(\omega)^*}, \tag{3}$$

$$U(\omega) = A(\omega)^{-1}T(\omega). \tag{4}$$

Jeżeli interesuje nas informacja o PDL, to w tym celu musimy obliczyć wartości własne macierzy $A(\omega)$, a PDL otrzymamy z wyrażenia [5]:

$$PDL = 10\log\left(\frac{\rho_1}{\rho_2}\right)[dB]. \tag{5}$$

Aby uzyskać DGD z macierzy $T(\omega)$ należy obliczyć wartości własne wyrażenia [6]:

$$\left(\frac{\partial \mathrm{T}(\omega)}{\partial\omega}\right)T(\omega)^{-1}, \tag{6}$$

które są zespolone, a DGD dane jest częścią rzeczywistą [6]. W celu obliczenia DGD można się również posłużyć wartościami własnymi wyrażenia:

$$\left(j\frac{\partial U(\omega)}{\partial\omega}\right)U(\omega)^{-1}, \tag{7}$$

Huttner i Gisin w [6] dowodzą, że sposób ten jest dokładniejszy, jeżeli mamy do czynienia z układem obarczonym zarówno PDL, jak i PMD. W praktyce, trudno jest uzyskać analityczne rozwiązanie 6 lub 7, ponieważ wymagają one wyliczenia pochodnej cząstkowej względem ω, stosuje się więc inna zależność w postaci [5]:

$$T(\omega + \Delta\omega)\cdot T(\omega)^{-1}, \tag{8}$$

a następnie wylicza wartość DGD z równania [5]:

$$DGD = \left(\frac{\rho_1}{\rho_2}\right)/\Delta\omega\,[ps]. \tag{9}$$

Ta metoda pozwala na dobre przybliżenie, pod warunkiem dobrania dostatecznie małej wartości $\Delta\omega$ [5]:

$$\Delta\omega\cdot DGD < \pi. \tag{10}$$

Pozostaje kwestia jak wyliczyć macierz Jonesa $T(\omega)$ posiadając założone parametry PDL oraz DGD. W niniejszym artykule posłużono się notacją Stokesa, wprowadzając

odpowiednie wektory opisujące PDL oraz PMD. Zdefiniujmy teraz te wektory jako: $\vec{\beta}$ wektor odpowiadający DGD oraz $\vec{\alpha}$ odpowiadający PDL. Powyższe wektory można rozpisać w następujący sposób:

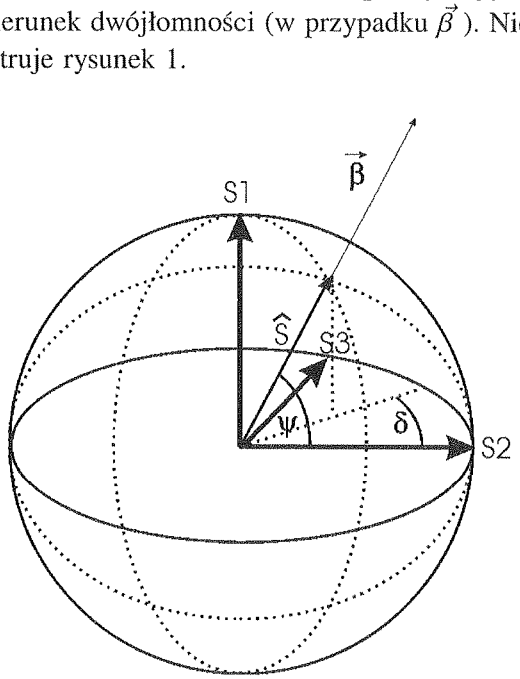
$$\vec{\beta} = \beta \cdot \hat{S}_\beta, \tag{11}$$

$$\vec{\alpha} = \alpha \cdot \hat{S}_\alpha. \tag{12}$$

Zmienna β, przy założeniu, że dwójłomność nie zależy od pulsacji ω (co jest dobrym przybliżeniem w wypadku światłowodu — patrz [6]) jest równa DGD elementu, który jest opisany wektorem $\vec{\beta}$. Wartość skalarna α jest współczynnikiem związanym z tłumieniem różnicowym [5]:

$$e^{2\alpha} = \frac{I_{\max}}{I_{\min}}, \tag{13}$$

i jest ona równa PDL wyrażonemu w skali liniowej. Wektory $\hat{S}_\alpha$ i $\hat{S}_\beta$ są rzeczywistym jednostkowymi wektorami Stokesa (w postaci znormalizowanej — 3 wymiary [4]) pokazującymi na sferze Poincaré'go odpowiednio polaryzację najmniej tłumioną (w przypadku $\vec{\alpha}$) oraz kierunek dwójłomności (w przypadku $\vec{\beta}$). Niech sposób tworzenia wektorów $\vec{\alpha}$ i $\vec{\beta}$ zilustruje rysunek 1.



Rys. 1. Sposób tworzenia wektora $\vec{\beta}$ za pomocą wektora jednostkowego $\hat{S}_\beta$ i wartości skalarnej β na sferze Poincaré'go

Fig. 1. The way of constructing vector $\vec{\beta}$ out of a unit vector $\hat{S}_\beta$ and scalar β on the Poincaré sphere

Mając ustalone wektory $\vec{\beta}$, $\vec{\alpha}$ oraz pulsację ω możemy wyliczyć dla nich macierz $T(\omega)$ korzystając z następującej zależności (patrz [6]):

$$T(\omega) = C \cdot \exp\left[\frac{(-j\vec{\beta}\omega + \vec{\alpha}) \cdot \vec{\sigma}}{2}\right], \tag{14}$$

gdzie: $\vec{\sigma} = [\sigma_1\ \sigma_2\ \sigma_3]$ jest wektorem Pauliego zdefiniowanym poprzez macierze:

$$\sigma_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \sigma_2 = \begin{bmatrix} 0 & -j \\ j & 0 \end{bmatrix}, \sigma_3 = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix},$$

natomiast C jest stałą zespoloną związaną z tłumieniem.
Jak już wspomniano wcześniej, C.D Poole wprowadził opis zjawiska PMD za pomocą PSP w [2]. Informacja o PSP jest zawarta w macierzy Jonesa i można ją uzyskać posługując równaniem własnym macierzy $T(\omega)$ [6]:

$$\left[\frac{\partial T(\omega)}{\partial \omega}\right] T(\omega)^{-1} = -\frac{i}{2}\vec{\Omega} \cdot \vec{\sigma}, \tag{15}$$

w przypadku braku PDL wektor $\vec{\Omega}$ (określany mianem wektora PSP) jest rzeczywisty i wskazuje na sferze Poincaré'go główny stan polaryzacji PSP, w wypadku wystąpienia PDL wektor $\vec{\Omega}$ staje się zespolony i jego powiązanie z głównym stanem polaryzacji nie jest już takie bezpośrednie (zainteresowanego tym tematem czytelnika odsyłam do [7]). Na potrzeby niniejszego artykułu nie będzie potrzebna znajomość wektora $\vec{\Omega}$, który jest 3 elementowym rzeczywistym wektorem Stokesa, wystarczy znajomość wektora w notacji Jonesa. Oznaczmy wektory PSP w notacji Jonesa jako $\vec{\psi}_+$ oraz $\vec{\psi}_-$, są one wektorami własnymi wyrażenia:

$$\left[\frac{\partial T(\omega)}{\partial \omega}\right] \cdot T(\omega)^{-1}. \tag{16}$$

Wektory $\vec{\psi}_+$ oraz $\vec{\psi}_-$ są ortogonalne w wypadku światłowodu obarczonego wyłącznie PMD, jeżeli jednak PDL jest niezerowy tak nie musi być. Zdefiniujmy więc współczynnik korelacji owych wektorów jako [7]:

$$\gamma = \left|\vec{\psi}_+^* \cdot \vec{\psi}_-\right|, \tag{17}$$

przy czym $\gamma \in < 0; 1 >$ i jest równe 0 wypadku braku korelacji, czyli w wypadku pełnej ortogonalności wektorów. Jeżeli wektory opisujące stany polaryzacji są ortogonalne, wtedy nie jest możliwa interferencja fal elektromagnetycznych, których polaryzacje opisane są przez te wektory. Tak dzieje się w wypadku modów polaryzacyjnych w światłowodzie, kiedy mamy do czynienia jedynie ze zjawiskiem PMD. Jeżeli natomiast wektory $\vec{\psi}_+$ i $\vec{\psi}_-$ nie są ortogonalne, to wtedy mody światłowodowe mogą ze sobą interferować, co będzie się przejawiało w kształcie impulsu świetlnego na wyjściu

światłowodu. W artykule [7] autorzy wyprowadzili zależność na poszerzenie impulsu gaussowskiego w światłowodzie o określonym DGD i współczynniku korelacji γ wektorów PSP. Wspomniana zależność wygląda następująco:

$$< t >_{\max/\min} = \frac{\pm 0,5 \cdot DGD}{\sqrt{1-\gamma^2}}, \tag{18}$$

i definiuje najkrótszy i najdłuższy czas w jakim impuls gaussowski pokona światłowód. W przypadku, gdy wektory $\vec{\psi}_+$ i $\vec{\psi}_-$ są ortogonalne ($\gamma = 0$) poszerzenie impulsu wyniesie:

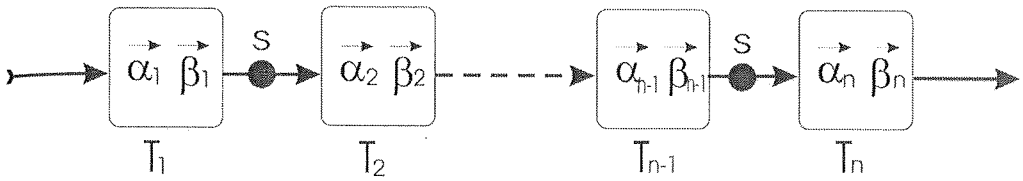
$$\delta_{\max} = < t >_{\max} - < t >_{\min} = DGD, \tag{19}$$

i równe będzie wyłącznie DGD.

3. MODEL KASKADOWY ŚWIATŁOWODU

W poprzednim punkcie podana została metoda opisu PDL oraz PMD za pomocą macierzy Jonesa. Jak powiedziano, macierz taka opisywać może zarówno własności poszczególnych elementów jak i całego układu. Tę właśnie cechę wykorzystano do zasymulowania światłowodu jako połączenia kaskadowego wielu składowych elementów. Światłowód jest specyficznym elementem sieci optycznych, znajduje się on w środowisku, które nie da się kontrolować. Ponadto rzeczywiste światłowody odbiegają od idealnych projektów, chociaż w ostatnich latach dał o sobie znać znaczący postęp technologiczny. Wszystko to sprawia, że światłowód jest elementem niestacjonarnym, poza tym tę niestacjonarność da się jedynie ograniczać, nie da jej się wyeliminować. Światłowód poddawany jest ciągłej zmianie temperatur i wahaniom naprężeń. To wszystko sprawia, że zmianie ulega lokalna dwójłomność światłowodu, co wpływa na wartość PMD. Dodatkowym aspektem jest sprzęganie pomiędzy modami polaryzacyjnymi LP_{01x} i LP_{01x}, które występuje na skutek fluktuacji składu szkła kwarcowego, z którego wykonane są rzeczywiste światłowody, przy czym sprzęganie pomiędzy modami polaryzacyjnymi należy rozumieć jako przelewanie się energii pomiędzy nimi. Należy zaznaczyć, że w procesie tym nie zostaje zaburzona ortogonalność modów polaryzacyjnych [8]. Na skutek wspomnianych tu problemów różnicowe opóźnienie grupowe DGD ma zmienny charakter i opisane jest rozkładem prawdopodobieństwa.

Aby zasymulować fluktuacje lokalnej dwójłomności oraz sprzęganie pomiędzy modami zaproponowano model kaskadowy światłowodu, przedstawiony na rysunku 2. Każdy segment został opisany zespoloną macierzą Jonesa, przy czym wektory $\vec{\beta}$ i $\vec{\alpha}$ opisujące poszczególne macierze posiadały różny kierunek i zwrot dla każdej macierzy. Przy takim modelu światłowodu, sprzęganie modowe następuje na styku poszczególnych elementów kaskady (punkty „s" na rysunku 2). Nie ma sprzęgania pomiędzy modami w obrębie jednego elementu kaskady; to założenie prowadzi do przyjęcia zasady, że wektory $\vec{\beta}$ i $\vec{\alpha}$ są zawsze zgodne co do kierunku i zwrotu dla danej macierzy $T(\omega)$ (patrz [9]).



Rys. 2. Schemat kaskadowego modelu światłowodowego składającego się z n elementów; Sprzęganie pomiędzy modami następuje w punktach oznaczonych „s"

Fig. 2. Scheme of the cascade fibre model consisting of n elements. Mode coupling takes place at "s" points

4. ROZKŁADY PRAWDOPODOBIEŃSTWA DGD PRZY RÓŻNYM PDL

Przeprowadzono serię symulacji bazujących na modelu matematycznym przedstawionym w punkcie 3. Ustalono następujące parametry symulacji: ilość elementów w kaskadzie –250, liczba iteracji –30000, wartość β poszczególnych segmentów –0, 2ps, wartość α ulegała zmianie dla różnych symulacji. Pulsację ω dobrano tak, aby odpowiadała ona środkowej długości fali dla trzeciego okna transmisyjnego. Położenie wektorów $\vec{\beta}_n$ i $\vec{\alpha}_n$ opisujących poszczególne segmenty kaskady było losowane odrębnie dla każdej symulacji i dla każdego segmentu oddzielnie, przy czym przyjęto rozkład jednostajny ich rozmieszczenia na sferze Poincaré'go.

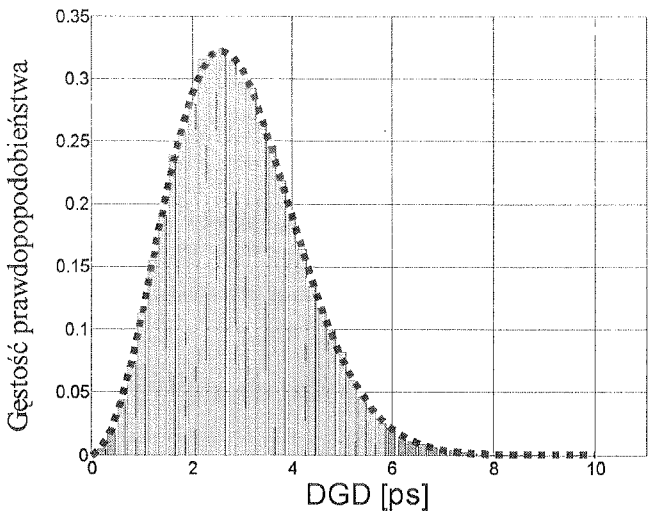
Wykresy 3 i 4 przedstawiają rozkład gęstości prawdopodobieństwa DGD, odpowiednio dla światłowodu charakteryzującego się tylko PMD oraz dla światłowodu zarówno z PMD jak i PDL. W tym drugim przypadku ustalono parametr $\alpha = 0,03$, co dało wypadkową średnią wartość PDL dla światłowodu 3,98 dB. Na wykresie 3 można zaobserwować, że rozkład prawdopodobieństwa dla DGD światłowodu obarczonego jedynie PMD dobrze opisuje rozkład prawdopodobieństwa Maxwella (wzór 20) (ta cecha PMD jest dobrze znana i opisana w literaturze [8; 10]).

$$p(x) = \sqrt{\frac{2}{\pi}} \cdot a^{\left(\frac{3}{2}\right)} \cdot x^2 \cdot e^{-a \cdot \frac{x^2}{2}} \tag{20}$$

Parametr rozkładu a w tym wypadku wynosi 0,301. Został on wyliczony za pomocą 1 momentu rozkładu Maxwella, który odpowiada wartości średniej:

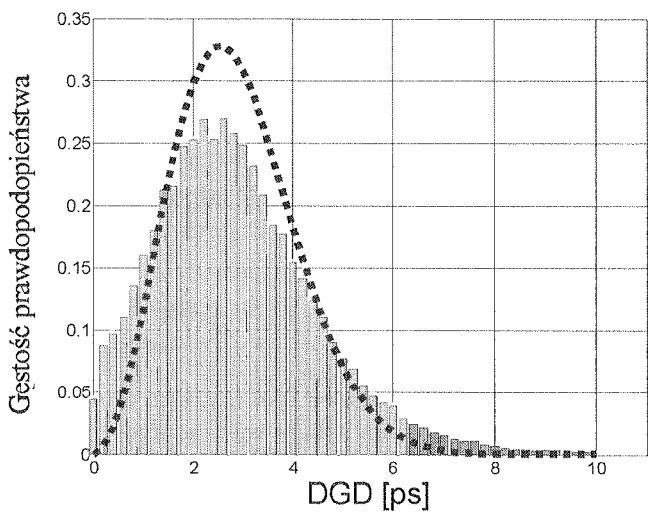
$$\mu_1 = 2\sqrt{\frac{2}{\pi a}}. \tag{21}$$

W przypadku światłowodu obarczonego zarówno PDL, jak i PMD można zaobserwować wyraźne odbieganie charakterystyki rozkładu prawdopodobieństwa od krzywej Maxwella (rys. 4.). Jest to dowód na oddziaływanie PDL na DGD światłowodu, przy czym charakter zmian nie powoduje zmian wartości średniej DGD (tabela 1). Rozkład



Rys. 3. Wykres gęstości prawdopodobieństwa DGD przy zerowym PDL światłowodu, przerywaną linią zaznaczono charakterystykę rozkładu Maxwella (wzór 20)

Fig. 3. Probability density function of fibre's DGD with zero PDL. The dotted line shows the Maxwellian characteristics



Rys. 4. Wykres gęstości prawdopodobieństwa DGD przy całkowitym średnim PDL światłowodu wynoszącym 3,9 dB ($\alpha = 0,03$), przerywaną linią zaznaczono charakterystykę rozkładu Maxwella (wzór 20)

Fig. 4. Probability density function of fibre's DGD with 3.9dB PDL ($\alpha = 0,03$). The dotted line shows the Maxwellian characteristics (equation 20)

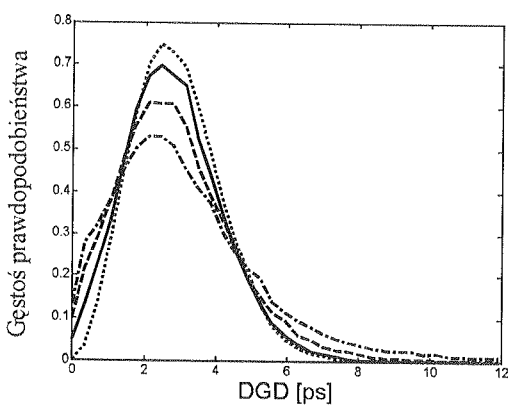
prawdopodobieństwa ulega zmianie w taki sposób, że część centralna ulega spłaszczeniu, a prawdopodobieństwa dla mniejszych i większych wartości od średniego DGD ulegają zwiększeniu. Efekt jest taki, że DGD ma tendencje do większych niestabilności, a co gorsza z większym prawdopodobieństwem może osiągać większe wartości. Sposób ewolucji rozkładu gęstości prawdopodobieństwa pod wpływem PDL można zaobserwować na wykresie 5, przy czym oznaczenia są następujące: linia kropkowana $\alpha = 0$; linia ciągła $\alpha = 0,02$; linia przerywana $\alpha = 0,03$; linia typu „-.-" $\alpha = 0,04$.

Tabela 1

Średnie wartości PDL i DGD symulowanego światłowodu, przy zmieniających się wartościach α segmentów

Average values of PDL and DGD of a numerically simulated fibre for different α values

α	średnie PDL [dB]	średnie DGD [ps]
0	0	2,9
0,005	0,63	2,9
0,01	1,28	2,9
0,015	1,92	2,9
0,02	2,59	2,8
0,025	3,28	2,8
0,03	3,98	2,9
0,035	4,72	3
0,04	5,5	3,2



Rys. 5. Wykresy gęstości prawdopodobieństw DGD. Linia kropkowana $\alpha = 0$; linia ciągła $\alpha = 0,02$; linia przerywana $\alpha = 0,03$; linia typu „-.-" $\alpha = 0,04$

Fig. 5. Probability density functions of DGD. The dotted line $\alpha = 0$; the solid line $\alpha = 0.02$; the dashed line $\alpha = 0.03$; the dashed-dotted line $\alpha = 0.04$

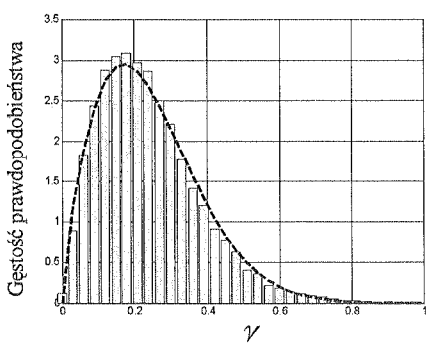
5. WPŁYW ORTOGONALNOŚCI PSP NA DGD PRZY NIEZEROWYM PDL

Przy opisie matematycznym w punkcie 2 powiedziano, że wystąpienie PDL w światłowodzie może prowadzić do utraty ortogonalności głównych stanów polaryzacji PSP, a co za tym idzie powodować interferencje między modami polaryzacyjnymi, co z kolei powoduje większe poszerzenie impulsu, które zostało wyrażone wzorem 18. Aby zilustrować to zjawisko wykorzystano znów ten sam model kaskadowy światłowodu opisany w punkcie 3. Ustalono następujące parametry symulacji: ilość elementów w kaskadzie – 250, liczba iteracji – 30000, wartość β poszczególnych segmentów – 0,2ps, wartość $\alpha = 0,015$ co dało wypadkowy średni PDL=1,92dB. Miarą ortogonalności głównych stanów polaryzacji był współczynnik γ zdefiniowany równaniem 17. Na wykresie 6 przedstawiono wykres gęstości prawdopodobieństwa współczynnika korelacji γ. Rozkład ten można opisać rozkładem Weibulla, w niniejszej symulacji parametry rozkładu opisanego wzorem 22 przyjmują wartości: a=10,31 i b=1,79. Wzór opisujący rozkład Weibulla ma postać:

$$P(x) = ab \cdot x^{b-1} \cdot e^{-ax^b} \cdot I_{(0,\infty)}(x), \tag{22}$$

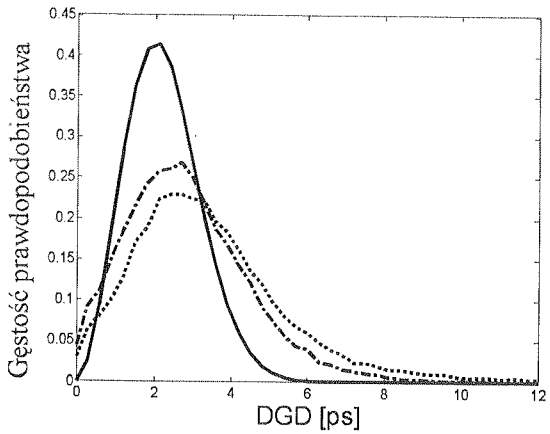
gdzie: a i b są parametrami rozkładu.

Na wykresie 7 przedstawiono porównanie rozkładów gęstości prawdopodobieństw w dla przypadku samego PMD (linia ciągła), dla przypadku z PDL z parametrem $\alpha = 0,135$ (linia typu „-.-„) oraz z uwzględnieniem współczynnika korelacji γ i zależności na poszerzenie impulsu opisanej wzorem 18 (linia kropkowana). Widać wyraźnie, że utrata ortogonalności ma wpływ na dodatkowe pogorszenie sytuacji dla światłowodu obarczonego PMD i PDL. Widać, że prawdopodobieństwa dużych DGD są jeszcze większe niż wypadku rozpatrywania samego oddziaływania PDL na PMD bez uwzględnienia utraty ortogonalności przez mody polaryzacyjne.



Rys. 6. Wykres gęstości prawdopodobieństwa współczynnika γ, linią przerywaną zaznaczono charakterystykę rozkładu Weibulla

Fig. 6. Probability density functions of γ coefficient, the dashed line shows the Weibull characteristics

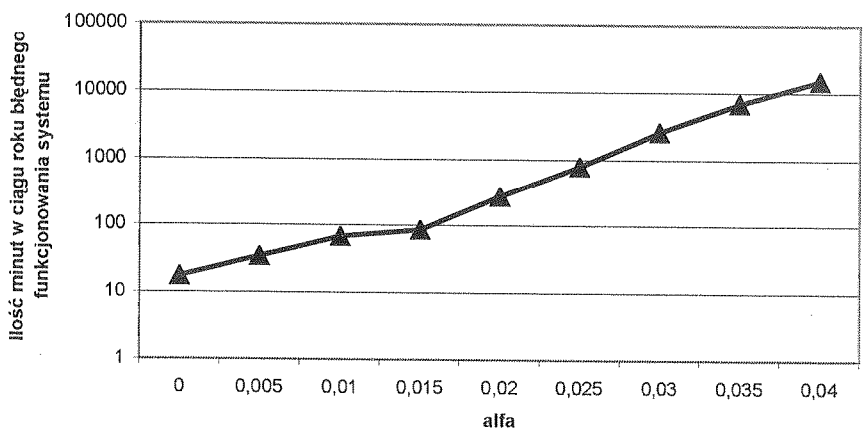


Rys. 7. Wykres przedstawiający gęstość prawdopodobieństwa DGD. Linia ciągłą odpowiada przypadkowi $\alpha = 0$; linia typu „-.-" $\alpha = 0,035$; linia kropkowana dla przypadku $\alpha = 0,035$ z uwzględnieniem utraty ortogonalności modów polaryzacyjnych

Fig. 7. Probability density functions of DGD. The solid line corresponds to the case when $\alpha = 0$; the dotted-dashed line corresponds $\alpha = 0,035$; the dotted line corresponds to $\alpha = 0,035$ including orthogonality loss of polarization modes

6. DEGRADACJA SYSTEMU POD WPŁYWEM PDL

W tym punkcie przedstawiono wyniki symulacji, które obrazują jak zmienia się liczba minut w roku, w ciągu których nie działa poprawnie w zależności od wartości PDL, przy ustalonej wartości PMD. Do tego celu posłużono się rozkładami prawdopodobieństw DGD symulowanego światłowodu uzależnionych od PDL.



Rys. 8. Wykres przedstawiający liczbę minut w ciągu roku, kiedy system nie działa poprawnie

Fig. 8. Plot showing the number of minutes of system malfunction within a one-year period

Parametry symulacji były takie same jak przy wyliczaniu rozkładów prawdopodobieństw w punkcie 4 (dla przypomnienia: ilość elementów w kaskadzie –250, liczba iteracji –30000, wartość β poszczególnych segmentów – 0,2, wartość β ulegała zmianie dla różnych symulacji).

Przyjmuje się, że średnia dopuszczalna wartość DGD dla danego systemu nie powinna być większa niż 10% trwania impulsu (np. dla systemu 40 Gbit/s, średnie DGD < 2,5ps). Znając własności rozkładu Maxwella przyjmuje się, że przy takiej wartości średniej DGD definiującej ten rozkład, wartości DGD wywołujące błędy w systemie mają na tyle małe prawdopodobieństwa, że można uznać system za wolny od wpływu PMD. W niniejszej analizie przyjęto 9ps jako wartość graniczną DGD, przy której pewien system przestaje poprawnie funkcjonować. Następnie opierając się na wyliczonych w symulacjach rozkładach prawdopodobieństw DGD dla różnych PDL wyliczono liczbę minut w roku, w ciągu których system nie działa poprawnie. Wyniki przedstawiono na wykresie 8. Wynika z niego, że przy liniowym wzroście PDL światłowodu (przy ustalonej wartości PMD) ekspotencjalnie wzrasta liczba minut błędnego funkcjonowania systemu. Ma to związek ze zmianami rozkładu prawdopodobieństwa DGD światłowodu, które zostały przedstawione w punkcie 4. Zwiększanie się prawdopodobieństw dla większych wartości DGD pod wpływem PDL powoduje gwałtowną degradację systemu, mimo że średnie wartości DGD nie ulegają zmianie (tabela 1). Ta ostatnia właściwość poddaje w wątpliwość sens stosowania średniego DGD jako jedynej miary potencjalnych problemów związanych z PMD.

7. WNIOSKI

Symulacje dowiodły, że istnieje związek pomiędzy tłumieniem zależnym od polaryzacji a rozkładem różnicowego opóźnienia grupowego. Rozkład DGD w światłowodzie bez PDL ma charakter rozkładu Maxwella. W wypadku niezerowego PDL rozkład prawdopodobieństwa zmienia się, zmiany te mają niekorzystny charakter, powodują większe rozchwianie wartości DGD i znaczący wzrost prawdopodobieństwa dla dużych wartości DGD. Zmiany w charakterystyce rozkładu prawdopodobieństwa DGD wpływają na degradację systemu transmisyjnego, co zostało pokazane na bazie symulacji komputerowych. Przy czym, przy liniowym wzroście PDL następuje ekspotencjalny wzrost czasu, przez który system błędnie funkcjonuje. Zmiany rozkładu prawdopodobieństwa DGD, a zwłaszcza większe fluktuacje wartości DGD, spowodowane PDL, stawiają większe wyzwania dla systemów kompensujących PMD [11]. Systemy takie muszą być tak zaprojektowane, aby reagować szybciej na zmiany DGD, niż by to było konieczne w przypadku występowania samego zjawiska PMD. Należy dodać, że wartości PDL użyte w symulacji są relatywnie duże i nie występują w normalnych światłowodach używanych do transmisji. Niemniej jednak na całkowity PDL toru transmisyjnego mają wpływ także inne elementy (np. sprzęgacze światłowodowe), to może spowodować, że całkowity PDL układu składającego się z odcinków

światłowodu poprzedzielanych sprzęgaczami może okazać się wystarczająco duży, aby spowodować zmiany rozkładu prawdopodobieństwa DGD.

Istnieje jeszcze jeden bardzo ważny aspekt oddziaływania zjawiska PDL na PMD. Jak zostało pokazane, dyspersja polaryzacyjna jest scharakteryzowana dwoma głównymi stanami polaryzacji PSP. Stany te są ortogonalne jedynie w wypadku braku PDL; jeśli on wystąpi, warunek ortogonalności nie musi być spełniony. To powoduje, że mody polaryzacyjne mogą ze sobą interferować i impuls doznaje dodatkowego poszerzenia. W artykule pokazano też, że współczynnik γ będący miarą korelacji PSP ma rozkład gęstości prawdopodobieństwa opisany charakterystyką rozkładu Weibulla. Utrata ortogonalności PSP ma również kluczowe znaczenie dla problemów związanych z kompensacją PMD, której algorytmy opierają się na założeniu o ortogonalności głównych stanów polaryzacji. Zjawisko PDL może spowodować błędne działanie kompensatorów dyspersji polaryzacyjnej.

8. BIBLIOGRAFIA

1. International Telecommunication Union (ITU-T): *Transmission media characteristics — Characteristics of optical components and subsystems — G.671.*
2. C. D. Poole, R. E. Wagner: *Phenomenological approach to polarization dispersion in long single mode fibers.* Electronic Letters, vol.22, 1986, pp.1029-1030.
3. N. Gisin: *Statistics of polarization dependent losses.* Optics Communications 114, 15 February 1995, pages 399-405.
4. F. Ratajczyk: *Dwójłomność i polaryzacja optyczna.* Oficyna Wydawnicza Politechniki Wrocławskiej, 2000.
5. E. Collett: *Polarized Light in Fiber Optics.* The PolaWave Group 2003, ISBN 0-9677167-1-3.
6. B. Huttner, C. Geiser, N. Gissin: *Polarization- Induced Distorsions in Optical Fiber Networks with Polarization-Dependent Losses.* IEEE Journal of Selected Topics in Quantum Eletronics, vol6., no2, March/April 2000.
7. N. Gisin, B. Huttner: *Combined effects of polarization mode dispersion and polarization dependent losses in optical fibers.* Optics Communications 142, 1 October 1997, pages 119-125.
8. J.P.Gordon, H.Kogelnik: PMD fundamentals: *Polarization mode dispersion in optical fibers.* Contributed by H.Kogelnik, February 2, 2000.
9. Ping Lu, Liang Chen, Xiaoyi Bao: *Pulse width dependence of polarization mode dispersion and polarization dependent loss for a pulse ant their impacts on pulse broadening.* Physics Departments, University Ottawa, 2001.
10. N.Gisin: *Polarization Mode Dispersion: Definitions, Measurements and Statistics.* Group of Applied Physics, University of Geneva.
11. Na Yong Kim, Duckey Lee, Hosung Yoon: *Limitation of PMD Compensation Due to Plarization-Dependent Loss in High-Speed Optical Transmission Links.* IEEE Photonics Technology Letters, vol.14, no. 1, January 2002.

Ł. MAKSYMIUK

IMPACT OF POLARIZATION DEPENDENT LOSS ON POLARIZATION MODE DISPERSION IN SINGLE-MODE FIBRE

Summary

The topic of the article is interdependence of PMD (Polarization Mode Dispersion) and PDL (Polarization Dependent Loss). First of all, the proper mathematical feedback is presented. Description of PMD and PDL by the 2×2 Jones complex transmission matrix is provided. Ways of calculating PDL and DGD (Differential Group Delay) out of a matrix are also presented (Jones matrix eigenanalysis JME). After that, the cascade fibre model (model of a fibre that consist of many e.g. 200 pieces) is introduced that reflects all the real system subtleties such as: varying birefringence through changes of temperature and stress, mode coupling and varying of PDL. Then, a set of numerical simulation results (based on the above mentioned model) is presented; these are: DGD probability density functions for different PDL and a plot illustrating the out-of- work time during one-year period as a function of PDL (with fixed PMD). On probability density functions plots the process of evolving distribution function from Maxwellian distribution to a new one due to PDL is shown. The new probability distribution function results in greater probabilities of high DGD values when some amount of PDL is present in the system. On other plots it is shown that linearly increasing PDL of a system produces exponential growth in the amount of out-of-work time during one-year period. The other problem mentioned in the article is the loss of orthogonality of PSP due to the PDL. As an orthogonality measure, a γ coefficient is provided which is a correlation of two PSPs. Then the probability density function of a γ coefficient of a system with PDL is shown. The presented characteristics fits well within the Weibull distribution. At the end the impact of non-orthogonality of PSP on increasing distortions due to PDL on the system with PMD are shown. The distortion manifests with grater probabilities of high DGD than in the case of considering PDL to PMD influence without taking into account the loss of orthogonality of two PSPs. The article outlines the problems which may occur in fibre networks due to the interaction of PMD and PDL and that may not be neglected when introducing high-speed optical networks. It is shown that the measure of average DGD is not sufficient parameter of describing the PMD problem when PDL is taken into consideration.

Keywords: PMD, PDL, polarization, combined effects, fibre

KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJI, 2005, 51, z. 2, ss. 319–334

Pompa kapilarna do chłodzenia układów elektronicznych — stany statyczne

TOMASZ WAJMAN, BOGUSŁAW WIĘCEK, BARTOSZ OSTROWSKI, REMIGIUSZ DANYCH

Instytut Elektroniki, Politechnika Łódzka
Ul. Wólczańska 223, 90-924 Łódź, Polska
E-mail: twajman@p.lodz.pl, wiecek@p.lodz.pl,
bostrow@p.lodz.pl, danych@p.lodz.pl

Otrzymano 2004.02.25
Autoryzowano 2004.04.07

Artykuł dotyczy zagadnień projektowania i konstrukcji pomp kapilarnych do chłodzenia układów elektronicznych. Omówiono rodzaje, budowę oraz zasadę działania pomp kapilarnych. Przedstawiono analizę transportu ciepła i masy w stanach ustalonych dla pomp kapilarnych typu CPL (Capillary Pumped Loops). Omówiono zaprojektowany i wykonany prototyp pompy kapilarnej. Przedstawiono wyniki chłodzenia układów elektronicznych przy pomocy prototypu pompy dla dwóch typów chłodzenia: konwekcji naturalnej i wymuszonej.

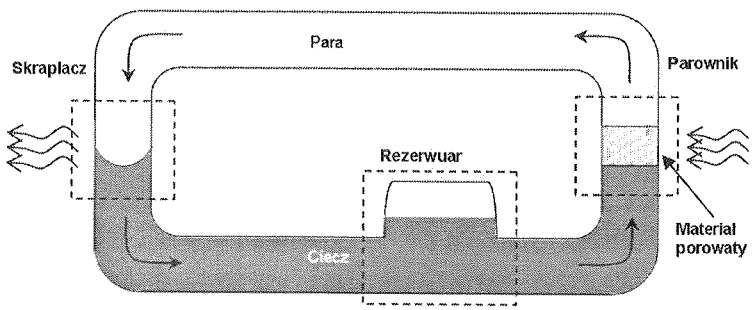
Słowa kluczowe: pompa kapilarna, chłodzenie układów elektronicznych

1. WSTĘP

Układy chłodzenia we współczesnej elektronice odgrywają ogromną rolę. Wzrost mocy wydzielanej w elementach elektronicznych wraz z postępującą miniaturyzacja wymusza stosowanie coraz wydajniejszych układów chłodzenia [1]. Parametry klasycznych układów chłodzenia, takich jak radiatory, wentylatory czy nawet ogniwa Peltier zaczynają ograniczać ich zastosowanie do nowych elementów elektronicznych, wymagających wydajniejszego odprowadzania ciepła. Poszukiwanie efektywnych metod odprowadzania energii cieplnej ze struktury półprzewodnikowej do otoczenia jest jednym z podstawowych zagadnień w dziedzinie elektroniki. Obecnie coraz szersze zastosowania mają układy chłodzące z cieczą, w szczególności układy wykorzystujące przemiany fazowe płynu chłodzącego. W wydajnych rozwiązaniach tego typu można wyróżnić kapilarne układy chłodzenia. Do tej grupy należą rury cieplne (ang. Heat

Pipes) [2, 3, 4], pompy kapilarne typu LHP (ang. Loop Heat Pipes) oraz typu CPL (ang. Capillary Pumped Loops) [5, 6].

Kapilarne układy chłodzenia łączą element elektroniczny, z którego chcemy odprowadzić ciepło z odbiornikiem ciepła (radiatorem, chłodnicą itp.). W części układu stykającej się z elementem, w którym wydzielają się znaczne ilości ciepła, następuje odparowanie cieczy. Ciepło wraz z parą przenoszone jest w kierunku chłodnicy, gdzie następuje oddanie ciepła i skroplenie pary. Cyrkulację zapewnia ciśnienie kapilarne wytwarzane w materiale porowatym. Należy podkreślić, że dla celów chłodzenia układów i przyrządów elektronicznych stosuje się układy z różnym medium, np. z wodą, amoniakiem, freonem, acetonem czy metanolem. Im większe ciepło parowania tym większa skuteczność odprowadzania ciepła.



Rys. 1. Schemat budowy pompy kapilarnej

Fig. 1. Capillary pump structure

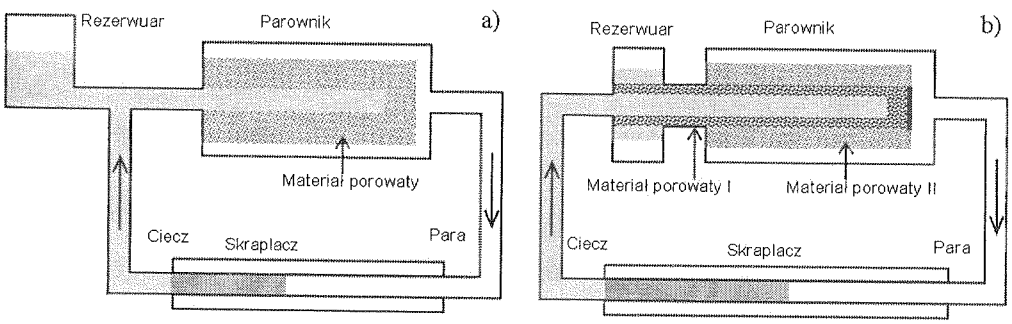
W artykule opisana została pompa kapilarna typu CPL [5, 6]. Schemat budowy pompy jest przedstawiony na rys. 1. Wyróżnia się w niej parownik, skraplacz, rezerwuar oraz przewody łączące. Parownik jest elementem bezpośrednio odbierającym ciepło od układu chłodzonego. Posiada wlot cieczy schłodzonej oraz wylot gorącej pary. Wewnątrz umieszczony jest materiał porowaty, który pełni kluczową rolę w działaniu pompy. Siły kapilarne występujące w tym materiale powodują powstanie podciśnienia, które zasysa ciecz doprowadzaną przez wlot. Ciecz jest transportowana na powierzchnię materiału porowatego, która znajduje się bezpośrednio w sąsiedztwie obudowy. Pod wpływem doprowadzonego ciepła następuje odparowanie czynnika chłodzącego. Para jest odprowadzana z powierzchni materiału porowatego specjalnymi kanałami do wyjścia. Następnie przez przewód transportowana jest do skraplacza. W skraplaczu następuje odebranie ciepła od pary, co powoduje skroplenie i ochłodzenie czynnika chłodzącego. Schłodzona ciecz powraca osobnym kanałem do parownika.

W pompie kapilarnej sekcja parowania jest wyraźnie oddzielona od sekcji skraplania. Dzięki rozdzieleniu kanału pary i schłodzonej cieczy wyeliminowane jest ich bezpośrednie oddziaływanie, co jest jedną z podstawowych wad rury cieplnej. Drugim elementem odróżniającym pompę kapilarną od rury cieplnej jest materiał porowaty. W

pompie nie występuje on na całej długości, lecz tylko w miejscu parowania. Dzięki temu występujące opory ruchu są niewielkie, przy zachowaniu ciśnienia kapilarnego niezbędnego do pracy układu. Dzięki swoim własnościom pompa kapilarna może być stosowana do transportu cieczy na duże odległości (nawet do kilku metrów). Rozdzielenie mechaniczne parownika i skraplacza daje nam możliwość użycia elastycznych przewodów łączących te elementy.

Istnieją dwa podstawowe typy pomp kapilarnych: CPL i LHP [6]. Pompa typu CPL (rys. 2a) posiada oddzielny rezerwuar, w którym istnieje możliwość regulacji ciśnienia, a tym samym regulacja punktu pracy układu. Ciśnienie można zmieniać i dostosowywać zarówno do mocy rozpraszanej w układzie elektronicznym jak i zmiennych warunków odprowadzania ciepła do otoczenia. Parownik zawiera jedną warstwę materiału porowatego. Wadą pompy kapilarnej typu CPL jest konieczność regulacji ciśnienia w rezerwuarze zarówno w trakcie rozruchu jak i podczas pracy. Z drugiej strony daje to możliwość pracy układu chłodzenia dla różnej temperatury elementu chłodzonego.

Pompa typu LHP posiada rezerwuar zintegrowany z parownikiem, oraz dwie warstwy materiału porowatego o różnych średnicach porów (rys. 2b). Materiał stykający się z cieczą ma większą średnicę porów i służy do wstępnego zasysania cieczy. Właściwe ciśnienie kapilarne niezbędne do pracy pompy uzyskuje się dzięki zastosowaniu drugiego materiału porowatego o mniejszej średnicy porów. Podwójny materiał porowaty zapewnia wyparcie pęcherzyków pary i powietrza z parownika do rezerwuaru w trakcie rozruchu. Ponadto w pompie LHP istnieje termiczne sprzężenie między parownikiem i rezerwuarem, co pozwala uzyskać autoregulację punktu pracy pompy w trakcie jej działania przy zmiennych warunkach generacji mocy w układzie i oddawania ciepła do otoczenia. Pompa kapilarna LHP nie wymaga zadawania warunków początkowych (ciśnienia) w celu jej rozruchu, ale jest budowana dla jednego punktu pracy (założona moc odprowadzana z układu przy założonej temperaturze).



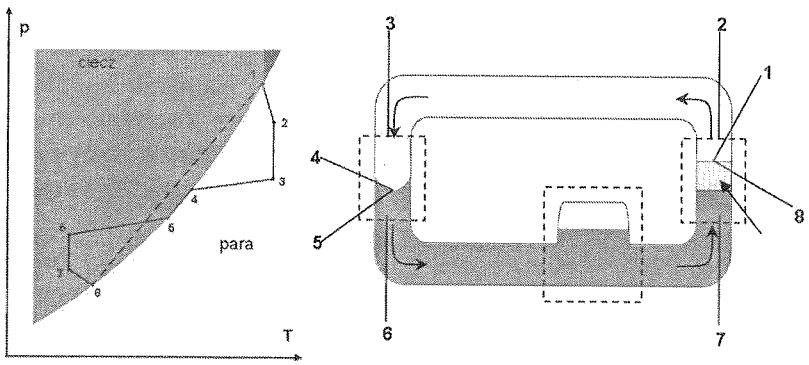
Rys. 2. Pompa kapilarna typu: a) CPL i b) LHP

Fig. 2. Capillary Pump Loop and Loop Heat Pipe

2. ANALIZA TRANSPORTU CIEPŁA I MASY W POMPACH KAPILARNYCH W STANIE USTALONYM

W pompie kapilarnej zachodzi jednoczesny przepływ ciepła i masy. Podstawową rolę odgrywają trzy zjawiska: parowanie cieczy z powierzchni materiału porowatego, skraplanie pary w skraplaczu pod wpływem ochłodzenia płynu, oraz wytwarzanie ciśnienia kapilarnego przez materiał porowaty.

Pracę pompy kapilarnej najłatwiej opisać korzystając z wykresu fazowego dla płynu wewnątrz pompy. Na rys. 3 przedstawiony jest wykres fazowy oraz schemat pompy kapilarnej z zaznaczonymi odpowiadającymi sobie punktami. Analizując przepływ ciepła i masy w poszczególnych elementach pompy, oraz interakcje między nimi można wyznaczyć warunki pracy pompy.



1, 8 – powierzchnia materiału porowatego; 2 – wylot pary z parownika; 3 – wlot do skraplacza; 4, 5 – powierzchnia skraplania pary; 6 – wylot wody ze skraplacza; 7 – wlot wody do parownika.

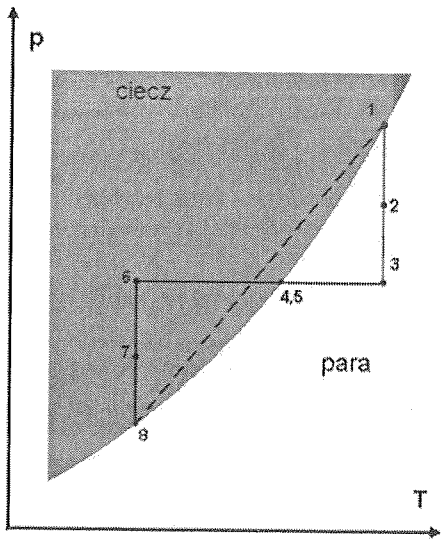
Rys. 3. Wykres fazowy dla płynu w pompie kapilarnej

Fig. 3. Phase diagram for the fluid in the capillary pump

Całkowity spadek cionienia medium chłodzącego w pompie (od punktu 1 do 8) jest sumą spadków cionienia w poszczególnych elementach. Począwszy od punktu 1 jest to spadek cionienia pary podczas przepływu przez kanały odprowadzające, następnie spadek cionienia pary podczas przepływu od wylotu pompy do wlotu skraplacza, spadek cionienia pary do momentu skraplania medium, różnica cionien na odcinku, na którym wystepuje skraplanie medium chłodzącego, spadek cionienia w drodze do wylotu skraplacza, różnica cionienia pomiedzy wylotem skraplacza a wejociem do parownika, oraz spadek cionienia przy przepływie cieczy przez materiał porowaty.

Przyjmując kilka założeń można stworzyć uproszczony model przepływu medium chłodzącego. Po pierwsze można założyć, że temperatura pary wychodzącej z parownika jest bliska temperatury panującej na powierzchni parowania ($T_2 = T_1$). Następnie

można założyć pomijalną zmianę temperatury przy przepływie medium przez linię pary i wody ($T_2 = T_3, T_6 = T_7 = T_8$). Przyjmijmy także, że spadek ciśnienia w skraplaczu jest mały w porównaniu ze spadkiem ciśnienia w długiej linii pary i cieczy ($p_3 = p_4 = p_5 = p_6$), oraz że skraplanie zachodzi na bardzo małym odcinku ($T_4 = T_5$). Upraszcza to wykres fazowy do przedstawionego na rys. 4.



Rys. 4. Uproszczony wykres fazowy dla płynu w pompie kapilarnej

Fig. 4. Simplified phase diagram for the fluid in the capillary pump

Bazując na tym wykresie fazowym można wyznaczyć spadek ciśnienia przepływającego medium w pompie kapilarnej (Δp_{tot}).

Spadek ciśnienia podczas parowania (Δp_{12}) można wyznaczyć z równania Hertza-Knudsena [7], opisującego szybkość ubytku masy cieczy podczas parowania z jednostki powierzchni w jednostce czasu:

$$\frac{\dot{m}}{A_{par}} = \varepsilon \sqrt{\frac{M}{2\pi R}} \left(\frac{RH \cdot p_o(T_o)}{\sqrt{T_o}} - \frac{p_o(T_{par})}{\sqrt{T_{par}}} \right). \tag{1}$$

gdzie: $\dot{m}$ — strumień masy w jednostce czasu, A_{par} — powierzchnia parowania, RH – wilgotność względna, T_o — temperatura otoczenia, T_{par} — temperatura powierzchni parowania, M — masa molowa cieczy, p_o — ciśnienie pary nasyconej, R – uniwersalna stała gazowa, ε — współczynnik parowania. Współczynnik parowania ε jest wyznaczany doświadczalnie dla danej geometrii układu.

Przekształcając (1) i korzystając z założonych uproszczeń można wyznaczyć spadek ciśnienia przy parowaniu cieczy ($p_2 = RH \cdot p_o$):

$$\Delta p_{12} = \frac{1}{\varepsilon}\sqrt{\frac{2\pi RT_1}{M}}\frac{\dot{m}}{A_{par}}. \tag{2}$$

Spadki ciśnienia w linii pary i cieczy (włączając w linię pary część wlotową skraplacza, oraz w linię wody cześć wlotową parownika i wylotową skraplacza) przyjmują postać:

$$\Delta p_{23} = \dot{m}\frac{8\nu_{par}L_{par}}{\pi R_{par}^4}, \Delta p_{67} = \dot{m}\frac{8\nu_{ciecz}L_{ciecz}}{\pi R_{ciecz}^4} \tag{3}$$

gdzie: ν_{par}, ν_{ciecz} — lepkość kinematyczna pary i cieczy; L_{par}, L_{ciecz} — długość linii pary i cieczy; R_{par}, R_{ciecz} — promień linii pary i cieczy.

Zostaje jeszcze policzyć spadek ciśnienia przy przepływie cieczy przez materiał porowaty. Podstawową wielkością opisującą materiał porowaty jest porowatość κ. Jest to stosunek objętości przestrzeni porów (kanałów kapilarnych) do objętości całego materiału porowatego. Różnica $1-\kappa$ odpowiada objętości stałej części matrycy. Definiuje się także przepuszczalność materiału porowatego $K = d_p^2\kappa^3/150(1-\kappa)^2$, gdzie d_p jest średnią średnicą ziaren materiału porowatego. Spadek ciśnienia przy przepływie cieczy przez materiał porowaty jest opisany przez prawo Darcy'ego:

$$\Delta p_{78} = \frac{\dot{m}\nu_{ciecz}L}{A_{por}K}, \tag{4}$$

gdzie: ν_{ciecz} — lepkość kinematyczna cieczy, L — grubość materiału porowatego.

Niech P będzie mocą odprowadzana przez pompę z powierzchni układu chłodzonego, a T_s żądaną temperaturą na powierzchni układu chłodzonego. Temperatura parowania (powierzchni materiału porowatego) wynosi

$$T_1 = T_s - \frac{Pd}{A_{par}\lambda}, \tag{5}$$

gdzie: A_{par} — powierzchnia parowania, d — odległość pomiędzy powierzchnią odbierania ciepła a powierzchnią materiału porowatego (grubość elementu pompy bezpośrednio przylegającego do elementu chłodzonego), λ — przewodność cieplna elementu pompy przenoszącego ciepło od układu chłodzonego do materiału porowatego.

Znając T_1 i P można w przybliżeniu (pomijając energię kinetyczną) wyznaczyć strumień masy medium chłodzącego, który jest odparowywany w parowniku:

$$\dot{m} = P/C_p, \tag{6}$$

gdzie C_p — ciepło parowania. Sumując ze sobą cztery spadki ciśnienia (2, 3, 4) i podstawiając (6) otrzymujemy wzór na całkowity spadek ciśnienia medium chłodzącego w pompie:

$$\Delta p_{tot} = \frac{P}{C_p}\left[\frac{1}{\varepsilon A_{par}}\sqrt{\frac{2\pi RT_1}{M}} + \frac{8}{\pi}\left(\frac{\nu_{par}L_{par}}{R_{par}{}^4} + \frac{\nu_{ciecz}L_{ciecz}}{R_{ciecz}{}^4}\right) + \frac{\nu_{ciecz}L_{por}}{A_{par}K}\right]. \tag{7}$$

Warunkiem działania pompy kapilarnej jest, co najmniej równoważenie tego spadku ciśnienia przez ciśnienie kapilarne materiału porowatego $\Delta p_{kap} \geq \Delta p_{tot}$. Ciśnienie kapilarne dla kapilar o przekroju okrągłym wyraża się przez:

$$\Delta p_{kap} = \frac{2\sigma \cos\theta}{r},$$

gdzie σ — napięcie powierzchniowe, θ — kąt zetknięcia cieczy ze ściankami kapilar, r – promień kapilar. Stąd mamy:

$$\frac{2\sigma \cos\theta}{r} \geqslant \frac{P}{C_p}\left[\frac{1}{\varepsilon A_{par}}\sqrt{\frac{2\pi R T_1}{M}} + \frac{8}{\pi}\left(\frac{\eta_{par} L_{par}}{R_{par}{}^4 \rho_{par}} + \frac{\eta_{ciecz} L_{ciecz}}{R_{ciecz}{}^4 \rho_{ciecz}}\right) + \frac{\eta_{ciecz} L_{por}}{A_{par} K \rho_{ciecz}}\right]. \quad (8)$$

W ostatniej nierówności występują trzy niewiadome: promień kapilar w materiale porowatym r, przepuszczalność materiału porowatego K oraz współczynnik parowania ε. Parametry materiału porowatego należy dobrać tak, aby była spełniona nierówność (8). Wartość współczynnika parowania musi być wyznaczona doświadczalnie dla konkretnych geometrii parownika.

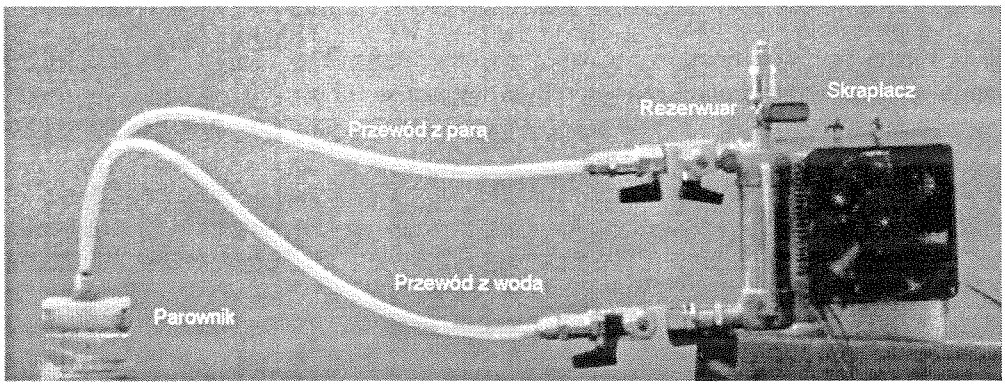
Znając wszystkie parametry pompy, włącznie z własnościami materiału porowatego i współczynnikiem parowania, można szukać zależności pomiędzy temperaturą na powierzchni układu chłodzonego i na wylocie skraplacza. Można, zakładając temperaturę skraplacza, szukać zależności pomiędzy temperaturą powierzchni układu a mocą wydzielaną w układzie chłodzonym.

3. PROTOTYP POMPY KAPILARNEJ TYPU CPL — BADANIA W STANACH STATYCZNYCH

3.1. BUDOWA POMPY

Na rys. 5 pokazano zdjęcie oraz strukturę prototypowej pompy kapilarnej typu CPL wykonanej w Instytucie Elektroniki Politechniki Łódzkiej. Po lewej stronie widać parownik umieszczony na elemencie grzejnym. Po prawej stronie skraplacz z radiatorami i wiatrakami oraz rezerwuar. Parownik z rezerwuarem połączony jest dwoma przewodami. Jeden odprowadzający parę z parownika do skraplacza, drugi wodę w kierunku przeciwnym. Na przewodach założone są kulowe zawory odcinające, niezbędne do napełniania pompy wodą oraz wytworzenia odpowiedniego podciśnienia.

Parownik ma wymiary 50mm×50mm×15mm. Można w nim wyróżnić trzy podstawowe części. Pierwszym elementem jest płytka mosiężna, która bezpośrednio styka się z elementem chłodzonym i odbiera od niego ciepło. Na płytce tej umieszczony jest materiał porowaty zasysający wodę ze zbiornika umieszczonego nad materiałem. Całość zamknięta jest elementem z tworzywa, z którego wychodzą dwa przewody: doprowadzające wodę i odprowadzające parę. W płytce mosiężnej stykającej się z



Rys. 5. Prototyp pompy typu CPL

Fig. 5. Capillary Pumped Loop prototype

materiałem porowatym wyfrezowana jest siatka rowków odprowadzających parę powstającą podczas pracy pompy na powierzchni materiału porowatego. Para przepływa do przewodu odprowadzającego ją do skraplacza.

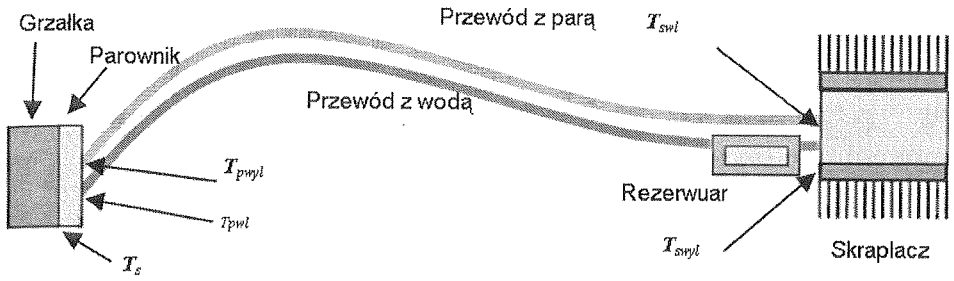
Materiałem porowatym jest płytka ze spieku szklanego o porowatości 0,7273 i przepuszczalności z zakresu: $3,44966 \cdot 10^{-10} - 1,37987 \cdot 10^{-9}$ m^2. Skraplacz ma wymiary 100mm×100mm×20mm. Jest to blok mosiężny z wydrążonym kanałem o średnicy 4mm. Miedziany rezerwuar ma 90mm wysokości i 30mm średnicy. W górnej części znajduje się wlot rozruchowy, zamknięty zaworem kulowym, niezbędny do napełnienia i ustalenia warunków początkowych w pompie. Z boku rezerwuaru zamontowano wskaźnik poziomu cieczy wykonany z przeźroczystego przewodu z tworzywa sztucznego. Skraplacz z parownikiem połączone są przeźroczystymi przewodami z tworzywa o średnicy wewnętrznej 4mm.

3.2. WYNIKI POMIARÓW POMPY W STANACH STATYCZNYCH

Do pomiarów wybrano pięć punktów pomiarowych (rys. 6): miejsce styku powierzchni parownika z układem chłodzonym (T_s), wlot wody (T_{pwl}) i wylot pary (T_{pwyl}) parownika, oraz wlot pary (T_{swl}) i wylot wody (T_{swyl}) skraplacza.

Wykonano dwie serie pomiarów. W pierwszej serii stosowano konwekcję naturalną (niewymuszoną) radiatorów skraplacza. W drugiej chłodzenie odbywało się drogą konwekcji wymuszonej przy użyciu wentylatorów. Wyniki pomiarów obu serii są przedstawione i omówione poniżej.

W pierwszej serii pomiarów moc grzałki regulowano w zakresie 75-160W. Dokonano pomiarów temperatury w punktach pomiarowych oraz obserwacji zachowania się cieczy i pary wewnątrz pompy kapilarnej. Wyniki pomiarów temperatury zostały przedstawione na rys. 7.



Rys. 6. Rozmieszczenie punktów pomiaru temperatury

Fig. 6. Temperature test points location

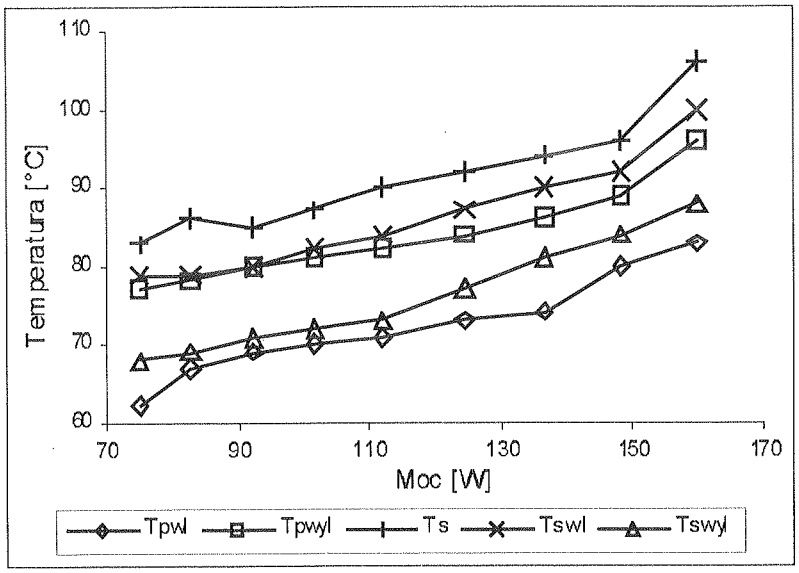
W założeniach pompa miała utrzymywać temperaturę powierzchni układu chłodzonego na poziomie poniżej 100°C. Pompa spełnia założenia do mocy nieco ponad 150W. Powyżej tej granicy temperatura powierzchni gwałtownie wzrasta i pompa kapilarna przestaje działać. Odpowiada to osiągnięciu na powierzchni materiału porowatego temperatury wrzenia wody. Powoduje to powstawanie dużej ilości pary i zaburzenie cyrkulacji medium chłodzącego w pompie. Podczas obserwacji można było zauważyć gwałtowne pojawienie się pary w przewodzie doprowadzającym wodę do parownika, zatrzymanie przepływu wody i skok temperatury powierzchni układu.

Najlepsza cyrkulacja medium zachodzi dla mocy z zakresu 110-150W. Para z dużą szybkością wydostaje się z parownika do przewodu i wpływa do skraplacza. Po ochłodzeniu o ok. 10°C (rys. 8) wypływa ze skraplacza i poprzez rezerwuar płynie do parownika.

W drugiej serii pomiarów dodano chłodzenie konwekcyjne wymuszone poprzez zastosowanie zamontowanych na radiatorach skraplacza. Moc grzałki regulowano w zakresie 40-290W. Tak jak w poprzednim przypadku, dokonano pomiarów temperatury w punktach pomiarowych oraz obserwacji zachowania się cieczy i pary wewnątrz pompy kapilarnej. Wyniki pomiarów temperatury zostały przedstawione na rysunku 9.

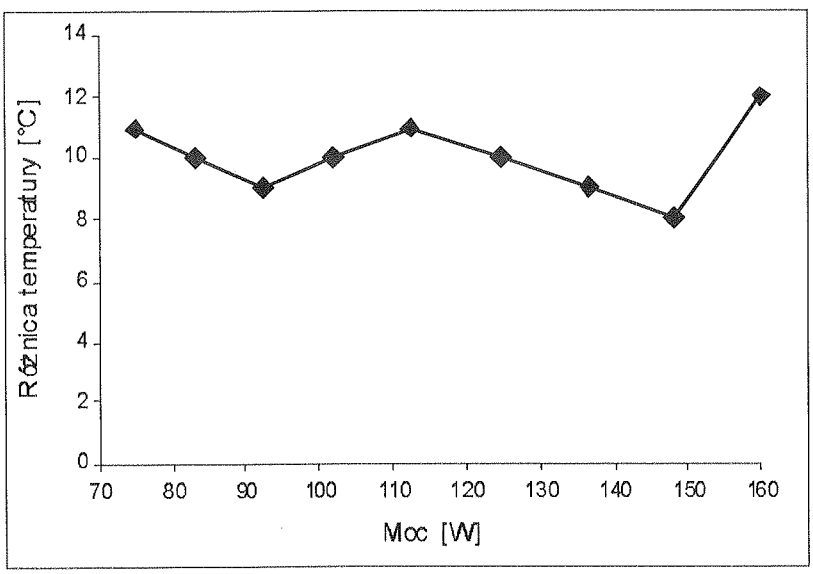
Przy włączonym chłodzeniu z konwekcją wymuszoną pompa utrzymuje temperaturę powierzchni układu chłodzonego na poziomie poniżej 100°C dla mocy do ok. 230W. Powyżej tej mocy temperatura powierzchni jest wyższa od założonego poziomu. Dla mocy 230W osiągana jest na powierzchni materiału porowatego temperatura wrzenia wody. Tak jak poprzednio, powoduje to powstawanie dużej ilości pary i zaburzenie cyrkulacji medium chłodzącego w pompie. Proces nie ma tak gwałtownego przebiegu jak dla pompy z chłodzeniem za pomocą konwekcji naturalnej. Najlepsza cyrkulacja medium zachodzi dla mocy z zakresu 130-230W. Para z dużą szybkością wydostaje się z parownika do przewodu i wpływa do skraplacza. Po ochłodzeniu o ok. 40°C (rys. 10) woda wypływa ze skraplacza i poprzez rezerwuar płynie do parownika.

Na rys. 11 pokazano porównanie wyników pomiarów temperatury w punktach kontrolnych dla układu z chłodzeniem konwekcyjnym wymuszonym i niewymuszonym



Rys. 7. Zależność temperatury w punktach pomiarowych od mocy dla chłodzenia z konwekcją naturalną

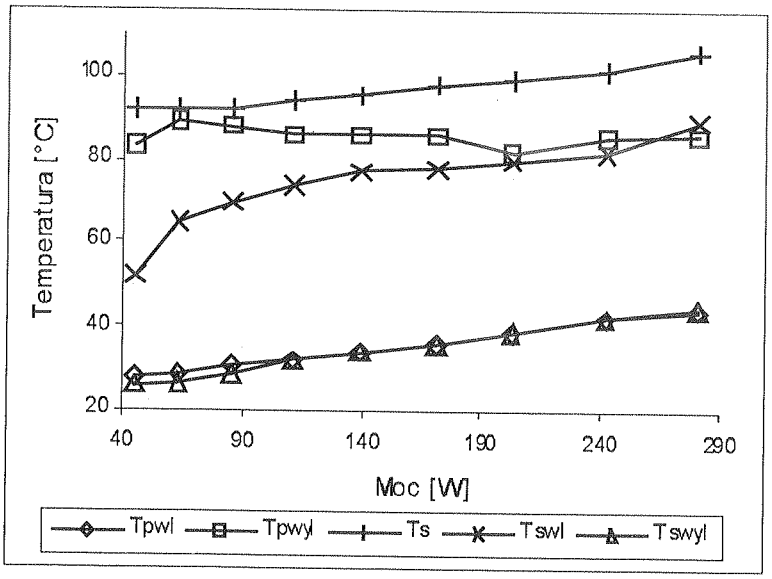
Fig. 7. Test points temperature as the power function for natural convection cooling



Rys. 8. Zależność różnicy temperatury na wylocie i wlocie skraplacza w funkcji mocy

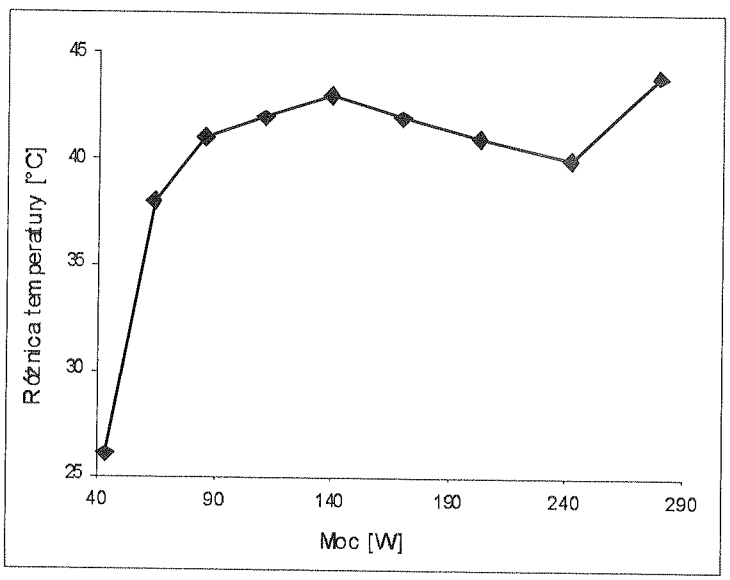
Fig. 8. Temperature difference between condenser outlet and inlet as the power function for natural convection cooling

Fig.



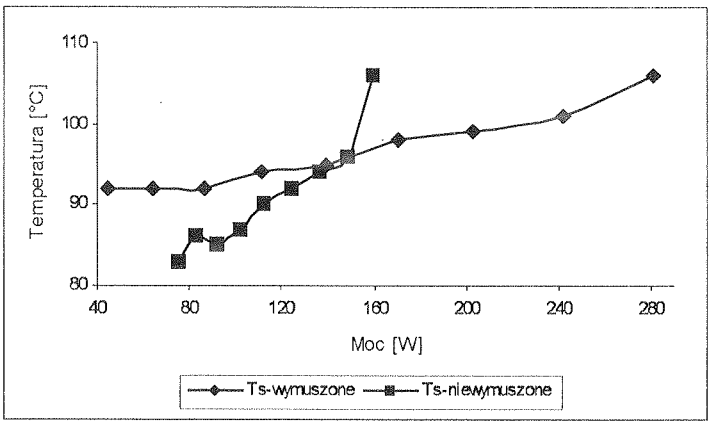
Rys. 9. Zależność temperatury w punktach pomiarowych od mocy dla chłodzenia z konwekcją wymuszoną

Fig. 9. Test points temperature as the power function for forced convection cooling



Rys. 10. Zależność różnicy temperatury na wylocie i wlocie skraplacza w funkcji mocy

Fig. 10. Temperature difference between condenser outlet and inlet as the power function for forced convection cooling



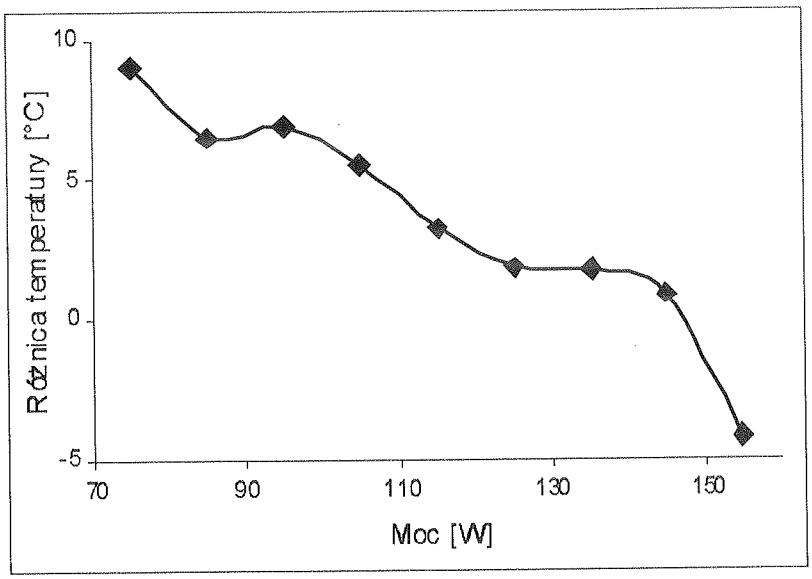
Rys. 11. Porównanie wyników dla pompy z chłodzeniem konwekcyjnym wymuszonym i niewymuszonym

Fig. 11. Results comparison for natural and forced convection cooling

(powierzchnia układu chłodzonego). Widać, że układ z chłodzeniem konwekcyjnym wymuszonym pracuje poprawnie w szerszym zakresie mocy wydzielanej w układzie chłodzonym. Utrzymuje on temperaturę powierzchni układu na założonym poziomie przy znacznie większych mocach niż układ z chłodzeniem konwekcyjnym niewymuszonym.

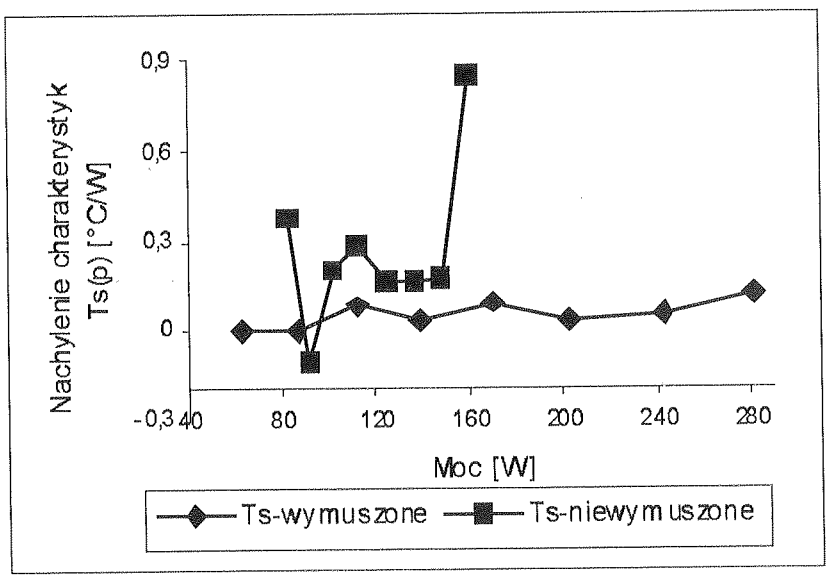
Układ z chłodzeniem konwekcyjnym niewymuszonym utrzymuje żądaną temperaturę dla mniejszego zakresu mocy. Jednak w porównaniu z chłodzeniem wymuszonym dla mocy do ok. 150W utrzymuje temperaturę niższą. Różnica ta maleje ze wzrostem mocy. Na rys. 12 pokazano różnicę temperatury na powierzchni dla układu z chłodzeniem konwekcyjnym wymuszonym i niewymuszonym. Różnica ta została wyznaczona z pomierzonych zależności temperatury od mocy, aproksymowanych wielomianami szóstego stopnia.

Temperatura układu z chłodzeniem konwekcyjnym wymuszonym wykazuje się mniejszą zależnością zmian w funkcji mocy. Na rys. 13 pokazano nachylenie charakterystyki $T_s(p)$ w funkcji mocy wydzielanej w układzie, rozumianej jako stosunek zmian temperatury powierzchni układu chłodzonego do zmiany mocy wydzielanej w układzie. Widać, że dla układu z chłodzeniem konwekcyjnym wymuszonym wzrost temperatury w stosunku do wzrosty mocy utrzymuje się dla całego zakresu prawie na jednym poziomie 0-0,2° C/W. Dla układu z chłodzeniem konwekcyjnym niewymuszonym stromość charakterystyki waha się znacznie w badanym zakresie mocy (-0,2-0,8° C/W). Dla zakresu 110- 150W można zauważyć małą zmienność nachylenia charakterystyki. Jest to zakres, w którym w pompie obserwowano stabilną cyrkulację czynnika chłodzącego. Przy chłodzeniu konwekcją wymuszoną stabilna cyrkulacja miała miejsce w szerszym zakresie mocy, co uwidacznia się w lepszej stabilizacji temperatury układu.



Rys. 12. Różnica temperatury powierzchni układu dla pompy z chłodzeniem konwekcyjnym wymuszonym i niewymuszonym

Fig. 12. Device surface temperature difference for the pump with forced and natural convective cooling



Rys. 13. Nachylenie charakterystyk $T_s(p)$ w funkcji mocy

Fig. 13. $T_s(p)$ characteristics slope as the power function

3.3. WARUNEK DZIAŁANIA DLA PROTOTYPU POMPY

Dla przedstawionego prototypu pompy mamy: zakładana moc odbierana przez pompę 150W, medium – woda, długości linii cieczy i wody 0,5m, średnica przewodów 0,004m, powierzchnia materiału porowatego $1,256 \cdot 10^{-3}$ m^2 (promień 0,02m), grubość materiału porowatego 4mm, temperatura powierzchni materiału porowatego 95°C, temperatura cieczy wypływającej ze skraplacza 40°C. Podstawiając te wartości do nierówności (8) otrzymujemy warunek

$$1 \geqslant r\left[\frac{1561,6}{\varepsilon} + 765,79 + \frac{6,164 \cdot 10^{-10}}{K}\right].$$

Do budowy pompy kapilarnej został użyty materiał porowaty o standardo- wych parametrach spieków szklanych wykorzystywanych w chemii analitycznej do filtracji (np. w lejkach Shotta). Dla spieku o standardowej porowatości G2 pod- stawowe parametry podane są poniżej:

Porowatość κ	0,727
Przepuszczalność K	$3,45 \cdot 10^{-10} - 1,38 \cdot 10^{-9}$ m^2
Średnica porów 2r	$4 \cdot 10^{-5} - 1 \cdot 10^{-4}$ m
Średnica kulek d_p	$1 \cdot 10^{-4} - 2 \cdot 10^{-4}$ m

Po policzeniu warunku (8) dla krańcowych wartości parametrów materiału porowatego G2 otrzymujemy warunki na wartość współczynnika parowania: $\varepsilon \geq 0,169$ dla minimalnej przepuszczalności $K = 3,45 \cdot 10^{-10}$m^2 i $\varepsilon \geq 0,0812$ dla maksymalnej przepuszczalności materiału porowatego $K = 1,38 \cdot 10^{-9}$m^2. Wyliczona wartość współczynnika musi być spełniona przez parownik. Jeśli konstrukcja parownika będzie miała odpowiednie parametry to przy założeniach przyjętych przy konstrukcji pompy, spełniony będzie warunek (8).

4. PODSUMOWANIE

W ramach badań opisanych w artykule stworzony został uproszczony model statyczny dla medium chłodzącego w pompie kapilarnej. Zaprojektowana została i wykonana pompa kapilarna typu CPL z pojedynczym materiałem porowatym i zbiornikiem wyrównawczym do regulacji ciśnienia pary i punktu pracy układu. Realizowana pompa kapilarna zapewnia odprowadzanie mocy na założonym poziomie 150W. Przy

powierzchni $2,5 \cdot 10^{-3}\text{m}^2$ daje to strumień ciepła odbieranego z elementu chłodzonego równy 60 kW/m^2. Dla porównania, płaska pompa kapilarna typu CPL firmy Astrium [6] może odprowadzać moce z zakresu od ok. 0,5 W do 100 W. Co przy powierzchni równej $0,9 \cdot 10^{-3}$ m^2 daje strumień ciepła ok. 0,55-110 kW/m^2.

Pompa kapilarna typu CPL jest pompą wymagającą sterowania. Podczas rozruchu i zmiany warunków pracy należy kontrolować i sterować warunkami panującymi wewnątrz pompy. Zazwyczaj kontrola odbywa się poprzez regulację temperatury rezerwuaru, co przekłada się na regulację ciśnienia panującego wewnątrz zbiornika. Zmiana ciśnienia powoduje zmianę punktu pracy pompy. Ciśnienie w rezerwuarze można także regulować bezpośrednio. W ramach dalszych badań przewiduje się badanie procesów dynamicznych zachodzących w pompie kapilarnej.

5. BIBLIOGRAFIA

1. B. Więcek: *Odprowadzanie ciepła w układach elektronicznych ze szczególnym uwzględnieniem konwekcji naturalnej i promieniowania.* Zeszyty Naukowe Politechniki Łódzkiej, Nr 820, 1999.
2. J. Legierski, B. Więcek: *Heat pipes cooling fins for VLSI integrated circuits.* XXII KKTOiUE Warszawa, 1999.
3. S. D. Garner: *Heat pipes for electronics cooling applications.* Thermacore Inc, 1999.
4. B. Więcek: *Heat Pipe Cooling Fins for Integrated Circuits.* Workshop: The Seminar Thermic'98, Microtherm'98, Zakopane, 1998, pp. 90-92.
5. Pin-Chin Chen, Wei-Keng Lin: *The application of capillary pumped loop for cooling of electronic components.* Applied Thermal Engineering 21, 2001, pp. 1739-1754.
6. C. Figus, L. Ounougha, P. Bonzom, W. Supper, C. Puillet: *Capillary fluid loop developments in Astrium.* Applied Thermal Engineering 23, 2003, pp. 1085-1098.
7. I. W. Eames, N. J. Marr and H. Sabir: *The evaporation coefficient of water: a review.* Int. J. Heat Mass Transfer, Vol. 40. No. 12, 1997, pp. 2963-2973.

T. WAJMAN, B. WIĘCEK, B. OSTROWSKI, R. DANYCH

CAPILLARY PUMP FOR COOLING OF ELECTRONIC DEVICES — STEADY STATES SUMMAMRY

Summary

Capillary pumps (Capillary Pumped Loops and Loop Heat Pipes) are passive, two-phase heat transport systems that utilize the capillary pressure developed in a fine pore evaporator wick to circulate the system's working fluid. When heat is applied to the evaporator, the liquid evaporates from the saturated wick and flows through the vapor line to the condenser. In the condenser heat is removed and the vapor condenses. The cooled liquid returns to the evaporator through the liquid line.

The essence of the project is to apply cooling that uses capillary pump as an element taking heat away from power elements. The capillary pump is capable of taking away up to a few hundred watts of power from a square centimeter and transporting it on long distances, even up to a few meters (e.g. beyond the converter system). It gives us the possibility of taking away considerable power and separating

the heat source from the place of giving up the heat to the environment, which solves the problems of standard cooling systems.

Analysis of heat and mass flow in CPL capillary pump based on phase diagram was presented. Some simplifications have been assumed and steady state pump working condition has been derived. Using this condition, a capillary pump prototype has been building and testing. The pump has been projected for

Two series of measurements with natural and forced convection cooling of condenser have been executed. Results of steady state experiments have been presented.

Keywords: capillary pump, electronic devices cooling

KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJI, 2005, 51, z. 2, ss. 335–359

CONTINUOUS-TIME ACTIVE-RC FILTER MODEL FOR COMPUTER-AIDED DESIGN AND OPTIMIZATION

SŁAWOMIR KOZIEŁ

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

Otrzymano 2004.03.02
Autoryzowano 2004.04.20

In the paper, a general model of integrator-based continuous-time Active-RC filters is presented. More specifically, filter structures containing inverting amplifiers and passive feedback network are considered. The structure is analyzed using matrix description. The explicit transfer function formula is given in general case. The extensions of the model that take into account finite DC gain and finite bandwidth of operational amplifiers as well as other non-ideal effects are presented. We also perform a noise analysis of the Active-RC filters in general setting. The matrix-based approach is formulated especially for efficient use in computer-aided analysis and design of Active-RC filters. For the sake of illustration, a few examples of application of the proposed approach are given.

Keywords: Active-RC filters, continuous-time filters, noise analysis

1. INTRODUCTION

Continuous-time analog filters based on operational amplifiers (OPAMPs) and RC elements (Active-RC filters) provide solutions for various signal-processing tasks. Many synthesis and design methods for different types and architectures of this class of filters have been reported [1],[2]. Although other techniques, particularly transconductance-capacitor (OTA-C) filters [3], are now dominant in high-frequency range, Active-RC filters are suitable for numerous medium-frequency applications, especially those which require high linearity and low noise, such as ADSL [4], VDSL [5], [6] WCDMA [7],[8], RFIC receivers for PANs [9], GSM baseband transmitters [10]. Recently, Active-RC filter for frequencies beyond 300MHz has been successfully implemented [11].

In this paper, we consider a general matrix-based approach suitable for computer-aided analysis, design and optimization of continuous-time Active-RC filters. In Section 2, we present a general Active-RC filter topology and develop a matrix description of this structure. The model is not the most general one, because it is restricted to filter topologies that contain only inverting amplifiers (especially integrators) and RC feedback network. Thus, it is not suitable for analyzing classical structures such as Sallen-Key biquad [3]. However, it covers a number of practical Active-RC filters, especially most multiple-loop feedback topologies including the ones based on RLC ladder simulation, leap-frog, follow-the-leader [1], [2], [12] and many others. It also makes it possible to generate and investigate properties of new filter topologies with possibly attractive properties. The general model that covers all conceivable Active-RC topologies will be given elsewhere. In Section 3 we consider extensions of the presented filter model that take into consideration non-ideality of OPAMPs, such as finite gain-bandwidth product (GBW) and non-zero output resistance. In Section 4 we present a noise analysis of Active-RC filters in general setting. Examples of application of the presented approach are given in Section 5. Section 6 contains remarks and conclusions.

2. GENERAL STRUCTURE OF INTEGRATOR-BASED ACTIVE-RC FILTER

Fig.1 shows the single-ended version of the general topology of integrator-based Active-RC filter. Note that we consider filter topologies containing somewhat more general building blocks than lossy or lossless integrators, i.e. inverting amplifiers that may contain nonzero capacitance at the input. The filter in Fig. 1 contains n operational amplifiers denoted as OA_i, $i = 1, \ldots, n$, n input resistors R_{bi} and capacitors C_{bi}, $i = 1, \ldots, n$, the set of internal feedback resistors R_{ij} and capacitors C_{ij}, $i, j = 1, \ldots, n$, as well as output summer consisting of amplifier OA_o, resistors R_d (direct feedforward path from input), R_{ci}, $i = 1, \ldots, n$ and R_o. Output nodes of OPAMPs are denoted as x_i, $i = 1, \ldots, n$. We shall also use z_i and x_i to denote voltages at OPAMP input and output nodes, respectively, which should not lead to confusion. It is seen that any integrator-based Active-RC filter is a particular case of the structure in Fig. 1, i.e. any particular filter can be obtained from the one in Fig. 1 by removing appropriate number of elements. In this section we shall assume that operational amplifiers are ideal (we relax this assumption in Section 3, where filters with non-ideal OPAMPs are considered). The analysis below is carried out in the Laplace transform domain.

Let us start the analysis of the circuit in Fig. 1. The total current flowing into each node z_i, $i = 1, \ldots, n$ is zero and the voltage at z_i is also zero, due to the assumed ideality of OPAMP. This allows us to write the following equations:

$$y_{bi}U_{in} + \sum_{j=1}^{n} y_{ij}x_j = 0, \quad i = 1, \ldots, n \tag{1}$$

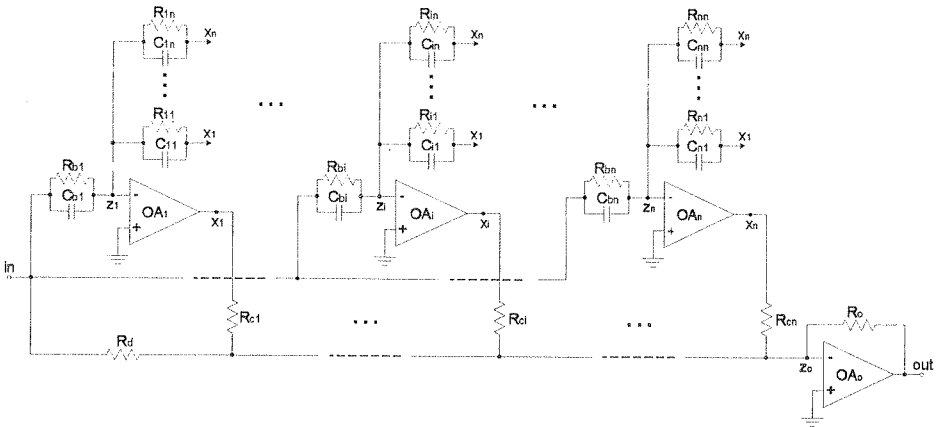


Fig. 1. General structure of integrator-based Active-RC filter

Rys. 1. Ogólna struktura filtru aktywnego RC ze wzmacniaczami odwracającymi

$$g_o U_{out} + \sum_{j=1}^{n} g_{cj} x_j + g_d U_{in} = 0 \tag{2}$$

where U_{in} and U_{out} denote Laplace transforms of input and output voltages, respectively, and

$$\begin{gathered} y_{bi} = g_{bi} + sC_{bi},\ g_{bi} = 1/R_{bi},\ i = 1, \ldots, n,\ y_{ij} = g_{ij} + sC_{ij}, \\ g_{ij} = 1/R_{ij},\ i, j = 1, \ldots, n \end{gathered} \tag{3}$$

$$g_{ci} = 1/R_{ci},\ i = 1, \ldots, n,\ g_d = 1/R_d,\ g_o = 1/R_o \tag{4}$$

Let us introduce the following notation

$$\begin{gathered} \boldsymbol{T} = \begin{bmatrix} C_{11} & \cdots & C_{1n} \\ \vdots & \ddots & \vdots \\ C_{n1} & \cdots & C_{nn} \end{bmatrix},\ \boldsymbol{G} = \begin{bmatrix} g_{11} & \cdots & g_{1n} \\ \vdots & \ddots & \vdots \\ g_{n1} & \cdots & g_{nn} \end{bmatrix},\ \boldsymbol{B} = \begin{bmatrix} g_{b1} + sC_{b1} \\ \vdots \\ g_{bn} + sC_{bn} \end{bmatrix},\ \boldsymbol{X} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \\ \boldsymbol{C} = \begin{bmatrix} c_1 & \cdots & c_n \end{bmatrix} = \begin{bmatrix} \dfrac{g_{c1}}{g_o} & \cdots & \dfrac{g_{cn}}{g_o} \end{bmatrix},\ \boldsymbol{D} = d = \frac{g_d}{g_o} \end{gathered} \tag{5}$$

Using (5) we can rewrite (1), (2) in the following form

$$\begin{gathered} (s\boldsymbol{T} + \boldsymbol{G})\,\boldsymbol{X} + \boldsymbol{B}U_{in} = 0 \\ U_{out} + \boldsymbol{C}\boldsymbol{X} + \boldsymbol{D}U_{in} = 0 \end{gathered} \tag{6}$$

This allows us to calculate the transfer function $H(s)$ of the filter

$$H(s) = \frac{U_{out}(s)}{U_{in}(s)} = \boldsymbol{C}(s\boldsymbol{T} + \boldsymbol{G})^{-1}\boldsymbol{B} - \boldsymbol{D} \tag{7}$$

Now, let us denote adjoint matrix of $sT + G$ as M where

$$M(s) = adj(sT_C + G) = \left[M_{ij}(s)\right]_{i,j=1}^{n} \tag{8}$$

This allows us to rewrite transfer function in the form:

$$H(s) = \frac{1}{det(sT + G)} \sum_{i,j=1}^{n} c_i(g_{bj} + sC_{bj})M_{ij}(s) - d \tag{9}$$

Note that in many filter structures, input signal is provided to only one internal node, say z_k (i.e. there is no input signal distribution), and there is no output summer (i.e. one of the internal nodes, say x_l, is the output of the filter), and no input capacitors. This means that $B = \begin{bmatrix} 0 & \cdots & 0 & g_{bk} & 0 & \cdots & 0 \end{bmatrix}^T$, $C = \begin{bmatrix} 0 & \cdots & 0 & -1 & 0 & \cdots & 0 \end{bmatrix}$ -1 at l-th position and $C_{bi} = 0$ for $i = 0, 1, \ldots, n$ (note that we have -1 because of the fact that if there is no output summer, there is no signal inversion due to OA_o either). In such case, expression (9) reduces to the form:

$$H(s) = -\frac{g_{bk}M_{lk}(s)}{\det(sT + G)} \tag{10}$$

Similar expression can be written for slightly more general case, with no input capacitors, input signal distribution (i.e. $B = [g_{b1} \cdots g_{bn}]$), and a trivial output summer ($C = [0 \cdots 0 \ -1 \ 0 \cdots 0]$, with -1 at l-th position). Then, we have

$$H(s) = -\frac{1}{\det(sT - G)} \sum_{i=1}^{n} g_{bi}M_{li}(s) \tag{11}$$

On the basis of the above expressions one can easily calculate the transfer function of any particular structure of integrator-based Active-RC filter. It is also worth noting that there is one-to-one correspondence between integrator-based Active-RC filters and matrices (5). Thus, using presented matrix description we can consider filter design, analysis and optimization in purely algebraic domain. The main advantage is that this can be easily handled by a suitable computer software.

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.

Based on the algebraic properties of the matrix T, we can distinguish an important subclass of integrator-based Active-RC filters, i.e. state-space filters. Suppose that the matrix T is invertible, i.e. T^{-1} exists. Then we can rewrite (6) as

$$\begin{aligned} sX &= -T^{-1}GX - T^{-1}BU_{in} \\ U_{out} &= -CX - DU_{in} \end{aligned} \tag{12}$$

Let us denote

$$A_s = -T^{-1}G,\ B_s = -\mathrm{T}^{-1}B,\ C_s = -C,\ D_s = -D \tag{13}$$

With this notation (12) takes the form

$$\begin{aligned} sX &= A_sX + B_sU_{in} \\ U_{out} &= C_sX + D_sU_{in} \end{aligned} \tag{14}$$

which is nothing else but the state-space description of the filter in Fig. 1 with node voltages x_i, $i = 1, \ldots, n$ being the state variables. Thus, the state-space Active-RC filter is the one for which the matrix T is invertible which is a necessary and sufficient condition for the existence of the state matrices. For state-space filters we can apply many useful matrix transformations which can be used, for example, to performs efficient parameter optimization.

It should be emphasized at this point that the algebraic description of the filter structure in Fig. 1 exhibits a lot of similarities to description of the general OTA-C filter topology introduced in [13]. It is beyond the scope of this paper to exploit these similarities, however, we would like to point out that the main equation for the OTA-C filter model is almost the same as equation (7). On the other hand, the matrix T for OTA-C filter model is always symmetric and positively semi-definite [13], which is not the case for the considered class of Active-RC filters and clearly indicates that the set of all OTA-C filter structures is isomorphic to the proper subset of the Active-RC filter topologies considered here (in the sense of existence a one-to-one correspondence).

In practice, Active-RC filters are mostly implemented in fully differential structures. Due to this we may assume that matrix entries in (5) can take both positive and negative values, which can be accomplished by cross-coupling of corresponding physical elements. More specifically, if the element, say R_{ij}, is cross-coupled (i.e. put between positive [negative] output of an amplifier and positive [negative] input of another one, see e.g. resistor R_1 in Fig. 2), this reflects in equation (1) so that the appropriate term has the form $g_{ij}(-x_j) = -g_{ij}x_j$, i.e. the original '-' from node voltage is moved into filter element, here g_{ij}. Obviously, the physical element remains positive. Negative value of the corresponding matrix entry is equivalent to cross-coupling. In case of single-ended implementation, negative elements are realized using inverters. The presented approach is primarily intended to be used as a basis for creating computer-aided design and optimization software. However, let us consider a simple example as an illustration how it can be used for hand filter design. Suppose that we want to synthesize an all-pole biquad filter, i.e. we intend to implement the following transfer function:

$$H(s) = \frac{H_o\omega_0^2}{s^2 + \frac{\omega_0}{Q}s + \omega_0^2} \tag{15}$$

If one wants to develop minimal implementation, two OPAMPs are needed, so we have $n = 2$ in formulas (1)-(11). Assume that our filter has no input signal distribution with input signal injected to the first OPAMP through conductance g_b, i.e. $B = [g_b\ 0]^T$,

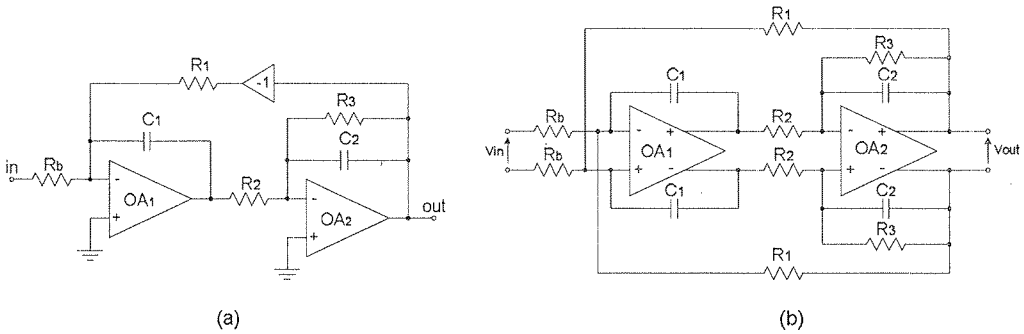


Fig. 2. All-pole Active-RC biquad corresponding to matrices (17): single-ended (a) and fully-differential version (b)

Rys. 2. Filtr aktywny RC odpowiadający macierzom (17): wersja z wyjściem niesymetrycznym (a) oraz wersja w pełni różnicowa (b)

and no output summer with output signal taken from the second OPAMP, i.e. $\boldsymbol{C} = [0 \ -1]^T$ and $\boldsymbol{D}$. Now we have to choose matrices $\boldsymbol{T}$ and $\boldsymbol{G}$ having in mind that in our case the transfer function of the filter is given by (10). More specifically, denominator of the transfer function is just $\det(s\boldsymbol{T} + \boldsymbol{G})$, while its numerator is $g_b M_{21}(s)$, where $M_{21}(s) = -[s\boldsymbol{T} + \boldsymbol{G}]_{21}$, i.e. element $_{21}$ of the matrix $s\boldsymbol{T}+\boldsymbol{G}$ (multiplied by -1). Although there are still many possibilities, the most straightforward choice is

$$\boldsymbol{T} = \begin{bmatrix} C_1 & 0 \\ 0 & C_2 \end{bmatrix}, \boldsymbol{G} = \begin{bmatrix} 0 & -g_1 \\ g_2 & g_3 \end{bmatrix} \tag{16}$$

which gives all-pole second-order transfer function:

$$H(s) = \frac{g_b g_2}{s^2 C_1 C_2 + s C_1 g_3 + g_1 g_2} = \frac{\frac{g_b g_2}{C_1 C_2}}{s^2 + s\frac{g_3}{C_2} + \frac{g_1 g_2}{C_1 C_2}} \tag{17}$$

Corresponding filter topology (both in single-ended and fully differential structures) is shown in Fig. 2. Obviously, we can obtain many more equivalent filter topologies that implement transfer function (15) by different choice of matrices $\boldsymbol{T}$, $\boldsymbol{G}$, $\boldsymbol{B}$, $\boldsymbol{C}$ and $\boldsymbol{D}$.

3. ACTIVE-RC FILTER WITH NON-IDEAL OPAMPs

In this section we shall extend the model presented in Section 2 so as to take into account non-ideal behavior of operational amplifiers. In particular, we shall consider finite gain-bandwidth product of OPAMP as well as its non-zero output resistance.

In case of finite gain of OPAMP we can no longer assume that node voltages z_i, $i = 1, \ldots, n$ are equal to zero. Let us denote the gain of i-th operational amplifier OA_i

as $A_i(s)$. It follows from Fig. 3 that i-th integrator of the filter can be described by the following equations:

$$\begin{aligned} y_{bi}(U_{in} - z_i) + \sum_{j=1}^{n} y_{ij}\left(x_j - z_i\right) = 0 \qquad & i = 1, \ldots, n \\ x_i = -z_i A_i(s) & \end{aligned} \tag{18}$$

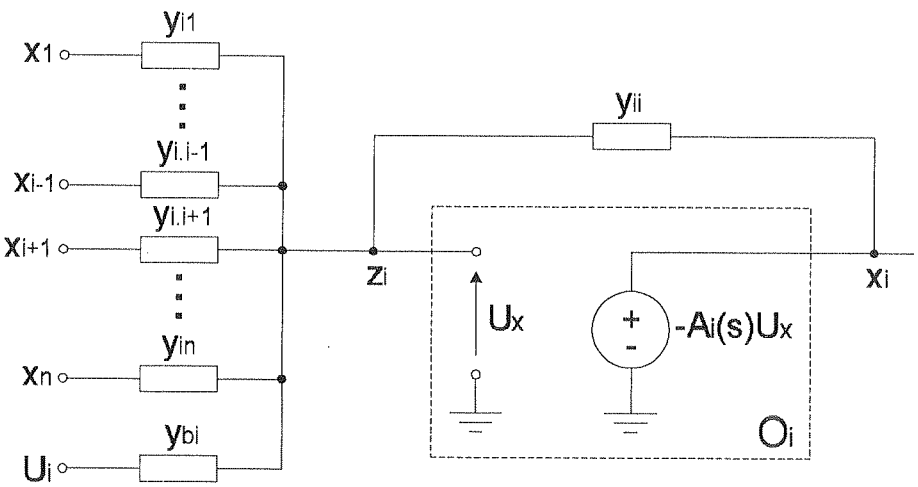


Fig. 3. Diagram of i-th integrator of the filter in Fig. 1 with finite gain OPAMP

Rys. 3. Schemat i-tego integratora filtru z Rys. 1 ze wzm. operacyjnym o skończonym wzmocnieniu

The corresponding equation for output summer is the following

$$\begin{aligned} g_o(U_{out} - z_o) + \sum_{j=1}^{n} g_{cj}\left(x_j - z_o\right) + g_d(U_{in} - z_o) = 0 \\ U_{out} = -z_o A_o(s) \end{aligned} \tag{19}$$

where $A_o(s)$ is the gain of the output operational amplifier OA_o. Equations (18) and (19) can be rewritten in matrix form as follows

$$\begin{aligned} (s\boldsymbol{T} + \boldsymbol{G} + \boldsymbol{Y}_A)\boldsymbol{X} + \boldsymbol{B}U_{in} = 0 \\ (1 + C_A)U_{out} + \boldsymbol{C}\boldsymbol{X} + \boldsymbol{D}U_{in} = 0 \end{aligned} \tag{20}$$

where

$$\boldsymbol{Y}_A = \begin{bmatrix} \dfrac{\sum_{j=1}^{n} \bar{y}_{1j} + \bar{y}_{b1}}{A_1(s)} & & 0 \\ & \ddots & \\ 0 & & \dfrac{\sum_{j=1}^{n} \bar{y}_{nj} + \bar{y}_{bn}}{A_n(s)} \end{bmatrix} \tag{21}$$

$$C_A = \frac{1 + \sum_{j=1}^{n} \bar{c}_j + \bar{d}}{A_o(s)} \quad \text{(or } C_A = 0 \text{ if there is no output summer)} \tag{22}$$

Here, $\bar{y}_{bi} = |g_{bi}| + s|C_{bi}|$, $\bar{y}_{ij} = |g_{ij}| + s|C_{ij}|$, $\bar{c}_i = |c_i|$, $i, j = 1, \ldots, n$, $\bar{d} = |d|$, which allows us to take into consideration the case when some entries of the matrices $\boldsymbol{T}$, $\boldsymbol{G}$, $\boldsymbol{B}$, $\boldsymbol{C}$ and $\boldsymbol{D}$ are negative. In such a case (which is equivalent to cross-coupling of the corresponding filter elements), negative sign of the element value corresponds in fact to the negative sign of appropriate node voltage x_j (see discussion in Section 2). However, from the point of view of input node voltage z_i element is still positive and we must take absolute value of negative elements to maintain correctness of equations. Using (20) we can calculate transfer function of the filter, which is:

$$H(s) = \frac{U_{in}(s)}{U_{out}(s)} = (1 + C_A)^{-1} \left(\boldsymbol{C} (s\boldsymbol{T} + \boldsymbol{G} + \boldsymbol{Y}_A)^{-1} \boldsymbol{B} - \boldsymbol{D} \right) \tag{23}$$

Note that even if $A_i(s)$ is modeled using single pole approximation, matrix elements of $\boldsymbol{Y}_A$ are, in general, of second order in s. This makes it difficult to directly evaluate the transfer function formula (23). However, if for frequencies of operation of the filter we have $|A_i(s)| \gg 1$, $i = 1, \ldots, n$, $|A_o(s)| \gg 1$, which means that $\|\boldsymbol{Y}_A\| \ll \|s\boldsymbol{T} + \boldsymbol{G}\|$ for any reasonable matrix norm, we can use the following approximation:

$$(\boldsymbol{I} + \boldsymbol{A})^{-1} \cong \boldsymbol{I} - \boldsymbol{A} \tag{24}$$

which is valid for any matrix $\boldsymbol{A}$ as far as $\|\boldsymbol{A}\| \ll \|\boldsymbol{I}\|$ ($\boldsymbol{I}$ stands for an identity matrix). Using (24) we obtain the following formula for $H(s)$:

$$\begin{aligned} H(s) &\cong (1 + C_A)^{-1} \left(\boldsymbol{C} \left(\boldsymbol{I} - (s\boldsymbol{T} + \boldsymbol{G})^{-1} \boldsymbol{Y}_A \right) (s\boldsymbol{T} + \boldsymbol{G})^{-1} \boldsymbol{B} - \boldsymbol{D} \right) = \\ &= (1 + C_A)^{-1} \left(\boldsymbol{C} (s\boldsymbol{T} + \boldsymbol{G})^{-1} \boldsymbol{B} - \boldsymbol{D} \right) - (1 + C_A)^{-1} \boldsymbol{C} (s\boldsymbol{T} + \boldsymbol{G})^{-1} \boldsymbol{Y}_A (s\boldsymbol{T} + \boldsymbol{G})^{-1} \boldsymbol{B} \end{aligned} \tag{25}$$

which can be rewritten as

$$H(s) \cong (1 + C_A)^{-1} H_0(s) - \Delta H(s) \tag{26}$$

where $H_0(s)$ is the nominal transfer function of the filter with ideal OPAMPs (eqn. (7)), and $\Delta H(s)$ is deviation from $H_0(s)$ due to finite OPAMP gains given by

$$\Delta H(s) = (1 + C_A)^{-1} \boldsymbol{C} (s\boldsymbol{T} + \boldsymbol{G})^{-1} \boldsymbol{Y}_A (s\boldsymbol{T} + \boldsymbol{G})^{-1} \boldsymbol{B} \tag{27}$$

In general case, however, we have to evaluate formula (23) to obtain accurate results.

Now, we shall consider the effect of non-zero output impedance of the filter OPAMPs. Fig. 4 shows i-th OPAMP of the filter, where ν_{oi} denotes its output admittance, $\tilde{x}_i$ and x_i are the internal and external output voltages of the OPAMP, respectively. Since in practice we always have $|z_j| \ll |x_i|$, $i, j = 1, \ldots, n$, we can write the following equation:

$$(\tilde{x}_i - x_i) y_{oi} = x_i \left(|g_{ci}| + \sum_{j=1}^{n} \bar{y}_{ji} \right), \quad i = 1, \ldots, n \tag{28}$$

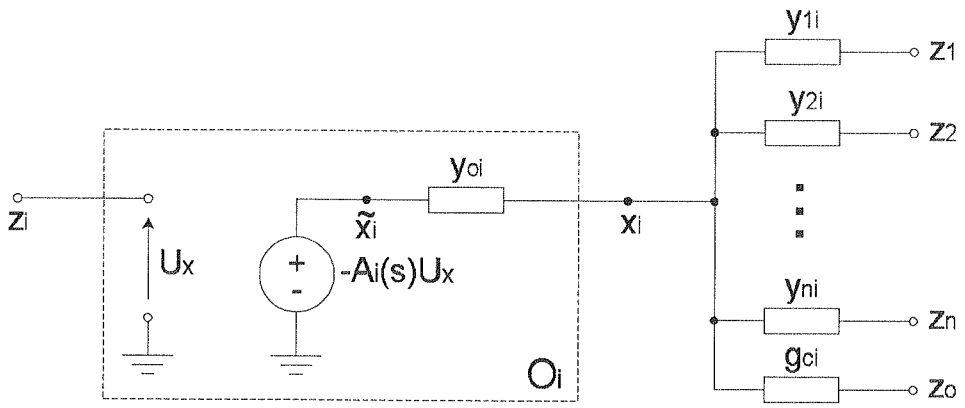


Fig. 4. Diagram of i-th OPAMP of the filter in Fig.1 with non-zero output admittance

Rys. 4. Schemat i-tego wzm. operacyjnego filtru z Rys.1 z niezerową admitancją wyjściową

where we wrote $|g_{ci}|$ and $\bar{y}_{ji}$ for the same reason as in (21). Hence, we have

$$x_i = \tilde{x}_i \frac{y_{oi}}{y_{oi} + |g_{ci}| + \sum_{j=1}^{n} \bar{y}_{ji}}, \quad i = 1, \ldots, n \tag{29}$$

This means that effective gain of the amplifier is reduced by the factor $\left(1 + y_{oi}^{-1}\left(|g_{ci}| + \sum_{j=1}^{n} \bar{y}_{ji}\right)\right)^{-1}$, which is, in general, frequency dependent. In order to take it into account in our previous formulas, we have to replace $A_i(s)$ in (21) by $\tilde{A}_i(s)$ given by

$$\tilde{A}_i(s) = \frac{A_i(s)}{1 + y_{oi}^{-1}\left(|g_{ci}| + \sum_{j=1}^{n} \bar{y}_{ji}\right)} \tag{30}$$

In a similar way, one can treat output impedance of OA_o — OPAMP of the output summer of the filter. We have

$$\left(\tilde{U}_{out} - U_{out}\right) y_o = U_{out}|g_o|, \tag{31}$$

where $\tilde{U}_{out}$ and U_{out} are internal and external output voltages of OA_o, while y_o is its output admittance. Hence, we get

$$U_{out} = \tilde{U}_{out} \frac{y_o}{y_o + |g_o|}, \tag{32}$$

which results in replacing $A_o(s)$ in (22) by $\tilde{A}_o(s)$:

$$\tilde{A}_o(s) = \frac{A(s)}{1 + y_o^{-1}|g_o|} \tag{33}$$

4. NOISE ANALYSIS OF ACTIVE-RC FILTERS

Noise performance is one of the very important characteristics of Active-RC filters. A number of papers have been published dealing with noise in Active-RC filters, mostly some particular topologies such as single amplifier filters, state-space filters (eg. [14]-[16]) but also in general setting [17]. In this section, we present a procedure for evaluating noise in Active-RC filters, which is based on the matrix description of the general Active-RC filter model presented in Section 2.

We shall carry out the noise analysis of the filter structure in Fig. 1. The output noise of any Active-RC filter is a combination of the noise contribution of its all operational amplifiers and resistors (we treat filter capacitors as noiseless). The noise of operational amplifier can be described by the equivalent input referred noise voltage source v_n as shown in Fig. 5. Spectral density $S_n(f)$ of the noise source can be modeled as:

$$S_n(f) = S_t + \frac{S_f}{f} \tag{34}$$

where both S_t (thermal noise component) and S_f (flicker noise component) depend on amplifier topology, however, in general we do not need to restrict ourselves to this model. Noise of the resistor of value R will be represented by the corresponding spectral density $4KTR$, where k is Boltzmann's constant, T is absolute temperature. We shall assume that noise sources associated to different OPAMPs and resistors are statistically independent.

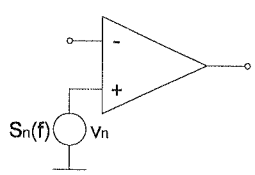


Fig. 5. Equivalent input voltage noise source representation of noise in OPAMP

Rys. 5. Reprezentacja szumów wzmacniacza operacyjnego za pomocą równoważnego źródła napięciowego na wejściu wzmacniacza

Our immediate goal is to obtain the explicit formula for output (and/or input) noise spectrum of the general Active-RC filter in Fig. 1. In order to do this, one has to consider what is the contribution of the noise of each individual operational amplifier and resistor to the output noise spectrum of the filter. We start from the noise contribution of filter OPAMPs OA_k, $k = 1, \ldots, n$. Let v_{Ok} denote the input referred noise voltage of the amplifier OA_k, whose spectral density is $S_{Ok}(f)$. Since we assume that, except their noise, OPAMPs are otherwise ideal, we can calculate output noise

U_{Ok} of the filter due to amplifier OA_k by rewriting equations (1), (2) with $U_{in} = 0$ and $z_k = v_{Ok}$ (noise voltage of OA_k appears at its negative input).

$$-y_{bi}\delta_{ik}v_{Ok} + \sum_{j=1}^{n} y_{ij}\left(x_j - \delta_{ik}v_{Ok}\right) = 0, \quad i,k = 1,\ldots,n \tag{35}$$

$$g_o U_{Ok} + \sum_{j=1}^{n} g_{cj}x_j = 0 \tag{36}$$

where δ_{ik} stands for the Kronecker's delta (i.e. $\delta_{ik} = 1$ for $i = k$ and 0 otherwise). Let us denote by $\boldsymbol{H}_{cv}$ the $1 \times n$ vector defined as follows

$$\boldsymbol{H}_{cv}(s) = \left[\begin{array}{ccc} H_1(s) & \cdots & H_n(s) \end{array}\right] = \boldsymbol{C}(s\boldsymbol{T} + \boldsymbol{G})^{-1} \tag{37}$$

Using this notation we can calculate the noise component U_{ok} as follows

$$U_{Ok} = H_k\left(\bar{y}_{bk} + \sum\nolimits_{j=1}^{n} \bar{y}_{kj}\right)v_{Ok} \tag{38}$$

where $\bar{y}_{bk} = |g_{bk}| + s|C_{bk}$, $\bar{y}_{kj} = |g_{kj}| + s|C_{kj}|$, $k,j = 1,\ldots,n$. We use absolute values to take into consideration the case when some entries of the matrices $\boldsymbol{T}$, $\boldsymbol{G}$, $\boldsymbol{B}$, are negative (see discussion after (22)). Corresponding spectral density $S_{o.Ok}(f)$ is given by the formula:

$$S_{o.Ok}(f) = S_{Ok}(f)\,|H_k(j2\pi f)|^2 |\bar{y}_{bk} + \sum\nolimits_{j=1}^{n} \bar{y}_{kj}|^2 \tag{39}$$

If the filter contains the output summer, we have to take into account the noise contribution of the amplifier OA_o as well. Let v_{Oo} denote the input referred noise voltage of the amplifier OA_o, whose spectral density is $S_{Oo}(f)$. We can calculate output noise of the filter due to OA_o from the following equation (note that v_{Oo} appears at the negative input of OA_o so $z_o = v_{Oo}$)

$$g_o(U_{Oo} - v_{Oo}) - \sum\nolimits_{j=1}^{n} g_{cj}v_{Oo} - g_d v_{Oo} = 0 \tag{40}$$

which gives

$$U_{Oo} = H_o\left(|g_o| + \sum\nolimits_{j=1}^{n} |g_{cj}| + |g_d|\right)v_{Oo} \tag{41}$$

where $H_o = (g_o)^{-1}$ was introduced for consistency of notation. Again, taking absolute values in (41) allows us to cover the case when some of the elements are negative. Corresponding spectral density $S_{o.Oo}(f)$ is given by the formula:

$$S_{o.Oo}(f) = S_{Oo}(f)\,|H_o|^2\left(|g_o| + \sum\nolimits_{j=1}^{n} |g_{cj}| + |g_d|\right)^2 \tag{42}$$

The total output noise voltage spectrum $S_O(f)$ of the filter in Fig.1 due to its operational amplifiers can be now calculated using (39), (42) and the assumption of statistical independence of noise sources as

$$S_O(f) = \sum_{k=1}^{n} S_{Ok}(f)\,|H_k(j2\pi f)|^2 |\bar{y}_{bk} + \sum_{j=1}^{n} \bar{y}_{kj}|^2 + S_{Oo}(f) \left| \left(|g_o| + \sum_{j=1}^{n} |g_{cj}| + |g_d| \right) / |g_o|^2 \right| \tag{43}$$

As a next step, we shall calculate the output noise of the filter in Fig.1 due to its resistors. Denote by v_{bk} the noise voltage of resistor R_{bk} (corresponding spectral density is $4kTR_{bk}$). Output noise voltage U_{bk} due to this resistor can be calculated from the following equations (note that resistor R_{bk} injects its noise current $g_{bk}v_{bk}$ into node z_k)

$$g_{bi}\delta_{ik}v_{bk} + \sum_{j=1}^{n} y_{ij}x_j = 0, \quad i,k = 1,\ldots,n \tag{44}$$

$$g_o U_{bk} + \sum_{j=1}^{n} g_{cj}x_j = 0 \tag{45}$$

Using (37) we get

$$U_{bk} = H_k g_{bk} v_{bk} \tag{46}$$

Corresponding spectral density $S_{bk}(f)$ is given by the formula:

$$S_{bk}(f) = 4kTR_{bk}|H_k(2\pi f)|^2 g_{bk}^2 = 4kT|g_{bk}| \cdot |H_k(2\pi f)|^2 \tag{47}$$

The noise due to feedback resistors R_{kl}, $k,l = 1, \ldots, n$ can be calculated in a similar way. Denote by v_{kl} the noise voltage of resistor R_{kl} (corresponding spectral density is $4kTR_{kl}$). Output noise voltage U_{kl} due to this resistor can be calculated from the following equations (note that resistor R_{kl} injects its noise current $g_{kl}(x_l - v_{kl})$ into node z_k)

$$\sum_{j=1}^{n} y_{ij}\left(x_j - \delta_{ik}\delta_{jl}v_{kl}\right) = 0, \quad i,\, k,\, l = 1, \ldots, n \tag{48}$$

$$g_o U_{kl} + \sum_{j=1}^{n} g_{cj}x_j = 0 \tag{49}$$

Using (37) we get

$$U_{kl} = H_k g_{kl} v_{kl} \tag{50}$$

Corresponding spectral density $S_{kl}(f)$ is given by the formula:

$$S_{kl}(f) = 4kTR_{kl}|H_k(2\pi f)|^2 g_{kl}^2 = 4kT|g_{kl}| \cdot |H_k(2\pi f)|^2 \tag{51}$$

If the filter contains non-trivial output summer, we have to take into account the noise of the resistors R_{ci}, $i = 1, \ldots, n$, R_d and R_o. Denote by v_{ci}, v_d, and v_o the noise voltage of resistors R_{ci}, R_d and R_o, respectively (corresponding spectral densities are $4kTR_{ci}$, $4kTR_d$, and $4kTR_o$, respectively). It can be easily shown that output noise voltages U_{ci}, U_d, and U_o due to these resistors can be calculated as

$$U_{ci} = H_o g_{ci} v_{ci}, \quad i = 1, \ldots, n \tag{52a}$$

$$U_d = H_o g_d v_d \tag{52b}$$

$$U_o = H_o g_o v_o \tag{52c}$$

and the corresponding spectral densities $S_{ci}(f)$, $S_d(f)$, $S_o(f)$ are

$$S_{ci}(f) = 4kTR_{ci}|H_o|^2 g_{ci}^2 = 4kT|g_{ci}| \cdot |H_o|^2 = 4kTR_o|c_i|, \quad i = 1, \ldots, n \tag{53a}$$

$$S_d(f) = 4kTR_d|H_o|^2 g_d^2 = 4kT|g_d| \cdot |H_o|^2 = 4kTR_o|d| \tag{53b}$$

$$S_o(f) = 4kTR_o|H_o|^2 g_o^2 = 4kT|g_o| \cdot |H_o|^2 = 4kTR_o \tag{53c}$$

Using (47), (51) and (53) one can easily calculate the total output noise voltage spectrum $S_R(f)$ of the filter in Fig. 1 due to its resistors

$$S_R(f) = \sum_{k=1}^{n} \left[S_{bk}(f) + \sum\nolimits_{l=1}^{n} S_{kl}(f) \right] + \sum\nolimits_{k=1}^{n} S_{ci}(f) + S_d(f) + S_o(f) \tag{54}$$

that is

$$\begin{aligned} S_R(f) = 4kT \Bigg[& \sum_{k=1}^{n} \left(|g_{bk}| + \sum\nolimits_{l=1}^{n} |g_{kl}| \right) |H_k(j2\pi f)|^2 \\ & + \left(\sum\nolimits_{k=1}^{n} |g_{ci}| + |g_d| + |g_o| \right) |H_o|^2 \Bigg] \end{aligned} \tag{55}$$

Finally, the total output noise voltage spectrum $S_{no}(f)$ of the filter in Fig. 1 can be calculated as a sum of $S_O(f)$ and $S_R(f)$:

$$\begin{aligned} S_{no}(f) &= S_O(f) + S_R(f) = \\ &= \sum_{k=1}^{n} \left(S_{Ok}(f) |\bar{y}_{bk} + \sum\nolimits_{j=1}^{n} \bar{y}_{kj}|^2 + 4kT \left(|g_{bk}| + \sum\nolimits_{j=1}^{n} |g_{kj}| \right) \right) \\ &|H_k(j2\pi f)|^2 + \left(S_{Oo}(f) \left(|g_o| + \sum\nolimits_{j=1}^{n} |g_{cj}| + |g_d| \right) + 4kT \right) \\ &\left(|g_o| + \sum\nolimits_{j=1}^{n} |g_{cj}| + |g_d| \right) / g_o^2 \end{aligned} \tag{56}$$

Equivalent input referred noise voltage spectrum $S_{ni}(f)$ can be calculated by dividing $S_{no}(f)$ by $|H(j2\pi f)|^2$ — the square of modulus of filter's transfer function given by (7).

Note that in equations (44)-(56) absolute values were taken where necessary to include the case when some of the matrix elements take negative values (see discussion after equation (38)). Note also that if formula (56) is to be applied to fully differential filter structure, noise spectrum of the filter resistors have to be counted twice, i.e. we have

$$\begin{aligned} S_{no}(f) &= S_O(f) + S_R(f) = \\ &= \sum_{k=1}^{n} \left(S_{Ok}(f) |\bar{y}_{bk} + \sum\nolimits_{j=1}^{n} \bar{y}_{kj}|^2 + 8kT \left(|g_{bk}| + \sum\nolimits_{j=1}^{n} |g_{kj}| \right) \right) \\ &|H_k(j2\pi f)|^2 \left(S_{Oo}(f) \left(|g_o| + \sum\nolimits_{j=1}^{n} |g_{cj}| + |g_d| \right) + 8kT \right) \\ &\left(|g_o| + \sum\nolimits_{j=1}^{n} |g_{cj}| + |g_d| \right) / g_o^2 \end{aligned} \tag{57}$$

Although formula (56) is quite complex, it can be easily evaluated numerically and can be used as a basis for automated noise analysis and optimization software. For the sake of illustration we perform a hand analysis of the noise of the second-order filter in Fig. 2. We assume for simplicity that both operational amplifiers are the same with input referred noise spectrum $S(f) = S_t$ (thermal noise only). We need transfer functions H_1 and H_2 which are (cf. (16)):

$$\begin{bmatrix} H_1(s) & H_2(s) \end{bmatrix} = \mathbf{C}(s\mathbf{T} + \mathbf{G})^{-1} = \frac{1}{D(s)} [g_2 - sC_1] \tag{58}$$

where $D(s) = s^2 C_1 C_2 + sC_1 g_3 + g_1 g_2$ is denominator of the transfer function of filter in Fig. 2. Now it follows from (56) and (58) that the output noise spectrum of the filter in Fig. 2 is

$$\begin{aligned} S_{no}(f) &= \frac{1}{|D(j2\pi f)|^2} \left(S_t |j2\pi f C_1 + g_b + g_1|^2 + 8kT(g_b + g_1) \right) g_2^2 \\ &+ \frac{1}{|D(j2\pi f)|^2} \left(S_t |j2\pi f C_2 + g_2 + g_3|^2 + 8kT(g_2 + g_3) \right) |j2\pi f C_1|^2 \end{aligned} \tag{59}$$

We can see that even in this simple case the output noise formula is quite complicated. We can, however, attempt to get some insight at least for low frequencies by setting $f = 0$ in (58). We obtain

$$S_{no}(0) = \frac{S_t (g_b + g_1)^2 + 8kT(g_b + g_1)}{g_1^2} = (R_1/R_b + 1)[S_t (R_1/R_b + 1) + 4kTR_1] \tag{60}$$

This result means that besides the obvious fact that in order to minimize the noise, resistor values should be minimized, we learn also that the ratio R_1/R_b should be kept as small as possible. This ratio, however, is equal to DC gain of the filter. If this is fixed, the only feasible optimization strategy is simultaneous reduction of R_1 and R_b (having in mind other design constraints such as power consumption and chip area

as well as the fact that the transfer function is to be kept unchanged). Obviously, if we consider the noise in the whole pass-band (e.g. noise integral) not just for $f = 0$, we have more degrees of freedom and we can possibly find different optimization strategies. The more general approach to the noise optimization is discussed in the next section.

5. VERIFICATION AND APPLICATION EXAMPLES

For the sake of verification we have compared theoretical results with SPICE simulation using two example filters: a 3rd order low-pass Chebyshev filter in leap-frog structure with 3dB frequency equal to 12MHz (specifications typical for VDSL filters [5]) shown in Fig. 6, and a 5th order low-pass Butterworth filter in leap-frog structure with 3dB frequency equal to 5MHz shown in Fig. 7. The filters are realized using simple two-stage class AB OPAMP with Miller compensation shown in Fig. 8. The circuit is implemented in standard 0.35μm CMOS process. Simulated OPAMP parameters are: DC gain −71dB, open-loop 3dB frequency - 430kHz, phase margin 45°, input referred noise spectrum 14.5nV/Hz$^{1/2}$ (only thermal noise is considered), output resistance r_o = 15kΩ. Filter elements are C_1 = 2.28pF, C_2 = 2.01pF, C_3 = 1.53pF, $R_i = R$ = 10kΩ, $i = 1,\ldots,6$ (Chebyshev filter), and C_1 = 5.10pF, C_2 = 5.59pF, C_3 = 4.56pF, C_4 = 2.95pF, C_5 = 1.02pF, , $R_i = R$ = 10kΩ, $i = 1,\ldots,10$ (Butterworth filter). Figs. 9 and 10 show theoretical and simulated frequency responses of the filters in Figs. 6 and 7, respectively. We can observe nominal (ideal) response as well as actual response which is distorted due to the finite gain-bandwidth product (GBW) and non-zero output resistance of OPAMPs. The agreement between theoretical and simulated data is very good. Figs. 11 and 12 show theoretical and simulated input referred noise spectrum of the filters in Figs. 6 and 7, respectively. Also in this case, the agreement between both sets of data is excellent.

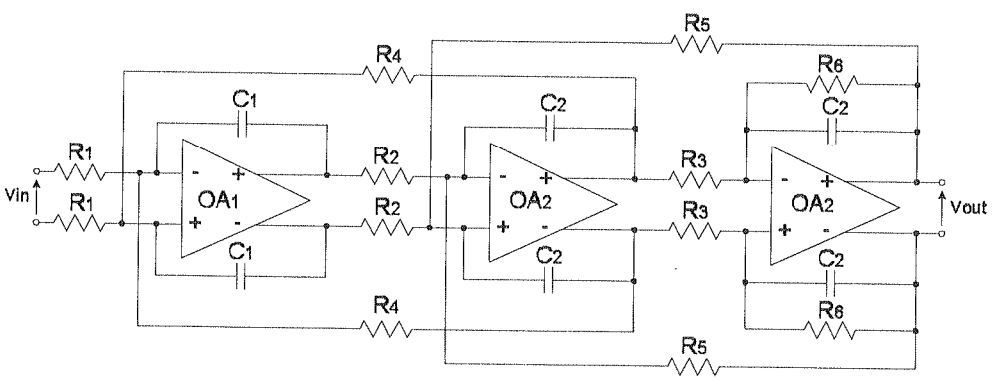


Fig. 6. Fully differential 3rd order leap-frog Active-RC filter

Rys. 6. W pełni różnicowy filtr aktywny RC trzeciego rzędu w konfiguracji leap-frog

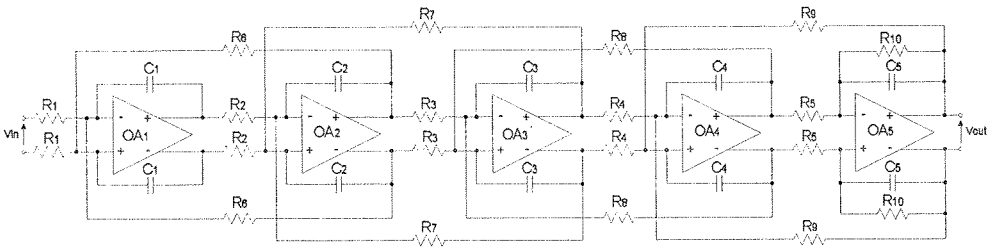


Fig. 7. Fully differential 5th order leap-frog Active-RC filter

Rys. 7. W pełni różnicowy filtr aktywny RC piątego rzędu w konfiguracji leap-frog

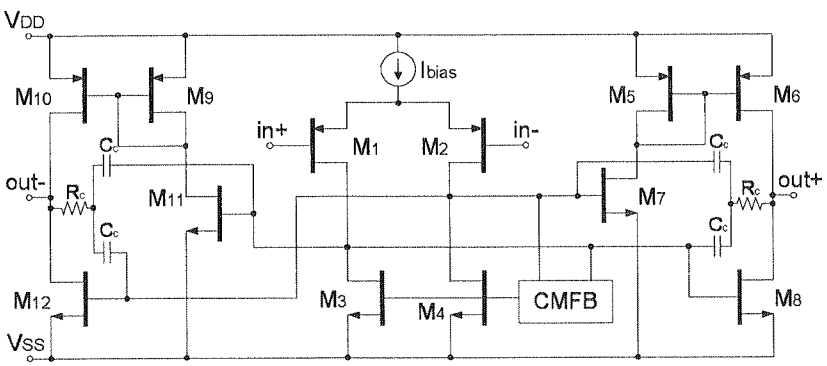


Fig. 8. Simple class AB fully-differential OPAMP

Rys. 8. Prosty wzmacniacz operacyjny ze stopniem wyjściowym klasy AB

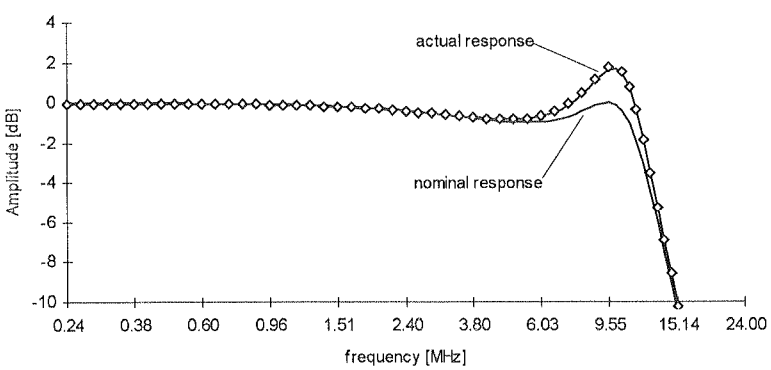


Fig. 9. Frequency response of 3rd order 1dB Chebyshev filter in Fig.6: nominal and actual response; theory (solid line), and simulation (points)

Rys. 9. Charakterystyka częstotliwościowa filtru Czebyszewa 3-go rzędu z Rys.6: odpowiedź nominalna oraz rzeczywista; teoria (linia ciągła), symulacje (punkty)

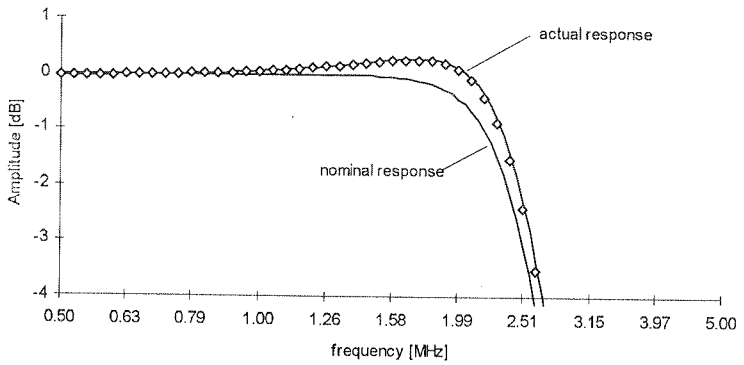


Fig. 10. Frequency response of 5th order Butterworth filter in Fig.7: nominal and actual response; theory (solid line), and simulation (points)

Rys. 10. Charakterystyka częstotliwościowa filtru Butterworth'a 5-go rzędu z Rys.7: odpowiedź nominalna oraz rzeczywista; teoria (linia ciągła), symulacje (punkty)

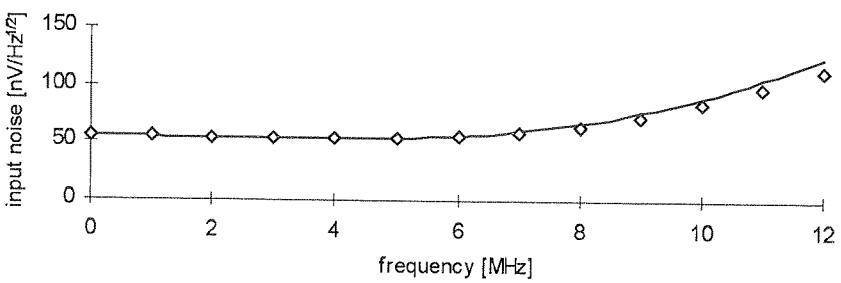


Fig. 11. Input referred noise spectrum of the filter in Fig.6; theory (solid line), and simulation (points)

Rys. 11. Odniesione do wejścia widmo szumów filtru z Rys.6; teoria (linia ciągła), symulacje (punkty)

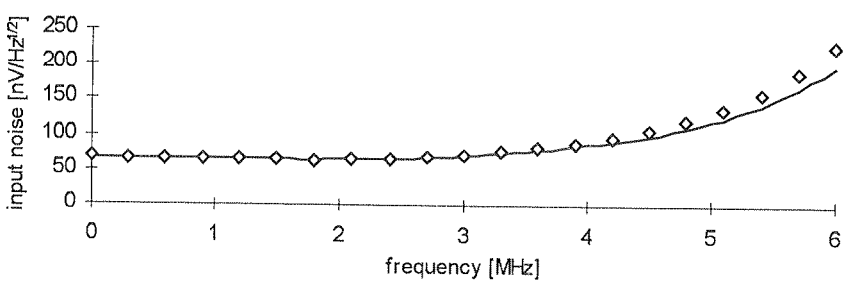


Fig. 12. Input referred noise spectrum of the filter in Fig.7; theory (solid line), and simulation (points)

Rys. 12. Odniesione do wejścia widmo szumów filtru z Rys.7; teoria (linia ciągła), symulacje (punkty)

As the first application example of the proposed model we shall obtain OPAMP specifications for use in the Chebyshev filter in Fig.6 assuming application to VDSL Analog Front-End. Design specifications for VDSL filters are very tough, especially with respect to noise and linearity. In this example, we are merely interested in the transfer function distortion due to finite gain-bandwidth product and output resistance of filter OPAMPs. We assume again that all resistors in the filter are the same and equal to the common value R. We aim at estimating required OPAMP gain-bandwidth product (GBW) and output resistance r_o so that distortion of the filter transfer function due to these non-idealities is within acceptable range.

Fig. 13 shows again the ideal frequency response of the filter in Fig.6, and the response for GBW of OPAMPs equal to 200MHz and $r_o/R = 0.5$. We can observe both amplitude distortion (picking) ΔA and 3dB frequency error Δf_{3dB}. Figs.14 and 15 show ΔA and Δf_{3dB}, respectively, versus GBW of OPAMP, for different values of the ratio r_o/R. All the results have been obtained in a single run of the Matlab program implementing formulas presented in Sections 2 and 3. It is seen that Δf_{3dB} is not a big problem since its value is relatively small and can be easily compensated by calibration. However, ΔA is large and heavily dependent on both GBW and r_o/R.

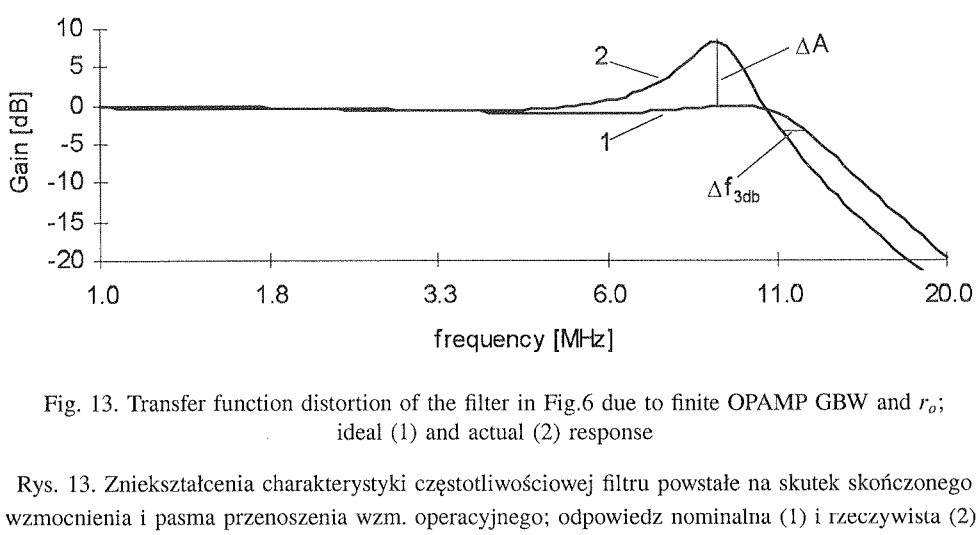


Fig. 13. Transfer function distortion of the filter in Fig.6 due to finite OPAMP GBW and r_o; ideal (1) and actual (2) response

Rys. 13. Zniekształcenia charakterystyki częstotliwościowej filtru powstałe na skutek skończonego wzmocnienia i pasma przenoszenia wzm. operacyjnego; odpowiedz nominalna (1) i rzeczywista (2)

Table 1 shows the value of GBW required to keep $\Delta A < 2$dB for different r_o/R ratios. The data in Table 1 represents trade-off between GBW and r_o of the OPAMP. Although we can relax requirements concerning GBW it is always at the expense of decreasing r_o/R ratio. It is not possible to keep this ratio small just by increasing resistor value R because this results in increasing filter noise, which is undesirable. Thus, we have an alternative: either try to obtain large GBW or decrease OPAMP output resistance (this, however, influences OPAMP architecture — very small values of r_o can be obtained either using output buffer which, in turn, degrades both circuit linearity and frequency performance, or at the expense of increased power consumption

in output stage). Discussion of this problem is, however, beyond the scope of this paper. Our goal was just to estimate GBW and r_o/R for which the transfer function distortion is under desired level.

Table 1

GBW of OPAMP for which $\Delta A < 2$dB for the filter in Fig.6

Wartości GBW wzmacniacza operacyjnego, dla których $\Delta A < 2$dB dla filtru z Rys.6

r_o/R	0	0.25	0.5	1.0	1.5
GBW [MHz]	165	350	550	970	1400

As the second example we shall investigate the trade-off between input referred noise spectrum of the filter in Fig. 6, input noise spectrum of its OPAMPs and the value of the filter resistors R. Fig. 16 shows input noise spectrum of the filter in Fig. 6 versus resistor value R with input noise spectrum S_n of filter OPAMPs as a parameter (we consider here thermal noise only), for $S_n = 2, 4, 6, \ldots, 20$ [nV/Hz$^{1/2}$]. Using the data presented in Fig. 16 and given filter noise design specifications one can easily determine OPAMP noise requirements for given value of R. For example, if the input noise spectrum requirement for the filter is 30nV/Hz$^{1/2}$ (the value typical for VDSL applications), it can be met for OPAMP's input noise not larger than 6nV/Hz$^{1/2}$, assuming $R = 5$kΩ, which is quite tough. This can be relaxed if we allow smaller resistors, e.g. for $R = 3$kΩ [$R = 1$kΩ], filter noise requirement is met for OPAMP's input noise not larger than about 8nV/Hz$^{1/2}$[10nV/Hz$^{1/2}$]. Obviously, this is a trade-off since keeping the noise small by decreasing the value of R leads to larger transfer function distortion (as indicated in the previous paragraph) or/and to other undesirable effects such as increased power consumption and degraded linearity (e.g. if we try to reduce output resistance of the OPAMP to counteract the transfer function distortion). These problems can be partially solved by noise optimization of the filter as shown in the sequel.

As the last example we shall consider noise optimization of the filter in Fig. 6. Let us start from some general remarks. Noise optimization task can be performed in different ways. For example, given filter transfer function and some initial topology, we can allow both the filter structure and element values to be changed while looking for the optimal solution. Or we can assume the filter topology to be fixed and change just element values. Some 'intermediate' approaches are also possible, e.g. we can assume that some features of the filter topology are fixed (e.g. no output summer) but others (e.g. input signal distribution) can be changed along with the element values. The Active-RC filter model in Fig. 1 and its matrix description allow us to cover all possible approaches and perform the task of noise optimization in full generality so that derived methodology (and software) can be used for any filter integrator-based Active-RC filter.

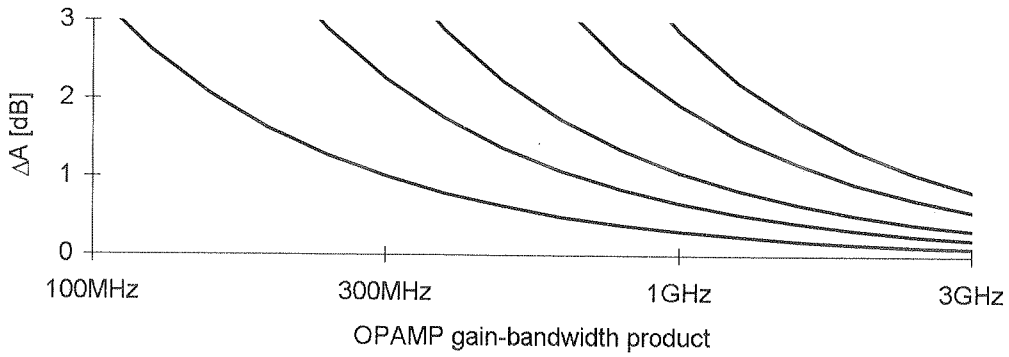


Fig. 14. Amplitude distortion ΔA versus GBW of filter OPAMPs for filter in Fig.6 and different values of r_o/R ratio

Rys. 14. Zniekształcenia amplitudy ΔA dla filtru z Rys.6 w funkcji GBW wzmacniacza operacyjnego dla rożnych wartości stosunku r_o/R

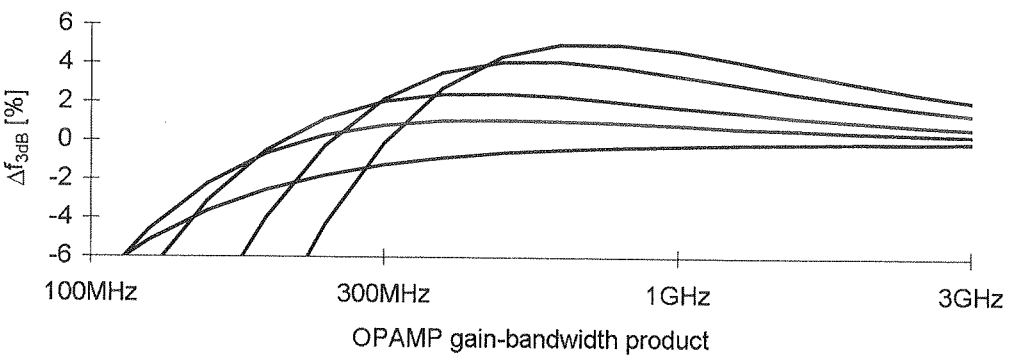


Fig. 15. Corner frequency error Δf_{3dB} versus GBW of filter OPAMPs for filter in Fig.6 and different values of r_o/R ratio

Rys. 15. Błąd częstotliwości 3-dB Δf_{3dB} dla filtru z Rys.6 w funkcji GBW wzmacniacza operacyjnego dla rożnych wartości stosunku r_o/R

The basic assumption is that the filter transfer function is to be unchanged while minimizing the noise. We start from the transfer function formula (7). The most general transform that can be applied to the filter is as follows. Let $\boldsymbol{P}$ and $\boldsymbol{Q}$ be two invertible $n \times n$ matrices. Then we have

$$\begin{aligned} H(s) &= \boldsymbol{C}(s\boldsymbol{T} + \boldsymbol{G})^{-1}\boldsymbol{B} - \boldsymbol{D} = \boldsymbol{C}\boldsymbol{P}\boldsymbol{P}^{-1}(s\boldsymbol{T} + \boldsymbol{G})^{-1}\boldsymbol{Q}^{-1}\boldsymbol{Q}\boldsymbol{B} - \boldsymbol{D} = \\ &= \boldsymbol{C}\boldsymbol{P}(s\boldsymbol{Q}\boldsymbol{T}\boldsymbol{P} + \boldsymbol{Q}\boldsymbol{G}\boldsymbol{P})^{-1}\boldsymbol{Q}\boldsymbol{B} - \boldsymbol{D} \end{aligned} \tag{61}$$

Let us define the following matrices

$$\bar{\boldsymbol{T}} = \boldsymbol{Q}\boldsymbol{T}\boldsymbol{P},\ \bar{\boldsymbol{G}} = \boldsymbol{Q}\boldsymbol{G}\boldsymbol{P},\ \bar{\boldsymbol{C}} = \boldsymbol{C}\boldsymbol{P},\ \bar{\boldsymbol{B}} = \boldsymbol{Q}\boldsymbol{B},\ \bar{\boldsymbol{D}} = \boldsymbol{D} \tag{62}$$

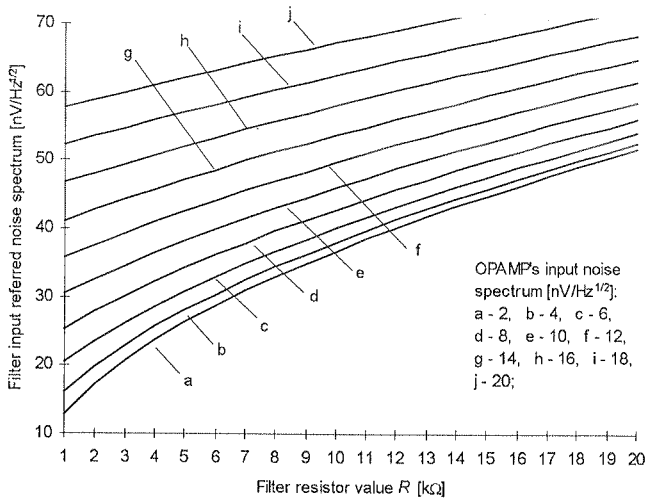


Fig. 16. Input noise spectrum of the filter in Fig.6 versus resistor value R with input noise spectrum of filter OPAMPs as a parameter

Rys. 16. Odniesione do wejścia widmo szumów filtru z Rys.6 w funkcji wartości rezystancji R z wartością widma szumów wzmacniaczy operacyjnych filtru jako parametrem

It is seen from (61) that $\bar{\boldsymbol{T}}$, $\bar{\boldsymbol{G}}$, $\bar{\boldsymbol{C}}$, $\bar{\boldsymbol{B}}$ and $\bar{\boldsymbol{D}}$ define a new filter, which has the same transfer function but different element values (and, possibly, topology) and, of course, other performance parameters, especially noise. The task is now to look for the matrices $\boldsymbol{P}$ and $\boldsymbol{Q}$ so that noise of the filter is minimized. More generally, we aim at minimizing a target function F consisting of one or more performance specifications, where F is a function of both initial filter and matrices $\boldsymbol{P}$ and $\boldsymbol{Q}$, i.e.

$$F = F(\boldsymbol{T},\ \boldsymbol{G},\ \boldsymbol{B},\ \boldsymbol{C},\ \boldsymbol{D};\ \boldsymbol{P},\ \boldsymbol{Q}) \tag{63}$$

Matrices $\boldsymbol{T}$, $\boldsymbol{G}$, $\boldsymbol{B}$, $\boldsymbol{C}$ and $\boldsymbol{D}$ are parameters, while $\boldsymbol{P}$ and $\boldsymbol{Q}$ are optimization variables. Performance function F depends on what we are going to achieve. Typically, in the case of noise optimization, it is noise of the filter integrated over suitable frequency band.

Usually, we also have some design constraints such as the maximum value of total capacitance of the filter (this affects chip area), maximum power consumption of the filter which depends, among others, on resistor values, both directly (because of the power dissipated in resistors) and indirectly (using small resistors influences OPAMP architecture enforcing low output resistance topologies that usually spend more power), allowable resistance [capacitance] ratio, i.e. the ratio of maximum to minimum resistance [capacitance] values in the filter (too large ratio may cause matching problems), and so on. All these constraints have to be taken into account in the optimization process. In general, constraints can be written as follows

$$m_j \leqslant c_j(\boldsymbol{T},\ \boldsymbol{G},\ \boldsymbol{B},\ \boldsymbol{C},\ \boldsymbol{D};\ \boldsymbol{P},\ \boldsymbol{Q}) \leq M_j, \quad j = 1, \ldots, N_c \tag{64}$$

where N_c is the number of constraints, c_j is the j constraint function (e.g. total capacitance of the filter) which is dependent on matrices describing the filter and optimization variables, while m_j and M_j are minimum and maximum values of c_j (which may be finite or infinite).

The optimization itself can be carried out using any available numerical procedure embedded into the optimization system. Choice of the optimization procedure depends on the complexity of the problem as well as constraints.

In this section, for the sake of brevity, we shall restrict our discussion to the case when the filter topology is fixed and what we are doing is just changing element values. In this case, we have to assume that both $\boldsymbol{P}$ and $\boldsymbol{Q}$ are real and diagonal, i.e.

$$\boldsymbol{P} = \begin{bmatrix} p_1 & & 0 \\ & \ddots & \\ 0 & & p_n \end{bmatrix}, \quad \boldsymbol{Q} = \begin{bmatrix} q_1 & & 0 \\ & \ddots & \\ 0 & & q_n \end{bmatrix}, \tag{65}$$

Matrix entries p_i and q_i, $i = 1, \ldots, n$ are optimization variables.

Let us now turn back to the filter in Fig.6. We would like to optimize its noise performance using the original design discussed before as a reference. We assume that the filter topology is fixed and we shall only change RC element values. We shall employ transformation (62) with diagonal matrices $\boldsymbol{P}$ and $\boldsymbol{Q}$. In our case $n = 3$ and, since we want the matrix $\boldsymbol{C}$ to be invariant under transformation (62) we put $p_3 = 1$. In this example, we aim at minimizing integrated input noise, i.e.

$$V_{n.RMS} = \sqrt{\int_0^{f_{3dB}} S_{ni}(f)\, df} \tag{66}$$

We also impose four design constraints: (a) maximum conductance g_{max} connected to the output of each OPAMP (as discussed before, it cannot be too large because of the non-zero output resistance of the real OPAMP as well as power consumption constraints), (b) maximum total capacitance C_{max} (because of the chip area limit), and (c) maximum resistance [capacitance] ratio $R_{r.max}$ [$C_{r.max}$] (too large ratio causes matching problems). These constraints can be written for the filter in Fig. 6 as follows (denote $g_i = 1/R_i$, $i = 1, \ldots, 6$):

$$\begin{aligned} &g_2 \le g_{\max},\ g_3 + g_4 \le g_{\max},\ g_5 + g_6 \le g_{\max} \\ &2(C_1 + C_2 + C_3) \le C_{\max} \\ &\max\{C_1, C_2, C_3\} / \min\{C_1, C_2, C_3\} \le C_{r.\max} \\ &\max\{R_i : i = 1, ..., 6\} / \min\{R_i : i = 1, ..., 6\} \le R_{r.\max} \end{aligned} \tag{67}$$

The constraints can be easily verified due to the matrix formulation of the filter model, e.g. (a) is verified by comparing sums of columns of the matrix $\bar{\mathbf{G}}$ to g_{max}, etc.

Table 2 shows optimization results for different constraint settings. The results are obtained with the software written in C language and implementing the filter model

of Section 2, the noise evaluation formulas (Section 4) as well as proper optimization routines (we do not discuss them here for the sake of brevity). The chosen values of g_{max} and C_{max} correspond to actual values of maximal conductance connected to outputs of OPAMPs and total capacitance in the reference filter. Integrated noise for the reference filter is $V_{n.RMS} = 211\mu V$. The results in Table 2 show that by proper scaling of element values we can obtain significant improvement in the noise performance of the filter but also relax somehow design specifications for filter OPAMPs. The results also show the trade-off between design constraints and noise performance improvement: the more relax constraints, the better improvement. Usually, constraints such as g_{max}, C_{max} and others are set up according to the available OPAMPs as well as designer budget concerning power consumption, chip area, etc.

Table 2

Noise optimization results for the filter in Fig.6

Wyniki optymalizacji parametrów szumowych filtru z Rys.6

Constraints			Optimized filter data		
g_{max} [mS]	C_{max} [pF]	$R_{r.max}$	$C_{r.max}$	$V_{n.RMS}$ [μV]	Improvement over reference filter [dB]
0.2	no limit	5	5	139	3.6
0.2	no limit	10	10	129	4.3
0.2	no limit	20	20	121	4.8
no limit	5.82	5	5	114	5.3
no limit	5.82	10	10	102	6.3
no limit	5.82	20	20	98	6.7

6. CONCLUSIONS

In the paper, a general model of integrator-based Active-RC filter is introduced and analyzed using algebraic description. As a result, a matrix-based framework for creating efficient computer-aided analysis and design tools for Active-RC filters is developed. The extensions of the model are discussed that allow us to consider filters

with non-ideal OPAMPs with finite gain and bandwidth as well as non-zero output impedance. A procedure for evaluating noise in Active-RC filters in general setting is also given. The presented methods and procedures are verified by comparing theoretical results to SPICE simulations. Several application examples are discussed that indicate usefulness of the model to automated Active-RC filter analysis, design and optimization. The goal of the future work is to develop — within the presented approach — tools for investigating nonlinear effects in Active-RC filters.

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. K. Su: *Analog Filters.* Kluwer Academic Publishers, 2002
3. Y. Sun (Editor): *Design of high frequency integrated analogue filters.* The Institution of Electrical Engineers, London, 2002
4. S.-S. Lee: *Integration and System Design Trends of ADSL Analog Front Ends and Hybrid Line Interfaces.* Proc. IEEE Custom Integrated Circuits Conf., 2002, pp. 37-44
5. H. Weinberger, A. Wiesbauer, M. Clara, C. Fleischhacker, T. Potscher, B. Seger: *A 1.8V 450mW VDSL 4-Band Analog Front End IC in 0.18μm CMOS.* Proc. IEEE Solid-State Circuits Conf. ISSCC, 2002, Vol. 1, pp. 326-471
6. N. Tan, F. Caster, C. Eichrodt, S.O. George, B. Horng, J. Zhao: *A Universal Quad AFE with Integrated Filters for VDSL, ADSL, and G.SHDSL.* Proc. IEEE Custom Integrated Circuits Conf., 2003, pp. 599-602
7. J. Jussila, K. Halonen: *A 1.5 V active RC filter for WCDMA applications. Proc. IEEE Int. Conf. Electronics Circuits.* Syst. ICECS, 1999, Vol. 1, pp. 489-492
8. W. Khalil, T.-Y. Chang, X. Jiang, S.R. Naqvi, B. Nikjou, J. Tseng: *A highly integrated analog front-end for 3G.* IEEE J. Solid-State Circuits, 2003, Vol. 38, pp. 774-781
9. C. Frost, G. Levy, B. Allison: *A CMOS 2 MHz self-calibrating bandpass filter for personal area networks.* Proc. Int. Symp. Circuits Syst. ISCAS, 2003, Vol. 1, pp. 485-488
10. C. S. Wong: *A 3-V GSM baseband transmitter.* IEEE J. Solid-State Circuits, 1999, Vol. 34, pp. 725-730
11. J. Harrison, N. Weste: *350MHz opamp-RC filter in 0.18μm CMOS.* Electron. Lett., 2002, Vol. 38, pp. 259-260
12. R. Schaumann, M. A. Van Valkenburg: *Design of Active Filters.* Oxford University Press, New York, 2001
13. S. Kozieł, S. Szczepański, R. Schaumann: *General Approach to Continuous-Time* $G_m - C$ *Filters.* Int. J. Circuit Theory Appl., 2003, Vol. 31, pp. 361-383
14. P. J. Biliam: *Practical optimization of noise and distortion in Sallen and Key filter sections.* IEEE J. Solid-State Circuits, 1979, Vol. SC-14, pp. 768-771
15. P. Bowron, K. A. Mezher: *Noise and sensitivity optimization in the design of second-order single-amplifier filters.* Int. J. Circuit Theory Appl., 1991, Vol. 19, pp. 389-402
16. D.T. Nguyen, B.D. Miller: *State-space noise calculation for single-operational-amplifier RC filters.* IEE Proc. G, 1984, Vol. 131, pp. 114-118
17. L. Toth, G. Efthivoulidis, V. Gopinathan, Y. P. Tsividis: *General Results for Resistive Noise in Active RC and MOSFET-C Filters.* IEEE Trans. Circuits Syst.II, 1995, Vol. 42, pp. 785-793

S. KOZIEŁ

OGÓLNY MODEL FILTRU AKTYWNEGO RC CZASU CIĄGŁEGO DLA KOMPUTEROWO WSPOMAGANEGO PROJEKTOWANIA I OPTYMALIZACJI

Streszczenie

W pracy przedstawiono ogólny model filtru aktywnego RC opartego na wzmacniaczach odwracających fazę. Model analizowany jest przy użyciu równań macierzowych. Podano jawna formule określająca funkcje przenoszenia filtru w ogólnym przypadku. Przedstawiono rozszerzenia modelu pozwalające uwzględniać nieidealność wzmacniaczy operacyjnych, w tym skończone wzmocnienie i pasmo przenoszenia oraz niezerową rezystancję wyjściową. Wyprowadzono również ogólne formuły pozwalające wyznaczyć odniesione do wejścia widmo szumów dla dowolnego filtru rozważanej klasy. Otrzymane wyniki teoretyczne zostały zweryfikowane poprzez porównanie z wynikami otrzymywanymi za pomocą symulatora SPICE dla wybranych przykładów filtrów dolnoprzepustowych: filtru Czebyszewa 3-go rzędu oraz filtru Butterwortha 5-go rzędu.

Dzięki sformułowaniu modelu w języku algebry liniowej może on być łatwo zaadoptowany do użycia we wspomaganym komputerowo projektowaniu i optymalizacji filtrów aktywnych RC. Jako ilustracje, przedstawiono zastosowanie modelu do rozwiązania praktycznych problemów: określenia wymagań projektowych dla wzmacniacza operacyjnego jako elementu aktywnego w dolnoprzepustowym filtrze Czebyszewa trzeciego rzędu dla zastosowań w części analogowej modemu VDSL, oraz optymalizacji parametrów szumowych tegoż filtru.

Słowa kluczowe: filtry aktywne RC, filtry czasu ciągłego, analiza szumowa

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 powinno charakteryzować oryginalne ujęcie zagadnienia, własna klasyfikacja, krytyczna ocena (teorii lub metod), omówienie aktualnego stanu, lub postępu danej gałęzi techniki oraz omówienie perspektyw rozwojowych. Artykuły publikowane w innych czasopismach nie mogą być kierowane do druku w Kwartalniku Elektroniki i Telekomunikacji w drugiej kolejności zgłoszenia.
Objętość artykułu nie powinna przekraczać 30 stron po około 1800 znaków na stronie, 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. BIBLIIOGRAFIA
- 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 spacjowania, podkreślania i pisania tekstu dużymi literami z wyjątkiem wyrazów, które umownie pisze się dużymi literami (np. FORTRAN). Proponowane wyróżnienia Autor może zaznaczyć w maszynopisie zwykłym ołówkiem za pomocą przyjętych znaków adjustacyjnych, np. podkreślenie linią przerywaną oznacza spacjowanie (rozstrzelenie), podkreślenie linią ciągłą – pogrubienie, podkreślenie wężykiem — kursywa.

Tekst powinien być napisany z podwójnym odstępem między wierszami, tytuły i podtytuły małymi literami. Marginesy z każdej strony powinny mieć około 35 mm. 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 Electronical Commision) oraz ISO (International Organization of Standarization).

Powołania.
Powołania na publikacje powinny być umieszczone na ostatnich stronach tekstu pod tytułem „Bibliografia", opatrzone numeracją kolejną bez nawiasów. Numeracja ta powinna być zgodna z odnośnikami w tekście artykułu. Przykłady opisu publikacji:

- periodycznej — 1. F. Valdoni: A new milimetre wave satelite. 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 powinny być umieszczone 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.

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.

INFORMATION FOR AUTHORS OF K.E.T.

The editorial stuff 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 Scince, 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. TThe 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 stuff 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)
  - Extenstive summary, hawever not exceeching 3600 characters (along with keywords) in Polish, if artide wos 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 ar attributed to. Formulas should be numbered with Arabic numbers place 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 Electronical Commission) and ISO (International Organisation of Standardisation).

References

References should be placed at the and of the main text with th subtitle „References". Refrences should be numbered (without brackets) adequately to references placed in the text. Examples of periodical [@], non-periodical [2] and book [3] references:

1. F. Valdoni: A new milimetre wave satelite. 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. Tvidis: Operation and modeling of the MOS transistors. Nw York. McGraw-Hill. 1987. p. 553

Figures

Figures should be clearly drawn on plain or milimetre 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 whit format not gratd than 10×15 cm. On th margin of each drawing and on th back side of each ohoto authorp name and abbreviation of the title of article should be placed. Figure's captures should be given in two longuaqes (first in the language the article is writtes in and then in additional language). Figure's captures 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 entiled to free of charg 20 copies of articl. Additional copies or the whole magazine can be ordered at publisher at th one's expense.

Author is obliges to perform the author's correction, which should be accomplished within 3 days starting from the date of receiving of the text from the ditorial stuff. Corrected text should be return to the editorial stuff personally or by mail. Correction marks should be placed on th margin of copies received from the editorial stuff 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 stuff of the publisher.

In case of changing of workplace or home address Authors are asked to inform the editorial stuff.