POLSKA AKADEMIA NAUK
KOMITET ELEKTRONIKI I TELEKOMUNIKACJI

PL ISSN 0035-938
Nr Indeksu 36318

# KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJ

# ELECTRONICS AND TELECOMMUNICATIONS

# QUARTERLY

dawniej

ROZPRAWY ELEKTROTECHNICZNE

TOM XXXVIII — ZESZYT 1

WYDAWNICTWO NAUKOWE PWN
WARSZAWA 1992

POLSKA AKADEMIA NAUK
KOMITET ELEKTRONIKI I TELEKOMUNIKACJI

PL ISSN 0035-938
Nr Indeksu 36318

# KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJI

# ELECTRONICS AND TELECOMMUNICATIONS QUARTERLY

dawniej

ROZPRAWY ELEKTROTECHNICZNE

TOM XXXVIII — ZESZYT 1

WYDAWNICTWO NAUKOWE PWN

WARSZAWA 1992

ADRES REDAKCJI

*00-665 Warszawa, ul. Nowowiejska 15/19 Politechnika,*
*Instytut Telekomunikacji, Gmach im. prof. JANUSZA GROSZKOWSKIEGO*

*Dyżury Redakcji: środy i piątki, godz. 14—16*
*tel. 628 89 81 21007-737*

*Telefony domowe: Redaktora Naczelnego: 12 17 65*
*Zast. Red. Naczelnego: 26 83 41*
*Sekretarza Odpowiedzialnego: 25 29 18*



WYDAWNICTWO NAUKOWE PWN
Warszawa, ul. Miodowa 10

| | |
|---|---|
| Ark. wyd. 19,75 Ark. druk. 16,5 | Podpisano do druku w maju 1992 r. |
| Papier offsetowy kl. III 70 g. B-1 | Druk ukończono w czerwcu 1992 r. |

Skład: GP-BIT W-wa ul. Marymoncka 34. Druk i oprawa: Drukarnia — W-wa, ul. Hajoty 50

KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJI, 1992, 38, z. 1, ss. 3-15

(621.382.066: 681.3.04)

# Opis języka formalnego do projektowania analogowych układów CMOS

KRZYSZTOF WAWRYN

*Instytut Elektroniki, Wyższa Szkoła Inżynierska w Koszalinie*

*Otrzymano 1991.02.08*

*Autoryzowano do druku 1991.07.10*

W pracy przedstawiono nowe podejście do projektowania układów analogowych. Wiedza o układach analogowych CMOS jest opisana językiem formalnym. Zdaniami języka są klauzule zbudowane ze zdań logicznych nazywanych predykatami. Zdania reprezentują projektowany układ na określonym poziomie hierarchicznego podziału. Projektowanie jest przeprowadzane w procesie przekształcania zdań za pomocą gramatyki języka. Proces ten rozpoczyna się od zdania reprezentującego wymagania wejściowe, a kończy się na zdaniu reprezentującym finalny projekt. Przedstawiono prosty przykład projektowania inwertera z wykorzystaniem wprowadzonego języka. Pokazano także aplikację metody w projektowaniu wzmacniacza operacyjnego CMOS.

## WPROWADZENIE

Wzrost gęstości upakowania elementów w układach scalonych spowodował zwiększenie zapotrzebowania na opracowanie nowych, wydajniejszych metod projektowania wspomaganego komputerem. Duży stopień skomplikowania struktur scalonych układów analogowych zmusza projektantów do hierarchicznego podziału projektowanego układu na mniejsze łatwiejsze w projektowaniu podukłady. Proces projektowania układu o zadanych wymaganiach może odbywać się z dołu do góry (od elementów do układu) lub z góry na dół (od ogólnej struktury układu do połączenia elementów o określonych wartościach parametrów). Hierarchiczny proces projektowania z góry na dół składa się z wyboru struktury połączeń i wyznaczeniu wymagań struktur niższego poziomu na każdym poziomie podziału. Ten sposób projektowania jest wykorzystywany we wszystkich najnowszych rozwiązaniach opartych na systemach z bazą wiedzy [1-6]. Podstawową wadą tych rozwiązań jest brak jednolitej reprezentacji wiedzy o układach analogowych. Pod tym pojęciem rozumie się zbiór konwencji opisujących układy elektroniczne (np. schemat ideowy i charakterystyki elektryczne w reprezentacji graficznej). W programach projektowania układów cyfrowych reprezentacja układu jest zwykle

przedstawiana za pomocą języka opisującego obiekty i sposób przetwarzania przez te obiekty sygnału cyfrowego [7]. Język taki nie jest odpowiednią reprezentacją układu i sygnału analogowego. Dotychczasowe próby zbudowania uniwersalnego języka opisu układów analogowych nie dały zadowalających rezultatów [8-10]. Większość języków w narzędziowych programach projektowych używa nazw elementów oraz numerów węzłów do opisu struktury układu. Układ scalony CMOS jest więc reprezentowany jako lista tranzystorów typu *n* oraz *p* i innych elementów, oraz numerów węzłów do których są podłączone ich końcówki. Np. każdy tranzystor w symulatorze SPICE [11] jest reprezentowany przez następującą listę:

<nazwa> <bramka> <źródło> <dren> <podłoże> <typ>
<lista parametrów>

gdzie <nazwa> jest symbolem tranzystora w schemacie ideowym, cztery kolejne argumenty podają numery węzłów do których są podłączone: bramka, źródło, dren i podłoże, <typ> określa typ tranzystora (*n* albo *p*), a <lista parametrów> parametry technologiczne tranzystora. Ta reprezentacja jest prosta lecz nieprzydatna w systemach z bazą wiedzy, ponieważ nie dostarcza wszystkich informacji koniecznych do zaprojektowania układu tj. jak układ działa czy jak należy wyznaczać wymagania podukładów i wartości elementów z zadanych wymagań, aby uzyskać końcowy projekt układu. Informacje takie jak to pokazano w pracy mogą być uwzględnione w języku formalnym układów analogowych opartym na reprezentacji predykatowej [12] i oznaczanym w dalszych rozważaniach $L_{ACDL}$ [13].

## 2. DEFINICJA JĘZYKA FORMALNEGO DO OPISU UKŁADÓW ANALOGOWYCH

W tym rozdziale wprowadzono notację języków formalnych w opisie języka $L_{ACDL}$. Język $L_{ACDL}$ posiada dwa słowniki końcowy i pomocniczy. Każdy słownik składa się ze słów zwanych symbolami. Słownik końcowy oznaczany $\Sigma$ składa się z symboli reprezentujących elementy takie, jak rezystory, kondensatory i tranzystory, ich wartości i sposób ich połączenia. Słownik pomocniczy oznaczany N składa się z symboli reprezentujących struktury z wyższych poziomów hierarchicznego podziału takich jak system, blok funkcjonalny, układ elementarny, podukład oraz informacji określających ich charakterystyki elektryczne.

Ciąg symboli z powtórzeniami ze słownika $V=\Sigma+N$ nazywa się zdaniem. Zdanie reprezentuje projektowany układ na określonym poziomie hierarchicznego podziału. Hierarchiczny proces wyboru struktury i wyznaczania wymagań prowadzący do układu zbudowanego z sieci połączeń indywidualnych elementów jest reprezentowany jako przekształcanie zdania początkowego w zdanie końcowe. Zdanie początkowe zwane symbolem startowym *S*, składa się w większości lub całkowicie z symboli pomocniczych opisujących projektowany układ przy pomocy jego nazwy i wymagań wejściowych. Zdanie końcowe składa się jedynie z symboli końcowych ze słownika

$\Sigma$ i reprezentuje żądany rezultat tj. projektowany układ opisany przez elementy, ich połączenia i wartości ich parametrów.

Podstawowymi elementami języka są symbole ze słownika $V=\Sigma+N$ i zdania zbudowane z tych symboli. Słownik końcowy $\Sigma$ zawiera tylko stałe, podczas gdy słownik pomocniczy N zawiera zmienne. Stałe są symbolami końcowymi, a zmienne pomocniczymi pojawiającymi się w zdaniach pośrednich prowadzących do zdań zbudowanych jedynie z symboli końcowych przez zastąpienie wszystkich zmiennych stałymi. Zbiór wszystkich zdań składających się z symboli ze słownika $V$ jest oznaczany przez $V^*$. Jeżeli zdanie $y \in V^*$ jest wyprowadzalne ze zdania $x \in V^*$ to zapisuje się to $x^* \rightarrow y$.

Do przekształcania zdań użyto gramatykę struktur zdaniowych nazywaną także gramatką kombinatoryczną [14]. Ogólnie gramatyka jest mechanizmem służącym do przekształcania zdań. Używa ona zbioru reguł nazywanych listą produkcji $P$.

*Gramatyką* struktur zdaniowych nazywa się system

$$G = <V, \Sigma, P, S> \qquad (1$$

gdzie:

$V$ – skończony niepusty zbiór nazywany słownikiem,

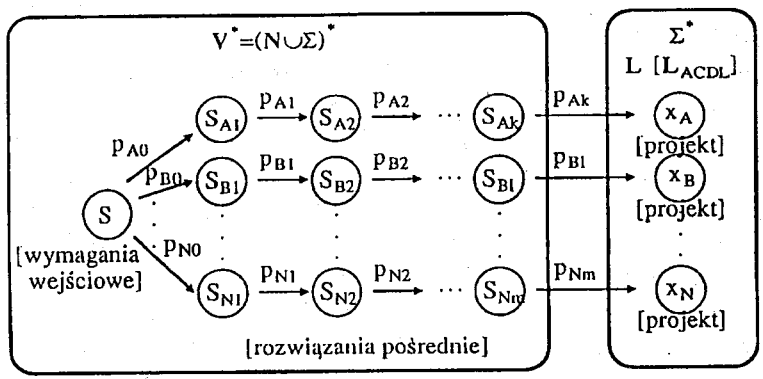
$\Sigma \subseteq V$ – skończony niepusty zbiór nazywany słownikiem końcowym,

$N = V - \Sigma$ – skończony niepusty zbiór nazywany słownikiem pomocniczym.

$S \in N$ – wyróżniony symbol pomocniczy nazywany symbolem początkowym,

$P$ – skończony zbiór reguł nazywany listą produkcji.

Językiem definiowanym lub generowanym przez daną gramatykę G nazywa się zbiór zdań (rys. 1.), który oznacza się przez L (G) i który składa się ze wszystkich zdań




Rys. 1. Graficzna reprezentacja generacji języka L(G) [$L_{ACDL}$]; wszystkie reguły p z indeksami [reguły przekształcania projektu] należą do zbioru reguł produkcji P

zbudowanych z symboli słownika końcowego $\Sigma$ wyprowadzalnych z symbolu $S$ przez reguły $P$ gramatyki G. Stąd otrzymuje się:

$$L(G) = \{x: x \in \Sigma^* \text{ i } S \overset{*}{\underset{G}{\Rightarrow}} x\} \qquad (2)$$

W tym miejscu można zauważyć rolę słownika pomocniczego $N$. Symbole z $N$ nie pojawiają się w zdaniach języka L (G), występują natomiast w wyprowadzeniach tych zdań. Obecność słownika pomocniczego $N$ tworzy pewnego rodzaju metajęzyk, w którym język L(G) jest definiowany.

Każdy język posiada dwie kategorie: syntaktykę i semantykę. Syntatyka czyli składnia definiuje elementy języka (symbole i zdania) i jego strukturę przy pomocy gramatyki G, podczas gdy semantyka definiuje znaczenie elementów języka. Zdania przyjętej reprezentacji są zbudowane z predykatów.

Predykat, nazywany funkcją zdaniową, jest wyrażeniem w postaci zdania logicznego składającego się z nazwy predykatu określającego relację między argumentami reprezentującymi obiekty z ustalonego zbioru. Argumenty są nazywane termami. Termami mogą być stałe (symbole), zmienne i wyrażenia. W wyniku podstawienia stałych za zmienne albo związania zmiennych kwantyfikatorami otrzymuje się zdanie prawdziwe, lub fałszywe. Przyporządkowanie obiektom symboli, a nazwie relacji między obiektami definiuje semantykę języka rachunku predykatów. Zaletą rachunku predykatów są proste i zrozumiałe interpretacje wyrażania zdań.

## 3. ZASTOSOWANIE JĘZYKA $L_{ACDL}$ DO OPISU UKŁADÓW ANALOGOWYCH

Język formalny do opisu układów analogowych zbudowano w oparciu o logikę predykatów i zastosowano do projektowania układów analogowych CMOS [12-13]. Obie kategorie języka: syntaktyka i semantyka są wprowadzone jednocześnie. Dla uproszczenia rozważań początkowo będą dotyczyły one jedynie struktury, następnie są rozszerzone o informacje dotyczące wyznaczania wymagań struktur niższego poziomu podziału.

W opisie struktur układów CMOS [15] argumentom predykatów przyporządkowuje się obiekty, którymi są symbole oznaczające wejścia, wyjścia oraz doprowadzenia napięć zasilania układu. Zostały one opisane symbolami używanymi dla opisu sygnałów wejściowych, wyjściowych, napięć i prądów źródeł i napięć zasilania np. $V_{in}$, $V_{out}$, $V_{gg}$, $i_p$, $V_{ss}$ i $V_{dd}$, ale nie mogą one być interpretowane jako napięcia czy prądy. Przyczyną użycia notacji stosowanej w opisie sygnałów do opisu predykatowego jest ułatwienie zrozumienia interpretacji zdań zbudowanych z predykatów jako alternatywnej formy przedstawiania struktury układu analogowego. Dla rozróżnienia znaczenia opisu symbole będące argumentami predykatów są opisane pogrubioną czcionką, podczas gdy rzeczywiste napięcia czy prądy będą opisane zwykłą czcionką. Elementy indywidualne takie jak tranzystory MOS typu $n$ oraz $p$ oraz komplementarna struktura CMOS zostały opisane następującymi predykatami:

***dn***$(a, b)$ – dren tranzystora typu $n$ ($a$ – bramka, $b$ – źródło),
***dp***$(a, b)$ – dren tranzystora typu $p$ ($a$ – bramka, $b$ – źródło).
***out***$(x, y)$ – wyjście $x$ z podukładu typu $n$ połączone z wyjściem $y$ z podukładu typu $p$.

Gramatykę $G = <V, \Sigma, P, S>$ wraz z semantyką predykatów języka $L_{ACDL}$ wykorzystano w opisie struktury układów CMOS. Dla wygody symbole słownika

końcowego są pisane pogrubioną trzcionką, a symbole słownika pomocniczego są zawarte między symbolami: < i >. Gramatyka zdań tworzy system reguł $P$ pokazujących jak symbole pomocnicze są przekształcane w celu zbudowania zdania reprezentującego strukturę i składającego się jedynie z symboli końcowych, jakimi są tranzystory MOS, wejścia, wyjścia i doprowadzenia napięć zasilania.

Słownik końcowy składa się z symboli: $V_{ss}$ i $V_{dd}$ oznaczających doprowadzenia napięć zasilania, $I$ zbiór wejść, $A$ zbiór obiektów pomocniczych, $d$ diodowe połączenie tranzystora, ***dn*** i ***dp*** dreny tranzystorów typu $n$ oraz $p$, ***out*** wyjście struktury komplementarnej CMOS, a także nawiasów "(" i ")", oraz przecinka ",". Stąd

$$\Sigma = \{V_{ss}, V_{dd}, I, A, d, (, , ,), dn, dp, out\}.$$

Obiekt pomocniczy najczęściej reprezentuje wyjście z układu lub podukładu (np. inwerter, źródło prądowe, układ napięcia odniesienia itp.). W sytuacji, gdy używamy makromodeli do symulacji projektowanego układu, obiekt pomocniczy traktuje się jako element indywidualny układu, a więc jest elementem słownika końcowego gramatyki. Natomiast, gdy wymaga się, aby także obiekty pomocnicze były opisane przy pomocy połączeń jedynie tranzystorów typu $n$ i $p$, wówczas obiekty pomocnicze stają się zbiorem zmiennych $<A>$ i wchodzą do słownika pomocniczego a nie końcowego. Taka elastyczność doboru sygnałów pomocniczych ułatwia opis połączeń układów na różnych poziomach hierarchicznego podziału układu.

Zmiennymi pomocniczymi są <układ_CMOS>, <podukład_$n$> i <podukład_$p$>, które opisują struktury sieci tranzystorów. Zmienne <podukład_$n$> i <podukład_$p$> opisują sieci tranzystorów typu $n$ i typu $p$, a zmienna <układ_CMOS> jest ich połączeniem i jednocześnie symbolem początkowym. Stąd:

$$N = \{<\text{układ\_CMOS}>, <\text{podukład\_}n>, <\text{podukład\_}p>\},$$
$$S = \{<\text{układ\_CMOS}>.$$

Dla wygody interpretacji znaczenie zmiennych jest określone przez ich nazwy. Ta konwencja jest przyjęta we wszystkich dalszych rozważaniach.

Alternatywną formą opisu reguł produkcji $P$ gramatyki G jest system równań normalnych Backusa (BNF). Ta forma używa cztery meta symbole, które nie wchodzą do słownika: "<", ">", "::=" i "|". Łańcuchy zawarte między symbolami < i > oznaczają zmienne. Symbol ::= oznacza operator podstawiania, a symbol | oznacza alternatywę. Lista produkcji $P$ dla układów CMOS w opisie BNF przedstawia się następująco:

<*układ*_CMOS> ::= **out** (<podukład_$n$>, <podukład_$p$>)

<*podukład*_$n$> ::= **dn** (I,$V_{ss}$) | **dn** (I, <podukład_$n$>) |
**dn** (I, A) | **dn** (d, <podukład_$n$>) |
**dn** (A, $V_{ss}$) | **dn** (A, <podukład_$n$>) |
**dn** (<układ_CMOS>, <podukład_$n$>)

<*podukład*_$p$> ::= **dp** (I,$V_{dd}$) | **dp** (I, <podukład_$p$>) |
**dp** (I, A) | **dp** (d, <podukład_$p$>) |
**dp** (A, $V_{dd}$) | **dp** (A, <podukład_$p$>) |
**dp** (<układ_CMOS>, <podukład_$p$>).

Językiem tej gramatyki jest zbiór zdań tj. struktur układów CMOS. Jest on nieskończony, ponieważ lista produkcji zawiera rekurencyjne reguły tworzenia zdań języka. Ponadto gramatyka ta umożliwia proste powiększanie listy produkcji $P$.

## 4. PROJEKTOWANIE UKŁADÓW ANALOGOWYCH CMOS

Projektowanie układów w oparciu o przedstawiony język $L_{ACDL}$ polega na przekształcaniu zdania początkowego będącego symbolem początkowym w zdanie końcowe będące finalnym projektem układu. Jednakże zdania wygenerowane gramatyką G interpretowane jako struktura układu CMOS na poziomie tranzystora nie gwarantują, że otrzymana struktura jest układem realizowanym praktycznie ze względu na zadane wymagania. Problem ten można rozwiązać poprzez każdorazowe testowanie spełnienia wymagań dla wygenerowanej struktury, rozpoczynając generację od struktury najprostszej. Niestety ten sposób poszukiwania nazywany generacją i testowaniem, poza wyznaczeniem najprostszej struktury ralizującej zadane wymagania, nie ma cech rozumnego działania i szybko prowadzi do eksplozji kombinatorycznej. Stąd zastosowanie tej metody poszukiwania jest ograniczone do najprostszych układów.

Wynika stąd wniosek, że tak zdefiniowany język jest nieprzydatny do projektowania dopóki, sposób generacji struktur jest niezależny od zadanych wymagań i nie określa parametrów lub wartości elementów końcowych. Nie istotny jest fakt, czy wartości te będą wyznaczane w trakcie poszukiwania, czy też później przez prostą procedurę numeryczną. Przy zadanej technologii projektowanie układu CMOS sprowadza się do wyznaczenia struktury i określenia stosunku W/L, a listę produkcji o reguły wyboru struktury w zależności od zadanych wymagań.

Uwzględniając powyższe rozważania słowniki $\Sigma$ i $N$ oraz symbol początkowy $S$ przedstawiają się następująco:

$\Sigma$ = {$\mathbf{V_{ss}}$, $\mathbf{V_{dd}}$, **I**, **A**, **d**, (, , ,), **dn**, **dp**, **out**, **W/L**},

$N$ = {<układ_CMOS (wymagania)>,
<podukład_$n$ (wymagania_$n$)>, <podukład_$p$ (wymagania_$p$)>}

$S$ = <układ_CMOS (wymagania)>

gdzie wymagania, wymagania_$n$ i wymagania_$p$ oznaczają odpowiednio wymagania układu, podukładu typu $n$ i podukładu typu $p$, a lista produkcji $P$ dla układów CMOS w opisie BNF przedstawia się następująco:

<układ_CMOS (wymagania)> ::=
**out** (<podukład_$n$ (wymagania_$n$)>,<podukład_p (wymagania_p)>)

<podukład_$n$ (wymagania_$n$)> ::=
**dn** (**I**, $\mathbf{V_{ss}}$, **W/L**) | **dn** (**I**, <podukład_$n$ (wymagania_$n$)>, **W/L**) |
**dn** (**I**, **A**, **W/L**) | **dn** (**d**, <podukład_$n$ (wymagania_$n$)>, **W/L**) |
**dn** (**A**,$\mathbf{V_{ss}}$, **W/L**) | **dn** (**A**, <podukład_$n$ (wymagania_$n$)>, **W/L**) |
**dn** (<układ_CMOS (wymagania)>, <podukład_$n$ (wymagania_$n$)>, **W/L**) |

<podukład_p (wymagania_p)>::=
**dp (I, $V_{dd}$, W/L)** | **dp (I,** <podukład_p (wymagania_p)>, **W/L)** |
**dp (I, A, W/L)** | **dp (d,** <podukład_p (wymagania_p)>, **W/L)** |
**dp (A,$V_{dd}$, W/L)** | **dp (A,** <podukład_p (wymagania_p)>, **W/L)** |
**dp (**<układ_CMOS ( wymagania)>, <podukład_p (wymagania_p)>, **W/L)** |

Przedstawiona gramatyka zawiera ogólne reguły tworzenia układów CMOS bez określania ich rodzaju. Szczegółowe reguły odnoszące się do konkretnych układów tworzy się w oparciu o wiedzę doświadczonego projektanta w postaci heurystyk oraz z analizy równań wiążących poszczególne parametry projektowanego układu. Reguły te wchodzą w skład listy produkcji, która stanowi dane zgromadzone w bazie wiedzy inteligentnego systemu projektowego układów CMOS.

## 5. PRZYKŁAD

Ilustracją procedury projektowej jest prosty inwerter CMOS. Procedura projektuje inwerter o zadanych wymaganiach: wzmocnieniu napięciowym $A$, rezystancji wyjściowej $R_{out}$ [kΩ] i prądzie drenu $I_D$ [$\mu A$]. Do opisu wykorzystuje się predykaty o nazwach: projekt, inwerter, podukład_sterujący_$n$, obciążenie_$p$, **out**, **dn** i **dp**. Gramatyka G wykorzystująca te predykaty jest następująca:

$\Sigma$ = {$V_{ss}$, $V_{dd}$, $V_{in}$, **d**, (,) ,,, **dn, dp, out, W/L**},
$N$ = { <projekt ($A$, $R_{out}$, $I_D$)>, <inwerter ($A$, $R_{out}$, $I_D$)>,
<podukład_sterujący_$n$ ($G_{m1}$, $I_1$)>, <obciążenie_$p$ ($G_{m2}$, $I_2$)>}
$S$ = <projekt ($A$, $R_{out}$, $I_D$)>.

a lista produkcji $P$

1) <projekt ($A$, $R_{out}$, $I_D$)> ::= <inwerter ($A$, $R_{out}$, $I_D$)>,
2) <inwerter ($A$, $R_{out}$, $I_D$)> ::=
**out (**<podukład_sterujący_$n$ ($A/R_{out}$, $I_D$)>,
<obciążenie_$p$ ($1/R_{out}$, $I_D$)>**)**,
3) <podukład_sterujący_$n$ ($G_{m1}$, $I_1$)> ::=
**dn (**$V_{in}$, $V_{ss}$, $G_{m1}*G_{m1}/(2*17*I_1)$**)**,
4) <obciążenie_$p$ ($G_{m2}$, $I_2$)> ::=
**dp (d,** $V_{dd}$, $G_{m2}*G_{m2}/(2*8*I_2)$**)**.

gdzie $G_{m1}$ i $G_{m2}$ [$\mu$S] oznaczają transkonduktancje tranzystorów, a liczby 17 i 8 oznaczają parametry transkonduktancyjne w nasyceniu $K'_{sat}$ [$\mu$A/V$^2$] tranzystorów typu $n$ i $p$.

Przykładowo jeżeli zadaniem jest zaprojektowanie inwertera opisanego przytoczoną gramatyką spełniającego następujące wymagania: $A = 6.52$, $R_{out} = 70$kΩ, $I_D = 50\mu$A, to symbol początkowy przedstawia się następująco:

$S$ = <projekt (6.52, 70, 50)>

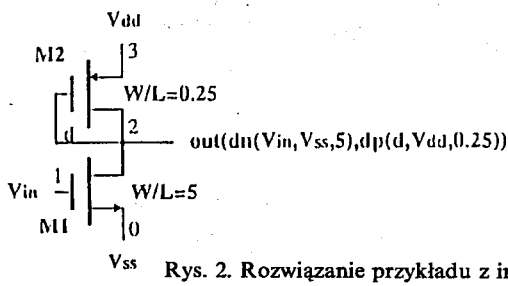
Zastosowanie kolejno reguł: nr1, nr2 i nr3 prowadzi do:

<projekt (6.52, 70, 50)> ::= <inwerter (6.52, 70, 50)>,
<projekt (6.52, 70, 50)> ::= **out** (<podukład_sterujący_$n$ (92, 50)>,
<obciążenie_$p$ (14.1, 50)>),
<projekt (6.52, 70, 50)> ::=
**out (dn** ($V_{in}$, $V_{ss}$, **5**), <obciążenie_$p$ (14.1, 50)>).

Zastosowanie reguły nr 4 daje rozwiązanie:

<projekt (6.52, 70, 50)> ::= **out (dn** ($V_{in}$, $V_{ss}$, 5), **dp (d,** $V_{dd}$, 0.25)).

Rozwiązanie jest pokazane na rys. 2.




Rys. 2. Rozwiązanie przykładu z inwerterem

Projektowane zwykle układy wymagają list produkcji zawierających setki lub tysiące reguł.

Przedstawiona w pracy reprezentacja układów analogowych CMOS do projektowania może być po niewielkich zmianach zaimplementowana w bazie wiedzy systemu ekspertowego napisanego w języku Prolog [16-17] w wersji Turbo. Przykładowo reguły produkcji przedstawione w przykładzie projektu inwertera mogą być przedstawione w notacji klauzulowej używanej w Turbo Prologu w sposób następujący:

projekt ($A$, $R_{out}$, $I_D$, $L$) :–
inwerter ($A$, $R_{out}$, $I_D$, $L$).

inwerter ($A$, $R_{out}$, $I_D$, $L$) :–
$G_{m2} = 1/R_{out}$,
$G_{m1} = A/R_{out}$,
podukład_sterujący_$n$ ($G_{m1}$, $I_D$, $L_n$),
obciążenie_$p$ ($G_{m2}$, $I_D$, $L_p$),
$L$=**out** ($L_n$, $L_p$).

podukład_sterujący_$n$ ($G_m$, $I$, $L_n$) :–
$S = G_m * G_m/(2*17*I)$,
$L_n$=**dn** ($V_{in}$, $V_{ss}$, $S$).

obciążenie_$p$ ($G_m$, $I$, $L_p$) :–
$S = G_m * G_m/(2*8*I)$,
$L_p$=**dp** (**d**, $V_{dd}$, $S$).

gdzie: pomocnicze listy $L$, $L_n$ i $L_p$ opisują struktury: inwertera, podukładu sterującego i obciążenia.

## 6. APLIKACJA

Metoda projektowania w oparciu o język $L_{ACDL}$ została wykorzystana do projektowania wzmacniaczy operacyjnych CMOS o zadanych wymaganiach projektowych. W opisie tranzystora MOS oraz parametrów elektrycznych charakteryzujących pracę wzmacniaczy używa się modeli elementów i zależności matematycznych wiążących parametry elektryczne różnych struktur wzmacniaczy szczegółowo opisanych w pracy [18]. Wiedza wynikająca z analizy tych zależności wraz z heurystyczną wiedzą projektanta zostały zgromadzone w bazie wiedzy. Szczegółowe reguły reprezentujące tę wiedzę są zbudowane na podstawie ogólnych reguł opisanych w języku $L_{ACDL}$ i przedstawionych w rozdziale 4. Przykładowe reguły projektowe dwustopniowego wzmacniacza operacyjnego CMOS (rys. 3) w opisie klauzulowym w Turbo Prologu przedstawiają się następująco:

```
wzmacniacz_operacyjny (Av, GB, CMR+, CMR-, Cl, Sr, Vout+, Vout-, Pd) :-
        wyznacz (Cc, I5, I2, Gm2),
        n_elementy_sterujące (Gm2, I2, L1, L2),
        wyznacz (Vds3),
        p_obciążenia (Vds3, I2, Vdd, L1, L4),
        wyznacz (Gm6, Vds6, I6),
        p_element_sterujący (Gm6, I6, L2, L4, L6),
        wyznacz (Vds5),
        n_źródło_prądowe (Vds5, I5, I6, L5, L7),
        L=out (L7, L6).

n_elementy_sterujące (Gm, I, L1, L2) :-
        oblicz_W_do_L ("gm", Gm, I, 17, S),
        L1=dn (S, Vin-, ip),
        L2=dn (S, Vin+, ip),

n_elementy_sterujące (Gm, I, L1, L2) :-
        wyznacz (Gm2),
        oblicz_W_do_L ("gm", Gm2, I, 17, S),
        n_elementy_sterujące (Gm, I, L1c, L2c),
        L1=dn (S, VBN, L1c),
        L2=dn (S, VBN, L2c).

p_obciążenia (Vsat, I, Lp, L1, L4) :-
        oblicz_W_do_L ("Vsat", Vsat, I, 8, S),
        L3=dp (d, Lp, S),
        L4=dp (out (L1, L3), Lp, S).

p_obciążenia (Vsat, I, Lp, L1, L4) :-
        wyznacz (Gm4, Vsat1),
        oblicz_W_do_L ("Vsat", Vsat, I, 8, S),
        p_obciążenia (Vsat, I, dp (VBP, Vdd, S), L1, L4c),
        L4=dp (VBP, L4c, S).
```

$p_$element$_$sterujący $(G_m, I, L_2, L_4, L_6)$ :–
oblicz$_W_$do$_L$ ("$g_m$", $G_m$, $I$, 8, $S$),
$L_6$ = **dp** (**out** $(L_2, L_4)$, $V_{dd}$, $S$).

$p_$element$_$sterujący $(G_m, I, L_2, L_4, L_6)$:–
wyznacz $(G_{mc})$,
oblicz$_W_$do$_L$ ("$g_m$", $G_m$, $I$, 8, $S$),
$p_$element$_$sterujący $(G_{mc}, I, L_2, L_4, L_{6c})$,
$L_6$ = **dp** ($V_{BP}$, $L_{6c}$, $S$).

$n_$źródło$_$prądowe $(V_{sat}, I_5, I_7, L_5, L_7)$ :–
oblicz$_W_$do$_L$ ("$V_{sat}$", $V_{sat}$, $I_5$, 17, $S$),
$L_5$ = **dn** ($V_{gg}$, $V_{ss}$, $S$),
wyznacz $(S_7)$,
$L_7$ = **dn** ($V_{gg}$, $V_{ss}$, $S_7$),

$n_$źródło$_$prądowe $(V_{sat}, I_5, I_7, L_5, L_7)$ :–
wyznacz $(G_{m7})$,
oblicz$_W_$do$_L$ ("$V_{sat}$", $V_{sat}$, $I_5$, 17, $S$),
$n_$źródło$_$prądowe $(V_{sat}, I_5, I_7, L_5, L_{7c})$,
$L_7$ = **dn** ($V_{BN}$, $L_{7c}$, $S$).

oblicz$_W_$do$_L$ ("$g_m$", $G_m$, $I$, $K$, $S$):–
$S = G_m * G_m$ / $(2 * K * I)$,
$W_$do$_L_$graniczne $(S)$.

oblicz$_W_$do$_L$ ("$V_{sat}$", $V_{sat}$, $I$, $K$, $S$):–
$S = 2 * I/(K * V_{sat} * V_{sat})$,
$W_$do$_$L$_$graniczne (S).

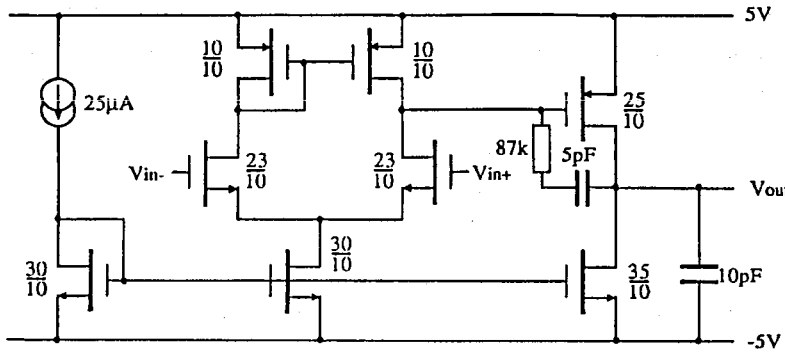
$W_$do$_L_$graniczne $(S)$ :–
$S < 100$,
$S > 1/10$.

gdzie: lista $L$ opisuje rozwiązanie, lista $L_p$ jest listą pomocniczą, listy $L_1,\ldots,L_7$ opisują pojedyncze tranzystory, a tworzone struktury kaskodowe są oznaczane indeksem $c$. Znaczenie wszystkich wielkości elektrycznych takich jak np. prądy można znaleźć w pracy [18]. Ponadto został wprowadzony predykat „*wyznacz*", aby nie pisać wzorów do obliczania wartości parametrów występujących wśród argumentów tego predykatu. W zbiorze reguł brakuje także reguł określających dopuszczalne wartości np. prądów, napięć i innych parametrów elektrycznych. Są one analogiczne do reguły określającej wartości graniczne parametru $W$ do $L$.

Baza wiedzy zawiera również wiedzę o strukturach innych wzmacniaczy CMOS. Ponadto w trakcie projektowania podukłady każdej ze struktur mogą być rekursyjnie powiększane przez stosowanie struktur kaskodowych. Zastosowanie ich powoduje niewielkie zmiany we wzorach parametrów wyznaczanych w sposób symboliczny.

Projektowanie wzmacniacza operacyjnego CMOS jest procesem wnioskowania rozwiązań na podstawie wiedzy o układach, podukładach wzmacniacza i tranzystorów

MOS i ich charakterystykach opisanych językiem $L_{ACDL}$ i zgromadzonych w bazie wiedzy. Rozpoczyna się ona od zadanych wymagań i kończy się na strukturze połączeń tranzystorów MOS i wyznaczeniu wartości parametru geometrycznego $W/L$ każdego tranzystora.




Rys. 3. Dwustopniowy wzmacniacz operacyjny CMOS

Projekt wzmacniacza operacyjnego CMOS został wykonany dla wymagań projektowych przedstawionych w tablicy. Są one spełnione przez strukturę wzmacniacza dwustopniowego pokazaną wraz z wartościami $W/L$ poszczególnych tranzystorów MOS na rys. 3. Wyniki symulacji otrzymanego projektu programem SPICE w celu porównania z zadanymi wymaganiami projektowymi są przedstawione w tablicy.

| Parametr | Jednostka | Wymagania | SPICE |
|---|---|---|---|
| $A_v$ | dB | 65 | 68 |
| GB | MHz | 1 | 1.06 |
| $\Phi$ | Deg | 80 | 93 |
| $S_r$ | V/μs | 5 | 5.2 |
| $P_d$ | mW | 0.8 | 0.6 |
| CMRR | dB | 80 | 80 |
| $CMR_+$ | V | 3.5 | 3.82 |
| $CMR_-$ | V | −3.5 | −4.17 |
| $C_L$ | pF | 10 | – |

## PODSUMOWANIE

W pracy przedstawiono reprezentację wiedzy o układach analogowych umożliwiającą całkowitą automatyzację procesu projektowania wykonywanego dotychczas w znacznej mierze bez użycia komputera. Zaproponowana nowa reprezentacja układów analogowych w postaci języka formalnego zbudowanego w oparciu o logikę predykatów daje możliwość jednolitego opisu pozwalającego na jednoczesne przekształcanie ogólnej struktury układu wraz z zadanymi wymaganiami w kolejne bardziej szczegółowe struktury. Dostarcza ona także prostej interpretacji zdań języka

wynikającej z zastosowania predykatów do ich budowy. Jak wykazano ten rodzaj reprezentacji dopuszcza stosowanie funkcjonalnych, analitycznych jak i heurystycznych opisów wiedzy o układach i ich projektowaniu i może być wykorzystany w programach pisanych w języku Prolog. Umożliwia to zbudowanie w pełni automatycznego systemu z bazą wiedzy do projektowania układów analogowych. Budowa prototypowego systemu do projektowania układów analogowych CMOS jest przedmiotem pracy autora [19], w którym proces projektowania jest realizowany jako proces wnioskowania projektów z wiedzy o układach, podukładach i ich parametrach elektrycznych.

## BIBLIOGRAFIA


1. DeGrauwe, M., Nys O., Vittoz E., Dijkstra E., Rijmenants J., Cserveny S., Sanchez J.: *An analog expert design system.* Proc. of the 1987 Int'l Solid State Circuit Conference, February 1987, pp. 212-213
2. Harjani R., Rutenbar R.A., Carley L.R.: *A prototype framework for knowledge based analog circuit synthesis.* Proc. 24th ACM/IEEE Design Automation Conf., June 1987, pp. 42-49
3. Koh H. Y., Sequin C. H., Gray P. R.: *OPASYN: A compiler for CMOS operational amplifiers,* IEEE Trans. on Computer Aided Design, Vol. 9, No. 2, 1990, pp. 113-125
4. El – Turky F. M., Nordin R. A.: *BLADES: An artificial intelligence approach to analog circuit design.* IEEE Trans. on Computer Aided Design, Vol. 8, No. 6, 1989, pp. 680-692
5. Sheu B.J., Lee J.C., Fung A.H.: *Flexible architecture approach to knowledge based analogue IC design,* IEE Proceedings, Vol. 137, Pt. G, No. 4, 1990, pp. 266-274
6. Toumazou C., Makris C.A., Berrah C.M.: *ISAID — A methodology for automated analogue IC design.* Proc. Int. Sym. on CAS, 1990, pp. 531-535
7. Lipsett R., Schaefer C., Ussery C.: *VHDL: hardware description and design.* Kluwer Academic Publishers, Boston, 1988
8. Boute R. T.: *System semantics, and formal circuit description.* IEEE Trans. on Circuits and Systems, Vol. CAS-33, No. 12, December 1986, pp. 1219-1231
9. Reeuvijk C. Middelhoek: *A new language to desccribe analog circuits.* Proc. Midwest Symp. on CAS, 1988, pp. 19-22
10. Kurker C.M., Paulos J.J., Cohen B.S., Cooley E.S.: *Developement of an analog hardware description language.* Proc. IEEE Custom Integrated Conference, 1990, pp. 5.4.1-5.4.6
11. Nagel L.: *SPICE 2: A computer program to simulate semiconductor circuits.* EECS/ERL M572, University of California, Berkeley, May 1975
12. Wawryn K.: *Predykatowa reprezentacja topologii w projektowaniu układów CMOS.* XIII KKTOi-UE, Bielsko-Biała, 1990, ss. 193-198
13. Wawryn K.: *Opis języka formalnego do reprezentacji układów CMOS.* Zeszyty naukowe, WSI Koszalin, 1991
14. Chomsky N.: *Syntactic structures.* Hague 1957
15. Wu C., Wójcik A., Ni L.: *A rule based circuit representation for automated circuit design.* Proc. IEEE Design Automation Conference, 1987, pp. 786-792
16. Clocksin W. F., Mellish C. S.: *Programming in PROLOG.* Springer Verlag, 1984
17. Kowalski R.: *Predicate logic as programming language.* Information Processing 74, North Holland Publishing Company, 1974, pp. 569-574
18. Allen P. E., Holberg D. R.: *CMOS analog circuit design.* Holt, Rinehart and Winston, Inc, 1987
19. Wawryn K.: *An artificial intelligence approach to analog circuit design.* International Journal of Circuits, System and Computers, 1991, vol. 1, No 2, pp. 149-176

K. WAWRYN

# FORMAL LANGUAGE DESCRIPTION FOR CMOS ANALOG CIRCUIT DESIGN


## Summary

A new approach to analog circuit design based on formal language analog CMOS circuit representation is proposed. This representation uses predicate sentences. The sentence represents designed circuit on certain level of hierarchy. The design process is performed as series of transformations done by a language grammar to obtain sentence representing final design from sentence representing input specifications. A design example and its Prolog implementation are presented.

KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJI, 1992, 38, z. 1, ss. 17-33

621.3.049
681.3.06

# Metoda komputerowej identyfikacji dla wybranej klasy układów nieliniowych rzędu drugiego

KRZYSZTOF BASZCZYŃSKI

*Centralny Instytut Ochrony Pracy, Łódź*

ANDRZEJ MATERKA

*Wydział Elektryczny, Politechnika Łódzka*

*Otrzymano 1991.05.17*

*Autoryzowano do druku 1991.07*

W pracy rozważono problem identyfikacji parametrów liniowych i nieliniowych obwodów RLC metodą dostrajania odpowiedzi modelu do pomierzonej odpowiedzi obwodu na wymuszenie skokowe. Założono, że chwila włączenia sygnału wymuszającego nie jest znana, co odpowiada warunkom badań złożonych struktur włókienniczych (linka bezpieczeństwa). Przebadano zaprojektowane i zbudowane obwody RLC odpowiadające elementom modelu mechanicznego struktury włókienniczej w warunkach statycznych i dynamicznych. Przeprowadzono badania zależności błędów identyfikacji od liczby próbek odpowiedzi obwodu oraz doboru punktów startowych programu optymalizacyjnego.

## 1. WSTĘP

Ciągle aktualnym problemem badawczym jest opracowanie urządzeń technicznych służących skutecznej ochronie człowieka w środowisku pracy. Problematyka ta obejmuje między innymi ochronę człowieka przed upadkiem z wysokości. Szczególne miejsce w przemysłowym sprzęcie chroniącym przed upadkiem z wysokości jak i sprzęcie alpinistycznym zajmuje linka bezpieczeństwa. W celu zbadania właściwości linek produkowanych w Polsce, poznania zjawisk w nich zachodzących podczas obciążania statycznego i dynamicznego oraz opracowania obiektywnych kryteriów ich oceny, w Centralnym Instytucie Ochrony Pracy w Łodzi zbudowano odpowiednie stanowisko badawcze [1]. Stanowisko to umożliwia dokonywanie pomiarów siły dynamicznej działającej w lince, w punkcie jej zamocowania do konstrukcji stałej, podczas zrzutów sztywnej masy (zaczepionej do drugiego końca linki) z zadanej wysokości. Przebieg siły w czasie jest rejestrowany w pamięci mikrokomputera.

Dla osiągnięcia wyżej wymienionych celów postawiono zadanie opracowania modelu złożonj struktury włókienniczej (linki), odwzorowującego właściwości linki w warunkach statycznych i dynamicznych oraz metody jego identyfikacji. Przed

identyfikację rozumiane jest tu: wyznaczenie najlepszego modelu matematycznego obiektu, z określonej klasy modeli, przy określonym kryterium oraz na podstawie danych:

– przed wykonaniem eksperymantu identyfikacji,

– pomiarowych (obserwacje wymuszenia i odpowiedzi).

Tak postawiony problem wymagał opracowania odpowiednich metod identyfikacji oraz ich zweryfikowania. Ponieważ badania dynamiczne i statyczne linek są badaniami niszczącymi, nie jest możliwe prowadzenie ich jednocześnie dla jednej próbki. W związku z tym utrudnione jest sprawdzenie czy przyjęta metoda identyfikacji daje poprawne rezultaty. Ponadto badania statyczne i dynamiczne są pracochłonne i wiążą się z poważnymi kosztami.

W związku z tym zadecydowano o posłużeniu się, dla celów weryfikacji metody identyfikacji, odpowiednio zaprojektowanymi nieliniowymi układami elektronicznymi, odpowiadającymi w działaniu strukturze włókienniczej [3], co jest tematem niniejszego artykułu.

## 2. METODA IDENTYFIKACJI

Opierając się na wiedzy o badanym obiekcie (linka bezpieczeństwa) i jej zachowaniu w warunkach dynamicznych sformułowano równanie (1).

$$m \cdot \frac{d^2 S}{dt^2} + C\left(\frac{dS}{dt}, S\right) \cdot \frac{dS}{dt} + k(S) \cdot S = Q \tag{1}$$

gdzie:

$m$ – masa sztywnego obciążnika dołączonego do końca linki bezpieczeństwa,

$C\left(\frac{dS}{dt}, S\right)$ – współczynnik tłumienia wiskotycznego,

$k(S)$ – współczynnik odkształcenia sprężysto-plastycznego,

$S$ – przemieszczenie sztywnego obciążnika przy powstrzymywaniu jego spadania przez linkę,

$Q$ – siła ciężkości działająca na obciążnik.

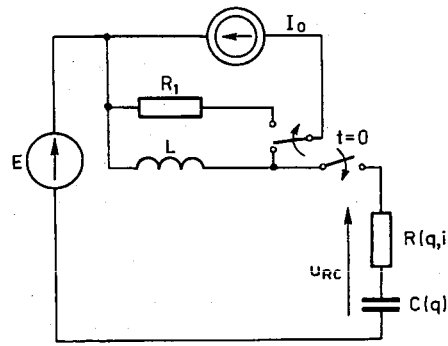
Równanie to opisuje ruch sztywnego obciążnika o masie m zaczepionego do badanej linki, podczas powstrzymywania jego spadania. Korzystając z równania (1) i zależności między wielkościami mechanicznymi i elektrycznymi [3] podanymi w tabeli 1 opracowano nieliniowy obwód elektryczny, którego schemat przedstawiony jest na rys. 1.

Obwód ten został opisany nieliniowym równaniem różniczkowym drugiego rzędu (2)

$$\mathrm{L} \cdot \mathrm{C}(q) \cdot \frac{d^2 u_C}{dt^2} + \mathrm{R}(i, q) \cdot \mathrm{C}(q) \cdot \frac{d u_C}{dt} + u_C = \mathrm{E} \tag{2}$$

z warunkami początkowymi dla chwili t=0

$$u_C(0) = 0, \quad i_L(0) = I_0$$




Rys. 1. Nieliniowy model elektryczny linki bezpieczeństwa

Tabela 1

| Wielkość mechaniczna | Wielkość elektryczna |
|---|---|
| S – droga | $q$ – ładunek |
| V – prędkość | $i$ – prąd |
| F – siła | $u$ – napięcie |
| $m$ – masa | L – indukcyjność |
| Q – siła ciężkości | E – źródło napięciowe |
| $V_0$ – prędkość początkowa | $I_0$ – źródło prądowe |
| $\frac{l}{k(S)}$ – współczynnik odkształcenia sprężysto-plastyczngo | C($q$) – nieliniowa pojemność |
| C(V, S) – współczynnik tłumienia wiskotycznego | R($i$, $q$) – nieliniowa rezystancja |
| $F_k(t)$ – siła pochodząca od elementu sprężysto-plastycznego | $U_C(t)$ – napięcie na nieliniowej pojemności |
| $F_C(t)$ – siła pochodząca od tłumika wiskotycznego | $U_R(t)$ – napięcie na nieliniowej rezystancji |
| $F(t) = F_U(t) + F_C(t)$ – siła wypadkowa mierzona podczas badania dynamicznego linki | $U_{RC}(t) = U_R(t) + U_C(t)$ – napięcia na nieliniowej rezystancji i pojemności |

Biorąc pod uwagę wiedzę o badanym obiekcie (lince) to znaczy jego zachowaniu się w warunkach statycznych i dynamicznych, sposób pomiaru odpowiedzi (próbkowany sygnał siły w funkcji czasu) oraz nieokreśloność w mierzonej odpowiedzi chwili podania wymuszenia zadecydowano o posłużeniu się metodą dostrajania modelu w dziedzinie czasu jako metodą identyfikacji. Metoda ta polegająca na takiej modyfikacji modelu (dostrajaniu), aby jego obliczona odpowiedź skokowa była

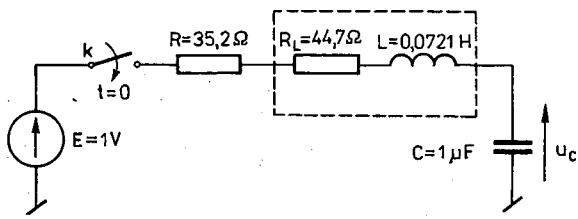
zgodna z odpowiedzią obiektu, w granicach założonego błędu uważana jest obecnie za uniwersalną i efektywną metodę identyfikacji, co znalazło odbicie w jej zastosowaniu w różnych dziedzinach techniki [4], [6], [10]. Dla przedstawionego obwodu elektrycznego zadanie identyfikacji polegało na znalezieniu współczynników określających przebieg funkcji gładkich $R(q,i)$, $C(q)$ na podstawie znajomości odpowiedzi $u_{RC}(t)$. Przybliżone przebiegi funkcji $R(q,i)$, $C(q)$ zostały określone na podstawie wstępnych badań statycznych i dynamicznych linek. Dla potrzeb identyfikacji zastosowano program optymalizacyjny napisany w języku Turbo Pascal, w którym wykorzystano metodę Powella, metodę Daviesa, Swanna, Campeya (metody bezgradientowe) [5] oraz metodę Stewarta ortogonalizacji kierunków poszukiwań. Zadaniem programu optymalizacyjnego była taka modyfikacja współczynników opisujących funkcje $R(q,i)$, $C(q)$ (dostrajanie modelu) aby funkcja celu określona jako

$$FC = \frac{1}{N} \sum_{i=1}^{N} [u_{RC}(t_i, R, L, C) - u_{pi}]^2 \tag{3}$$

gdzie:
$u_{pi}$ — wartości pomierzone (odpowiedź obwodu),
$u_{RC}(t_i,R,L,C)$ — wartości obliczone (odpowiedź modelu),
$t_i$ — chwile próbkowania odpowiedzi
osiągnęła wartość minimalną. Funkcja celu w tym wypadku określona jest jako błąd średniokwadratowy [9] między odpowiedzią obwodu a odpowiedzią jego modelu. Stosując powyższe kryterium zgodności odpowiedzi modelu z ciągiem zarejestrowanych próbek zakłada się, że za pomocą wybranego modelu można wystarczająco dokładnie aproksymować zachowanie się rzeczywistego obiektu.
Nieliniowe równanie różniczkowe drugiego rzędu (wzór 2) w programie optymalizacyjnym było rozwiązywane metodą Rungego-Kutty czwartego rzędu [2] natomiast równania nieliniowe opisujące $R(q,i)$, $C(q)$ metodą Newtona [2].

## 3. IDENTYFIKACJA LINIOWEGO SZEREGOWEGO OBWODU RLC

Pierwszym przybliżeniem modelu mechanicznego linki bezpieczeństwa, które opisuje wybrane własności badanej struktury włókienniczej za pomocą wielkości elektrycznych jest liniowy szeregowy obwód RLC. W związku z tym metoda identyfikacji z dostraja-

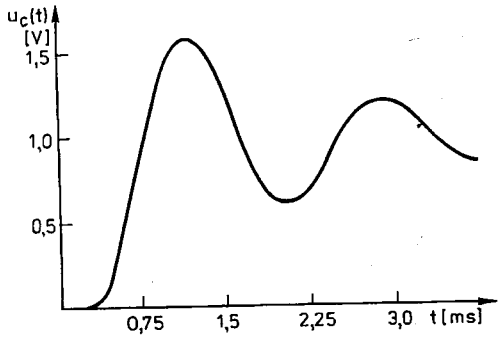



Rys. 2. Szeregowy liniowy obwód RLC

niem modelu została przebadana dla liniowego szeregowego obwodu RLC, przedstawionego na rys. 2. Istotą eksperymentu była identyfikacja parametrów $\alpha$, $\omega_n$ gdzie:

$$\alpha = \frac{R + R_L}{2 \cdot L}; \quad \omega_n = \frac{1}{\sqrt{L \cdot C}}. \tag{4}$$

Obwód RLC (rys. 2) został zbudowany z użyciem wybranych elementów elektronicznych, o wartościach zaznaczonych na rysunku. Skok wymuszającego napięcia wejściowego wynosił E = 1 V. Próbki odpowiedzi $u_C(t)$ (rys. 3) obwodu, pobrane w odstępach $T_p = 15\,\mu s$, przetworzono na postać cyfrową za pomocą 8-bitowego przetwornika A/C i zapisano w pamięci mikrokomputera.




Rys. 3. Odpowiedź $u_C(t)$ liniowego szeregowego obwodu RLC

Można wykazać, że odpowiedź $u_C(t)$ obwodu zależy od wielkości $\alpha$ i $\omega_n$ danych przez (4), a nie zależy jawnie od parametrów R,L,C obwodu. Nie jest zatem możliwa identyfikacja wielkości R,L,C przez minimalizację błędu średniokwadratowego między odpowiedzią zmierzoną a obliczoną, albo też uzyskiwane wyniki mogą być obciążone dużymi błędami. W celu oceny błędów identyfikacji pomierzono wartości elementów R,L,C (za pomocą standardowych przyrządów pomiarowych), na podstawie których (według 4) obliczono wartości parametrów $\alpha$, $\omega_n$, które dalej będą oznaczane przez $\alpha_s$, $\omega_{ns}$.
Obwód RLC w programie optymalizacyjnym został opisany liniowym równaniem różniczkowym drugiego rzędu

$$\frac{d^2 u_C}{dt^2} + 2 \cdot \alpha \cdot \frac{d u_C}{dt} + \omega_n^2 \cdot u_C = \omega_n^2 \cdot E \tag{5}$$

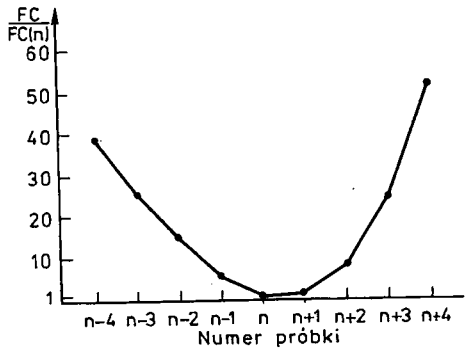
z warunkami początkowymi dla chwili $t = 0$

$$u_C(0) = 0; \quad i_L(0) = 0.$$

Próbki sygnału $u_C(t)$ z pomiaru zostały wprowadzone do programu optymalizacyjnego, gdzie były porównywane z odpowiedzią modelu obliczoną według wzoru (5). Uzyskane wyniki identyfikacji wykazywały silne uzależnienie od wyboru tej próbki sygnału $u_C(t)$, która traktowana jest jako początek odpowiedzi na wymuszanie E czyli odpowiada chwili $t = 0$ (zamknięcie klucza $k$, rys. 2).
Problem ten wynika z nieokreśloności w sygnale $u_C(t)$ chwili $t = 0$ podobnie jak w przypadku sygnału siły w badaniach linki gdzie nie jest znana faktyczna chwila początku powstrzymywania swobodnego spadku obciążnika.

Problem ten został rozwiązany w dwojaki sposób. Pierwszą metodą było poszukiwanie zbioru próbek sygnału $u_C(t)$ w kolejności ich pobierania i wybranie pierwszej (o numerze $n$), której wartość była większa od zera. Przyjmując, że próbka ta była pobrana w chwili $t=0$ wprowadzono do programu optymalizacyjnego $N$ próbek o numerach $n, n+1, n+2, \ldots, n+N$ i dokonywano identyfikacji parametrów $\alpha$, $\omega_n$. Badanie to przeprowadzono dla takiej samej liczby próbek N, ale zaczynając ich pobieranie od numerów

$$n-i, \ldots, n-3, n-2, n-1 \text{ oraz } n+1, n+2, n+3, \ldots, n+i, \text{ dla } i=4.$$




Rys. 4. Zależność minimalnego błędu średniokwadratowego między odpowiedzią obwodu a odpowiedzią modelu od numeru pierwszej próbki sygnału $u_C(t)$

Porównując wartości błędu średniokwadratowego identyfikacji FC (2) dla przeprowadzonych badań wybierano ten przypadek, dla którego FC przyjmuje wartość najmniejszą. Pierwsza próbka dla tego przypadku odpowiada chwili $t=0$. Wyniki badania dla $n=60$ przedstawia rys. 4. Błąd średniokwadratowy FC został tu odniesiony do błędu FC($n$), który występuje gdy $n$-ta próbka odpowiada chwili $t=0$. Dla przypadku, w którym pierwsza próbka wprowadzona do programu optymalizacyjnego odpowiada chwili $t=0$, błędy określone jako

$$\mathscr{E}_\alpha = \frac{\alpha - \alpha_s}{\alpha_s} \cdot 100\% \tag{6}$$

$$\mathscr{E}_{\omega_n} = \frac{\omega_n - \omega_{ns}}{\omega_{ns}} \cdot 100\% \tag{7}$$

przyjmują również najmniejsze wartości.

Poprawność przedstawionej metody została zweryfikowana doświadczalnie przez równoczesne zarejestrowanie przebiegu $u_C(t)$ oraz przebiegu wymuszenia E, między lewym zaciskiem rezystora R i masą obwodu z rys. 2.
Wynik doświadczenia potwierdził, że próbka sygnału $u_C(t)$ oceniona w przedstawionej metodzie jako odpowiadająca chwili $t=0$ odpowiadała czasowo próbce, dla której pojawiło się wymuszenie E na kluczu $k$, z dokładnością $\pm\frac{1}{2}T_p$ (gdzie $T_p$ jest okresem próbkowania wynoszącym 15 $\mu$s).

Drugą metodą pozwalającą na rozwiązanie problemu nieznajomości chwili $t=0$ było wprowadzenie dwóch dodatkowych parametrów do identyfikacji, będących warunkami początkowymi dla równania 5.

$$a = u_C(0); \quad b = i_L(0). \tag{8}$$

Zabieg ten pozwalał na identyfikację parametrów $\alpha$, $\omega_n$ bez zwiększenia błędów $\mathscr{E}_\alpha$, $\mathscr{E}_{\omega_n}$ w porównaniu z poprzednią metodą. Wadą tej metody jest wydłużenie czasu obliczeń oraz zawężenie możliwości jej stosowania do przypadków, w których próbki sygnału $u_C(t)$ pobierane są począwszy od numerów

$$n, n+1, n+2, \ldots \text{ (gdy } n \text{ odpowiada chwili } t=0).$$

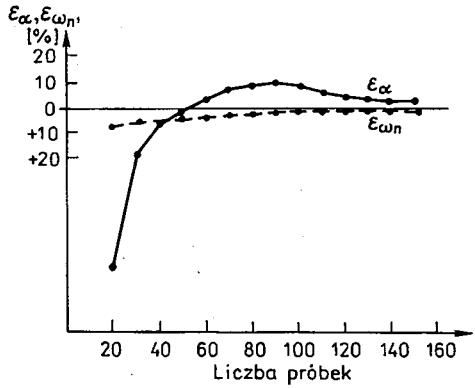
W przypadku wprowadzenia próbek sygnału $u_C(t)$ począwszy od numerów

$$n-1, n-2, n-3, \ldots$$

wyniki identyfikacji $\alpha$, $\omega_n$ były obciążone istotnymi błędami.

Podczas badania metody identyfikacji z dostrajaniem modelu stwierdzono, że dla zastosowanego algorytmu optymalizacyjnego uzyskuje się szybszą zbieżność po normalizacji optymalizowanych parametrów $\alpha$, $\omega_n$ [5]. Oznacza to, że wartości bezwzględne tych parametrów powinny być tego samego rzędu. Normalizację parametrów otrzymano dzieląc ich wartości odpowiednio przez $\alpha_s$, $\omega_{ns}$.

Dla zaproponowanej metody identyfikacji sprawdzono wpływ liczby $N$ próbek sygnału $u_C(t)$ (przy stałym okresie próbkowania), wprowadzonych do programu optymalizacyjnego, na błędy $\mathscr{E}_\alpha$, $\mathscr{E}_{\omega_n}$. Eksperyment ten miał swoje źródło w badaniach dynamicznych linek bezpieczeństwa. Ze względu na fakt, że opracowany model linki jest prawdziwy tylko w początkowym okresie odpowiedzi czasowej (do pierwszego maksimum), konieczne jest sprawdzenie, czy liczba próbek sygnału pokrywająca ten przedział jest wystarczająca do prawidłowej identyfikacji modelu.




Rys. 5. Wpływ liczby próbek wprowadzanych do programu optymalizacyjnego na błędy identyfikacji $\mathscr{E}_\alpha$, $\mathscr{E}_{\omega_n}$

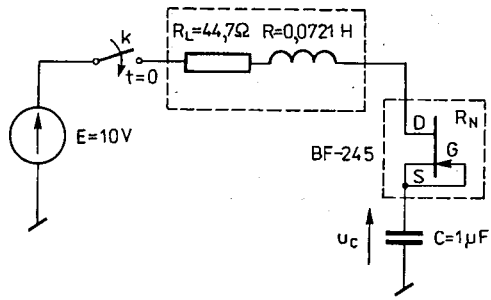
Otrzymane wyniki eksperymentu przedstawia rys. 5. Z badania tego wynika, że aby parametry $\alpha$, $\omega_n$ otrzymać z błędami względnymi $\mathscr{E}_\alpha$, $\mathscr{E}_{\omega_n}$ mniejszymi od 5% wystarczy do programu optymalizacyjnego wprowadzić $N=60$ próbek,

co odpowiada w przybliżeniu odcinkowi czasowemu, na którym sygnał $u_C(t)$ narasta od wartości 0 do wartości maksymalnej (rys. 3).

Obserwowane błędy identyfikacji wynikają z błędów pomiaru sygnału $u_C(t)$, a w tym błędów kwantyzacji, przyjęcia uproszczonego modelu obwodu RLC, w którym pominięto pasożytnicze parametry cewki L (pojemność międzyzwojowa) oraz błędów obliczeń w programie optymalizacyjnym (głównie błędy metody Rungego-Kutty rozwiązywania równania różniczkowego).

## 4. IDENTYFIKACJA SZEREGOWEGO OBWODU RLC Z NIELINIOWĄ REZYSTANCJĄ

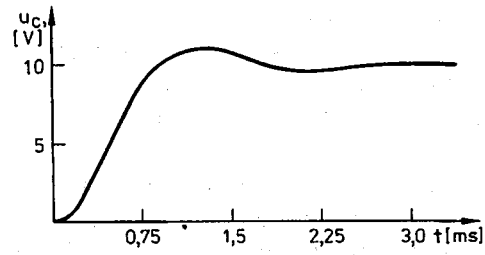
Jednym z elementów modelu mechanicznego linki jest tłumik wiskotyczny. Siła pochodząca od tego elementu zależy przede wszystkim od prędkości z jaką rozciągana jest struktura włókiennicza. W obwodzie przedsatwionym na rys. 1 tłumik wiskotyczny odpowiada nieliniowej rezystancji. W celu zweryfikowania zaproponowanej metody identyfikacji dla modelu z takim elementem wykonano eksperyment identyfikacji parametrów nieliniowej rezystancji $R_N$ wchodzącej w skład szeregowego obwodu RLC. Charakterystyka $u(i)$ zastosowanej nieliniowej rezystancji $R_N$ odpowiada w przybliżeniu charakterystyce $F\left(\frac{dS}{dt}\right)$ tłumika wiskotycznego, będącego fragmentem modelu mechanicznego linki, która została wyznaczona z pomiarów pośrednich. Obwód RLC został wykonany z wybranych elementów elektronicznych o wartościach przedstawionych na rys. 6. Rezystancją nieliniową




Rys. 6. Obwód RLC z nieliniową rezystancją $R_N(U_N)$

w obwodzie był tranzystor FET (BF−245), dla którego ustalono, że napięcie polaryzujące bramkę wynosi $U_{GS}=0$ V. Wartość napięcia $E=10$ V będącego wymuszeniem dla badanego obwodu została dobrana tak aby był wykorzystany cały zakres nieliniowości charakterystyki tranzystora FET. Istotą badania była identyfikacja parametrów $\alpha$, $\beta$, $\lambda$ określających charakterystykę i(u) nieliniowej rezystancji na podstawie odpowiedzi $u_C(t)$ (rys. 7) obwodu na wymuszenie skokiem napięcia E. Sygnał $u_C(t)$ został w tym celu spróbkowany ($T_p=15\,\mu$s), przetworzony na postać

cyfrową za pomocą 8-bitowego przetwornika A/C i zapamiętany w pamięci mikrokomputera.



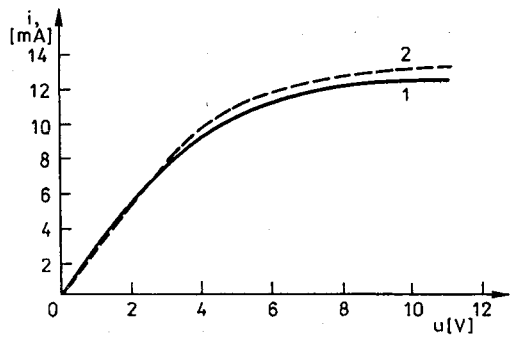

Rys. 7. Odpowiedź $u_C(t)$ szeregowego obwodu RLC z nieliniową rezystancją $R_N(U_N)$

W celu oceny błędów identyfikacji pomierzono charakterystykę $i(u)$ nieliniowej rezystancji (za pomocą standardowych przyrządów pomiarowych), która przedstawiona jest za pomocą linii ciągłej na rys. 8. Charakterystykę tę można opisać wzorem [7]:

$$i = \beta \cdot (1 + \lambda \cdot u_N) \cdot (U_{GS} - U_T)^2 \cdot \operatorname{tgh}(\alpha \cdot u_N), \tag{9}$$

gdzie:

$u_N$ – napięcie na nieliniowej rezystancji odpowiadające napięciu dren-źródło,
$U_{GS}$ – napięcie bramka-źródło,
$U_T$ – napięcie odcięcia,
$i$ – prąd w nieliniowej rezystancji odpowiadający prądowi drenu,
$\alpha, \beta, \lambda$ – współczynniki charakterystyczne dla danego tranzystora FET.




Rys. 8. Porównanie charakterystyk rezystancji nieliniowej utworzonych na podstawie: 1 – pomiarów, 2 – zidentyfikowanych parametrów $\alpha$, $\beta$, $\lambda$

Na podstawie wyznaczonej charakterystyki obliczono wartości parametrów $\alpha, \beta, \lambda$ dla zastosowanego tranzystora, które oznaczono symbolami $\alpha_s$, $\beta_s$, $\lambda_s$.

Obwód RLC w programie optymalizacyjnym został opisany równaniami:

$$\frac{d^2 u_C}{dt^2} + \frac{R}{L} \cdot \frac{d u_C}{dt} + \frac{1}{L \cdot C} \cdot (u_C + u_N) = \frac{1}{L \cdot C} \cdot E \tag{10}$$

z warunkami początkowymi dla chwili $t=0$, $u_C(0)=0$, $i_L(0)=0$

$$i = C \cdot \frac{d u_C}{dt} \tag{11}$$

oraz równaniem [9] dla którego $U_{GS}=0\,V$, $U_T=5\,V$.

Równanie algebraiczne (9) było rozwiązywane za pomocą procedury realizującej metodę Newtona [2], która była wywoływana w procedurze rozwiązywania nieliniowego równania różniczkowego (10) metodą Rungego-Kutty czwartego rzędu.

Dla odpowiedzi $u_C(t)$ (rys. 7) omawianego obwodu rozważano wpływ nieznajomości chwili $t=0$ podania wymuszenia na błędy identyfikacji. Problem ten został rozwiązany w analogiczny sposób jak dla liniowego obwodu RLC czyli przez wybór próbki $n$ sygnału $u_C(t)$, dla której błąd średniokwadratowy FC między odpowiedzią obwodu a jego modelu osiągnął wartość najmniejszą.

Podczas badania zaproponowanej metody identyfikacji sprawdzono również wpływ liczby próbek sygnału $u_C(t)$ wprowadzanych do programu optymalizacyjnego na błędy identyfikacji parametrów $\alpha$, $\beta$, $\lambda$ określone jako:

$$\begin{aligned} \mathscr{E}_\alpha &= \frac{\alpha - \alpha_s}{\alpha_s} \cdot 100\% \\ \mathscr{E}_\beta &= \frac{\beta - \beta_s}{\beta_s} \cdot 100\% \\ \mathscr{E}_\lambda &= \frac{\lambda - \lambda_s}{\lambda_s} \cdot 100\% \end{aligned} \tag{12}$$

oraz błąd określający rozbieżność charakterystyki rzeczywistej elementu nieliniowego i charakterystyki otrzymanej na podstawie zidentyfikowanych parametrów $\alpha$, $\beta$, $\lambda$, określony jako:

$$\mathscr{E} = \frac{\sum_{i=1}^{M} \frac{|i(U_i) - i_s(U_i)|}{i_s(U_i)}}{M} \cdot 100\% \tag{13}$$

gdzie:

$U_i$ — wartości napięcia na rezystancji nieliniowej,

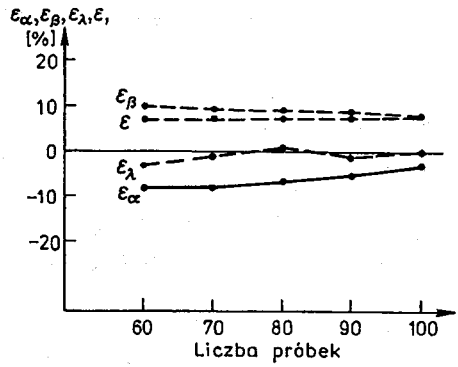
$i_s(U_i)$ — wartości prądu wynikające z charakterystyki rzeczywistej,

$i(U_i)$ — wartości prądu wynikające z charakterystyki otrzymanej na podstawie zidentyfikowanych parametrów $\alpha$, $\beta$, $\lambda$.

Wyniki tego badania przedstawiają wykresy na rys. 9. Wykazują one, że identyfikację parametrów $\alpha$, $\beta$, $\lambda$ uzyskuje się z błędami względnymi mniejszymi od 10% po wprowadzeniu więcej niż 60 próbek sygnału $u_C(t)$.

Dla 86 próbek, co odpowiada przedziałowi czasowemu od $t=0$ do osiągnięcia przez $u_C(t)$ wartości maksymalnej, błędy przyjmują wartości:

$$\mathscr{E}_\alpha = 5{,}9\%,\ \mathscr{E}_\beta = 9{,}4\%,\ \mathscr{E}_\lambda = 0{,}6\%,\ \mathscr{E} = 7{,}7\%.$$




Rys. 9. Wpływ liczby próbek sygnału $u_C(t)$ wprowadzonych do programu optymalizacyjnego na błędy identyfikacji $\mathscr{E}_\alpha$, $\mathscr{E}_\beta$, $\mathscr{E}_\lambda$, $\mathscr{E}$

Przy dalszym zwiększeniu liczby próbek błędy te nie ulegają istotnemu zmniejszeniu. Na podstawie wyników identyfikacji dla 86 próbek wykreślono charakterystykę rezystancji nieliniowej, która przedstawiona jest za pomocą linii przerywanej na rys. 8. Do głównych przyczyn błędów identyfikacji należą w tym przypadku:

– błędy pomiaru sygnału $u_C(t)$, a w tym błędy kwantyzacji,

– pominięcie w modelu obwodu RLC elementów pasożytniczych cewki,

– błędy obliczeń w programie optymalizacyjnym (głównie w metodzie Newtona i metodzie Rungego-Kutty).

Dla zastosowanego programu optymalizacyjnego sprawdzono jak czułe są wyniki identyfikacji na zmiany punktów startowych optymalizacji $\alpha_{p_\alpha}$, $\beta_{p_\beta}$, $\lambda_{p_\lambda}$. Punkty te dobierane są na podstawie wiedzy a priori o obiekcie. Ponieważ znane były wartości $\alpha_S$, $\beta_S$, $\lambda_S$ identyfikowanych parametrów, to we wcześniej opisanych eksperymentach przyjęto, że:

$$p_\alpha = \alpha_S, \quad p_\beta = \beta_S, \quad p_\lambda = \lambda_S \tag{14}$$

W celu zbadania wpływu tych punktów na błędy identyfikacji zmieniano ich wartości obserwując wartości błędu $\mathscr{E}$. Punkty startowe były zmieniane w granicach

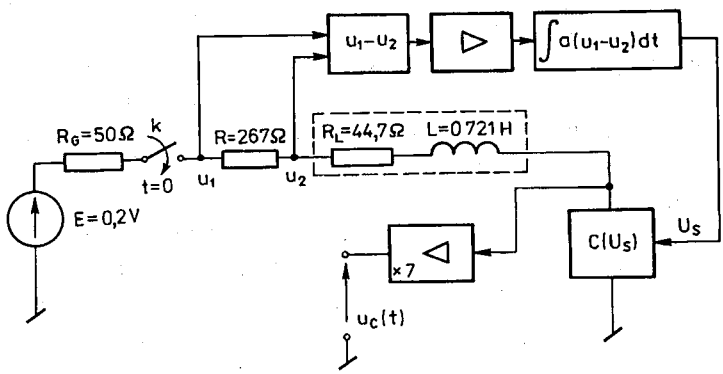
$$\begin{aligned} 0{,}5 \times \alpha_S &\leqslant p_\alpha \leqslant 1{,}5 \times \alpha_S \\ 0{,}5 \times \beta_S &\leqslant p_\beta \leqslant 1{,}5 \times \beta_S \\ 0{,}5 \times \lambda_S &\leqslant p_\lambda \leqslant 1{,}5 \times \lambda_S \end{aligned} \tag{15}$$

z krokiem wynoszącym odpowiednio $0{,}1 \times \alpha_S$, $0{,}1 \times \beta_S$, $0{,}1 \times \lambda_S$. Otrzymano w wyniku tego eksperymentu, że błąd identyfikacji zwiększa się nie więcej niż o 1% swojej wartości otrzymanej dla przypadku (14). W przypadku dalszej zmiany wartości punktów startowych program optymalizacyjny (metoda Newtona) tracił zbieżność nie dając żadnych wyników. Można więc stwierdzić, że o błędach identyfikacji w zaproponowanym programie optymalizacyjnym w małym stopniu decyduje wybór punktów startowych.

## 5. IDENTYFIKACJA SZEREGOWEGO OBWODU RLC Z NIELINIOWĄ POJEMNOŚCIĄ

Element sprężysto-plastyczny jest jednym z elementów modelu mechanicznego linki. Siła pochodząca od tego elementu zależy od wydłużenia badanej struktury włókienniczej. W obwodzie przedstawionym na rys. 1 element sprężysto-plastyczny odpowiada nieliniowej pojemności $C(q)$. Energia zgromadzona w tej pojemności odpowiada energii wykorzystanej na odkształcenie sprężyste i plastyczne struktury włókienniczej.

W celu zweryfikowania omawianej metody identyfikacji dla modelu z takim elementem wykonano eksperyment identyfikacji parametrów nieliniowej pojemności $C(q)$, wchodzącej w skład szeregowego obwodu RLC. Charakterystyka $C(q)$ tej pojemności odpowiada w przybliżeniu (zgodnie z relacjami podanymi w tabeli 1) wyznaczonej doświadczalnie w badaniach statycznych charakterystyce $k(s)$ elementu sprężysto-plastycznego będącego fragmentem modelu mechanicznego linki.

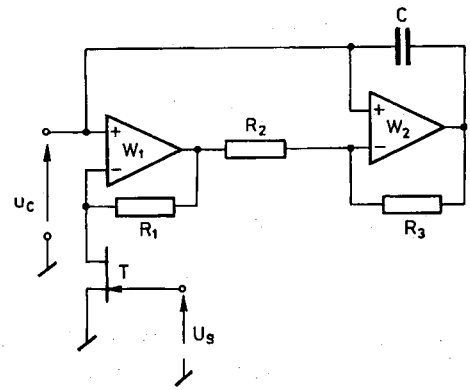



Rys. 10. Szeregowy obwód RLC z nieliniową pojemnością $C(q)$

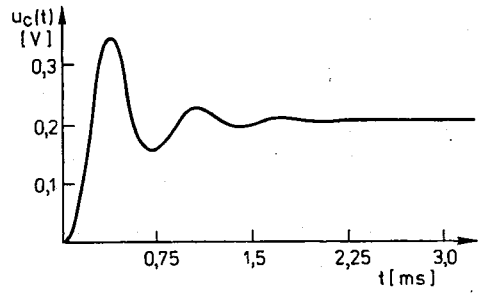
Schemat blokowy tego obwodu przedstawia rys. 10. Układem symulującym pojemność $C(q)$ był konwerter impedancji [8], którego pojemność wejściowa uzależniona jest od wartości napięcia sterującego $U_S$ podanego na bramkę tranzystora FET (rys. 11).

Aby pojemność ta była uzależniona od ładunku na niej zgromadzongo, na wejście sterujące konwertera impedancji podano scałkowany sygnał napięciowy, który jest wprost proporcjonalny do prądu płynącego przez tę pojemność (rys. 10). Konwerter impedancji został zaprojektowany tak aby jego charakterystyka $C(q)$ odwzorowywała właściwości statyczne struktury włókienniczej.

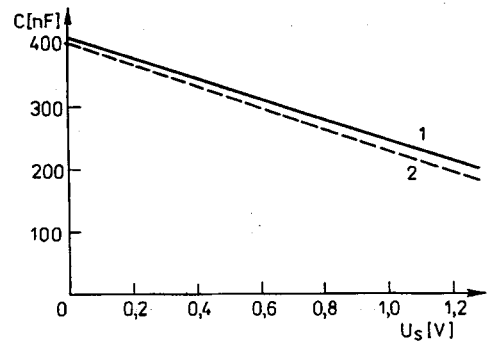
Celem eksperymentu była identyfikacja parametrów $C_0$, $k$, określających kształt charakterystyki $C(U_S)$ nieliniowej pojemności na podstawie odpowiedzi czasowej $u_C(t)$ przedstawionej na rys. 12, na wymuszenie skokiem napięcia E. Sygnał $u_C(t)$ został w tym celu przetworzony na postać cyfrową w ten sam sposób jak dla poprzednich obwodów RLC. Dla oceny błędów identyfikacji pomierzono charakterystykę $C(U_S)$




Rys. 11. Uproszczony schemat konwertera impedancji




Rys. 12. Odpowiedź $u_C(t)$ obwodu RLC z nieliniową pojemnością $C(q)$




Rys. 13. Porównanie charakterystyk $C(U_S)$ konwertera impedancji utworzonych na podstawie: 1 – pomiarów, 2 – zidentyfikowanych parametrów $C_0$, $k$

nieliniowej pojemności używając standardowych przyrządów pomiarowych. Charakterystyka ta przedstawiona jest na rys. 13 za pomocą linii ciągłej. Na jej podstawie obliczono wartości parametrów $C_0$, $k$, które dalej będą oznaczane $C_{0S}$, $k_S$.
Obwód został opisany równaniem różniczkowym

$$\frac{d^2q}{dt^2} + \frac{(R_G + R + R_L)}{L} \cdot \frac{dq}{dt} + \frac{l}{L \cdot C(q)} \cdot q = \frac{E}{L} \tag{16}$$

z warunkami początkowymi dla chwili $t=0$

$$q(0) = 0, \quad i_L(0) = 0,$$

gdzie $C(q)$ określa charakterystykę nieliniowej pojemności i opisane jest równaniem

$$C(q) = C_0 - k \cdot q. \tag{17}$$

Wybór q jako zmiennej w równaniu (16) podyktowany był łatwością opisu zjawisk zachodzących w obwodzie RLC i uproszczeniem obliczeń przy rozwiązywaniu tego równania.

Dla odpowiedzi $u_C(t)$ (rys. 12) obwodu rozważano jak poprzednio wpływ nieznajomości chwili $t=0$ podania wymuszenia na błędy identyfikacji. Problem ten został rozwiązany w sposób analogiczny jak dla liniowego obwodu RLC, czyli przez wybór próbki $n$ sygnału $u_C(t)$, dla której błąd średniokwadratowy FC między odpowiedzią obwodu i jego modelu jest najmniejszy. W eksperymencie tym stwierdzono ponadto, że próbką odpowiadającą chwili $t=0$ jest próbka o numerze o 2 lub 1 mniejszym od pierwszej próbki o wartości większej od 0. Efekt ten potwierdzają obliczenia, że przy okresie próbkowania sygnału $u_C(t)$ równym $T_p=15\,\mu s$, rozdzielczości kwantowania 10 mV i parametrach obwodu podanych na rys. 10 w czasie od chwili $t=0$ do $t=15\,\mu s$ wartość $u_C$ wzrośnie od 0 do około 5 mV czyli wynik konwersji A/C jest równy 0. Dodatkowo biorąc pod uwagę możliwość maksymalnego przesunięcia chwili próbkowania w stosunku do chwili zamknięcia klucza $k$ (rys. 10) $t=0$ o $\pm\frac{1}{2}T_p$ jasną staje się konieczność wprowadzenia do programu optymalizacyjnego jednej lub dwóch próbek sygnału $u_C(t)$ o wartości równej 0 przed ciągiem próbek o wartościach większych od 0.

Ze względów przedstawionych w poprzednich rozdziałach, dla omawianej metody identyfikacji przeprowadzono badanie wpływu liczby próbek sygnału odpowiedzi $u_C(t)$ na błędy określone jako:

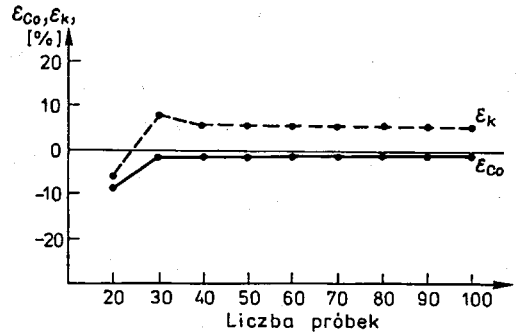
$$\mathscr{E}_{C_0} = \frac{C_{0S} - C_0}{C_{0S}} 100\% \tag{18}$$

$$\mathscr{E}_k = \frac{k_S - k}{k_S} 100\%.$$

Wartość błędów $\mathscr{E}_{C_0}$, $\mathscr{E}_k$ w zależności od liczby próbek zastosowanych w programie optymalizacyjnym przedstawiają wykresy na rys. 14.

Analizując te wykresy można zauważyć, że dla 30 próbek sygnału $u_C(t)$ wynoszą one $\mathscr{E}_{C_0}=1{,}4\%$, $\mathscr{E}_k=5{,}1\%$.

Przy dalszym zwiększeniu liczby próbek błędy $\mathscr{E}_{C_0}$, $\mathscr{E}_k$ nie ulegają istotnej zmianie. Ponieważ dla omawianego przypadku pierwsze maksimum sygnału $u_C(t)$ przypada na 26 próbkę (rys. 12) można przyjąć, że dla zapewnienia identyfikacji z minimalnymi




Rys. 14. Wpływ liczby próbek sygnału $u_C(t)$ wprowadzanych do programu optymalizacyjnego na błędy $\mathscr{E}_k$, $\mathscr{E}_{C_0}$

błędami wystarczą próbki z przedziału czasowego, w którym sygnał narasta od wartości 0 do wartości maksymalnej. Charakterystyka nieliniowej pojemności $C(U_S)$ utworzona na podstawie zidentyfikowanych parametrów $C_0$, $k$ przedstawiona jest za pomocą linii przerywanej na rys. 13.

W celu określenia rozbieżności charakterystyki rzeczywistej i pochodzącej z identyfikacji obliczono błąd określony jako

$$\mathscr{E} = \frac{\sum_{i=1}^{M} \frac{|C_S(U_i) - C(U_i)|}{C_S(U_i)}}{M} \cdot 100\% \tag{19}$$

gdzie:

$C_S(U_i)$ – wartości wynikające z charakterystyki rzeczywistej,

$C(U_i)$ – wartości wynikające z charakterystyki otrzymanej na podstawie zidentyfikowanych parametrów $C_0$, $k$.

Dla przypadku identyfikacji z 30 próbkami sygnału $u_C(t)$ wartość błędu $\mathscr{E}$ wynosi 5%. Błędy identyfikacji, widoczne w różnicy charakterystyk, wynikają głównie z:

– nieuwzględnienia w modelu obwodu RLC elementów pasożytniczych cewki,

– błędów pomiaru sygnału $u_C(t)$, a w tym kwantyzacji,

– przybliżenia charakterystyki $C(U_S)$ za pomocą linii prostej,

– błędów obliczeniowych metody Rungego-Kutty rozwiązywania równania różniczkowego.

## 6. PODSUMOWANIE

Podsumowując wyniki zaprezentowanych prac zmierzających do zweryfikowania metody identyfikacji z dostrajaniem modelu wykorzystującej minimalizację błędu średniokwadratowego między odpowiedzią obiektu a odpowiedzią modelu dla układów elektronicznych (obwodów RLC z elementami liniowymi i nieliniowymi) można stwierdzić, że zastosowane metody obliczeniowe dały zadowalające wyniki,

jeśli wziąć pod uwagę wielkość błędów identyfikacji. Opracowana została metoda identyfikacji na podstawie odpowiedzi czasowej obwodu bez jednoznacznie określonej chwili $t=0$ podania wymuszenia. Metoda ta polega na wyszukaniu początkowej próbki ciągu próbek, dla której błąd średniokwadratowy FC między odpowiedzią obiektu i odpowiedzią modelu jest najmniejszy. Próbka ta odpowiada chwili $t=0$ podania wymuszenia z dokładnością $\pm\frac{1}{2}T_p$. Błędy identyfikacji dla przedstawionego programu optymalizacyjnego uzależnione są przede wszystkim od błędów pomiaru odpowiedzi obwodu i od wyboru prawidłowej struktury modelu.

Zastosowany program optymalizacyjny, wykorzystujący metody bezgradientowe przy założeniu wyboru prawidłowego modelu, jest mało czuły na zmiany punktów startowych optymalizacji. Program ten pozwala na relatywnie szybką identyfikację, bo dla komputera klasy IBM PC/AT czas obliczeń wynosi od kilkudziesięciu sekund do kilku minut. Czas obliczeń może wydłużać się w przypadku stosowania procedury Newtona do rozwiązywania równań nieliniowych i dużej liczby próbek sygnału odpowiedzi (ponad 100). Analiza metody dla przypadku szeregowego, liniowego obwodu RLC wykazuje konieczność identyfikowania niezależnych parametrów modelu tzn. $\alpha$, $\omega_n$ a nie R,L,C w celu uniknięcia dużych błędów, które kompensują się w odpowiedzi modelu.

Dla przedstawionych przykładów obwodów RLC z elementami nieliniowymi odpowiadającymi elementom modelu mechanicznego można stwierdzić, że w przypadku badania ich odpowiedzi (sygnał $u_C(t)$) będącej oscylacjami gasnącymi uzyskuje się zadowalające wyniki identyfikacji (ze względu na błędy) badając próbki sygnału odpowiedzi obwodu obejmujące odcinek czasowy od chwili podania wymuszenia $t=0$ do osiągnięcia przez sygnał wartości maksymalnej (przy założeniu odpowiedniego okresu próbkowania $T_p$).

Wniosek ten jest szczególnie ważny dla badań linki, gdyż zmiany zachodzące w jej strukturze podczas obciążania dynamicznego są nieodwracalne. Po przekroczeniu wartości maksymalnej siły działającej w lince powstrzymującej spadanie sztywnej masy należy przyjąć inny model tej linki.

Ponieważ modele przebadanych obwodów elektrycznych odpowiadają elementom modelu mechanicznego linki, uzyskane wyniki analizy przedstawionej metody identyfikacji mogą być pomocne w badaniach struktur włókienniczych.

W pracy wykorzystano elektryczny model struktur włókienniczych, co umożliwia niezależne modelowanie takich zjawisk mechanicznych, jak sprężystość i tłumienie wiskotyczne, których oddzielny pomiar napotyka na znaczne trudności. Tą drogą każdy z elementów obwodu elektrycznego RLC był oddzielnie symulowany za pomocą odpowiednio zaprojektowanego i zbadanego doświadczalnie układu elektronicznego. Pozwala to na obiektywną ocenę jakości proponowanej metody identyfikacji poszczególnych zjawisk mechanicznych. Pełna weryfikacja tej metody wymagą dalszych badań z udziałem specjalistów z dziedziny włókiennictwa i mechaniki. Badania takie są prowadzone i będą opublikowane w przyszłości. Z tych względów przy interpretacji dotychczas uzyskanych wyników posłużono się wielkościami elektrycznymi.

## BIBLIOGRAFIA

1. K. Baszczyński, A. Materka: *Elektroniczny system pomiarowy do badań dynamicznych sprzętu chroniącego przed upadkiem z wysokości.* Prace Centralnego Instytutu Ochrony Pracy, Zeszyt 142, WNT, Warszawa 1989
2. L. O. Chua, Pen-Min Lin: *Komputerowa analiza układów elektronicznych.* WNT, Warszawa 1981
3. J. P. Den Hartog: *Drgania mechaniczne.* PWN, Warszawa 1971
4. P. Eykoff: *Identyfikacja w układach dynamicznych.* PWN, Warszawa 1980
5. W. Findeisen, J. Szymankowski, A. Wierzbicki: *Teoria i metody obliczeniowe optymalizacji.* PWN, Warszawa 1980
6. J. Giergiel, T. Uhl: *Identyfikacja układów mechanicznych.* PWN, Warszawa 1990
7. T. Kacprzak, A. Materka: *Compact dc Model of GaAs FET's for Large-Signal Computer Calculation.* JEEE Journal of Solid-State Circuits, vol. SC–18, NO. 2, April 1983
8. Z. Kulka, M. Nadachowski: *Analogowe układy scalone.* WKiŁ, Warszawa 1983
9. K. Mańczak, Z. Nahorski: *Komputerowa identyfikacja obiektów dynamicznych.* PWN, Warszawa 1983
10. J. Ogrodzki, L. Opalski: *Badanie własności numerycznych zadania ekstrakcji parametrów tranzystora bipolarnego.* Teksty referatów, tom 1. XII Krajowej konferencji teorii obwodów i układów elektronicznych. Wydział Elektryczny Politechniki Rzeszowskiej, Rzeszów 1989

K. BASZCZYŃSKI, A. MATERKA

## A METHOD OF COMPUTER IDENTIFICATION FOR SELECTED CLASS OF SECOND-ORDER SYSTEM

### Summary

The problem of identification of linear and nonlinear RLC circuits parameters, using the method of least-squares fit of the model response to the circuit step response, is discussed. There is the assumption that the moment of turning on the input signal is unknown. This corresponds to the conditions of complicated textile structures (lanyard) tests, which are modelled using the RLC circuits. Designed and built RLC circuits which correspond to mechanical phenomena in textile structure under static and dynamic conditions are investigated. The investigations of depedence of identification errors on number of circuit response samples and on the choice of values of the starting points in the optimisation program are presented.

KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJI, 1992, 38, z. 1, ss. 35-44

621.372.8:535.3

# Solitony w światłowodach: właściwości propagacyjne

ADAM MAJEWSKI

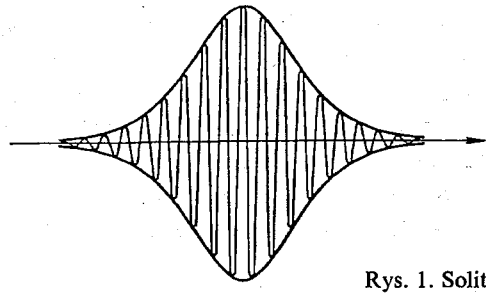
*Instytut Podstaw Elektroniki. Politechnika Warszawska*

*Otrzymano 1991.04.10*

*Autoryzowano do druku 1991.07.24*

Omówiono właściwości propagacyjne solitonów, solitonów wyższych rzędów i ciemnych solitonów w jednomodowych światłowodach włóknistych z uwzględnieniem tłumienności, dyspersji i nieliniowości wyższego rzędu oraz ich zastosowania głównie do przesyłania sygnałów na duże odległości bez zniekształceń.

W przypadku przesyłania w światłowodzie impulsów pikosekundowych a tym bardziej femtosekundowych o zniekształceniach decyduje nie tyle tłumienność co dyspersja. Niewielka bowiem dyspersja powoduje, że impulsy ulegają poszerzeniu i osłabieniu tak, że światłowód staje się praktycznie nieużyteczny do zastosowań w technice krótkich impulsów. Negatywne skutki dyspersji może skompensować nieliniowość światłowodu jeśli natężenie wprowadzonego światła jest dostatecznie duże. W wyniku oddziaływania samomodulacji fazy i dyspersji powstają bowiem solitony (rys. 1).



Rys. 1. Soliton obwiedni

Współczynnik załamania szkła wyraża się wzorem

$$\bar{n} = n + n_2 |E|^2 \tag{1}$$

przy czym $E$ oznacza natężenie pola elektrycznego a $n_2$ jest nieliniowym współczynnikiem załamania, który dla czystego szkła kwarcowego wynosi ok. $1{,}2\,10^{-22}$ m$^2$/V$^2$. Wartość ta jest bardzo mała i mimo, iż natężenie pola jest zwykle duże w światłowodzie

($E \approx 10^6$ V/m dla pola modu 100 $\mu m^2$ i mocy 1 W) całkowita zmiana współczynnika załamania jest rzędu $10^{-10}$ i mogłoby się wydawać, że może być pominięta. Przyczyną tego, że tak mała zmiana współczynnika załamania okazuje się istotna jest to, że szerokość widma, która w przybliżeniu jest odwrotnością szerokości impulsu, jest znacznie mniejsza niż częstotliwość fali świetlnej (np. dla długości fali $\lambda = 1550$ nm, $\omega = 1{,}2\,10^{15}$ l/s, a więc dla impulsu o szerokości 10 ps mamy, że stosunek szerokości widma do pulsacji fali świetlnej $\Delta\omega/\omega = 10^{-4}$) a po drugie dyspersja, która zależy od $\Delta\omega$ jest też mała. Jednak aby te efekty miały znaczenie, zniekształcenia sygnału na skutek strat mocy winny być od nich mniejsze. Zatem straty na odcinku równym długości fali nie powinny być większe niż $10^{-10}$. Oznacza to, że tłumienność światłowodu winna być mniejsza niż 1 dB/km. To wymaganie opóźniło przeprowadzenie pierwszej transmisji solitonowej (jednomodowy światłowód o dostatecznie małej tłumienności wykonano dopiero w 1979 r.). Nie było też odpowiednich laserów dla zakresu fal 1550 nm, a ponadto właściwości dyspersyjne światłowodu nie były dostatecznie poznane. Dopiero MOLLENAUER i inni [1] potwierdzili doświadczalnie istnienie solitonu w światłowodzie w 1980 r. Przesłali oni impulsy optyczne o szerokości 7 ps, długości fali 1550 nm i mocy szczytowej wynoszącej kilka W w światłowodzie o długości 700 m i pokazali, że szerokość impulsu zależy od wielkości natężenia wprowadzonego światła.

Soliton jako zjawisko fizyczne na powierzchni wody zaobserwował RUSSELL [2] w 1834 r. a opisane zostało równaniem przez KORTEWEGa i de VRIESa w 1895 r [3], które nazywa się KdV. Słowo soliton użyto po raz pierwszy w publikacji ZABUSKYego i KRUSKALa w 1965 r. [4]. Soliton optyczny różni się jednak od solitonu na powierzchni wody i innych. Jest on bowiem obwiednią fali świetlnej (rys. 1).

Rozchodzenie się światła w światłowodzie nieliniowym można opisać w przypadku, gdy impulsy mają szerokość $T_0 \geqslant 100$ fs równaniem

$$j\frac{\partial A}{\partial z} - \frac{1}{2}\beta_2 \frac{\partial^2 A}{\partial T^2} + |A|^2 A + \frac{j}{2}\alpha A = 0 \tag{2}$$

przy czym A jest obwiednią fali świetlnej, $\beta_2 = \partial^2\beta/\partial\omega^2$ jest dyspersją prędkości grupowej ($\beta_2$[ps$^2$/km] $\approx -$D[ps/km nm] w zakresie długości fal 1300 $\div$ 1550 nm), $\alpha$ jest tłumiennością, $z$ i $T$ oznaczają odpowiednio współrzędną i czas, $\gamma = n_2 k/A_{sk}$ jest współczynnikiem nieliniowym, $k = 2\pi/\lambda$, a $A_{sk}$ jest skutecznym przekrojem rdzenia. $W$ przypadku impulsu opisanego funkcją Gaussa $A_{sk} = \pi W_G^2$, przy czym $W_G$ jest promieniem modu. Skuteczny przekrój rdzenia światłowodu wynosi od 10 $\div$ 20 $\mu m^2$ dla światła widzialnego do 50 $\div$ 100 $\mu m^2$ w zakresie długości fal 1550 nm, a zatem $\gamma = 30 \div 1{,}3$ W$^{-1}$ km$^{-1}$.

Jeśli pominiemy tłumienność to zamiast (2) otrzymamy, że

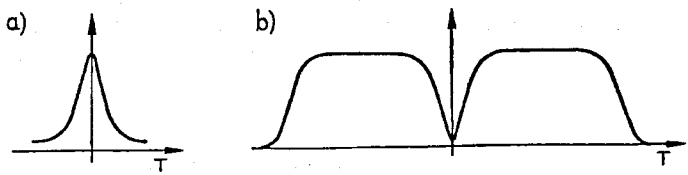
$$j\frac{\partial U}{\partial \xi} - \frac{1}{2}\operatorname{sgn}(\beta_2)\frac{\partial^2 U}{\partial \tau^2} + N^2|U|^2 U = 0 \tag{3}$$

które w literaturze poświęconej solitonom nazywa się Nieliniowym Równaniem Schrödingera (NRS). Zastosowano tu następujące oznaczenia: $U = A/\sqrt{P_0}$, $\xi = z/L_D$,

$\tau = T/T_0$, przy czym $P_0$ oznacza moc szczytową, $L_D = T_0^2/|\beta_2|$ – długość dyspersyjną, $L_{NL} = 1/\gamma P_0$ – długość nieliniową, zaś $N^2 = L_D/L_{NL}$. Parametr $N$ nazywa się rzędem solitonu. W przypadku gdy oznaczymy, że $u = NU$ a dyspersja jest anormalna, tj. $\mathrm{sgn}(\beta_2) = -1$, to NRS przyjmuje standardową postać

$$j\frac{\partial u}{\partial \xi} + \frac{1}{2}\frac{\partial^2 u}{\partial \tau^2} + |u|^2 u = 0 \tag{4}$$

Równanie to jak wykazali ZACHAROW i SZABAT [5] może być rozwiązywane metodą odwrotnego rozpraszania. HASEGAWA i TAPPERT [6] w 1973 wykazali, że w światłowodzie impuls świetlny tworzy soliton jasny (ang. bright soliton) gdy dyspersja jest anormalna ($\beta_2 < 0$), a ciemny soliton (ang. dark soliton), gdy dyspersja jest normalna ($\beta_2 > 0$) (rys. 2). EMPLIT i inni [7] uzyskali w światłowodzie o długości 52 m, ciemne solitony o szerokości 5 ps w impulsach o szerokości 25 ps i mocy 0,2 W dla fali o długości 595 nm.




Rys. 2. Solitony: a – jasny, b – ciemny

Minęło ćwierć wieku od czasu pojawienia się słowa soliton. W tym czasie odkryte zostały różne rodzaje solitonów prawie we wszystkich dziedzinach fizyki. Obecnie ponad 100 nieliniowych równań różniczkowych ma rozwiązanie solitonowe lub solitonopodobne. Jednym z zastosowań solitonów optycznych jest telekomunikacja.

Przy założeniu, że $u \to 0$ dla $\pm\infty$, gdy $N = 1$ rozwiązaniem (4) jest funkcja

$$u(\xi, \tau) = \eta \,\mathrm{sech}(\eta\tau)\, e^{j\eta^2\xi/2}. \tag{5}$$

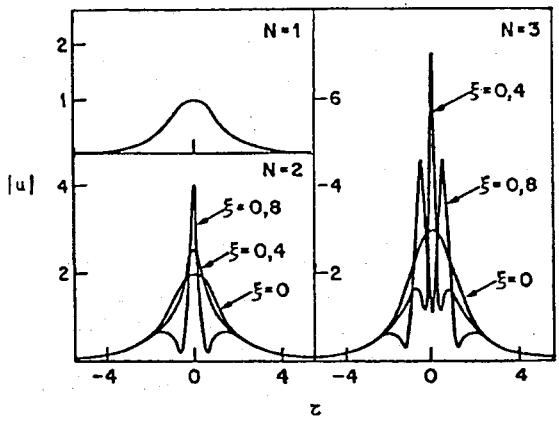
Widzimy tu, że szerokość impulsu jest odwrotnie proporcjonalna do jego amplitudy. Jeśli założymy, że $u(0,0) = 1$ to zamiast (5) otrzymamy, że

$$u(\xi, \tau) = \mathrm{sech}(\tau)\, e^{j\xi/2} \tag{6}$$

Równanie to opisuje soliton podstawowy. Wskazuje ono, że jeśli impuls w kształcie funkcji secans hiperboliczny ma szerokość $T_0$ i moc szczytową $P_0$ taką, że $N = 1$, to w światłowodzie bezstratnym będzie się on poruszał bez zmiany kształtu na dowolną odległość. Jest to właściwość solitonu podstawowego, która czyni go atrakcyjnym dla zastosowań w telekomunikacji optycznej. Moc progową wymaganą do tego by mógł powstać podstawowy soliton wyznacza się ze wzoru

$$P_{pr} = \frac{|\beta_2|}{\gamma T_0^2} = 3{,}11\,\frac{|\beta_2|}{\gamma T_{\mathrm{FWHM}}^2} \tag{7}$$

przy czym $T_{\mathrm{FWHM}} \approx 1{,}76\ T_0$ jest całkowitą szerokością impulsu określoną przez punkty, dla których natężenie światła wynosi połowę wartości szczytowej (ang. *full*




Rys. 3. Solitony: podstawowych i wyższych rzędów dla warunku początkowego $u(0,\tau)=N\mathrm{sech}(\tau)$ [8]

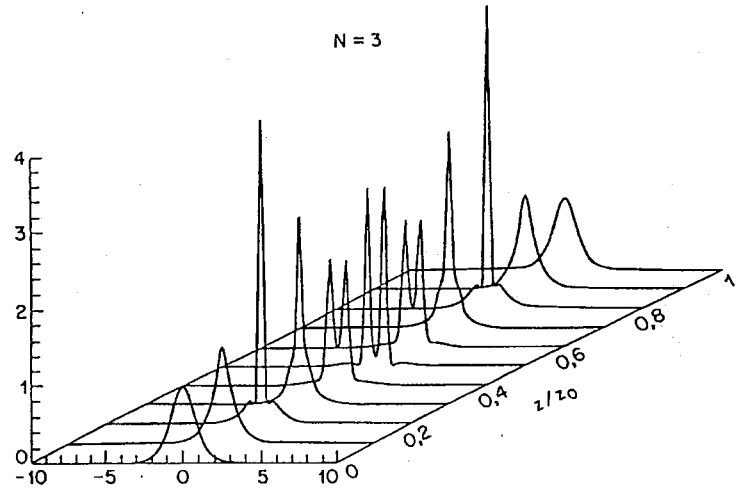
*width half maximum*). Dla standardowego światłowodu i $\lambda=1550$ nm, $P_{pr}\approx 5$ W, gdy $T_0=1$ ps i 50 mW dla $T_0=10$ ps. W przypadku światłowodu o przesuniętej charakterystyce $|\beta_2|=2$ ps$^2$/km i moc jest o rząd wielkości mniejsza. Szczególną rolę odgrywają solitony, których początkowy kształt opisuje funkcja

$$u(0,\tau) = N\,\mathrm{sech}(\tau). \tag{8}$$

Wykresy funkcji $|u|$ spełniającej ten warunek przedstawiono na rys. 3 dla wybranych wartości $\xi$ i $N=1$, 2 i 3.

Moc szczytową konieczną do wytworzenia solitonu rzędu $N$ wyznacza się z definicji $N$ i jest ona $N^2$ razy większa niż w przypadku solitonu podstawowego. Właściwością solitonów wyższych rzędów jest periodyczność [8], tzn. $|u|^2$ jest funkcją okresową o okresie $\xi_0=\pi/2$. Można zatem wyznaczyć okres solitonowy

$$z_0 = \frac{\pi}{2}L_D \approx 0{,}505\,T^2_{FWHM}/|\beta_2| \tag{9}$$




Rys. 4. Czasowa ewolucja solitonu rzędu trzeciego w ciągu okresu solitonowego $z_0$ [9]

Okres solitonowy jest odległością po przebyciu, której soliton wyższego rzędu ma kształt początkowy (rys. 4).

Periodyczność solitonów wyższych rzędów wynika z oddziaływania samomodulacji fazy (SMF) i dyspersji. W przypadku solitonu podstawowego te dwa efekty kompensują się natomiast w przypadku solitonów wyższych rzędów początkowo dominuje SMF a potem dyspersja. W rezultacie impuls w kształcie funkcji secans hiperboliczny doznaje periodycznej ewolucji o okresie $z_0$. Dla $\lambda = 1550$ nm, $\beta_2 = -20$ ps$^2$/km okres solitonowy wynosi $z_0 \approx 80$ m dla $T_0 = 1$ ps i $z_0 \approx 8$ km dla $T_0 = 10$ ps.

Jeśli moc szczytowa nie jest dokładnie taka jaka być powinna tzn. N nie jest liczbą całkowitą, to impuls dąży asymptotycznie do solitonu rzędu $N_c$ (liczba całkowita najbliższa wartości $N$). Część energii ulega rozporoszeniu. Matematycznie, $N = N_c + a$, przy czym $|a| < 1/2$. Dla podstawowego solitonu ($N_c = 1$) impuls staje się szerszy, gdy $a < 0$ i odwrotnie w przypadku $a > 0$; nie powstaje jednak soliton, gdy $N \leqslant 1/2$.

W przypadku dyspersji normalnej ($\beta_2 > 0$), przy założeniu, że $u \to \text{const}$, gdy $\tau \to \pm\infty$, rozwiązaniem NRS jest funkcja

$$u(\xi, \tau) = \text{th}(\tau)\, e^{j\xi}. \tag{10}$$

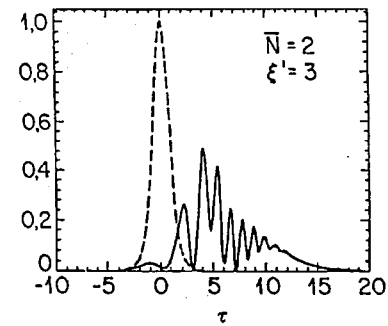
Funkcja $|u|^2$ opisuje wgłębienie w kształcie solitonu w impulsie świetlnym, które nazywa się solitonem ciemnym. Istnienie solitonów wyższych rzędów potwierdzono doświadczalnie [10].

Jeśli w światłowodzie rozchodzi się impuls o długości fali bliskiej długości fali, przy której dyspersja ma wartość zero to $\beta_2 \approx 0$. Należy wtedy w NRS uwzględnić człon zawierający $\beta_3 = \partial^3\beta/\partial\omega^3$ i wtedy otrzymamy równanie

$$j\frac{\partial u}{\partial \xi'} - \frac{j}{6}\frac{\partial^3 u}{\partial \tau^3} + |u|^2 u = 0 \tag{11}$$

przy czym $\xi' = z/L_D$, $u = \bar{N}U$, $L_D' = T_0^3/|\beta_3|$, $\bar{N}^2 = L_D'/L_{NL}$.

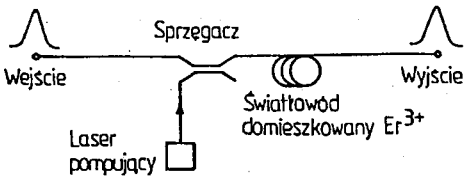
W zakresie długości fal 1300 ÷ 1550 nm $\beta_3$[ps$^3$/km] $\approx D_1$[ps/km nm$^2$], gdzie $D_1 = \partial D/\partial\lambda$, przy czym $D_1$ wynosi $6 \div 8\,10^{-2}$ ps/km nm$^2$.




Rys. 5. Kształt impulsu sech po przebyciu drogi $z = 3L_D'$, gdy $\beta_2 = 0$ i $\bar{N} = 2$

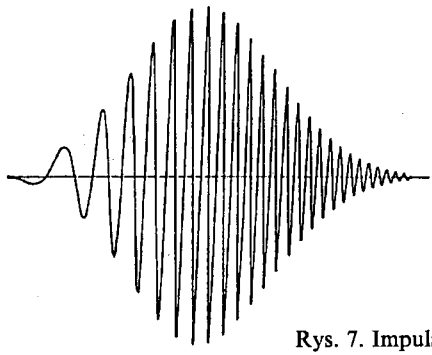
Na rys. 5 przedstawiono impuls sech dla $\beta_2 > 0$ (linia przerywana) i dla $\xi' = 3$, gdy $N = 2$. Widzimy tu oscylacje od strony opadania impulsu. Część energii od strony narastania tworzy soliton dla $\xi' = 5$. Pozostała energia ulega rozproszeniu. Należy

podkreślić, że z powodu rozszerzenia widma na skutek SMF, impuls nie rozchodzi się przy długości fali, przy której dyspersja wynosi zero nawet, gdy początkowo $\beta_2 = 0$. Faktycznie impuls wytwarza swą własną wartość $\beta_2$. Ogólnie, do powstania solitonu w przypadku gdy $\beta_2 = 0$, wymagana jest mniejsza energia niż wtedy, gdy $\beta_2 < 0$.



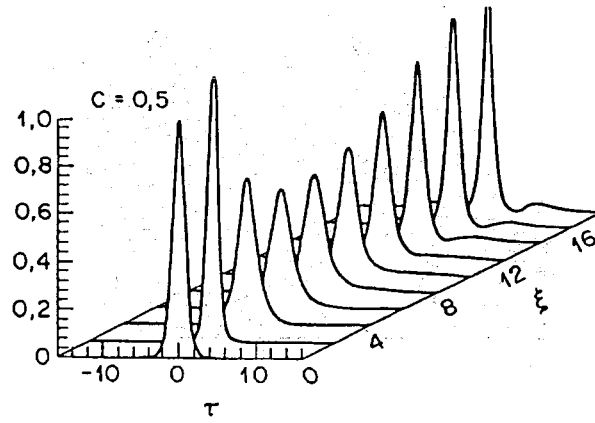

Rys. 6. Tor do przesyłania solitonów

Dotychczas zakładaliśmy, że światłowód jest bezstratny. W rzeczywistości oprócz tłumienności występują również straty energii na skutek mikrozgięć, makrozgięć i na połączeniach. Należy zatem uwzględnić w NRS człon zawierający $\alpha$. Moc impulsu na skutek strat maleje a więc zmniejsza się oddziaływanie SMF na dyspersję co powoduje, że soliton ulega rozszerzeniu. Zachodzi ono prawie liniowo lecz wolniej niż w przypadku światłowodu liniowego. Podobnie jest dla solitonów wyższych rzędów. Dla impulsów o szerokości 100 ps i $\lambda = 1550$ nm, okres solitonowy jest raczej duży (powyżej 500 km) a więc zwężenie się impulsu kompensuje jego poszerzenie na skutek strat. Pozwala to na zwiększenie odległości między regeneratorami nawet dwukrotnie. Wykorzystanie wzmocnienia Ramana do wzmacniania solitonu w sposób ciągły poprzez pompowanie światła o krótszej długości fali do światłowodu pozwala na zupełne skompensowanie strat. Potwierdzono to doświadczalnie. Przesłano impulsy o szerokości 50 ps na odległość 6000 km bez zauważalnej zmiany ich kształtu [11]. Sam światłowód może być również wykorzystywany do ciągłego wzmacniania sygnału jeśli zawiera domieszki metali ziem rzadkich np. $Er^{+3}$ (rys. 6).



Rys. 7. Impuls Gaussa z chirpem dodatnim

Impuls wchodzący do światłowodu nie powinien mieć chirpu (rys. 7). Początkowy chirp nakłada się bowiem na chirp powstający na skutek SMF i powoduje zakłócenie równowagi między dyspersją i SMF. Na rys. 8 podano ewolucję podstawowego solitonu ($N = 1$) *w* przypadku stosunkowo małej wartości chirpu $C = 0{,}5$. Początkowo impuls zwęża się głównie na skutek dodatniego chirpu, potem doznaje poszerzenia a następnie ponownie się zwęża, przy czym część opadająca oddziela się stopniowo od



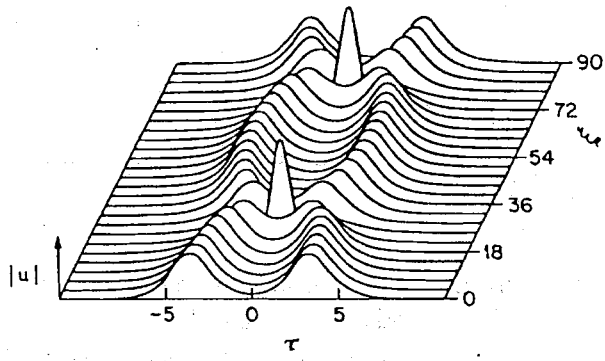

Rys. 8. Tworzenie się solitonu w przypadku chirpu początkowego, gdy $C=0{,}5$ i $N=1$

części głównej, która ewoluje do solitonu po przebyciu drogi $\xi=15$. Utworzenie się solitonu może mieć miejsce tylko dla małych wartości $|C|$. Po przekroczeniu krytycznej wartości $C$ soliton się rozpada. W przypadku $C<0$ impuls rozchodzi się podobnie z wyjątkiem fazy początkowej. Początkowy chirp jest raczej zjawiskiem szkodliwym.

Stwierdzono, że solitony oddziaływują ze sobą. Amplitudę pary solitonów można zapisać jako

$$u(0,\tau) = \operatorname{sech}(\tau - q_0) + r\operatorname{sech}[r(\tau + q_0)]\,e^{j\theta} \tag{12}$$

gdzie $r$ jest względną amplitudą, $\theta$ względną fazą a $q_0$ początkową separacją. W zależności od tych parametrów obserwuje się inne zachowanie takiej pary (rys. 8 i 9).

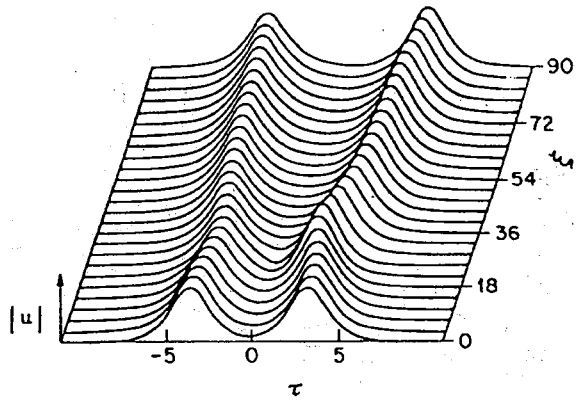



Rys. 9. Para solitonów dla $N=1$, $\theta=0$, $r=1$ i $q_0=3{,}5$ [12]

Jeśli impulsy są krótsze niż 100 fs to zamiast równania (3) należy stosować równanie

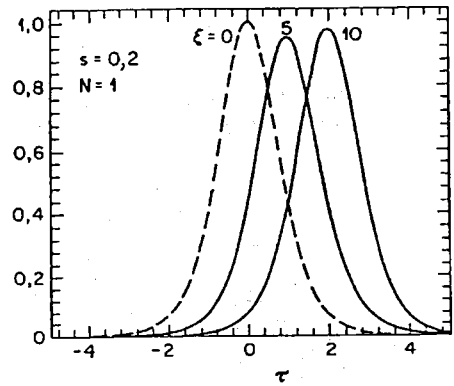
$$j\frac{\partial U}{\partial \xi} + \frac{1}{2}\frac{\partial^2 U}{\partial \tau^2} - j\frac{\partial^3 U}{\partial \tau^3} = -N^2\left[|U|^2 + js\frac{\partial}{\partial \tau}(|U|^2 U) - \tau_R U\frac{\partial}{\partial \tau}(|U|^2)\right] \tag{13}$$

przy czym występujące tu parametry: $\delta=\beta_3/6\,|\beta_2|\,T_0$, $s=2/\omega_0 T_0$, $\tau_R=T_R/T_0$ reprezentują odpowiednio efekty dyspersji wyższego rzędu, samopodnoszenie się zbocza impulsu (ang. self steepening) oraz opóźnioną odpowiedź nielinową. W standardowym światłowodzie $\delta=0{,}03$, $s=0{,}05$ i $\tau_R=0{,}1$ dla $T_0=30$ fs, $\lambda=1550$ nm, gdy



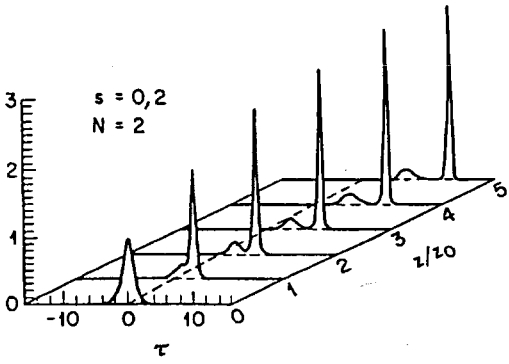

Rys. 10. Para solitonów dla N=1, $\theta=\Pi/2$ $r=1$ i $q_0=3{,}5$ [12]

$T_R=3$ fs. Samopodnoszenie się zbocza impulsu wynika z zależności prędkości grupowej od natężenia światła a, które objawia się w ten sposób, że wierzchołek porusza się wolniej niż zbocza. Prowadzi to do szoku optycznego, gdy nie występuje dyspersja. Dyspersja przeciwdziała temu, jednak obserwuje się przesunięcie środka impulsu (rys. 11).

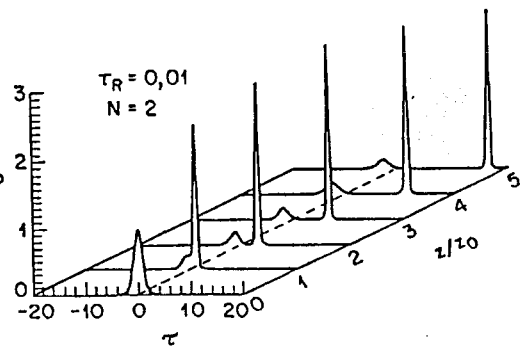



Rys. 11. Kształty solitonów dla $\xi=5$ i 10, gdy $s=0{,}2$. Dla s=0 krzywe te są takie jak w przypadku $\xi=0$

W przypadku solitonów wyższych rzędów zjawisko samopodnoszenia się zbocza impulsu powoduje, że rozpadają się one na solitony podstawowe. Dla $N=2$, gdy $s=0{,}2$ już przy $z=2\,z_0$ następuje rozpad na dwa solitony (rys. 12). Wpływ dyspersji wyższego rzędu na rozpad solitonu uwzględnia człon zawierający trzecią pochodną. Nawet, gdy nie występują efekty nieliniowe wyższego rzędu, następuje rozpad solitonu, gdy $\delta$ przekroczy wartość krytyczną. Dla N=2 wynosi ona 0,022 a dla $N=3$ $\delta=0{,}006$. Efekt opóźnienia nieliniowej odpowiedzi uwzględnia człon zawierający $\tau_R$. Nawet, gdy $\tau_R$ ma małą wartość następuje rozpad solitonów wyższych rzędów (rys. 13). Porównanie rys. 12 i 13 wskazuje, że efekty reprezentowane przez $\tau_R$ są dominujące. Ponadto widzimy, że na rys. 12 oba solitony są opóźnione podczas, gdy na rys. 13 soliton o mniejszej amplitudzie jest przyśpieszony. Trzeba podkreślić, że równanie (13) jest przybliżone. Należy bowiem uwzględnić również zależność nieliniowego współczynnika załamania od czasu.




Rys. 12. Rozpad solitonu rzędu $N=2$ dla $s=0,2$




Rys. 13. Rozpad solitonu rzędu $N=2$ dla $\tau_R=0,01$

Oprócz telekomunikacji solitony znalazły zastosowanie w technice krótkich impulsów do generacji i kompresji impulsów. Wykonano też laser solitonowy.

Solitony mimo iż są wytworem techniki komputerowej stały się czymś realnym w technice.

## BIBLIOGRAFIA

1. L. F. Mollenauer, R. H. Stolen, J. P. Gordon; Phys. Rev. Lett., 45, No 13, 1980, 1095–97
2. J. Scott-Russell; Fourteenth Report of the British Association, 1844, 319–20
3. D. J. Korteweg, G. de Vries; Phil. Mag. and J. of Sci., Ser. 5, 39, No 240, 1895
4. N. J. Zabusky, M. D. Kruskal; Phys. Rev. Lett., 15, No 6, 1965, 240–43
5. W. Je. Zacharow, A. B. Szabat; ŻETF, t 61, W 1, No 7, 1971, 118–34
6. A. Hasegawa, F. Tappert; Appl. Ph. Lett., 23, No 3, 1973, 142–44
7. P. Emplit, J. P. Hamaide, Reynaud F., Froehly C., Bartelemy A.; Opt. Comm., 62, No 6, 1987, 374–5
8. J. Satsuma, N. Yajima; Suppl. of the Progress of Theoretical Phys. No 55, 1974, 284–306
9. G. P. Agrawal; Nonlinear Fiber Optics, 1989, Acad. Press.
10. A. M. Weiner, J. P. Heritage, R. J. Hawkins, R. N. Thurston, E. M. Kirschner, D. E. Leaird; Phys. Rev. Lett., 61, No 21, 1988, 2445–8; Opt. Lett. 14, No 16, 1989, 868–70
11. L. F. Mollenauer, K. Smith; ECOC 89, Gothenburg, 10–14 Sep. 1989, 71–8
12. C. Dessem, P. L. Chu; IEE Proc., 134, Pt J, No 3, 1987, 145–51
13. A. Majewski; Solitony w światłowodach, W. P. W, 1991

A. MAJEWSKI

## SOLITONS IN OPTICAL FIBERS: PROPAGATION CHARACTERISTICS

### Summary

Propagation characteristics of the solitions: fundamental, higher order and dark solitons in single-mode optical fibers and their applications, especially to signal transmission without distortions over a long distance are discussed. The attenuation, chirp, higher order dispersion and nonlinearity are considered.

KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJI, 1992, 38, z. 1, ss. 45-57

621.372.852

# Analysis of correlation between pass-band filters outputs

BORYS KALININ

*Katedra Elektroniki i Teletransmisji, Politechnika Świętokrzyska*

*Otrzymano 1991.02.21*

*Autoryzowano do druku 1991.05.06*

Correlation investigation problem is discussed in association with the definition of the noise compensators characteristics. Methods of normalizing coefficient computation are discussed. Output signal forms for gaussian pass-band filter with presence of frequency difference are investigated. Correlation between the output signal envelopes for two pass-band filters tuned to different frequencies is analised. Parameters values areas where correlation is considerable are specified.

## 1. INTRODUCTION

The problem of the mutual correlation research arises in particular in investigations of the noise compensator potential chracteristics. The noise compensation is an effective method of the noise-immunity encreasing in variant signal transmissing systems. Autocompensators (AC) are used in radiolocation, in communication systems, in medicine. The compensation can be realized, if the noise signal current values are known. An additional compensating receaving channel can be used as a source of such information. For example the additional channels are used in radars for compensation of noise influence by the antenna characteristics side lobes. Required information is present in the main receiver in the case of slowly alternating or periodical noise compensation. Accordingly delayed main signal is used as a compensating signal in such a case. The local signal compensation in radars can be mentioned as an example.

The compensation is realized by the subtraction of the weighted compensating signal from the noisy useful signal in the main channel. The optimal weight coefficient value depends on correlation between the main channel noise signal and the noise signal in the compensating receiver. In principle the absolute noise compensation can be achieved, when the signals are completely correlated. The noise cannot be completely compensated, if the useful signal is present in the compensating channel.

### 1.1. BASIC RELATIONS

Let's denote:
- $y$ is the signal at the compensator output,
- $n_0$ is the noise signal in the main receiving channel,
- $n_k$ is the noise signal in the compensating receiver,
- $P_0$, $P_k$ are the noise power values in compensator channels,
- $K_{0k}$ is the covariance coefficient between the noise signals in the channels,
- $s_0$ is the useful signal,
- $W$ is the weight coefficient value.

Suppose that there is no useful signal in the compensating receiver. Suppose also that the useful signal and the noise signal are not correlated and the noise mean value equals to zero. In this case the noise mean power equals to its variance and the covariance coefficient may be described by the equation:

$$K_{0k} = \overline{n_0 n_k^*} / \sqrt{P_0 P_k},$$

when „ $\overline{\quad}$ " denotes the time-average operation and „*" denotes the complex conjugation. The AC output signal equals to

$$y = s_0 + n_r, \tag{1}$$

where

$$n_r = n_0 - W \times n_k. \tag{2}$$

The weight coefficient $W$ is calculated so, that the noise rest signal $n_r$ is not correlated with the noise signal $n_k$ in the compensating channel:

$$\overline{n_r n_k^*} = 0. \tag{3}$$

Correlation absence ensures the best noise suppression. It's easy to calculate the weight coefficient optimal value using Eqs. (2), (3):

$$W_{opt} = K_{0k} \sqrt{P_0 / P_k}. \tag{4}$$

If the weight coefficient value is optimal, the noise rest power equals to

$$P_r = P_0 (1 - |K_{0k}|^2). \tag{5}$$

As it follows from Eq. (5), the noise rest power equals to zero, if the noise signals $n_0$, $n_k$ are completely correlated. Note that $P_r = 0$ also in the case, when $K_{0k} = -1$; i.e., when the signals shapes are identical but the signals have opposite signs.

In above equations we used everywhere time average values. So the optimal weight coefficient value was estimated by means of time-average operation. Such a calculation method is acceptable if the compensation problem is solved, when it is necessary to suppress given concrete noise realizations. Therefore, the noise signal ergodicity is not required. On the other hand, the compensation figures are true for all the noise realizations in the ergodic case.

### 1.2. OBSERVATION WINDOW CHOISE

The signal observation time interval is usually called the observation window. In our case the window width equals to the smoothed samples number $N$. Let's consider the choise of this number. In general $N$ must be sufficiently large in order to minimize estimation errors. It can be easily ensured in the stationary case. If the noise signal is not stationary, the observation interval cannot be too large. Moreover, the calculated weight coefficient value becomes not optimal, *„obsolete"* at the moment of the caclulation ending. Hence, it is necessary to keep in a storage signals values while the calculation process is not finished. The delay lines can be used in the compensator channels for this purpose.

There ate two methods of the $N$ needed samples choise: the crawling window method and the method with a group (jumping) window. If the crawling window is used, the proper weight coefficient value is calculated for each compensating signal sample. For a new current value calculation a new signal sample is added to the $N$ samples set and the oldest sample is eliminated. In the group window case the same weight coefficient value is used for the $N$ samples signal set compensation. Then the weight coefficient value is calculated for the new group of signal samples etc. The group coefficient value must be kept in the compensator memory, if this calculation method is used.

In general, the number $N$ can be variable. For example the weight coefficient group value can be calculated separately for each noise fluctuation. In order to use this method the beginning and the ending of each fluctuation must be determined at first.

Another corresponding problem is the problem of the window shape choise, in other words, the problem of the integrating filters characteristics choise. If the rectangular window is used, each signal sample have the same influence on the calculation result. But it is intuitively evident, that a sample at the moment near to the compensation moment must have a greater influence on the weight coefficient value than a distant one. Non-rectangular windows can be used for this conception realisation. For example the window shape can be choosed close to the receiving channel pulse characteristics shape.

## 2. ANALYSIS OF CORRELATION BETWEEN PASS-BAND FILTERS OUTPUT ENVELOPES

### 2.1. PRELIMINARIES

As it follows from the above analysis the extention of the area of the compensator applications is connected with finding sources of additional information about the noise signal. The compensator potential quality investigation comes to the analysis of correlation between the noise signals in the compensator channels. Let the interfering signal be a radio pulse which carrier frequency can differ from the receiver center

frequency. Such a situation can arise when radio installations transmit in neibouring frequency bands. Pulse interference arises also when internal-combustion engines work etc. Consider two different types of enterference pulses: the rectangular pulse and the pulse with the gaussian envelope. The gaussian pulse has infinite duration and its amplitude spectrum falls gradually in infinity. For finite duration pulses such as a rectangular pulse the amplitude spectrum is an oscillating function. In order to analise the output signal characteristics immanent to the given input signal shape, it is necessary to choose an adequate receiver frequency characteristics (FC) description. The ideal pass-band filter with rectangular FC is not convenient for this purpose because its pulse characteristics (PC) is the function of $\mathrm{Sin}(x)/x$ type and each pulse at the filter input produces an infinite pulse sequence at the output. It's better to choose the FC description in the shape of a smoothing window such as Hamming window. Kaiser window etc. Let's describe FC with the help of the gaussian function:

$$K(f) = \exp[-b^2 (f - f_k)^2], \tag{6}$$

where $f_k$ is the filter central frequency and b depends on filter bandwidth. The gaussian filter is not causal, its PC does not equal to zero on the left half of absciss axis. But an approximate amplitude caracteristics of Eq. (6) type can be formed in a real filter having sufficiently large signal delay. It's known that the gaussian function (6) describes the resonant amplifier amplitude FC if the number of the amplifier stages is sufficiently large. The gaussian filter don't introduce phase distortions and each amplitude jump at the filter input produces only one pulse at the output.

Consider a set of two similar gaussian filters tuned to different frequencies. Let's denote:

- $t$ is a current time value,
- $f$ is a frequency,
- $T$ is the input signal duration (in case of gaussian pulse $T$ equals to the half power level duration),
- $T_{0.001}$ is the duration of the gaussian pulse at level 0.001 from its maximal value ($T_{0.001} = 4.6 \times T$),
- $B_f$ is the filter bandwidth (at the half power level),
- $B_s$ is the signal spectrum bandwidth,
- $f_k$ is the k-th filter pass-band central frequency ($k = 1,2$),
- $\Delta f = |f_1 - f_2|$ is the mutual frequency difference – the difference between central frequencies of the filters,
- $f_s$ – the signal carrier frequency,
- $\Delta_k = |f_k - f_s|$ is the $k$-th filter frequency difference with reference to the signal carrier frequency.

The signal spectrum bandwidth can be written as $B_s = a_s/T$, where $a_s$ is the coefficient depedent on the signal shape ($a_s = 0.44$ for the gaussian shape pulse and $a_s \cong 1$ for the rectangular pulse). Let the values $B_s$, $B_f$, $\Delta f$, $\Delta_k$ be small in comparsion

with the signal carrier frequency. In such a case the complex signal representation can be used.

As it is known the pulse characteristics envelope of the gaussian filter (7) is also gaussian. Let $M_f$ denote the PC envelope maximum value. Let's define the transient process duration $T_{tr}$ as the time interval necessary for the PC amplitude decreasing from $M_f$ to $0.01 \times M_f$. It follows from this definition that $T_{tr}=0.812/B_f$. Let's define the output signal duration $T_{out}$ as the sum $T_{out}=T_{inp}+2\times T_{tr}$, where $T_{inp}$ is the input signal duration; $T_{inp}=T$ for the rectangular pulse and $T_{inp}=T_{0.001}$ for the gaussian input signal.

In order to make the calculation results more general, let's introduce relative parameters values into consideration. For this purpose use the signal spectrum bandwidth as the basic value. In this way introduce the next relative values:

– $r_f=f/B_s=T\times f/a_s$ is the relative frequency,

– $T_f=B_f/B_s=T\times B_f/a_s$ is the relative filters bandwidth or else $T_f$ is the relative pulse duration,

– $r_t=t\times B_f/a_s$ is the relative time,

– $r_{fk}=f_k/B_s$ is the $k$-th filter relative central frequency,

– $r_{\Delta f}=\Delta f/B_s=T\times \Delta f/a_s$ is the relative mutual frequency difference,

– $\Delta r_k=\Delta_k/B_s=T\times \Delta_k/a_s$ is the $k$-th filter relative frequency difference (with reference to the signal carrier frequency).

The calculation results have been obtained with the help of a digital simulating model using FFT method. The finite duration FC is formed in the model from the gaussian function by means of two methods:

1) with the help of the rectangular window

$$K_1(f)=\begin{cases}K(f), & |f-f_k|\leqslant \Delta F,\\ 0, & \text{otherwise},\end{cases} \tag{7}$$

2) by using the Hamming window

$$K_2(f)=\{0.54+0.46\,\mathrm{Cos}[\pi(f-f_k)/\Delta F]\}\,K(f),\ \text{for}\ |f-f_k|\leqslant \Delta F. \tag{8}$$

The $\Delta F$ value is choosen so that

$$K(f\pm \Delta F)\leqslant 10^{-3} \tag{9}$$

that is, $\Delta F=2.3\times B_f$. $K(f)$ is the gaussian function defined by Eq. (6). Time-averages estimations are computed in the model on the time interval equal to the output signal duration $T_{out}$.

## 2.2. RECTANGULAR SIGNAL CASE

The calculations were performed for the filters with the FC of Eq. (8) type. At first the output signal shapes were analised. The case of a small pass-band value $T_f=0.25$ is illustrated by Fig. 1.

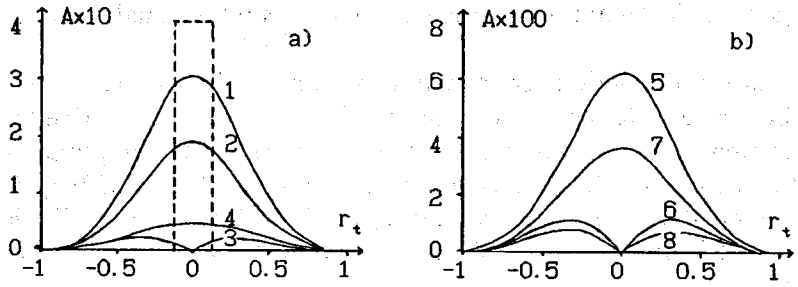



Fig. 1. The case of a small pass-band value $T_f=0.25$

The output signal envelopes are depicted for various values of the difference between the filter central frequency and the signal carrier frequency. Vlues of the relative frequency difference are given in Table 1.

Table 1

Values of the relative frequency difference

| N | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| $\Delta r$ | 0 | 0.5 | 1.0 | 1.25 | 1.5 | 2.0 | 2.5 | 3.0 |

Relative time values are put along the absciss axis. Signal amplitude values are multiplied by 10 or by 100. The input pulse shape is shown by dotted line. (The signal spectrum $S$ and the filter caracteristics $F_1$, $F_2$ for the frequency difference values $\Delta r=1.5$ and $\Delta r=2.0$ are shown in Fig. 4). As it is seen from Fig. 1, the output signal envelope has a distinctive shape with two maxima – a *„two-humped"* shape, if the relative frequency difference value is integer. If the frequency difference increases, the signal shape alters quasi-periodically being in turn two-humped or one-humped. The following explanation of this phenomenon may be given. If $\Delta r=k+1/2$, where $k$ is integer, the filter central frequency coincides with one of signal-spectrum maxima, and the output signal shape is similar to the gaussian pulse characteristics of the filter. If $\Delta r$ value is integer, the filter central frequency coincides with *„zero"* of the input signal spectrum (see Fig. 4). In this case the output spectrum is similar to the envelope-differentiator FC and the output signal shape is similar to the two-humped envelope of the differentiator pulse characteristics [4].

Output signal shapes for $T_f=1.0$ are shown in Fig. 2. Frequency difference values are given in Table 2.

In this case the signal shape is two-humped for all $\Delta r$ values greater than 1. As it is seen from Fig. 2, the relative minimum value in the center of pulse alters quasi-periodically, if $\Delta r$ increases.

Signal shapes for large pass-band value $T_f=2.0$ are shown in Fig. 3. Frequency difference values are given in Table 3.

In this case, if $\Delta r \geqslant 4$, the output signal is a sequence of two gaussian pulses. Output signal amplitude is maximum at points obeginning and ending of the input pulse.

# Szanowni Czytelnicy

**Technical University of Lodz**
**Institute of Electric Power Engineering**
**POLAND**

**Universite de Liège**
**Institut d'Electricite Montefiore**
**BELGIUM**

## 5th INTERNATIONAL SYMPOSIUM ON SHORT-CIRCUIT CURRENTS IN POWER SYSTEMS

**Warsaw, 8 - 10 September 1992**

W imieniu Komitetu Naukowego i Komitetu Organizacyjnego V Międzynarodowego Sympozjum nt. **PRĄDY ZWARCIOWE W SYSTEMACH ELEKTROENERGETYCZNYCH** informujemy, że sympozjum to odbędzie się w Warszawie w dniach 8-10 września 1991 r.

Adres organizatorów: Prof. Zbigniew Kowalski
Politechnika Łódzka, Łódź
ul. Stefanowskiego 18/22

tel. 48/42/361193, 48/42/312605
Fax: 48/42/368522, Telex: 886136

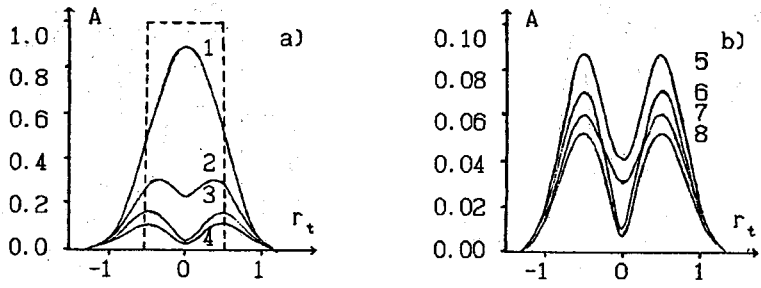



Fig. 2. Output signal shapes for $T_f=0.1$

Table 2

Frequency difference values for the curves of Fig. 2

| N | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| $\Delta r$ | 0 | 1.0 | 1.5 | 2.0 | 2.5 | 3.0 | 3.5 | 4.0 |

The next calculation series was performed for the set of two gaussian filters with fixed mutual frequency difference. Correlation between the output signal envelopes was investigated as a function of the first filter frequency difference.

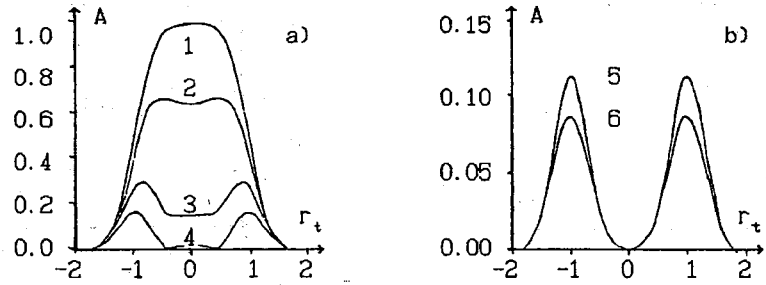



Fig. 3. Signal shapes for large pass-band value $T_f=2.0$

Table 3

Frequency difference values for the curves of Fig. 3

| N | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $\Delta r$ | 0 | 1.0 | 2.0 | 3.0 | 4.0 | 5.0 |

There are in Figs. 4 – 6 illustration to the case of a small pass-band value $T_f=0.25$. The fixed mutual frequency difference equals to $r_{\Delta f}=0.5$ for the curves depicted in Figs. 4.5. The input signal spectrum and the caracteristics of the filters are shown in Fig. 4. Correlation coefficient values are shown in Fig. 5 as a function of the first filter frequency difference $\Delta r_1$. The second filter frequency difference increases at the same

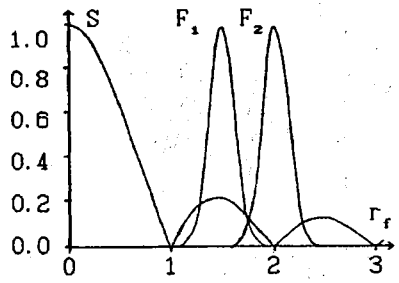



Fig. 4. The caracteristics of the filters and the signal spectrum

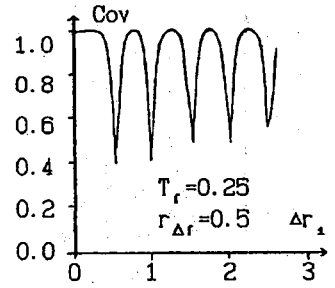



Fig. 5. The correlation coefficient values for the case of $T_f=0.25$, $r_{\Delta f}=0.5$

time as $\Delta r_1$, while mutual frequency difference $r_{\Delta f}$ is constant. If $\Delta r_1 \geqslant 0.5$, the covariance coefficient alters quasi-periodically, when $\Delta r_1$ increases. Correlation is minimal if $\Delta r_1 = 0.5 \times m$, where $m$ is integer and $m \geqslant 2$. In such a case output signals have different shapes: one of the signals has a two-humped envelope while the other is a gaussian pulse (see Fig. 1). If $r_{\Delta f}=1.0$, that is, the mutual frequency difference equals to the frequency difference between input signal spectrum side-lobes, the signal envelopes at filters outputs alter synchronously when $\Delta r_1$ increases, so that filters outputs shapes are similar for all $\Delta r_1$ values. As a result, the correlation coefficient values are close to 1 if $\Delta r_1 > 0.3$ (see Fig. 6).

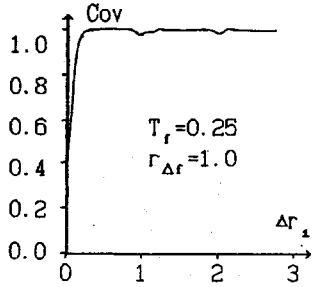



Fig. 6. The correlation coefficient values for the case of $T_f=0.25$, $r_{\Delta f}=1.0$

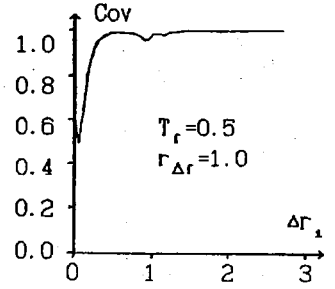



Fig. 7. The function cov ($\Delta r_1$) for the case of $T_f=0.5$, $r_{ff}=1.0$

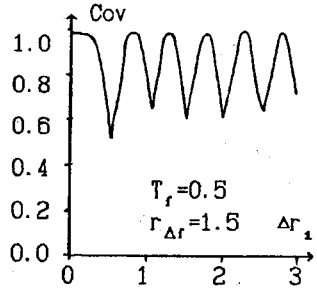



Fig. 8. The function cov ($\Delta r_1$) for $T_f=0.5$, $r_{\Delta f}=1.5$

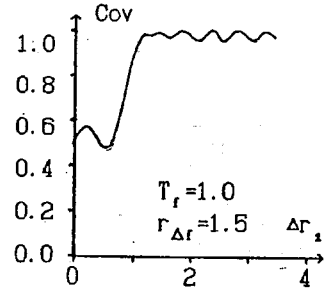



Fig. 9. The function cov ($\Delta r_1$) for $T_f=1.0$, $r_{\Delta f}=1.5$

The function $\mathrm{cov}(\Delta r_1)$ for $T_f = 0.5$ is depicted in Figs. 7,8. Curve shown in Fig. 7 corresponds to the mutual frequency difference value $r_{\Delta f} = 1.0$. In this case signal shapes at filters outputs alter synchronously and $\mathrm{cov}(\Delta r_1) \cong 1$ if $\Delta r_1 > 0.3$. If $r_{\Delta f} = 1.5$, signal shapes alter *„in opposite phases"*: when one of the signals has a two-humped shape, another signal envelope is one-humped. As a result, $\mathrm{cov}(\Delta r_1)$ is an oscillating function as it seen from Fig. 8.

The illustrations to the case of $T_f = 1.0$ are given in Figs. 9,10. If $r_{\Delta f} = 1.5$, covariance coefficient value increases to nearly 1, while $\Delta r_1$ alters from 0 to 1, and oscillates slightly near 1, when $\Delta r_1 > 1$ (see Fig. 9). If $r_{\Delta f} = 2.0$ covariance coefficient equals practically to 1 when $\Delta r_1 > 1.5$ (see Fig. 10).

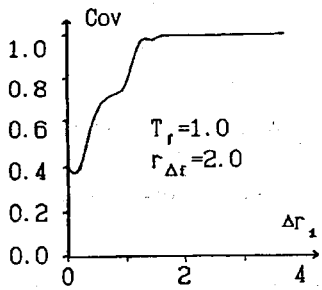



Fig. 10. The function cov $(\Delta r_1)$ for $T_f = 1.0$, $r_{\Delta f} = 2.0$

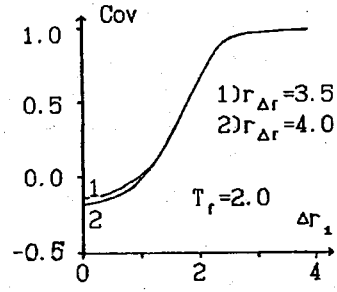



Fig. 11. The function cov $(\Delta r_1)$ for the case of $T_f = 2.0$

The case of a large pass-band value $T_f = 2.0$ is illustrated by Fig. 11 which shows the covariance coefficient alteration for two frequency difference value: $r_{\Delta f} = 3.5$ (curve 1) and $r_{\Delta f} = 4.0$ (curve 2). In this case $\mathrm{cov}(\Delta r_1)$ is a monotonous increasing function for both $r_{\Delta f}$ values.

### 2.3. CASE OF INPUT SIGNAL WITH GAUSSIAN ENVELOPE

If both the filter caracteristics and the input signal envelope are gaussian, than the output signal envelope also is a gaussian function. But in practice we can't use ideal gaussian functions. So both the filter FC and the input signal are finite-lenth sequences. As result, the output signal spectrum includes an oscillating part of relatively small amplitude and the output signal envelope is two-humped when the frequency difference is large. In order to minimise the oscillating part of spectrum, the filter FC was formed by use of the Hamming window Eq. (8) and the input signal enveloped was formed with the help of the Hann window

$$x(t) = 0.5[1 + \mathrm{Cos}(\pi t/\Delta T)]\exp(-\alpha^2 t^2), \; |t| \leqslant \Delta T. \tag{10}$$

In Eq. (10) the coefficient $\alpha$ is defined by the impulse duration at half-power level and the range $\Delta T$ is choosen satisfying the condition

$$\exp(-\alpha^2 \Delta T^2) \leqslant 10^{-3} \tag{11}$$

that is, $\Delta T = 0.5 \times T_{0.001} = 2.3 \times T$.

The output signal envelopes are depicted in Fig. 12,13 for various values of the difference between the filter central frequency and the signal carrier frequency. The relative filter bandwidth equals to $T_f = 1.0$. Values of the relative frequency difference $\Delta r$ are given in the Figs.

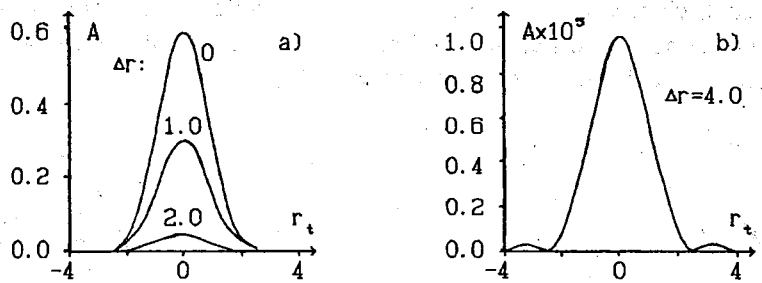



Fig. 12. The output signal envelopes for various frequency difference $\Delta r$

As it can be seen, the output signal envelope is gaussian when the relative frequency difference is small. If the frequency difference is sufficiently large, the envelope is two-humped with maxima at the points of the beginning and the ending of the input pulse: $t = \pm \Delta T$.

Calculations were also performed for the case when both the filter FC and the input signal were formed from gaussian functions with the help of rectangular windows. Both the signal and the frequency characteristics were truncated at level 0.001. The results are illustrated by Fig. 14, which shows the output signal envelope for two frequency difference values given in the Fig. The relative filter pass-band equals to $T_f = 1.0$. As it is seen gaussian shape distortions arise at smaller frequency difference values than in previous case. The envelope is two-humped already when $\Delta r_1 = 4.0$.

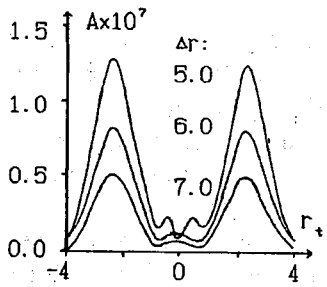



Fig. 13. The output signal envelopes

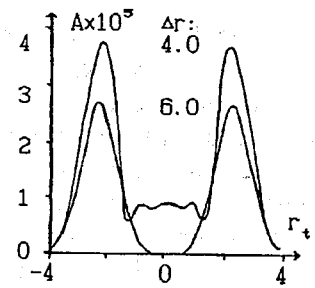



Fig. 14. The signal envelopes for the case of rectangular windows

Discussed signal shapes define correlation between signals at filters outputs. Calculations were performed for the set of two filters with fixed mutual frequency

difference. Covariance between the filters outputs envelopes was investigated. Calculations results are illustrated by Figs. 15,16,18–20, which show covariance coefficient values as a function of the first filter frequency difference $\Delta r_1$.

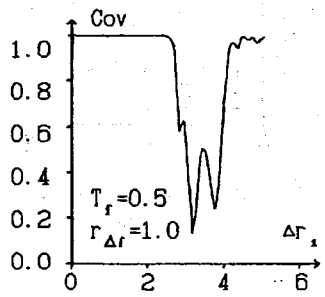



Fig. 15. The correlation coefficient values for the case of $T_f=0.5$, $r_{\Delta f}=1.0$

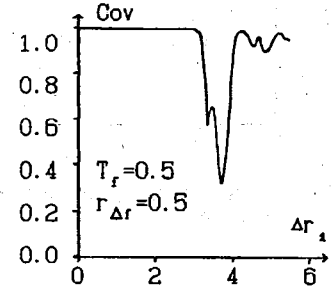



Fig. 16. Function cov ($\Delta r_1$) for the case of $T_f=0.5$, $r_{\Delta f}=0.5$

For the curves plotted in Figs. 15,16 the relative filters pass-band value equals to $T_f=0.5$. The mutual filters frequency difference equals to 1.0 for Fig. 15 and to 0.5 for Fig. 16. The covariance coefficient equals to 1 while the frequency difference $\Delta r_1$ is small and both output signals are gaussian. Covariance decreases rapidly when the second filter output envelope becomes a two-humped curve. If $\Delta r_1$ is sufficiently large, so that both output signals are two-humped functions, the covariance coefficient value is near 1. The input signal spectrum and the filters characteristics are shown in Fig. 17 for the case of $T_f=0.5$ and $r_{\Delta f}=1.0$.

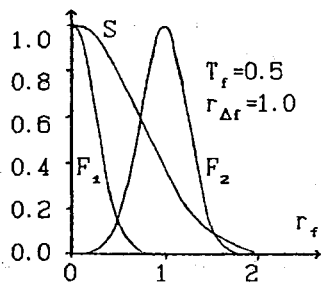



Fig. 17. Input signal spectrum and the filters characteristics for the case $T_f=0.5$ and $r_{\Delta f}=1.0$

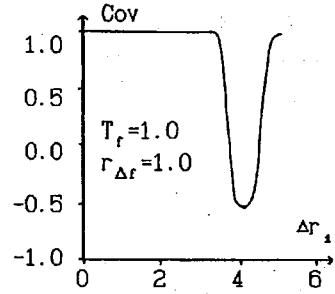



Fig. 18. Function cov ($\Delta r_1$) for the case of $T_f=1.0$ and $r_{\Delta f}=1.0$

The function cov($\Delta r_1$) for another values of the filters pass-band $T_f$ and the mutual frequency difference $r_{\Delta f}$ is plotted in Fig. 18–20. Values of $T_f$ and $r_{\Delta f}$ are given in the Figs. As it is seen, three areas of argument values can be specified for all shown curves. The first is the area of small $\Delta r_1$ values in which both output signals are gaussian and the covariance coefficient equals to 1.

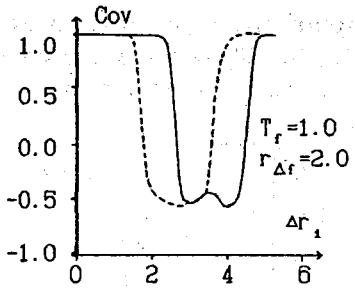



Fig. 19. Function cov ($\Delta r_1$) for the case of $T_f$=1.0 and $r_{\Delta f}$=2.0

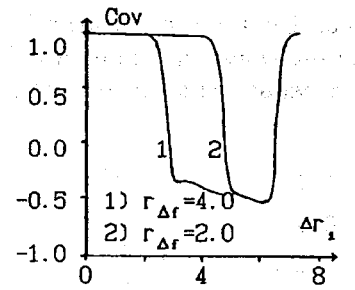



Fig. 20. Function cov ($\Delta r_1$) for the case of $T_f$=2.0

The second area is the area of negative covariance. In that area analysed signals have different shapes: the first filter output is gaussian while the output of the second filter has a two-humped shape. The second area width increases with increasing of the filters pass-band $T_f$ and the mutual frequency difference $r_{\Delta f}$. The third area is the area of large $\Delta r_1$ values. In that area both signals envelopes are two-humped and the covariance coefficient is near 1.

In order to compare, covariance values were also calculated for the case when both the filter FC and the input signal were truncated with the help of the rectangular window. Calculations results are shown in Fig. 19 by dotted line for $T_f$=1.0 and $r_{\Delta f}$=2.0. When rectangular windows are used, gaussian signal shape distortions arise at smaller values of the frequency difference. Accordingly, the area of negative covariance is displaced to the left.

Hence, there are two areas of frequency difference values in which the covariance coefficient equals to 1 or is near 1. Note that only the first of these areas is of importance in practice because the output signal amplitude is negligibly small in the area of large argument values.

## 3. CONCLUSION

Output signal shapes had been investigated for gaussian filters in presence of a difference between the signal carrier frequency and the filter central frequency. Two input signal types were analysed: the rectangular shape pulse and the signal with the gaussian envelope. Correlation between output signal envelopes was investigated for a set of two gaussian filters with a mutual frequency difference. As a result the areas of parameters values can be specified in which signals envelopes are correlated considerably and, hence, can be used for the mutual compensation. For gaussian input signals that is the area of small frequency difference values in which both output signals are gaussian. For rectangular input pulses two areas of considerable covariance values can be specified. The first is the area of large frequency difference values and large filters band-width values, when each of the filters passes more than one signal spectrum

side-lobes. The second is the area of small frequency differencies and small filters band-width, when both filters pass-bands are put completely within the boundaries of the signal spectrum main lobe.

## REFERENCES

1. У и д р о у Б. и др.: *Адаптивные компенсаторы помех. Принципы построения и применения.* ТИИЭР, 1975, т. 63, N 12
2. Ю ж а к о в В. Д.: *Принципы построения автокомпенсаторов сигналов активных помех.* Зарубежная радиоэлектроника, 1986, N 2
3. К а л и н и н Б. Н.: *Характеристики систем автокомпенсации помех.* Krajowe sympozjum telekomunikacji'89. Bydgoszcz, 1989
4. K a l i n i n B.: *Digital simulation of envelope differentiation process for narrow-band signal with noise.* Proceedings of the 1989 COSMEX Meeting, World Scientific, Singapore, 1990

B. KALININ

## BADANIA KORELACJI WZAJEMNEJ MIĘDZY OBWIEDNIAMI SYGNAŁÓW WYJŚCIOWYCH FILTRÓW PASMOWOPRZEPUSTOWYCH

S t r e s z c z e n i e

Badania korelacji rozpatrywane są w związku z określeniem charakterystyk autokompensatorów zakłóceń. Przeanalizowane są metody wyliczania wartości optymalnej współczynnika normującego. Rozpatrywane są kształty sygnałów wyjściowych filtru pasmowoprzepustowego w obecności pobudzeń impulsowych. Przebadana jest korelacja między obwiedniami sygnałów wyjściowych dwóch filtrów o wzajemnie przesuniętych pasmach przepustowości. Znalezione są obszary znacznej korelacji sygnałów.

KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJI, 1992, 38, z. 1, ss. 59-65

621.391.332.1

# Procedura filtracji liniowej obrazu

JÓZEF MODELSKI, ANDRZEJ BUCHOWICZ

*Instytut Radioelektroniki, Politechnika Warszawska*

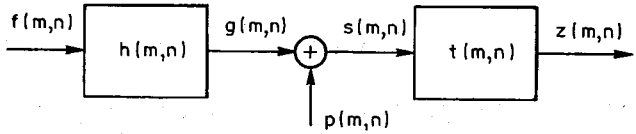
*Otrzymano 1991.04.26*

*Autoryzowano do druku 1991.06.04*

W pracy opisano sposób realizacji procedury filtracji liniowej obrazu w języku C. Zastosowana organizacja danych redukuje liczbę przesłań zawartości pamięci obrazu do i z buforów w pamięci operacyjnej komputera jak również liczbę przesłań pomiędzy buforami, co zapewnia znaczne skrócenie czasu wykonywania programu.

## 1. WPROWADZENIE

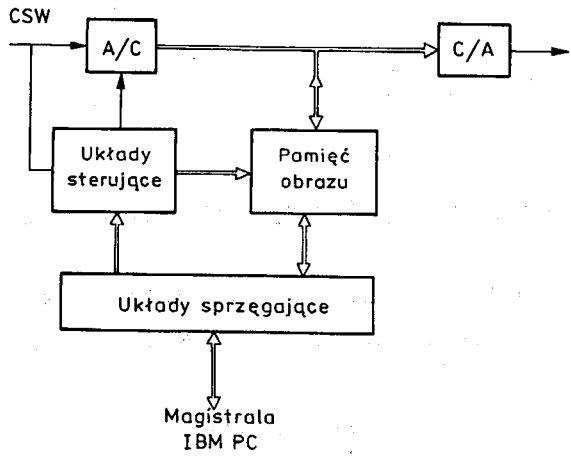
Schemat toru przetwarzania obrazu przedstawiono na rys. 1. Opracowanie algorytmu poprawy jakości obrazu polega na znalezieniu takiej funkcji t[m,n], która kompensowałaby zniekształcenia wnoszone przez układy formowania i przesyłania obrazu opisane funkcją h[m,n], eliminowała zakłócenia opisywane funkcją p[m,n] i nie powodowała przy tym pogorszenia jakości obrazu (np. pogorszenia rozdzielczości). W szeregu stosowanych algorytmów filtracja obrazu polega na przyporządkowaniu




Rys. 1. Tor przetwarzania obrazu: f[m,n] – funkcja reprezentująca obraz wejściowy, h[m,n] – funkcja reprezentująca zniekształcenia wnoszone przez układ formowania i tor przesyłania obrazu, p[m,n] – funkcja reprezentująca zakłócenia (np. szumy) powstające w torze przesyłania obrazu, t[m,n] – funkcja opisująca realizowany algorytm z[m,n] – obraz wyjściowy, m=0..M−1, n=0..N−1, M, N – rozmiary obrazu

każdemu punktowi obrazu wyjściowego liczby będącej pewną funkcją punktów otoczenia odpowiadającego mu punktu w obrazie wejściowym:

$$z[m,n]=t(\ldots, s[m+i, n+j],\ldots)$$
$$\text{gdzie: } -K/2 < i < K/2 \quad -L/2 < j < L/2$$
$$K/2 < m < M-K/2 \quad L/2 < n < N-L/2,$$




Rys. 2. Schemat blokowy karty przetwornika obrazu

liczby $K$, $L$ określają wielkość otoczenia, które jest uwzględniane przy obliczeniach odpowiedzi filtru.

Obraz wejściowy dla procedur filtracji jest uzyskiwany za pomocą karty przetwornika obrazu (*frame grabber*), której uproszczony schemat blokowy pokazano na rys. 2. Wejściowy sygnał wizyjny pochodzący, np. z kamery telewizyjnej jest próbkowany, a następnie kwantowany w przetworniku analogowo-cyfrowym. Częstotliwość próbkowania jest całkowitą wielokrotnością częstotliwości impulsów synchronizujących linii występujących w sygnale wejściowym. Skwantowane próbki są zapisywane w pamięci obrazu i równocześnie przesyłane do wyjściowego przetwornika cyfrowo-analogowego. Zawartość pamięci obrazu może być odczytywana i modyfikowana przez komputer klasy IBM PC współpracujący z kartą przetwornika obrazu jak również przesyłana do przetwornika C/A w celu jej wyświetlenia.

## 2. SPOSÓB REALIZACJI ALGORYTMU FILTRACJI OBRAZU

W celu wybrania optymalnego sposobu komunikacji programu filtreacji z pamięcią obrazu karty zostały przetestowane trzy metody realizacji algorytmu polegające na:

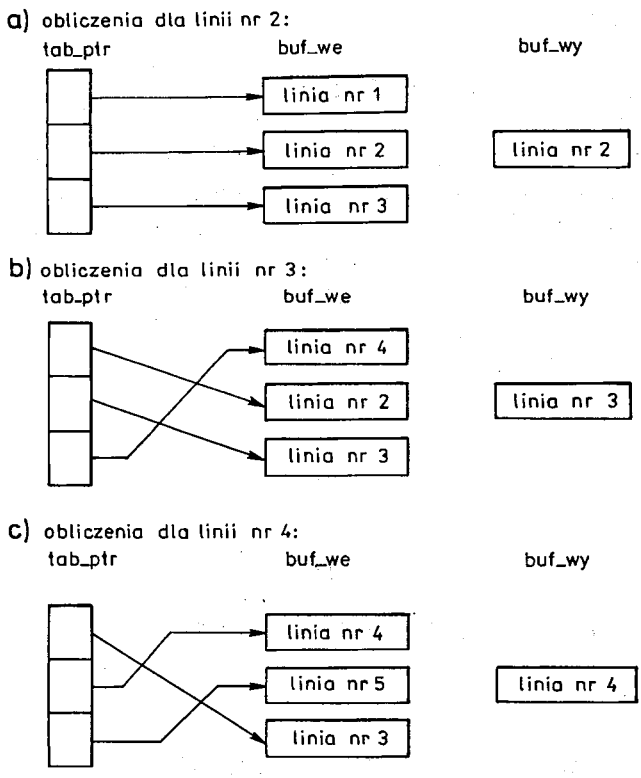
– bezpośrednim odczycie i zapisie poszczególnych punktów w pamięci obrazu

– podziale całego ekranu na szereg prostokątnych okien, kopiowaniu poszczególnych okien do pamięci programu i niezależnym przetwarzaniu każdego z tych okien

– kopiowaniu odpowiedniej liczby linii obrazu telewizyjnego do pamięci programu i filtracji poszczególnych linii.

Uzyskane czasy filtracji obrazu były w przypadku metody bezpośredniego odczytu i zapisu poszczególnych punktów w pamięci obrazu dziesięciokrotnie większe w stosunku czasów uzyskiwanych w dwóch pozostałych metodach. Filtracja metodą podziału obrazu na szereg okien i niezależnej filtracji każdego z nich ma poważną wadę polegającą na tym, że pozostawia znaczną liczbę nieprzetworzonych punktów na

krawędziach styku okien, gdzie operacja filtracji jest niezdefiniowana. Punkty te tworzą wyraźnie widoczną sieć obejmującą całą powierzchnię ekranu. Ostatecznie zdecydowano się na realizację filtru metodą kopiowania linii, w której zastosowano strukturę danych umożliwiającą minimalizację liczby przesłań pomiędzy pamięcią obrazu i buforami linii w pamięci programu. Zostanie ona omówiona dla przypadku, w którym do obliczenia odpowiedzi filtru wykorzystana jest zawartość trzech kolejnych linii obrazu. Schematycznie została ona przedstawiona na rysunku 3. W procedurze zadeklarowane są trzy bufory przechowujące trzy kolejne linie obrazu wejściowego, jeden bufor przechowywujący jedną linię obrazu wyjściowego oraz tablica wskaźników. W początkowej fazie przetwarzania kopiowane są do buforów trzy pierwsze linie obrazu i inicjalizowana jest tablica wskaźników w ten sposób, że pierwszy z nich wskazuje na początek pierwszego bufora, drugi na początek pierwszego bufora, drugi na początek drugiego i.t.d. Następnie obliczana jest odpowiedź filtru dla drugiej linii obrazu (pierwsza linia pozostaje nie zmieniona), wynik jest umieszczony w buforze wyjściowym. Po zakończeniu obliczeń dla linii nr 2 zawartość bufora wyjściowego jest kopiowana do pamięci obrazu, do bufora przechowującego dotychczas linię nr 1 kopiowana jest linia nr 4 i zamieniane są wskaźniki zgodnie z rys. 3b. Powyższe operacje są powtarzane aż do zakończenia




Rys. 3. Schemat struktur danych zastosowanych w procedurach filtracji

filtracji całego obrazu. Wskaźniki w tablicy wskaźników są tak ustawiane, że drugi z nich zawsze wskazuje na bufor przechowujący linię aktualnie przetwarzaną, pierwszy na linię poprzednio przetwarzaną, trzeci na linię dla której odpowiedź będzie wyznaczana w następnym kroku programu, dzięki czemu możliwy jest dostęp do wszystkich elementów obrazu niezbędnych do obliczenia odpowiedzi filtru. Zaletą takiej organizacji programu jest to, że poszczególne linie obrazu są kopiowane tylko raz z pamięci obrazu do odpowiedniego bufora wejściowego, a następnie z bufora wyjściowego do pamięci obrazu, przy przejściu do obliczeń kolejnych linii zamieniane są tylko wartości wskaźników, a nie są wymieniane między sobą zawartości buforów. Powyższy schemat może być stosowany także w przypadku, gdy do obliczeń niezbędna jest większa liczba linii, wymaga to jedynie powiększenia liczby buforów i zwiększenia pojemności tablicy wskaźników.

Procedura filtracji została zaprogramowana w języku C. W podanym poniżej listingu realizuje ona filtrację liniową, której współczynniki zawarte są w tablicy wskazywanej przez zmienną *a*. W analogiczny sposób może być zaprogramowany dowolny inny algorytm

```
#include <dos.h>
#include <memory.h>
#include <malloc.h>
#include <vist_lib.h>
        /* kody bledow procedury filtracji */
#define NO_MEM -1
int filtracja(int *a,           /*wspólcz. filtru w kolejnych wierszach*/
              int M1, int N1, /*wspólcz. lewego-dolnego wierzchołka obrazu*/
              int M2, int N2, /*wspólcz. prawego-górnego wierzołka*/
              int K,          /*liczba kolumn okna filtru*/
              int L)          /*liczba linii okna filtru*/
{
register unsigned char far **tab_ptr;
register int k,l;
unsigned char far *buf_wy;
int m,n;
int K_2=K/2,
    L_2=L/2,
    szer=M2-M1+1; /*ilosc pikseli w jednej linii */
int suma_wsp;
        /* obliczenie sumy współczynników filtru */
for(suma_wsp=0,k=0;k<K*L;k++)
    suma_wsp+=a[k];

        /* przydzielenie pamięci dla tablicy wskaźników */
tab_ptr=malloc(L*sizeof(char**));
tab_ptr+=L_2;           /* przesunięcie początku tablicy wskaźników*
```

```
        /* przydzielenie pamięci dla buforów linii */
for(l=L_2;l<=L_2;l++)}
    tab_ptr[l]=malloc(szer);
    if(tab_ptr[l]==NULL)
      return NO_MEM;}

        /* przydzielenie pamięci dla bufora wyjściowego */
if((buf_wy=malloc(szer))==NULL)
   return NO_MEM;

        /* wczytanie pierwszych L-1 linii */
for(l=-L_2,n=N1;l<L_2;l++,n++){
   read_line(tab_ptr[l],n,M1,szer);}

        /* filtracja obrazu */
for(n=N1+L_2;n<N2-L_2;n++){
    read_line(tab_ptr[L_2],n+L_2,M1,szer); /* wczytanie kolejnej linii */
    for(m=0;m<K_2;m++){
    buf_wy[m]=(tab_ptr[0])[m]; /* pierwsze K/2 punkty każdej linii bez zmian */

}
for(m=K_2;m<=M2-M1-K_2;m++){
    register int suma=0;
    register int *temp_a;
    temp_a=a;
    for(l=-L_2;l<=L_2;l++){
        for(k=-K_2;k<=K_2;k++){
            suma+=(tab_ptr[l])[m+k]*(*temp_a++);}
    }
    buf_wy[m]=(unsigned char)(suma/suma wsp);}

for(m=M2-M1-K_2;m<=M2-M1;m++){
    buf_wy[m]=(tab_ptr[0])[m];} /* ostatnie K/2 punkty każdej linii bez zmian*/
        /* zapis przetworzonej linii */
write_line(buf_wy,n,M1,szer);
 {      /* zamiana wskaźników */
    char *temp_ptr;
    temp_ptr=tab_ptr[-L_2];
    for(l=-L_2;l<L_2;l++)
        tab_ptr[l]=tab_ptr[l+1];
    tab_ptr[L_2]=temp_ptr;
 }
}   /* koniec for(n=... */
```

```
/*zwolnienie pamięci zajmowanej przez bufory linii i tablice wskaźników*/
    free(buf_wy);
    for(l=-L_2;l<=L_2;l++){
        free(tab_ptr[l]);}
    tab_ptr-=L_2;
    free(tab_ptr);

    return 0;
}
```

zbiór nagłówkowy vist_lib.h zawiera deklaracje funkcji odczytu i zapisu wskazanej linii obrazu:

```
void read_line(char far *buf,      /* bufor przechowujący linie */
               int nr_lin,         /* numer linii */
               int pocz_lin,       /* współrzędna pierwszego punktu linii */
               int dlug_lin);      /* długość linii */
void write_line(char far *buf,
               int nr_lin,
               int pocz_lin,
               int dlug_lin);
```

Funkcje te muszą być dostosowane do wykorzystywanej karty przetwornika obrazu.

## 3. CZASY WYKONANIA PROGRAMU

Czas wykonania filtracji pojedynczego obrazu zależy od jego rozmiarów, wielkości okna filtru (tj. wielkości otoczenia punktu uwzględnianego do obliczeń odpowiedzi filtru) oraz szybkości zastosowanego komputera i karty przetwornika obrazu. Przykładowo czasy filtracji obrazu o rozmiarze 512*512 piksli na komputerze IBM PC/AT – 386 wyposażonym w pamięć typu Cache, współpracującym z kartą VIST firmy VIKOM są równe:

okno 3*3 – 12.14 s
5*5 – 28.18 s
7*7 – 51.63 s
9*9 – 82.05 s

## PODSUMOWANIE

Zaprezentowana w pracy procedura filtracji może być z łatwością dostosowana do dowolnej karty przetwornika obrazu – wymaga to tylko opracowania procedur zapisu i odczytu linii obrazu odpowiednich dla danego typu karty. Również zmiana algorytmu fitracji wymaga stosunkowo niewielkich zmian w procedurze – zmieniony

musi być fragment zawarty wewnątrz instrukcji for($m=K\_2$;... i ewentualnie dodane deklaracje dodatkowych zmiennych niezbędnych w tym algorytmie. Zastosowana struktura danych redukuje liczbę przesłań pomiędzy pamięcią karty przetwornika obrazu i buforami w pamięci operacyjnej komputera realizującego procedurę filtracji – każda linia jest tylko raz odczytywana i raz zapisywana. Ponadto nie są wymieniane zawartości buforów między sobą, zmianie ulegają jedynie zmienne wskazujące poszczególne bufory. Z przeprowadzonych doświadczeń wynika, że taka organizacja procedur fitracji znacznie redukuje czas ich wykonania.

## BIBLIOGRAFIA

1. N. K. S a c h: *Poprawa jakości obrazów telewizyjnych metodami cyfrowymi.* Rozprawa habilitacyjna, Warszawa: WPW 1990
2. W. K. P r a t t: *Digital Image Processing*, wydanie rosyjskie Moskwa: MIR 1982
3. R. F. G o n z a l e s: *Digital Image Processing.* Addison-Wesley Publishing Company 1987
4. B. K e r n i g h a m, D. R i t c h i e: *Język C*, Warszawa: WNT 1988
5. J. B i e l e c k i: *Język C – interpretacja standardu.* Warszawa: WNT 1987

J. MODELSKI A. BUCHOWICZ

## PROCEDURE OF LINEAR IMAGE FILTRATION

### S u m m a r y

Procedure of linear image filtration written in C programming language was described in this article. The applied structure of data decreases the number of transfers between the screen memory and buffers in the computer operational memory and eliminates transfers between these buffers. This method considerably reduces filtration execution time.

KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJI, 1992, 38, z. 1, ss. 67-76

621.372.55
621.391.883

# Algorytm obliczania wielokanałowego korektora minimalizującego błąd średniokwadratowy

JERZY KISILEWICZ

*Otrzymano 1991.04.12*

*Autoryzowano do druku 1991.08.25*

Opisano interferencję międzysymbolową występującą w kanale transmisji danych i pomiędzy kanałami. Rozpatrzono przypadek, gdy interferencja opisuje się macierzową transmitancją Z, której elementami są wielomiany. Dla takiej interferencji wielokanałowej zaproponowano wielokanałowy prosty korektor transwersalny. Wyprowadzono nowy, wygodny do obliczeń numerycznych układ równań, którego rozwiązaniem jest korektor minimalizujący błąd średniokwadratowy. Tworzenie tych równań nie wymaga zakładania wielkości opóźnienia wnoszonego przez skorygowane kanały. Zamieszczono przykładowe obliczenia korektora dwukanałowego.

## 1. WSTĘP

Zadanie obliczania nastaw korektora interferencji międzysymbolowej występującej w pojedynczym kanale transmisji danych jest znane literaturowo [1,2,3,8]. W wielu sytuacjach [4,5,6,7,9] należy się liczyć z występowaniem interferencji międzykanałowej oraz z koniecznością jej korekcji. W niniejszej pracy równoczesne występowanie interferencji międzysymbolowej i międzykanałowej nazywa się interferencją wielokanałową oraz stawia i rozwiązuje się zadanie korekcji takiej interferencji wielokanałowej.

Postawione zadanie zostanie rozwiązane przy założeniu, że macierzowa transmitancja opisująca korektor zawiera tylko funkcje wielomianowe ograniczonego stopnia. Przedmiotem optymalizacji są tu współczynniki tych wielomianów. Praktyczną realizacją takich funkcji są filtry transwersalne proste. Optymalizowane współczynniki są wartościami wzmocnień na odczepach linii opóźniających użytych do realizacji tych filtrów.

Takie samo zadanie stawia i rozwiązuje się w pracy [7]. Przedstawione tam rozwiązanie zakłada aprioryczną znajomość opóźnienia wnoszonego przez skorygowane kanały. Opóźnienie to najczęściej nie jest znane. Zadanie korekcji rozwiązuje się więc dla wszystkich możliwych opóźnień i wybiera rozwiązanie dające najmniejszy błąd średniokwadratowy. W tej pracy zostanie przedstawiony taki algorytm, w którym

w efekcie rozwiązania odpowiedniego układu równań z wieloma prawymi stronami otrzymuje się komplet rozwiązań dla wszystkich możliwych opóźnień. Proste wymnożenie odpowiednich macierzy dostarcza macierz wartości próbek interferencji skorygowanych kanałów. Z tej macierzy w prosty sposób otrzymuje się oceny wszystkich rozwiązań i wybiera rozwiązanie najlepsze. Przedmiotem pracy jest konstrukcja odpowiednich macierzy do utworzenia układu równań i do wyznaczenia skorygowanej interferencji wielokanałowej.

## 2. INTERFERENCJA WIELOKANAŁOWA I JEJ KOREKCJA

Niech $n$ oznacza liczbę kanałów, pomiędzy którymi występuje wzajemna interferencja. Niech podobnie jak w [7] $y_k^{ij}$ $(i,j=1,2,\ldots,n;\ k=0,1,\ldots,g)$ będą wartościami próbek odpowiedzi w kanale $i$ na pojedynczy impuls symbolu transmitowanego w kanale $j$. Funkcja

$$\bar{y}_{ij}(z) = y_0^{ij} + y_1^{ij} z^{-1} + y_2^{ij} z^{-2} + \ldots + y_g^{ij} z^{-g} \tag{1}$$

opisuje wpływ symboli transmitowanych w kanale $j$ na symbole transmitowane w kanale $i$, przy czym $g$ jest maksymalną liczbą rozważanych próbek interferencyjnych. Macierz

$$\bar{\mathbf{Y}}(z) = [\bar{y}_{ij}(z)], \quad (i,j = 1, 2, \ldots, n), \tag{2}$$

opisuje interferencję $n$ kanałową.

Zadaniem korektora jest eliminacja interferencji czyli sprowadzenie $\bar{\mathbf{Y}}(z)$ do macierzy jednostkowej $\mathbf{I}$. Przyjmując, że macierze

$$\bar{\mathbf{C}}(z) = [\bar{c}_{ij}(z)], \quad \bar{\mathbf{E}}(z) = [\bar{e}_{ij}(z)], \quad (i,j = 1, 2, \ldots, n), \tag{3}$$

opisują kolejno korektor i skorygowaną interferencję żąda się, aby $\bar{\mathbf{E}}(z)$ było zbliżone do macierzy jednostkowej, czyli aby [1,3]

$$\bar{\mathbf{E}}(z) = \bar{\mathbf{C}}(z)\,\bar{\mathbf{Y}}(z) \simeq z^{-h}\mathbf{I}_n$$

gdzie $z^{-h}$ oznacza opóźnienie wprowadzane przez kanał $z$ korektorem. Dąży się zatem aby

$$\bar{\mathbf{C}}(z) \simeq z^{-h}\bar{\mathbf{Y}}^{-1}(z).$$

Postać funkcji $\bar{c}_{ij}(z)$ zależy od przyjętej realizacji korektora. W dalszej części zakłada się prostą korekcję transwersalną. W tym przypadku $\bar{c}_{ij}(z)$ są wielomianami. Zakłada się, że maksymalny stopień tych wielomianów wynosi $m$, a więc

$$\bar{c}_{ij}(z) = c_{h0}^{ij} + c_{h1}^{ij} z^{-1} + c_{h2}^{ij} z^{-2} + \ldots + c_{h,m-1}^{ij} z^{-(m-1)},$$
$$\bar{e}_{ij}(z) = e_{h0}^{ij} + e_{h1}^{ij} z^{-1} + e_{h2}^{ij} z^{-2} + \ldots + e_{h,g+m-1}^{ij} z^{-(g+m-1)},$$

gdzie indeks $h$ przy współczynnikach powyższych wielomianów oznacza, wielomiany te zostały wyznaczone dla opóźnienia równego $h$ taktów.

## 3. BŁĄD ŚREDNIOKWADRATOWY KOREKCJI

W celu wyznaczenia współczynników wielomianów $\bar{c}_{ir}(z)\bar{y}_{rj}(z)$ tworzy się $m$ wierszowe macierze konwolucji $Y_{rj}$, których p-ty wiersz ($p=1,2,\ldots,m$) zawiera $m+g$ elementów kolejno: $p-1$ zer, $g+1$ próbek interferencyjnych $y_0^{rj},\ldots,y_g^{rj}$, $m-p$ zer. Wiersze macierzy $Y_{rj}$ mają postać

$$[0, \ldots, 0, y_0^{rj}, y_1^{rj}, \ldots, y_g^{rj}, 0, \ldots, 0].$$

Oznaczając jednowierszowe wektory

$$\begin{aligned} E_{ij}^h &= [e_{h0}^{ij}, e_{h1}^{ij}, \ldots, e_{h,m+g-1}^{ij}], \\ \mathscr{E}_{ij}^h &= [\varepsilon_{h0}^{ij}, \varepsilon_{h1}^{ij}, \ldots, \varepsilon_{h,m+g-1}^{ij}], \\ c_{ij}^h &= [c_{h0}^{ij}, c_{h1}^{ij}, \ldots, c_{h,m-1}^{ij}], \end{aligned} \tag{4}$$

gdzie

$$\varepsilon_{hk}^{ij} = \begin{cases} 1 \text{ dla } j=i \text{ oraz } k=h \\ 0 \text{ dla } j\neq i \text{ lub } k\neq h \end{cases},$$

można zapisać

$$E_{ij}^h = \sum_{r=1}^{n} C_{ir}^h Y_{rj}. \tag{5}$$

W [7] podano wzór na błąd średniokwadratowy w $i$-tym kanale, który po uwzględnieniu powyższych oznaczeń można zapisać w postaci

$$\delta_i^2(h) = \sum_{j=1}^{n}\left[ s^2\left(\sum_{r=1}^{n} C_{ir}^h Y_{rj} - \mathscr{E}_{ij}^h\right)\left(\sum_{r=1}^{n} C_{ir}^h Y_{rj} - \mathscr{E}_{ij}^h\right)^T + \sigma_j^2 C_{ij}^h (C_{ij}^h)^T\right], \tag{6}$$

gdzie indeks $T$ oznacza transpozycję,

$S^2$ jest mocą sygnału,

$\sigma_j^2$ jest mocą zakłóceń w $j$-tym kanale.

W pracy [7] pokazano, że parametry korektora minimalizującego kryterium

$$Q_h = \sum_{i=1}^{n} \delta_i^2(h) \tag{7}$$

muszą spełniać równania

$$[C_{i1}^h, C_{i2}^h, \ldots, C_{in}^h]\,[r\text{-te } m \text{ kolumn } YY^T] + \mathscr{S}^{-2} C_{ir}^h \sigma_r^2 = \mathscr{E}_{ii}^h Y_{ri}^T, \tag{8}$$

gdzie $i, r = 1, 2, \ldots, n$,

$$\mathbf{Y} = [\mathbf{Y}_{ij}] = \begin{bmatrix} Y_{11} & \ldots & Y_{1n} \\ \cdot & \ldots & \cdot \\ Y_{n1} & \ldots & Y_{nn} \end{bmatrix}. \tag{8a}$$

W pracy [7] dla ustalonego $h$ tworzy się układ równań zawierający macierz $\mathbf{E}^h$, którego rozwiązaniem jest macierz $\mathbf{C}^h$ gdzie

$$\mathbf{C}^h = [\mathbf{C}_{ij}^h], \quad \mathbf{E}^h = [\mathbf{E}_{ij}^h], \quad (i, j = 1, 2, \ldots, n).$$

W praktyce obliczeń numerycznych wygodniej jest zastosować metodę, w której nie występuje macierz $\mathscr{E}^h$ wymagająca założenia wartości opóźnienia $h$, metodę dającą wszystkie możliwe rozwiązania $\mathbf{C}^h$ za jednym cyklem obliczeń. Metoda ta zostanie opisana poniżej.

Przepisując wzór (8) dla $h=0,1,\ldots,m+g-1$ oraz oznaczając

$$\mathbf{C}_{ij} = \begin{bmatrix} C_{ij}^0 \\ C_{ij}^1 \\ . \\ . \\ . \\ C_{ij}^{m+g-1} \end{bmatrix}, \quad \tilde{\mathbf{C}} = \begin{bmatrix} C_{11} & \ldots & C_{1n} \\ . & & . \\ . & & . \\ . & & . \\ C_{n1} & \ldots & C_{nn} \end{bmatrix},$$

gdzie $\mathbf{C}_{ij}$ są macierzmi o $m+g$ wierszach oraz $m$ kolumnach, a $\tilde{\mathbf{C}}$ jest macierzą o $n(m+g)$ wierszach oraz $nm$ kolumnach, otrzymuje się

$$[C_{i1}, C_{i2}, \ldots, C_{in}]\ [r\text{-te } m \text{ kolumn } YY^T] + S^{-2} C_{ir} \sigma_r^2 = Y_{ri}^T,$$

dla $i, r = 1, 2, \ldots, n$.

Łącząc powyższy zapis dla wszystkich $i$ oraz $r$ otrzymuje się następującą postać równań liniowych

$$\tilde{C}(YY^T + S^{-2}\Sigma^2) = Y^T.$$

gdzie:

$$\Sigma^2 = \begin{bmatrix} \Sigma_1^2 & 0 & \ldots & 0 \\ 0 & \Sigma_2^2 & \ldots & 0 \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ 0 & 0 & \ldots & \Sigma_n^2 \end{bmatrix} \text{ ma } nm \text{ wierszy oraz } nm \text{ kolumn},$$

$$\Sigma_i^2 = \mathbf{I}_m \sigma_i^2 \quad (i = 1, \ldots, n).$$

Po rozwiązaniu powyższego układu wiersze $C_{ij}^h$ podmacierzy $\mathbf{C}_{ij}$ $(i,j=1,\ldots,n)$ macierzy $\tilde{\mathbf{C}}$ zawierają po $m$ wartości wzmocnień na odczepach linii opóźniających liniowych korektorów transwersalnych z wejściem w kanale $j$ oraz wyjściem w kanale $i$ dla kolejnych wartości $h$ $(h=0,1,\ldots,m+g-1)$. Za rozwiązanie należy wybrać taką wartość $\mathscr{H}$, dla której otrzyma się najmniejszą wartość kryterium błędu średniokwadratowego. Aby określić tę wartość, należy obliczyć macierze $\mathbf{E}_{ij}^h$ $(i,j=1,2,\ldots,n)$ skorygowanych interferencji dla każdego $h$. Oznaczając

$$\mathbf{E}_{ij} = \begin{bmatrix} E_{ij}^0 \\ E_{ij}^1 \\ \cdot \\ \cdot \\ \cdot \\ E_{ij}^{m+g-1} \end{bmatrix}, \quad \tilde{\mathbf{E}} = \begin{bmatrix} E_{11} & \cdots & E_{1n} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ E_{n1} & \cdots & E_{nn} \end{bmatrix},$$

gdzie $\mathbf{E}_{ij}$ są macierzami o $m+g$ wierszach oraz $m+g$ kolumnach, a $\tilde{\mathbf{E}}$ jest macierzą o $n(m+g)$ wierszach oraz $n(m+g)$ kolumnach, otrzymuje się

$$\tilde{\mathbf{E}} = \tilde{\mathbf{C}}\mathbf{Y}. \tag{9}$$

Wiersze $E_{ij}^h$ podmacierzy $\mathbf{E}_{ij}$ $(i,j=1,\ldots,n)$ macierzy $\tilde{\mathbf{E}}$ zawierają po $m+g$ wartości próbek skorygowanej interferencji $n$ kanałowej pomiędzy kanałem $j$ a kanałem $i$ dla kolejnych wartości $h$ $(h=0,1,\ldots,m+g-1)$.

## 4. ROZWIĄZANIE ZADANIA KOREKCJI WIELOKANAŁOWEJ

Problem wielokanałowej korekcji transwersalnej minimalizującej błąd średniokwadratowy można teraz rozwiązać wg niżej opisanego algorytmu.

Zadanie korekcji interferencji wielokanałowej

Dla zadanej interferencji n kanałowej danej macierzą $\bar{\mathbf{Y}}(z)$ (2), której elementy $\bar{y}_{ij}(z)$ są wielomianami o współczynnikach $y_0^{ij}, y_1^{ij}, \ldots, y_g^{ij}$, $(i,j=1,2,\ldots,n)$, ewentualnie danych mocach szumu kanałach $\sigma_1, \sigma_2, \ldots, \sigma_n$ i mocy sygnału $S^2$ wyznaczyć macierz $\bar{\mathbf{C}}(z)$ (3) korektora, której elementami są wielomiany o współczynnikach $c_0^{ij}, c_1^{ij}, \ldots, c_{m-1}^{ij}$, $(i,j=1,2,\ldots,n)$ taką, aby zminimalizwoać kryterium błędu średniokwadratowego wyrażone wzorami (6) i (7).

Algorytm:

1. Wybrać liczbę $m$ odczepów linii opóźniających korektora (zaleca się $m \geqslant 3g$ – tym większe im wymagana jest większa dokładność obliczeń).

2. Utworzyć macierz daną wzorem (8a) złożoną z macierzy konwolucji $Y_{ij}$ $(i,j=1,2,\ldots,n)$. Macierze $Y_{ij}$, zawierają po $m$ wierszy zbudowańych tak, że $p=$ty wiersz $(p=1,2,\ldots,m)$ zawiera $m+g$ elementów kolejno: $p$-1 zer, $g+1$ próbek interferencyjnych $y_0^{rj}, \ldots, y_g^{rj}$, $m-p$ zer. Wiersze macierzy $Y_{rj}$ mają więc postać

$$[0, \ldots, 0, y_0^{rj}, y_1^{rj}, \ldots, y_g^{rj}, 0, \ldots, 0].$$

3. Obliczyć macierz $\tilde{\mathbf{C}}$ o $n(m+g)$ wierszach oraz $nm$ kolumnach rozwiązując układ nm równań liniowych z $n(m+g)$ prawymi stronami

$$(YY^T + S^{-2}\Sigma^2)^T(\tilde{\mathbf{C}})^T = Y, \tag{10a}$$

lub układ

$$(YY^T)^T(\tilde{\mathbf{C}})^T = Y, \tag{10b}$$

gdy $S^2 \gg \sigma_i^2$ dla każdego $i=1,2,\ldots,n$.

Każdy wiersz $C_{ij}^h$ $(h=0,1,\ldots,m+g-1)$ podmacierzy $\mathbf{C}_{ij}$ macierzy $\tilde{\mathbf{C}}$ zawiera rozwiązanie zadania korekcji średniokwadratowej dla określonej wartości $h$.

4. Obliczyć macierz $\tilde{\mathbf{E}}=\tilde{\mathbf{C}}\mathbf{Y}$ o $n(m+g)$ wierszach oraz $n(m+g)$ kolumnach. Każdy wiersz $E_{ij}^h$ podmacierzy $\mathbf{E}_{ij}$ macierzy $\tilde{\mathbf{E}}$ zawiera prążki interferencji skorygowanego kanału dla danego $h$.

5. Obliczyć macierz $\hat{\mathbf{E}}$ wg wzoru

$$\hat{\mathbf{E}} = \tilde{\mathbf{E}} - \mathbf{I}_{n(m+g)},$$

gdzie $\mathbf{I}_{n(m+g)}$ jest macierzą jednostkową rzędu $n(m+g)$.

6. Obliczyć średniokwadratowe błędy $\delta_i^2(h)$ $(h=0,1,\ldots,m+g-1)$ dla każdego kanału $(i=1,2,\ldots,n)$ w.g wzoru

$$\delta_i^2(h) = S^2 \sum_{k=1}^{n(m+g)} (\hat{e}_{rk})^2 + \sum_{j=1}^{n} \sigma_j^2 \sum_{k=(j-1)m+1}^{jm} (\tilde{c}_{rk})^2 \tag{11a}$$

lub gdy w każdym kanale moc szumu jest dużo mniejsza od mocy sygnału $(\sigma_j^2 \ll S^2)$ wg wzoru

$$\delta_i^2(h) = S^2 \sum_{k=1}^{n(m+g)} (\hat{e}_{rk})^2, \tag{11b}$$

gdzie $\hat{e}_{rk}$ i $\tilde{c}_{rk}$ są elementami odpowiednio macierzy $\hat{\mathbf{E}}$ i $\tilde{\mathbf{C}}$ oraz

$$r = (i-1)(m+g)+1.$$

7. Dla każdej wartości $h$ obliczyć wartość kryterium błędu średniokwadratowego $Q_h$ ze wzorów (7) i (11) a następnie wybrać takie $h$, dla którego kryterium to jest minimalne, czyli

$$Q_{\mathscr{H}} = \min_h \{Q_h\}.$$

8. Wiersze

$$C_{ij}^{\mathscr{H}} = [c_{\mathscr{H}0}^{ij}, c_{\mathscr{H}1}^{ij}, \ldots, c_{\mathscr{H},m-1}^{ij}] \quad (i,j=1,2,\ldots,n)$$

zawierają szukane rozwiązanie zadania n kanałowej korekcji średniokwadratowej.

9. Jeśli otrzymana wartość kryterium jest za duża to zwiększyć $m$ o $g$ oraz przejść do punktu 2 chyba, że poprzednie zwiększenie wartości $m$ nie dało zadowalającego efektu.

## 5. PRZYKŁAD

W pracy [7] zamieszczono przykładowe obliczenia korektora 2 kanałowego dla $h=0$. W tym przykładzie zostanie rozpatrzony ten sam przypadek interferencji

2 kanałowej. Nie będzie się jednak zakładać znajomości wartości $h$. Ponieważ $n=2$, $g=1$ oraz $m=6$ macierz $\tilde{\mathbf{C}}$ ma 14 wierszy i 12 kolumn, natomiast macierze $\tilde{\mathbf{E}}$ i $\hat{\mathbf{E}}$ mają po 14 wierszy i 14 kolumn. Po jednej iteracji algorytmu otrzymano parametry korektora:

$$\mathbf{C}_{11} = \begin{bmatrix} 1.332 & 0.665 & 0.413 & 0.201 & 0.118 & 0.042 \\ -0.001 & 1.329 & 0.660 & 0.402 & 0.188 & 0.081 \\ -0.003 & -0.006 & 1.321 & 0.644 & 0.376 & 0.135 \\ -0.004 & -0.014 & -0.020 & 1.285 & 0.602 & 0.261 \\ -0.009 & -0.021 & -0.040 & -0.072 & 1.203 & 0.433 \\ -0.014 & -0.044 & -0.063 & -0.155 & -0.208 & 0.834 \\ -0.029 & -0.066 & -0.127 & -0.232 & -0.416 & -0.748 \end{bmatrix},$$

$$\mathbf{C}_{12} = \begin{bmatrix} -0.666 & -0.166 & -0.206 & -0.049 & -0.059 & -0.005 \\ 0.000 & -0.664 & -0.165 & -0.198 & -0.047 & -0.031 \\ 0.002 & 0.003 & -0.660 & -0.155 & -0.188 & -0.015 \\ 0.002 & 0.010 & 0.005 & -0.633 & -0.150 & -0.100 \\ 0.005 & 0.010 & 0.020 & 0.037 & -0.602 & -0.048 \\ 0.005 & 0.030 & 0.016 & 0.107 & 0.052 & -0.321 \\ 0.018 & 0.033 & 0.064 & 0.117 & 0.208 & 0.378 \end{bmatrix},$$

$$\mathbf{C}_{21} = \begin{bmatrix} -0.666 & 0.166 & -0.206 & 0.049 & -0.059 & 0.005 \\ -0.000 & -0.664 & 0.165 & -0.198 & 0.047 & -0.031 \\ 0.002 & -0.003 & -0.660 & 0.155 & -0.188 & 0.015 \\ -0.002 & 0.010 & -0.005 & -0.633 & 0.150 & -0.100 \\ 0.005 & -0.010 & 0.020 & -0.037 & -0.602 & 0.048 \\ -0.005 & 0.030 & -0.016 & 0.107 & -0.052 & -0.321 \\ 0.018 & -0.033 & 0.064 & -0.117 & 0.208 & -0.378 \end{bmatrix},$$

$$\mathbf{C}_{22} = \begin{bmatrix} 1.332 & -0.665 & 0.413 & -0.201 & 0.118 & -0.042 \\ 0.001 & 1.329 & -0.660 & 0.402 & -0.188 & 0.081 \\ -0.003 & 0.006 & 1.321 & -0.644 & 0.376 & -0.135 \\ 0.004 & -0.014 & 0.020 & 1.285 & -0.602 & 0.261 \\ -0.009 & 0.021 & -0.040 & 0.072 & 1.203 & -0.433 \\ 0.014 & -0.044 & 0.063 & -0.155 & 0.208 & 0.834 \\ -0.029 & 0.066 & -0.127 & 0.232 & -0.416 & 0.748 \end{bmatrix}.$$

oraz interferencję wynikową:

$$\mathbf{E}_{11} = \begin{bmatrix} 0.9994 & -0.0011 & -0.0020 & -0.0036 & -0.0064 & -0.0115 & -0.0205 \\ -0.0011 & 0.9978 & -0.0036 & -0.0070 & -0.0115 & -0.0224 & -0.0368 \\ -0.0020 & -0.0036 & 0.9936 & -0.0115 & -0.0205 & -0.0368 & -0.0658 \\ -0.0036 & -0.0070 & -0.0115 & 0.9776 & -0.0368 & -0.0716 & -0.1178 \\ -0.0064 & -0.0115 & -0.0205 & -0.0368 & 0.9342 & -0.1178 & -0.2104 \\ -0.0115 & -0.0224 & -0.0368 & -0.0716 & -0.1178 & 0.7708 & -0.3769 \\ -0.0205 & -0.0368 & -0.0658 & -0.1178 & -0.2104 & -0.3769 & 0.3266 \end{bmatrix},$$

$$\mathbf{E}_{12} = \begin{bmatrix} 0.0000 & 0.0002 & 0.0003 & 0.0005 & 0.0009 & 0.0017 & 0.0029 \\ -0.0002 & 0.0009 & -0.0005 & 0.0029 & -0.0017 & 0.0093 & -0.0055 \\ 0.0003 & 0.0005 & 0.0009 & 0.0017 & 0.0029 & 0.0055 & 0.0093 \\ -0.0005 & 0.0029 & -0.0017 & 0.0093 & -0.0055 & 0.0299 & -0.0176 \\ 0.0009 & 0.0017 & 0.0029 & 0.0055 & 0.0093 & 0.0176 & 0.0299 \\ -0.0017 & 0.0093 & -0.0055 & 0.0299 & -0.0176 & 0.0957 & -0.0564 \\ 0.0029 & 0.0055 & 0.0093 & 0.0176 & 0.0299 & 0.0564 & 0.0956 \end{bmatrix},$$

$$\mathbf{E}_{21} = \begin{bmatrix} 0.0000 & -0.0002 & 0.0003 & -0.0005 & 0.0009 & -0.0017 & 0.0029 \\ 0.0002 & 0.0009 & 0.0005 & 0.0029 & 0.0017 & 0.0093 & 0.0055 \\ 0.0003 & -0.0005 & 0.0009 & -0.0017 & 0.0029 & -0.0055 & 0.0093 \\ 0.0005 & 0.0029 & 0.0017 & 0.0093 & 0.0055 & 0.0299 & 0.0176 \\ 0.0009 & -0.0017 & 0.0029 & -0.0055 & 0.0093 & -0.0176 & 0.0299 \\ 0.0017 & 0.0093 & 0.0055 & 0.0299 & 0.0176 & 0.0957 & 0.0564 \\ 0.0029 & -0.0055 & 0.0093 & -0.0176 & 0.0299 & -0.0564 & 0.0956 \end{bmatrix},$$

$$\mathbf{E}_{22} = \begin{bmatrix} 0.9994 & 0.0011 & -0.0020 & 0.0036 & -0.0064 & 0.0115 & -0.0205 \\ 0.0011 & 0.9978 & 0.0036 & -0.0070 & 0.0115 & -0.0224 & 0.0368 \\ -0.0020 & 0.0036 & 0.9936 & 0.0115 & -0.0205 & 0.0368 & -0.0658 \\ 0.0036 & -0.0070 & 0.0115 & 0.9776 & 0.0368 & -0.0716 & 0.1178 \\ -0.0064 & 0.0115 & -0.0205 & 0.0368 & 0.9342 & 0.1178 & -0.2104 \\ 0.0115 & -0.0224 & 0.0368 & -0.0716 & 0.1178 & 0.7708 & 0.3769 \\ -0.0205 & 0.0368 & -0.0658 & 0.1178 & -0.2104 & 0.3769 & 0.3266 \end{bmatrix}.$$

Powyższe macierze posiadają po 7 wierszy związanych kolejno z wartościami $h=0,1,..,6$. Obliczane dla $S=1$ w punkcie 7 algorytmu wartości kryterium błędu średnia kwadratowego $Q_h$ zamieszczono w tabeli:

| $Q_0$ | $Q_1$ | $Q_2$ | $Q_3$ | $Q_4$ | $Q_5$ | $Q_6$ |
|---|---|---|---|---|---|---|
| 0.0013 | 0.0044 | 0.0128 | 0.0448 | 0.1315 | 0.4585 | 1.3468 |

Jak widać najmniejszą wartość kryterium otrzymano tu dla $h=0$, a zatem $\mathscr{H}=0$. Za rozwiązanie należy więc przyjąć pierwsze wiersze powyższych macierzy. Wyniki te są takie same jakie otrzymano w pracy [7].

## 6. PODSUMOWANIE

Opracowany tu algorytm jest uogólnieniem do przypadku wielokanałowego opisanego w [2] algorytmu rozwiązywania zadania korekcji średniokwadratowej. W porównaniu do algorytmu opisanego w [7] przedstawiony tu algorytm nie zakłada znajomości opóźnienia $h$ wnoszonego przez kanał z korektorem i jest prostszy obliczeniowo i wygodniejszy w zastosowaniu praktycznym. Równoczesne obliczanie nastaw korektora dla wszystkich możliwych opóźnień $h$ nie wymaga konstruowania macierzy $\mathscr{E}^h$, jak to ma miejsce w [7]. Należy też zauważyć, że czas obliczeń rozwiązania $nm$ równań liniowych (10) z $n(m+g)$ prawymi stronami jest mniejszy od czasu $m+g$ krotnego rozwiązania tych równań z $n$ prawymi stronami oddzielnie dla poszczególnych wartości $h$.

Macierz $\tilde{\mathbf{C}}$ nastaw korektora można wprawdzie wyrazić wzorami

$$\tilde{\mathbf{C}} = Y^T(YY^T + S^{-2}\Sigma^2)^{-1},$$

lub gdy $S^2 \ll \sigma_i^2$ $(i=1,\ldots,n)$

$$\tilde{\mathbf{C}} = Y^T(YY^T)^{-1},$$

ale praktycznie wygodniej jest obliczać $\tilde{\mathbf{C}}$ rozwiązując układ równań liniowych (10).

Opisany tu algorytm oprogramowano tworząc podprogram kor_n_k. Podprogram ten został napisany jako funkcja w języku C.

Podprogram kor_n_k

Podprogram kor_n_k realizuje punkty od 2 do 8 algorytmu i służy do obliczania parametrów $n$ kanałowego korektora transwersalnego minimalizującego błąd średniokwadratowy dla zadanej interferencji wielokanałowej, mocy białego szumu gaussowskiego i liczby parametrów korektora. Kor_n_k nie realizuje punktu 1 *i* punktu 9 algorytmu, pozostawiając użytkownikowi możliwość dowolnego wyboru liczby parametrów korektora.

Nagłówek podprogramu:

```
kor_n_k(n, g, Yij, sig, m, Cij, Q)
int n, g, m;
double (*Yij)[ ], (*sig)[ ], (*Cij)[ ], *Q;
```

Dane:

n — liczba kanałów,

g — stopień wielomianów (1) opisujących interferencję,

Yij — tablica $n^2(g+1)$ elementowa zawierająca współczynniki $y_k^{ij}$ wielomianów (1) w kolejności najszybciej zmieniającego się k, potem j, a następnie i,

sig — n elemntowa tablica zawierająca moce szumu w kolejnych kanałach odniesione do mocy sygnału (gdy szum nie jest uwzględniany należy podstawić sig=NULL),

m — założona liczba odczepów linii opóźniających korektora.

Wyniki:

C – wektor $n^2m$ parametrów $c^{ij}_{\mathscr{H}0}$, $c^{ij}_{\mathscr{H}1}$, ..., $c^{ij}_{\mathscr{H},m-1}$ korektora,

Q – wartość błędu średniokwadratowego $Q_h$,

kor_n_k – funkcja zwraca optymalne opóźnienie $\mathscr{H}$ lub – 1 gdy, wystąpił błąd obliczeń.

## BIBLIOGRAFIA

1. A. P. Clark: *Advanced Data-Transmission Systems.* London Press Limited, 1977
2. A. P. Clark: *Principles of Digital Data Transmission.* Pentech Press, London 1976
3. A. Dąbrowski: *Odbiór sygnałów transmisji danych w obecności zniekształceń interferencyjnych i zakłóceń.* Rozdz. 4 pracy zbiorowej pod kier. Z. Barana: Podstawy Transmisji Danych. WKiŁ, Warszawa 1982
4. B. Hirosaki: *An Analysis of Automatic Equalizers for Orthogonally Multiplexed QAM Systems.* IEEE Trans. on Com., January 1980, vol. COM–28, nr. 1, s. 73–83
5. I. Korn: *Simple Expression for Interchannel and Intersymbol Interference Degradation in MSK Systems with Application to Systems with Gaussian Filters.* IEEE Trans. on Com., August 1982, vol. COM–30, nr. 8, s. 1968–1972
6. J. Kisielewicz: *Eliminacja interferencji międzykanałowej.* Kwartalnik Elektroniki i Telekomunikacji, 1991, z. 1-2
7. J. Kisilewicz: *Korektor interferencji wielowymiarowej minimalizujący błąd średniokwadratowy.* Kwartalnik Elektroniki i Telekomunikacji, 1990, z. 3–4
8. R. W. Lucky, J. Salz, E. J. Weldon: *Principles of Data Communication.* Mc. Graw-Hill, New York 1968
9. D. P. Taylor, M. Shafi: *Decision Feedback Equalization for Multipath Induced Interference in Digital Microwave LOS Links.* IEEE Trans. on Com., March 1984, vol. COM–32, nr. 3, s. 267–279

J. KISILEWICZ

## AN ALGORITHM FOR DESIGN MULTICHANNEL EQUALIZER GIVING THE MINIMUM MEAN-SQUARE ERROR

### Summary

Intersymbol and interchannel interference is presented in this paper. The case when the interference is described by Z-transform matrix which contains only polynomials is considered. For this multichannel interference a linear feedforward transversal equalizer is suggested. A new system of equations is derived. The solution of this system allows to obtain minimum mean-square error of corrections in simple way. In order to illustrate the procedure proposed, an example is given. In the latter two channel equalizer is investigated.

KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJI, 1992, 38, z. 1, ss. 77-86

621.391.81

# Analiza efektywności pracy sieci lokalnej z dostępem rywalizacyjnym

JERZY HOSZA, TADEUSZ STACHOWIAK

*Instytut Podstaw Informatyki, Polska Akademia Nauk, Warszawa*

*Otrzymano 1991.05.17*

*Autoryzowano do druku 1991.06.16*

W pracy przedstawiono wyniki badań efektywności działania lokalnej sieci magistralowej z dostępem rywalizacyjnym. Efektywność sieci lokalnej określano za pomocą przepustowości i średniego czasu transmisji. W przyjętym modelu badanej sieci uwzględniono zasadnicze czynniki mające decydujący wpływ na pracę sieci. Należą do nich: czas dostępu do kanału, przetwarzanie w adapterach, sprzęg między adapterem i komputerem oraz podstawowe parametry protokołu transportowego jak rozmiar okna i czas retransmisji. Zbadano zmiany przepustowości sieci oraz średniego czasu transmisji (pakietów lub wiadomości) w zależności od liczby stacji, obciążenia sieci i długości pakietów. Przeprowadzono również eksperymenty pozwalające na określenie wpływu wielkości okna i czasu retransmisji na efektywność pracy sieci lokalnej.

## 1. WPROWADZENIE

W ostatnich latach obserwuje się szybki rozwój lokalnych sieci komputerowych służących różnorodnym celom przetwarzania danych. Znajdują one zastosowanie wszędzie tam gdzie istnieje potrzeba wymiany informacji między urządzeniami na niewielkim obszarze, obejmującym swym zasięgiem teren zakładu przemysłowego, ośrodka akademickiego czy urzędu. Za pomocą sieci lokalnych realizowane jest zarządzanie, zbieranie danych, przesyłanie wiadomości, rozdział funkcji obliczeniowych, sterowanie procesami produkcyjnymi. W sieci lokalnej mogą pracować komputery wszelkich typów poczynając od dużych zestawów obliczeniowych a kończąc na małych komputerach osobistych. Istotną cechą sieci lokalnych jest wspólne korzystanie wszystkich stacji z łączącego je kanału transmisyjnego. Prowadzi to do uzyskania dużej przepustowości podsystemu komunikacyjnego i efektywnego wykorzystania sieci. Umożliwia również równoważenie szybkości transferu danych i szybkości ich przetwarzania.

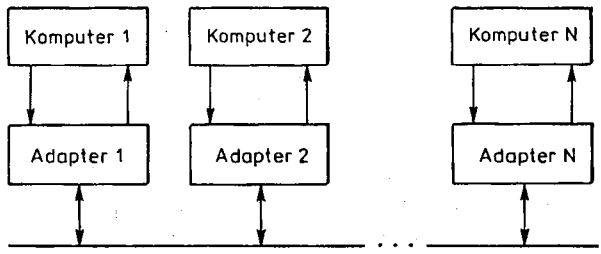
Sieć lokalną można traktować jako pewien złożony system masowej obsługi. Badanie właściwości tego systemu metodami analitycznymi napotyka często poważne

trudności. Uzyskanie związków między zmiennymi opisującymi sieć jest zazwyczaj wynikiem znacznych uproszczeń. Czasami nawet pojawiają się wątpliwości co do słuszności otrzymanych wyników. Alternatywą dla metod analitycznych mogą być badania pomiarowe. Jednak możliwości eksperymentowania w działających sieciach są ograniczone. Jeszcze mniejsze szanse realizacji ma budowa nowej sieci tylko dla celów badawczych. Koszty realizacji są często barierą nie do pokonania. Takich ograniczeń nie ma podczas symulacji komputerowej, do której konkretna sieć nie jest potrzebna. Badania symulacyjne umożliwiają odwzorowanie dynamiki sieci lokalnej, rozumianej jako kombinacja wielu działających łącznie różnorodnych procesów. Ta oraz inne zalety symulacji, zwłaszcza w porównaniu z odmiennymi technikami modelowania wynikają z możliwości budowy modeli symulacyjnych na dowolnym poziomie abstrakcji.

W pracy dokonano analizy efektywności działania sieci lokalnej z rywalizacyjnym dostępem do kanału. Badania przeprowadzono za pomocą symulacji cyfrowej. Ze względu na charakter działania sieci lokalnej zastosowano metodę kolejnych zdarzeń. Do określenia efektywności pracy sieci przyjęto dwa powszechnie stosowane wskaźniki: przepustowość i średni czas transmisji (pakietu lub wiadomości) [1].

## 2. MODEL BADANEJ SIECI

Przedmiotem badań symulacyjnych będzie sieć lokalna magistralowa typu CSMA/CD z dostępem rywalizacyjnym do kanału. Ogólny schemat takiej sieci przedstawiono na rys. 1.




Rys. 1. Podstawowy schemat modelowanej sieci

Jak widać komputery są dołączone do magistrali za pomocą adapterów. W przypadku sieci lokalnej, ze względu na duże szybkości, nie mogą to być klasyczne urządzenia stosowane do wymiany informacji między komputerem a otoczeniem, jak RS 232C lub Centronics. Są one po prostu zbyt wolne. Dlatego konstruuje się adaptery przeznaczone specjalnie do tego celu. Są to tak zwane adaptery dedykowane (służące do szybkiej wymiany informacji).

Wymianę informacji między adapterem dedykowanym a pamięcią operacyjną komputera zwykle realizuje się za pomocą jednego z trzech następujących sposobów [5]:

a) z użyciem portów,

b) z wykorzystaniem bezpośredniego dostępu do pamięci,

c) z zastosowaniem pamięci dwuwejściowej.

Spośród nich sposób trzeci jest najbardziej efektywny ze względu na szybkość przekazywania danych.

Zadaniem adaptera sieci lokalnej jest realizacja określonego zakresu usług oraz udostępnienie tych usług swojemu komputerowi. Zakres tych usług jest zwykle różnorodny. W wariancie minimalnym obejmuje usługi podwarstwy dostępu, a w wariancie maksymalnym – usługi warstwy sesji. Często spotyka się również rozwiązania pośrednie obejmujące usługi do warstwy transportowej (włącznie) [2].

Minimalny zakres usług realizuje się zwykle sprzętowo, z tym, że na ogół wprowadza się w adapterze pamięć buforową. Wynika to z bezwzględnej potrzeby nadawania i odbioru całości informacji stanowiącej pakiet w sposób ciągły z nominalną szybkością właściwą dla danej sieci. Tego typu adapter nazywamy nieinteligentnym. Nie może on realizować funkcji wyższych warstw komunikacyjnych. Są one wykonywane programowo w komputerze macierzystym i obciążają jego procesor. Osłabia to jednak moc obliczeniową komputera.

Rozszerzony zakres usług realizuje się za pomocą środków sprzętowo-programowych. Adaptery tego typu nazywamy inteligentnymi. Umożliwiają one odciążenie procesora komputera od realizacji części wyższych warstw. Adaptery takie są wyposażone w mikroprocesor zwany peryferyjnym, pamięć buforową i pamięć stałą z programem. Realizacja styku adaptera inteligentnego z magistralą komputera odbywa się według podobnych zasad jak dla adaptera nieinteligentnego. W obu przypadkach, w zależności od wykonania adaptera, mamy do czynienia z [5]:

– pamięcią dwuwejściową stanowiącą część przestrzeni adresowanej komputera,

– zbiorem portów obejmującym część przestrzeni adresowej portów wejścia-wyjścia,

– sygnałem przerwania, który musi być dołączony do odpowiedniego poziomu układu przerwań komputera.

Bez względu na rodzaj adaptera, skorzystanie z jego usług wymaga przygotowania dla komputera macierzystego odpowiedniego oprogramowania obsługi adaptera (stopień sterujący).

W modelu symulacyjnym przyjęto, że adapter realizuje szeroki zakres usług – do warstwy transportowej włącznie. Zgodnie z wcześniejszymi uwagami powinien to być adapter inteligentny. Taka konfiguracja jest często spotykana w praktyce [2]. Przeprowadzone eksperymenty symulacyjne dotyczyły jednosegmentowej sieci typu Ethernet. Dlatego w modelu adaptera założono brak warstwy sieciowej. Użytkownikami warstwy transportowej są procesy umieszczone w pamięci operacyjnej komputera macierzystego. Zlecenia generowane przez te procesy są jednostkami usług warswty transportowej (Transport Service Data Unit – TSDU). Przesłanie informacji w obrębie warstwy transportowej odbywa się przy użyciu jednostek danych protokołu transportowgo (Transport Protocol Data Unit – TPDU). Wiadomości generowane przez użytkowanika sieci mogą być zbyt długie, aby można było umieścić je w jednym pakiecie. Na przykład norma IEEE 802.3 LAN ogranicza się maksymalną długość pakietu do 1518 bajtów. W związku z tym jedną z usług protokołu transportowego

musi być dzielenie wiadomości na segmenty. Następnie każdy segment jest przekazywany kolejno do warstwy łącza danych. W architekturze logicznej sieci lokalnych w warstwie tej wyróżnia się dwie podwarstwy: dostępu do kanału i łącza logicznego. Ogólnie można powiedzieć, że warstwa łącza danych pełni funkcje związane z formowaniem ramki (serializacja i deserializacja, nadawanie preambuły, synchronizacja blokowa, adresowanie) oraz z kontrolą poprawnej transmisji (wykrywanie błędów). Ramki z błędami są odrzucane, a więc nie docierają do warstwy transportowej.

Warstwa tranportowa jest odpowiedzialna za bezbłędną transmisję. Osiąga się to przez wprowadzenie potwierdzeń. Mianowicie, po odebraniu nieuszkodzonego pakietu w stacji docelowej na poziomie warstwy transportowej wytwarzany jest sygnał potwierdzenia ACK. Nadajnik po odebraniu takiego sygnału może rozpocząć wysyłanie następnego pakietu. Jednak potwierdzenie każdego pakietu nie zawsze jest efektywne. Dotyczy to sieci lokalnych. W tego typu sieciach błędy pojawiają się rzadko. Można więc stosować potwierdzenie grupowe [51]. Takie też rozwiązanie przyjęto w modelu symulacyjnym. Nadajnik wysyła W pakietów i wstrzymuje pracę czekając na potwierdzenie. Taki mechanizm sterowania przepływem nosi nazwę okna. Liczbę W nazywamy wielkością okna. Kiedy stacja docelowa przyjmie W-ty pakiet, jej warstwa transportowa generuje potwierdzenie ACK i wysyła je do stacji źródłowej. Po odebraniu potwierdzenia stacja źródłowa może ponownie rozpocząć nadawanie kolejnej porcji pakietów. Oczywiście może się zdarzyć, że potwierdzenie dotrze zbyt późno do celu lub zostanie zgubione. Stacja źródłowa nie może więc czekać zbyt długo. Dlatego wprowadza się mechanizm time out'ów. Stacja rozpoczynająca nadawanie uruchamia w odpowiednim momencie zegar i oczekuje na potwierdzenie tylko przez określony czas zwany czasem retransmisji. Po jego upływie rozpoczyna się ponowne nadawanie tych samych danych. Jak widać mechanizm retransmisji zabezpiecza nie tylko przed błędami, ale również pozwala na realizację sterowania przepływem w sieci. Właściwie dobrany czas retransmisji zabezpiecza sieć przed przeciążeniami.

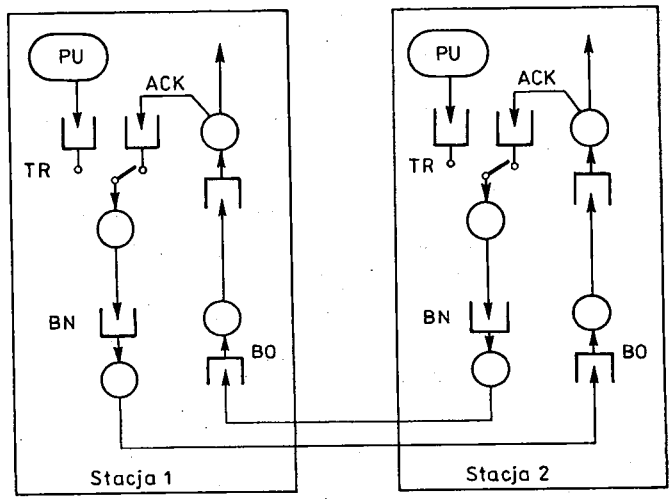
W badanej sieci dostęp do kanału odbywa się według protokołu 1-trwały CSMA/CD (1-persistent). Zgodnie z tym protokołem każda aktywna stacja w sposób ciągły prowadzi nasłuch łącza. Z punktu widzenia stacji łącze może być zajęte, wolne lub może trwać strefa buforowa (interframe gap). Strefa buforowa stanowi pewien odcinek czasowy po ustaniu zajętości łącza. Istnienie strefy buforowej ułatwia odbiornikowi poprawne przyjmowanie kolejnych ramek. Stacja może nadawać tylko wówczas, gdy łącze jest wolne. Jeżeli w momencie badania łącze jest zajęte, lub trwa strefa buforowa, to należy odłożyć rozpoczęcie nadawania do końca tej strefy. Gdy po rozpoczęciu nadawania zostanie wykryta kolizja, to po okresie wymuszenia kolizji stacja zawiesza swą aktywność na czas $t_i$, gdzie i jest numerem próby. Po pierwszej próbie stacja może podejmować co najwyżej 15 prób dodatkowych. Jeżeli żadna z nich się nie powiedzie to wstrzymuje się realizację danego żądania. Czas $t_i$ zawieszenia aktywności po $i$-tej próbie, ($i=1,2,\ldots,16$) jest wyznaczony z zależności $t_i=rt$, przy czym r jest liczbą naturalną wybieraną losowo z przedziału $<0,2k-1>$, $k=\min\{i,10\}$, a $t$ jest wartością szczeliny czasowej. Szczelina czasowa jest umowną wielkością wyznaczoną jako podwójny, maksymalny czas propagacji sygnału, powiększony o czas niezbędny do wykrycia kolizji i wymuszania kolizji. Dla podstawowego

wariantu normy ISO 8802.3 dostosowanego do szybkości transmisji 10 Mb/s obowiązują następujące dane liczbowe:

| | |
|---|---|
| minimalna strefa buforowa | 9,6 $\mu$s, |
| szerokości szczeliny czasowej | 51,2 $\mu$s |
| czas wymuszania kolizji | 3,2 $\mu$s, |
| maksymalna długość ramki | 1518 bajtów, |
| minimalna długość ramki | 64 bajty. |

Szerokość szczeliny czasowej ogranicza dopuszczalną wartość podwójnego czasu propagacji, a więc limituje geometryczną rozległość sieci.

Dotychczasowe rozważania pozwalają na sformułowanie modelu symulacyjnego badanej sieci. Przedstawiono go na rys. 2.




Rys. 2. Model symulacyjny dla dwóch wyróżnionych stacji

Zgłoszenia do sieci, w formie wiadomości, generowane są przez proces użytkownika PU. Wiadomość podzielona na segmenty umieszczona jest w warstwie transportowej (bufor transportowy BT). Kolejne segmenty, zgodnie z omówionym wcześniej mechanizmem retransmisji, są przesyłane kolejno, w miarę wolnych miejsc do bufora nadawczego (układ obsługi BN) z szybkością $v_{tn}$. Tu następuje formowanie pakietu do nadawania. Czas formowania określa parametr $t_n$. Dostęp do magistrali odbywa się zgodnie z wcześniej omówionym protokołem CSMA/CD. Po przejściu przez sieć pakiety umieszczone są w buforze odbiorczym (układ obsługi BO). Tu odbywa się przetworzenie informacji kontrolno-sterujących. Czas przetwarzania określa parametr $t_0$. Utworzone w ten sposób segmenty są przesyłane do warstwy transportowej z szybkością $v_{0t}$. Po odebraniu W segmentów z jednej stacji generowane jest potwierdzenie ACK. W modelu przyjęto pierwszeństwo dla pakietów potwierdzających na poziomie warstwy transportowej.

## 3. BADANIA SYMULACYJNE

Badania symulacyjne dotyczyły jednego segmentu sieci lokalnej typu Ethernet. Podstawę do obliczeń stanowiła magistrala o długości d = 500m. Założono, że stacje są równomiernie rozmieszczone wzdłuż kabla. W związku z tym odstęp między dwoma sąsiednimi stacjami wynosił $d/(n-1)$ gdzie $n$ — liczba stacji. Prędkość transmisji w magistrali $v_m = 10$ Mb/s. Szybkość przesyłania informacji w adapterze między buforem warstwy transportowej (BT) a buforami warstwy łącza danych (BN lub BO) wynosiła $v_{tn} = v_{0t} = 800$ kb/s, natomiast czasy przetwarzania pakietów w tych buforach były określone jako $t_n = t_0 = 1$ms. Są to typowe wielkości w obecnie spotykanych adapterach [2], [3]. Ruch generowany przez procesy użytkowe miał rozkład Poissona z parametrem $\lambda_i$ (dla $i$-tej stacji). Przyjęto, że średnia długość wiadomości wynosi $l_w = 80000$ bitów, a podstawowa długość pakietu (segmentu) bez bitów sterujących i kontrolnych równa się $l_p = 8000$ bitów. Założono też, że ruch w sieci jest symetryczny to znaczy każda stacja generuje zgłoszenia z taką samą intensywnością $\lambda_i, i = 1,2,..,n$. Rozpływ ruchu był równomierny. Oznaczało to, że prawdopodobieństwo, iż wiadomość ze stacji i jest kierowana do stacji j wynosiło $p_{ij} = 1/(n-1)$. Należy tu podkreślić, że założenie o równomiernym rozpływie nie stanowi żadnego ograniczenia dla opracowanego programu symulacyjnego. Program jest bowiem tak skonstruowany, że umożliwia prowadzenie badań symulacyjnych dla różnych strategii rozpływu ruchu. W świetle przeprowadzonych badań symulacyjnych [4], rozpływ równomierny jest najmniej korzystny z punktu widzenia średniego czasu transmisji pakietu. W badaniach przyjęto również, że liczba buforów nadawczych wynosi $b_n = 10$, a podstawowa liczba buforów odbiorczych równa się $b_0 = 20$. Ponadto założono, że podstawowy czas retransmisji $t_r = 100$ ms, natomiast wielkość okna W = 5. W przypadku wystąpienia kolizji wprowadza się losowe opóźnienie pakietów zgodnie z binarnym wykładniczym algorytmem wycofywania. W myśl tego algorytmu pakiety uczestniczące w kolizji szesnastokrotnie są usuwane z sieci. Ze względu na przyjęty mechanizm potwierdzenia i retransmisji takie pakiety nie giną bezpowrotnie. Docierają do celu jedynie z większym opóźnieniem.

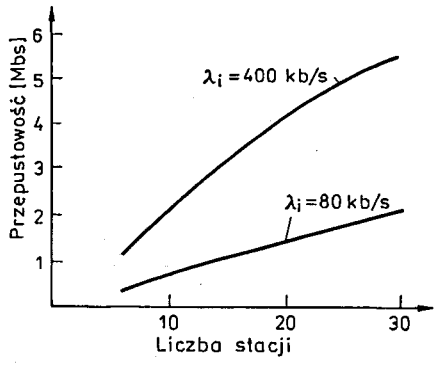
Przejdziemy teraz do omówienia wyników badania efektywności pracy sieci lokalnej za pomocą symulacji. Jako podstawowe wskaźniki efektywności przyjęto średni czas retransmisji (pakietu lub wiadomości) i przepustowość.

Pierwsze eksperymenty dotyczyły badania obciążenia w sieci lokalnej. Na rys. 3. przedstawiono zależność przepustowości sieci od liczby stacji. Przepustowość jest rozumiana jako liczba przesyłanych bitów (pakietów, wiadomości) w sieci w jednostce czasu. Założono, że każda stacja generuje taką samą ilość zgłoszeń do sieci. Rozpatrzono dwa przypadki:

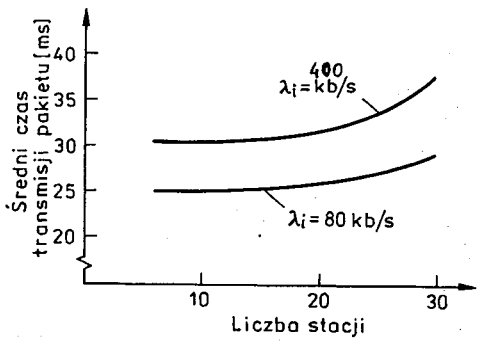
a) gdy $i$-ta stacja generuje zgłoszenia z intensywnością $\lambda_i = 80$ kb/s

b) gdy $i$-ta stacja generuje z intensywnością $\lambda_i = 400$ kb/s.

Jak widać z rys. 3. początkowo przepustowość rośnie liniowo, dalej w miarę zwiększania się liczby stacji wzrost przepustowości staje się coraz mniejszy. Oznacza to pogorszenie się warunków pracy sieci, co można zaobserwować na rys. 4. Przedstawiono tam związek między średnim czasem transmisji pakietu a liczbą stacji.

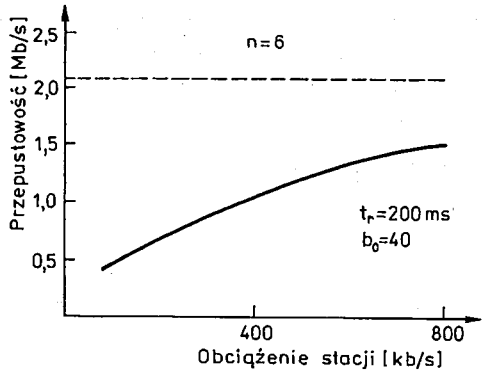



Rys. 3. Wpływ liczby stacji na przepustowość

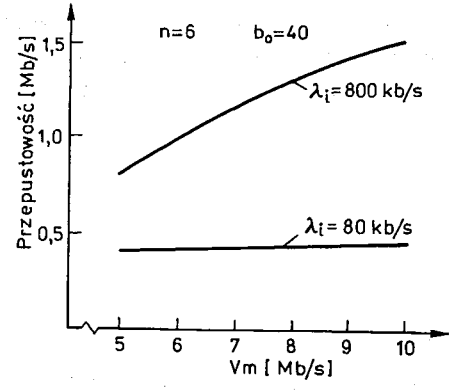



Rys. 4. Wpływ liczby stacji na średni czas transmisji pakietu

W modelu symulacyjnym przyjęto, że nowa wiadomość może zostać wygenerowana dopiero po pozytywnym wysłaniu poprzedniej. Jest to zrozumiałe z praktycznego punktu widzenia. W przypadku sieci słabo obciążonej istnieje zawsze wyraźny odstęp czasu między pojawieniem się nowej wiadomości a zakończeniem nadawania poprzedniej. Przykładowo zbadano zależność przepustowości sieci od szybkości pojawiania się nowych wiadomości w stacjach. Badanie to umożliwiło także wyznaczenie maksymalnej przepustowości rozpatrywanej sieci (oczywiście przy ustalonych pozostałych parametrach). Uzyskane wyniki przedstawiono na rys. 5. Linią przerywaną zaznaczono przepustowość maksymalną.




Rys. 5. Wpływ szybkości pojawiania się nowych zgłoszeń na przepustowość sieci



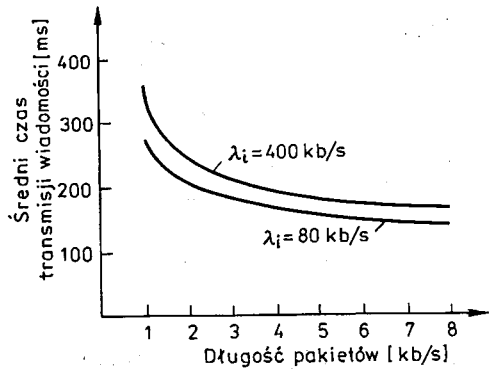

Rys. 6. Wpływ liczby stacji na przepustowość sieci

Kolejne badania dotyczyły wpływu prędkości transmisji w magistrali na przepustowość sieci. Uzyskane wyniki przedstawiono na rys. 6. Rozpatrzono dwa przypadki:

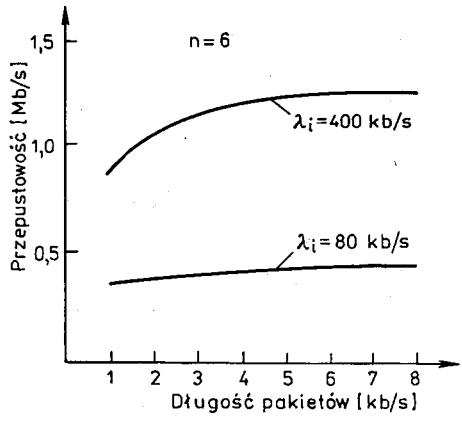
a) sieć słabo obciążona ($\lambda_i = 80$ kb/s)

b) sieć mocno obciążona ($\lambda_i = 800$ kb/s).

Jak widać prędkości transmisji w magistrali ma duży wpływ na przepustowość sieci w przypadku dużego ruchu.



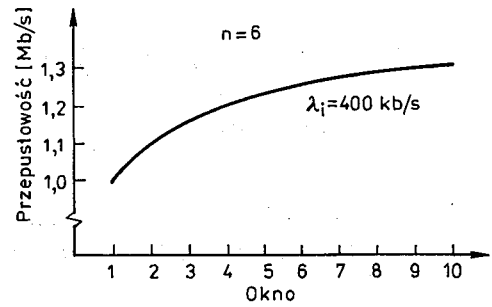

Rys. 7. Wpływ długości pakietów na średni czas transmisji wiadomości



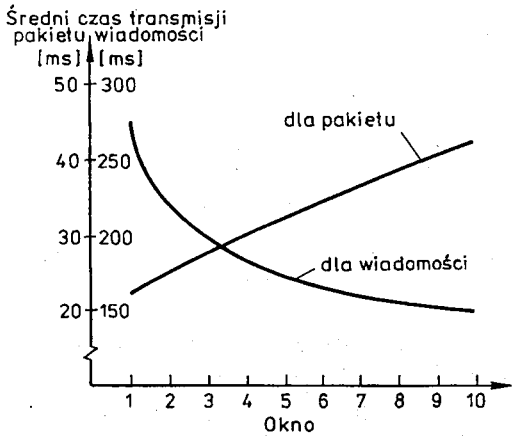

Rys. 8. Wpływ długości pakietów na przepustowość sieci

Z kolei zbadano wpływ długości pakietów na czas transmisji wiadomości. Dla dwóch różnych intensywności napływu zgłoszeń uzyskane wyniki przedstawiono na rys.7. Jak widać przez wydłużenie pakietu można wyraźnie zmniejszyć czas transmisji wiadomości. Jest to spowodowane przede wszystkim zmniejszeniem czasu przetwarzania pakietów tworzących jedną wiadomość (maleje liczba pakietów przypadających na jedną wiadomość). W taki sam sposób można wyjaśnić wpływ długości pakietów na przepustowość sieci, jak to pokazano na rys. 8.

Następnie zbadano wpływ parametrów protokołu transportowego (wielkość okna i czas retransmisji) na efektywność pracy sieci lokalnej. Na wstępie rozpatrzono wpływ wielkości okna. Na rys. 9. przedstawiono zależność przepustowości sieci od wielkości okna. Ze wzrostem okna rośnie liczba pakietów, które mogą jednocześnie znajdować



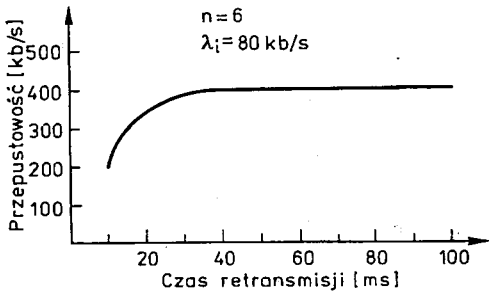

Rys. 9. Wpływ wielkości okna na przepustowość sieci



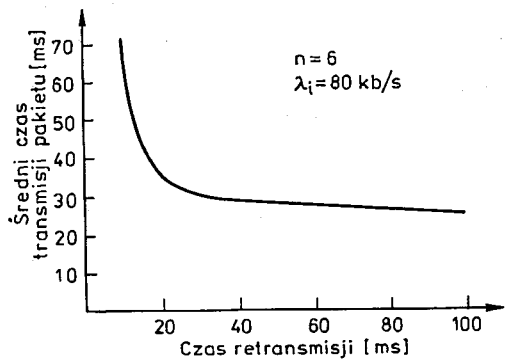

Rys. 10. Wpływ wielkości okna na średni czas transmisji pakietu

się w sieci. To stanowi główną przyczynę wzrostu przepustowości oraz zmniejszenia czasu transmisji wiadomości oczywiście przy założeniu, że sieć nie jest przeciążona, czyli, że oczekiwanie w buforach sieci nie trwa zbyt długo. Jednakże wzrost liczby pakietów w sieci powoduje, że rośnie nieco ich średni czas transmisji. Potwierdziły to badania symulacyjne, których wyniki przedstawiono na rys. 10.

Dalsze badania dotyczyły wpływu czasu retransmisji na efektywność pracy sieci. Trzeba zaznaczyć, że duża wartość czasu retransmisji w stosunku do czasu transmisji pakietów nie ma wpływu na pracę sieci. W normalnych warunkach wpływ ten zaznacza się dopiero wtedy, gdy wielkości te są porównywalne. Zaczynają wówczas pojawiać się duplikaty pakietów, już krążących w sieci. Dalsze zmniejszanie czasu retransmisji (*time out*) szybko przyspiesza ten proces. Obciążenie w sieci znacznie wzrasta. Wydłuża się przez to wyraźnie czas transmisji pakietów. Szybko zmniejsza się przepustowość sieci, gdyż maleje liczba pakietów efektywnie obsłużonych. Widać to wyraźnie na rysunkach 11 i 12. W skrajnym przypadku może nawet dojść do zablokowania się sieci. Takie przypadki miały miejsce w trakcie eksperymentów symulacyjnych.




Rys. 11. Wpływ czasu retransmisji na przepustowość sieci




Rys. 12. Wpływ czasu retransmisji na średni czas transmisji pakietu

Wszystkie badania symulacyjne zostały przeprowadzone przy pomocy programu komputerowego napisanego w języku Turbo Pascal. Program ten w wersji źródłowej zajmuje 50 kB. Czasy obliczeń były bardzo zróżnicowane. Dla sieci słabo obciążonej, do określenia jednego punktu krzywej, obliczenia trwały kilkanaście minut. Natomiast dla dużych obciążeń i dużej liczby stacji niektóre przebiegi trwały nawet kilkanaście godzin. Wszystkie obliczenia przeprowadzono przy założeniu, że poziom ufności wynosi 95%.

## ZAKOŃCZENIE

W pracy przedstawiono wyniki badań efektywności działania lokalnej sieci magistralowej z dostępem rywalizacyjnym. Efektywność sieci lokalnej określano za pomocą przepustowości i średniego czasu transmisji. W przyjętym modelu badanej

sieci uwzględniono zasadnicze czynniki mające decydujący wpływ na pracę sieci. Należą do nich: czas dostępu do kanału, przetwarzanie w adapterach, sprzęg między adapterem i komputerem oraz podstawowe parametry protokołu transportowego jak rozmiar okna i czas retransmisji. Zbadano zmiany przepustowości sieci oraz średniego czasu transmisji (pakietów lub wiadomości) w zależności od liczby stacji, obciążenia sieci i długości pakietów. Przeprowadzono również eksperymenty symulacyjne pozwalające na określenie wpływu wielkości okna i czasu retransmisji na efektywność pracy sieci lokalnej.

## BIBLIOGRAFIA

1. J. Hayes: *Modeling and Analysis of Computer Communications Networks.* Plenum Press, New York, 1984
2. S. Heatley, D. Stokesterny: *Analysis of Transport Measurements over a Local Area Network.* IEEE Communications Magazine, vol. 27, nr 6, 1989
3. B. Meister, P. Janson, L. Svobodova: *Connection-Oriented versus Connectionless Protocols: A Performance Study.* IEEE Trans. on Computers, vol. C–34, nr 12, 1985
4. T. Nowicki (red.): *Symulacja wybranych protokołów lokalnej sieci teleprzetwarzania np. dla potrzeb grafiki komputerowej.* IPI PAN, Warszawa, 1989
5. A. Wolisz: *Podstawy lokalnych sieci komputerowych. Sprzęt sieciowy.* Warszawa: WNT, 1990

J. HOSZA, T. STACHOWIAK

## PERFORMANCE ANALYSIS OF A LOCAL AREA NETWORK WITH CONTENTION ACCESS METHOD

### Summary

The paper presents performance analysis of a local area network with contention access method. The authors used computer simulation to carry out the investigations. The performance of the network is measured in terms of throughput and message transmission time. The model of the network includes four lower layers of the OSI architecture.

KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJI, 1992, 38, z. 1, ss. 87-119

# Optimal nonuniform sampling zero-crossing digital phase-locked loops for tracking the offsets in phase or in phase and frequency

MARIUSZ ŻÓŁTOWSKI

*Instytut Telekomunikacji, Politechnika Gdańska*

*Received 1991.01.04*

*Authorized 1991.03.20*

An approach to the design of optimal nonuniform sampling digital phase-locked loops (DPLL-s) by optimal controlling algorithm on circle is considered in systematic way. The offsets in phase or in phase and frequency of input waveform are of concern. Similar case has been considered by Willsky previously [1]. However the problem of estimation only was of concern before. The evolving densities are given in terms of their Fourier coefficients. Limiting cases and linear approximations are presented either. The approach of concern is in order to give rise to different controlling algorithms leading to improvement in performance of nonuniform sampling DPLL-s.

## 1. INTRODUCTION

The digital circuits are of increasing interest and importance in modern communication systems [3]. The phase locked loop or strictly digital phase-locked loop as implemented within digital circuitry is just one of these ones. It has been devised in order to make the performance of communication systems more reliable, flexible, accurate and almost optimal. In this paper the solution to the problem of optimal nonlinear control is used to design the optimal zero-crossing digital phase-locked loops. The probabilistic approach is used.

## 2. THE STATEMENT OF THE PROBLEM OF OPTIMAL NONLINEAR FILTRATION

Suppose $x_t$ is the vector of state of dynamical system at time instant $t$ while $Y_t$ is the set of attainable observations concerning the state up to time instant $t$. Then the conditional density of probability

$$p(x, t \mid Y_t) \tag{1}$$

is the solution to the problem of nonlinear filtration [2]. This is true because $p(x,t \mid Y_t)$ contains the whole statistical information about $x_t$ which can be gained from observations and from the initial condition $x_0$ given in terms of its density $p(x_0)$. Bayesian point view is expressed this way according to which a priori information about $x_0$ is taken into account. However if we succeed with getting the density of (1) in hand there is still problem what kind of estimate is required. The solution to the latter problem is given in two steps. The measure of estimation given in terms of estimation error

$$\epsilon_t = x_t - \hat{x}_t, \tag{2}$$

while $\hat{x}_t$ of (2) stands for the estimate of $x_t$ is introduced first.

Definition

Any Lebesgue measurable, nonnegative function $L$,

$$L : X \to R,$$
$$X \ni x \to L(x)$$

satysfying dim $X=n$, $X$ convex subset of $R^n$,

$$L(0) = 0$$

and which is convex is called a loss function.
Secondly, after the loss function $L$ has been defined, the estimate $\hat{x}_t$ of vector $x_t$ is looked for which minimizes the expected loss

$$\underset{X \times Y}{E} \{L(\epsilon_t)\}. \tag{3}$$

The expectation concerns the set of all possible observations Y and possible configurations of state $X$. Moreover this operation can be performed in two stages;

$$\underset{Y \times X}{E} \{L(\epsilon_t)\} = \underset{Y}{E}\{\underset{X|Y}{E}(\epsilon_t) \mid Y_t)\}. \tag{4}$$

Due to this last decomposition it is enough to consider the minimization of conditional expectation

$$\underset{X|Y}{E} \{L(\epsilon_t) \mid Y_t\}. \tag{5}$$

The result of optimization is well known in particular but important case.

Theorem 1 (The case of continous space of state)

Suppose $p(x,t) \mid Y_\tau)(\tau \leqslant t)$ is the unimodal density of probabilistic measure absolutely continous in respect to Lebesgue one of the state vector $x_t$ (of nx1 dimension). Let also the conditional expectation exist

$$\hat{x}_t^\tau = E\{x_t \mid Y_\tau\}. \tag{6}$$

Then the optimal estimate which minimizes the loss function $L$ symmetric is the conditional mean $\hat{x}_t^\tau$ of (6). The conditional mean is the minimal variance estimate in any case. The proof of this theorem is given in Appendix 1.

## 3. THE SOLUTION TO THE PROBLEM OF NONLINEAR FILTRATION IN THE CONTINOUS SPACE OF STATE AND DISCRETE TIME CASE

Suppose the state of dynamical system evolves according to

$$x_{k+1} = a(x_k, t_k, t_{k+1}) + b(x_k, t_k) w_{k+1} \qquad k = 0, 1, \ldots \tag{7}$$

$$x_{k_{nx1}},\ a_{nx1},\ b_{nxr},\ w_{k_{rx1}},$$

$\{w_k : k=1,\ldots\}$ – white sequence of gaussian random variables $w_k \sim N(O, Q_k)$. The latter notation stands for the random variable of gaussian distribution, zero mean value and $Q_k$ covariance matrix. Moreover the initial condition $x_0$ and $w_k (k=1,\ldots)$ are statistically independent.

On the other hand the observations change according to

$$y_k = h(x_k, t_k) + v_k \tag{8}$$

$$y_{k_{mx1}},\ h_{mx1},\ v_{k_{mx1}},$$

$v_k (k=1,\ldots)$ is the white sequence of random gaussian variables; $v_k \sim N(0, R_k)$. The sequences $\{w_k : k=1,\ldots\}$ and $\{v_k : k=1,\ldots\}$ are statistically independent whereas matrices $Q_k$ and $R_k$ are positively definite; $Q_k > 0$ and $R_k > 0$.

Now, the solution to optimal filtering problem, the evolution of conditional density of probability, is readily obtained by the rule of Bayes in view of Markov model under consideration (see Appendix II):

$$p(x, t_k \mid Y_k) = \frac{p(y_k|x_k) \int p(x_k|x_{k-1})\, p(x_{k-1}, t_{k-1} \mid Y_{k-1})\, \mathrm{d}x_{k-1}}{\iint p(y_k|x_k)\, p(x_k|x_{k-1})\, p(x_{k-1}|Y_{k-1})\, \mathrm{d}x_{k-1}\, \mathrm{d}x_k} \tag{9a}$$

$$p(x, t_0 \mid Y_0) = p(x_0) \tag{9b}$$

$$p(y_k|x_k) = \frac{1}{\sqrt{(2\Pi)^m |R_k|}} \exp\left\{-\frac{1}{2}(y_k - h(x,t_k))^T \times R_k^{-1} \times (y_k - h(x,t_k))\right\} \tag{9c}$$

$$p(x_k|x_{k-1}) = \frac{1}{\sqrt{(2\Pi)^n |Q_k'|}} \exp\left\{-\frac{1}{2}(x_k - a(x_{k-1}, t_{k-1}, t_k))^T \times \right.$$

$$\left. \times\ Q_k'^{-1} (x_k - a(x_{k-1}, t_{k-1}, t_k))\right\}, \tag{9d}$$

$$Q_k' \overset{\mathrm{df}}{=} b(x_k, t_k)\, Q_k\, b^T(x_k, t_k). \tag{9e}$$

## 4. THE SOLUTION TO THE PROBLEM OF OPTIMAL TRACKING OF ZERO-CROSSING INSTANTS OF HARMONIC WAVEFORM. OPTIMAL CONTROLLING ALGORITHM FOR NONUNIFORM SAMPLING DPLL

The dynamics of the nonuniform sampling DPLL which is a nonlinear feedback system (see fig. 1) is given in terms of equations

$$x_{k+1} = x_k + \Theta_{k+1} - \Theta_k - \omega_0 c_k \tag{10a}$$

$$y_k^1 = A_m \sin x_k + v_k^1 \tag{10b}$$

$$y_k^2 = A_m \cos x_k + v_k^2 . \tag{10c}$$

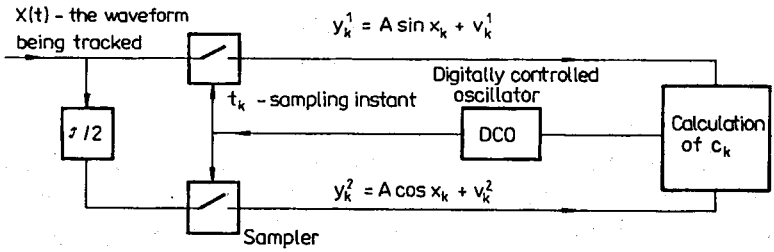



Fig. 1. Nonuniform sampling zerocrossing digital phase locked loop DPLL

The deviation from the state of phase entrainment (zerocrossings of input waveform) is given in terms of the instantenous phase error (shortly phase error) process $x_k{:}k=0,1,\ldots$, $A_m$ stands for amplitude, $\Theta_k$ for the phase and $\omega_0$ for the radian frequency of the input waveform. Index $k$ is to mark the sampling instant $t_k{:}k=0,1,\ldots$. The disturbed by noise observations are given in terms of (10b – 10c). The sequences $v_k^1{:}k=0,1,\ldots$, $v_k^2{:}k=0,1,\ldots$ are white, gaussian and statistically independent. The digital oscillator of the loop of fig. 1 is controlled according to the algorithm

$$t_{k+1} = t_k + T - c_k, \; k = 0, 1, \ldots \tag{11}$$

$T$ of (11) is the nominal period of sampling.
Also

$$T = \frac{2\Pi}{\omega_0} N \; (N \text{ being integer}) \tag{12}$$

while the perturbed $N$:1 phase entrainment is of interest.

Now the problem is to determine the optimal controlling sequence $c_k{:}k=0,1,\ldots$ . This sequence should be chosen according to the criterion which measures the fidelity of tracking of zerocrossing transitions of input waveform. The case of fidel tracking of zerocrossings is called the state of phase entrainment. The speed of acquisition of phase lock is very important from the practical point of view. That is why the criterion which is chosen to optimize is

$$\underset{X|Y}{E}\{L(x_{k+1})|Y_k\}, \; k = 0, 1, \ldots \tag{13}$$

In other words the control $c_k$:$k=0,1,\ldots$ is to assure the minimal loss on every stage of observations.

In view of previous chapter the general solution to the problem stated before is as follows:

1) The determination of conditional density of probability

$$p(x, t_k | Y_k) \; k = 0, 1, \ldots$$

2) With this density in hand the optimal prediction of $x_{k+1}$ which minimizes the expected loss of (13) can be obtained. Note that

$$p(x, t_{k+1} | Y_k) = \int p(x_{k+1} | x_k) p(x, t_k | Y_k) \, dx_k . \tag{14}$$

The density of (14) depends on control $c_k$: $k=0,1,\ldots$ and $\Theta_{k+1} - \Theta_k$ in view of (10). The optimal control which minimizes

$$\underset{X|Y}{E}\{L(x_{k+1}) | Y_k\} \tag{15}$$

is chosen. It is denoted as $c_k^{opt}$, $k=0,1,\ldots$ . Thus one should set $c_{k-1} = c_{k-1}^{opt}$ while $p(x_k|x_{k-1})$ of (9) is considered.

If the conditional mean

$$\hat{x}_{k+1}^k \stackrel{df}{=} E\{x_{k+1} | Y_k\} \tag{16}$$

is the solution to the problem of optimal prediction then

$$\begin{aligned} \hat{x}_{k+1}^k &= E\{x_{k+1} | Y_k\} = E\{x_k + \Theta_{k+1} - \Theta_k - \omega_0 c_k | Y_k\} = \\ &= E\{x_k | Y_k\} + E\{\Theta_{k+1} - \Theta_k | Y_k\} - \omega_0 c_k^{opt} = \\ &= \hat{x}_k^k + E\{(\Theta_{k+1} - \Theta_k) | Y_k\} - \omega_0 c_k^{opt} \end{aligned} \tag{17a}$$

and

$$\hat{x}_k^k \stackrel{df}{=} E\{x_k | Y_k\} \tag{17b}$$

Moreover the value of loss function for the optimal estimate of $x_{k+1}$ is

$$L(\hat{x}_{k+1}^k) = L(\hat{x}_k^k + E\{\Theta_{k+1} - \Theta_k | Y_k\} - \omega_0 c_k^{opt}). \tag{18}$$

Thus

$$c_k^{opt} = \hat{x}_k^k + E\{\Theta_{k+1} - \Theta_k | Y_k\} / \omega_0 : k = 0, 1, \ldots \tag{19}$$

is the optimal control while

$$\hat{x}_{k+1}^k = 0$$

is the unbiased estimate of $x_{k+1}$ for which

$$L(\hat{x}_{k+1}^k) = 0.$$

## 5. THE OPTIMAL TRACKING OF PHASE OFFSET

The case of optimal tracking of the offset of phase versus the local reference from DCO is considered in this chapter.
The equations of the DPLL are in this case (see appendix 3)

$$x_{k+1} = x_k - \omega_0 c_k + w_k \tag{20a}$$

$$y_k^1 = A_m \sin x_k + v_k^1 \tag{20b}$$

$$y_k^2 = A_m \cos x_k + v_k^2 \tag{20c}$$

$$w_k \sim N(0, Q_k) \tag{20d}$$

$$v_k^1 \sim N(0, R_k);\ v_k^2 \sim N(0, R_k). \tag{20e}$$

The optimal control $c_k: k=0,1,\ldots$ is obtained according to (19). Suppose the density of initial phase error is uniform in $<-\Pi,\Pi>$ closed interval;

$$p(x_0) = \frac{1}{2\Pi},\ x_0 \in <-\Pi, \Pi>. \tag{21a}$$

Then

$$c_k^{opt} = \hat{x}_k^k/\omega_0 \tag{21b}$$

$$c_0^{opt} = 0 \tag{21c}$$

$$E\{\Theta_{k+1} - \Theta_k\} = 0 \tag{21d}$$

and

$$\hat{x}_k^k = E\{x_k \mid Y_k\} = \int x\, p(x, t_k \mid Y_k)\, dx. \tag{21e}$$
$$k = 0, 1, \ldots$$

The equations (9) are for this case

$$p(x, t_0 \mid Y_0) = p(x_0) \tag{22a}$$

$$p(x, t_k \mid Y_k) = \frac{p(y_k \mid x_k) \int p(x_k \mid x_{k-1})\, p(x_{k-1}, t_{k-1} \mid Y_{k-1})\, dx_{k-1}}{\iint p(y_k \mid x_k)\, p(x_k \mid x_{k-1})\, p(x_{k-1}, t_{k-1} \mid Y_{k-1})\, dx_{k-1}\, dx_k}. \tag{22b}$$
$$k = 0, 1, \ldots$$

Whereas

$$p(y_k \mid x_k) = p(y_k^1, y_k^2 \mid x_k) = \frac{1}{2\Pi R_k} \exp\left\{ \frac{-\frac{1}{2}(y_k^1 - A_m \sin x_k)^2}{R_k} + \right.$$
$$\left. - \frac{1}{2} \frac{(y_k^2 - A_m \cos x_k)^2}{R_k} \right\} \tag{22c}$$
$$k = 0, 1, \ldots$$

and

$$p(x_k|x_{k-1}) = \frac{1}{\sqrt{2\Pi Q_{k-1}}} \exp\left\{-\frac{1}{2}\frac{(x_k - x_{k-1} + \omega_0 c_{k-1}^{opt})^2}{Q_{k-1}}\right\} \tag{22d}$$

$$k = 1, \ldots$$

However, the distributions given in hand should be considered on circle $S^1 = <-\Pi, \Pi>$ closed interval with equivalent end points in view of cyclic nature of phase error process. Thus only $x_k$ from $S^1$ cycle manifold are concerned while (22c) is considered and (22d) is modified according to

$$p(x_k|x_{k-1}) = \sum_{m=-\infty}^{\infty} \frac{1}{\sqrt{2\Pi Q_{k-1}}} \exp\left\{-\frac{1}{2}\frac{(x_k - x_{k-1} + \omega_0 c_{k-1}^{opt} + 2m\Pi)^2}{Q_{k-1}}\right\} \tag{23}$$

$$x_k, x_{k-1} \in S^1 \quad k = 1, \ldots$$

Moreover the integration of (22b) inudues the $S^1$ circle manifold only. Now, the practical way of computing the optimal control may be looked for. First of all the conditional density of (23) can be represented in terms of Fourier series;

$$p(x_k|x_{k-1}) = \sum_{n=-\infty}^{\infty} a_n e^{jn(x_k - x_{k-1} + \omega_0 c_{k-1}^{opt})} \tag{24}$$

$$x_k, x_{k-1} \in S^1.$$

Similar representation can be used while the density of (22c) is considered (see Appendix 4)

$$p(y_k|x_k) = \sum_{n=-\infty}^{\infty} b_n(y_k) e^{jnx_k} \tag{25}$$

$$x_k \in S^1, \; y_k = (y_k^1, y_k^2), \; k = 0, 1, \ldots$$

Secondly, the conditional density $p(x,t_k|Y_k)$ can be given Fourier representation also;

$$p(x, t_k|Y_k) = \sum_{n=-\infty}^{\infty} c_n(k) e^{jnx_k} \tag{26}$$

$$k = 0, 1, \ldots$$

Thus the conditional density for the prediction purpose is

$$p(x, t_k|Y_{k-1}) = \int p(x_k|x_{k-1}) p(x_{k-1}|Y_{k-1}) \, dx_{k-1} =$$

$$= \int_{-\Pi}^{\Pi} \sum_{n_1=-\infty}^{\infty} a_{n_1} e^{jn_1(x_k - x_{k-1} + \omega_0 c_{k-1}^{opt})} \sum_{n_2=-\infty}^{\infty} c_{n_2}(k-1) e^{jn_2 x_{k-1}} \, dx_{k-1} =$$

$$= \sum_{n_1=-\infty}^{\infty} \sum_{n_2=-\infty}^{\infty} e^{jn_1(x_k + \omega_0 c_{k-1}^{opt})} a_{n_1} c_{n_2}(k-1) \int_{-\Pi}^{\Pi} e^{jx_{k-1}(n_2 - n_1)} \, dx_{k-1} =$$

$$= \sum_{n=-\infty}^{\infty} 2\Pi a_n c_n(k-1)\, e^{jn(x_k+\omega_0 c_{k-1}^{opt})} =$$

$$= \sum_{n=-\infty}^{\infty} c_n(k\,|\,k-1)\, e^{jnx_k} \tag{27a}$$

while

$$c_n(k\,|\,k-1) = 2\Pi a_n c_n(k-1)\, e^{jn\omega_0 c_{k-1}^{opt}} \tag{27b}$$

$$n = \ldots, -1, 0, 1, \ldots$$

It means that Fourier coefficients of the conditional density being looked for are given in terms of previous ones according to (27b). Subsequently

$$p(y_k\,|\,Y_{k-1}) = \iint p(y_k\,|\,x_k)\, p(x_k\,|\,x_{k-1})\, p(x_{k-1}\,|\,Y_{k-1})\, \mathrm{d}x_{k-1}\, \mathrm{d}x_k =$$

$$= \int p(y_k\,|\,x_k)\, p(x_k\,|\,Y_{k-1})\, \mathrm{d}x_k =$$

$$= \int_{-\Pi}^{\Pi} \sum_{n_1=-\infty}^{\infty} b_{n_1}(y_k)\, e^{jx_k n_1} \sum_{n_2=-\infty}^{\infty} c_{n_2}(k\,|\,k-1)\, e^{jx_k n_2}\, \mathrm{d}x_k =$$

$$= \int_{-\Pi}^{\Pi} \sum_{n_1=-\infty}^{\infty} \sum_{n_2=-\infty}^{\infty} b_{n_1}(y_k)\, c_{n_2}(k\,|\,k-1)\, e^{j(n_1+n_2)x_k}\, \mathrm{d}x_k =$$

$$= 2\Pi \sum_{n=-\infty}^{\infty} b_n\, c_{-n}(k\,|\,k-1)\,. \tag{28}$$

Also

$$p(y_k\,|\,x_k)\, p(x_k\,|\,Y_{k-1}) = \sum_{n=-\infty}^{\infty} b_n(y_k)\, e^{jx_k n} \sum_{n=-\infty}^{\infty} c_n(k\,|\,k-1)\, e^{jx_k n} =$$

$$= \sum_{n=-\infty}^{\infty} d_n(y_k)\, e^{jx_k n}, \tag{29a}$$

while

$$d_n(y_k) = \sum_{m=-\infty}^{\infty} c_{n-m}(k\,|\,k-1)\, b_m(y_k) \tag{29b}$$

From (28) and (29b)

$$p(y_k\,|\,Y_{k-1}) = 2\Pi\, d_0(y_k) \tag{29c}$$

Finally, according to (9a)

$$c_n(k) = \frac{d_n(y_k)}{2\Pi\, d_0(y_k)} \qquad n = \ldots, -1, 0, 1, \ldots \tag{30}$$

$$k = 1, \ldots$$

It turns out that the evolving probabilities which are required for the solution of the problem of optimal control are given in the recursive way in terms of their Fourier coefficients;

$$c_n(k|k-1) = 2\Pi a_n c_n(k-1) e^{jn\omega_0 c_{k-1}^{opt}} \tag{31a}$$

$$d_n(y_k) = \sum_{m=-\infty}^{\infty} c_{n-m}(k|k-1) b_m(y_k) \tag{31b}$$

$$c_n(k) = \frac{d_n(y_k)}{2\Pi d_0(y_k)}, \quad n = \ldots, -1, 0, 1, \ldots \tag{31c}$$

$$k = 1, \ldots$$

Note, that in view of the cyclic nature of the error of phase it is enough also to consider the control $c_k^{opt}: k=0,1,\ldots$ restricted to $S^1$ circle manifold.

## 6. OPTIMAL TRACKING OF PHASE AND FREQUENCY OFFSETS

Now the case of optimal tracking of phase and frequency offsets can be considered. The equations of the loop while the incoming waveform is detuned in phase and frequency versus the local reference are

$$\begin{bmatrix} x_{k+1}^1 \\ x_{k+1}^2 \end{bmatrix} = \begin{bmatrix} 1 & T-c_k \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_k^1 \\ x_k^2 \end{bmatrix} - \begin{bmatrix} \omega_0 c_k \\ 0 \end{bmatrix} + \begin{bmatrix} w_k^1 \\ w_k^2 \end{bmatrix}. \tag{32a}$$

On the other hand the equation of observations is the same as before

$$\begin{bmatrix} y_k^1 \\ y_k^2 \end{bmatrix} = \begin{bmatrix} A_m \sin x_k^1 \\ A_m \cos x_k^1 \end{bmatrix} + \begin{bmatrix} v_k^1 \\ v_k^2 \end{bmatrix}. \tag{32b}$$

The first equation is different from the previous one while it contains additional coordinate variable due to unknown offset in frequency.
Moreover

$$w_k^1 \sim N(0, Q_k^1); \; w_k^2 \sim N(0, Q_k^2) \tag{33a}$$

and the sequences $w_k^1: k=0,1,\ldots$ and $w_k^2: k=0,1,\ldots$ are statistically independent.
The initial distribution of probability are

$$p(x_0^1) = \frac{1}{2\Pi}, \; x_0^1 \in <-\Pi, \Pi> \; = S^1, \tag{33b}$$

$$p(x_0^2) = \frac{1}{2\Delta\omega}, \; x_0^2 \in <-\Delta\omega + \omega_0, \omega_0 + \Delta\omega> \tag{33c}$$

or

$$p(x_0^2) \sim N(\omega_0, R_\omega). \tag{33d}$$

In view of statistical independence of phase and frequency offsets

$$p(x_0) = p(x_0^1, x_0^2) = p(x_0^1)p(x_0^2). \tag{33e}$$

Now, the optimal estimates of phase and frequency are

$$\hat{x}^{1k}_{k+1} = E\{x^1_{k+1} \mid Y_k\} = \hat{x}^{1k}_k + \hat{x}^{2k}_k T - (\hat{x}^{2k}_k + \omega_0)c_k \tag{34a}$$

and

$$\hat{x}^{2k}_{k+1} = E\{x^2_{k+1} \mid Y_k\} = \hat{x}^{2k}_k. \tag{34b}$$

Thus the optimal control is

$$c^{opt}_k : k = 0, 1, \ldots \tag{35a}$$

$$c^{opt}_k = (\hat{x}^{1k}_k + \hat{x}^{2k}_k T)/(\hat{x}^{2k}_k + \omega_0) \tag{35b}$$

and

$$c^{opt}_0 = 0. \tag{35c}$$

The required densities of probabilities are subsequently

$$p(x, t_0 \mid Y_0) = p(x_0) \tag{36a}$$

$$p(y_k \mid x_k) = \frac{1}{2\Pi R_k} \exp\left\{-\frac{1}{2}\frac{(y^1_k - A_m \sin x^1_k)^2}{R_k} - \frac{1}{2}\frac{(y^2_k - A_m \cos x^1_k)^2}{R_k}\right\} \tag{36b}$$

$$p(x_k \mid x_{k-1}) = p(x^1_k, x^2_k \mid x^1_{k-1}, x^2_{k-1}) =$$

$$= \frac{1}{2\Pi\sqrt{Q^1_{k-1}Q^2_{k-1}}} \exp\left\{-\frac{1}{2}(x^1_k - x^2_{k-1}(T - c^{opt}_{k-1}) + \omega_0 c^{opt}_{k-1})^2 / Q^1_{k-1}\right\} \times$$

$$\times \exp\left\{-\frac{1}{2}(x^2_k - x^2_{k-1})^2 / Q^2_{k-1}\right\}. \tag{36c}$$

The evolution of the density $p(x,t_k|Y_k)(k=0,1,\ldots)$ is described by the general expressions (9). Moreover the marginal densities are used for calculations of $\hat{x}^{1k}_k$ and $\hat{x}^{2k}_k$.

Now, let consider practical method of calculation of optimal control in this case. The phase error $x^1_k : k=0,1,\ldots$ will be restricted to $S^1$ — circle manifold as before. Respectively

$$p(x_k \mid x_{k-1}) = \sum_{n=-\infty}^{\infty} a_n e^{jn(x^1_k - x^1_{k-1} - x^2_{k-1}T + (x^2_{k-1} + \omega_0)c^{opt}_{k-1})} \times$$

$$\times \frac{1}{\sqrt{2\Pi Q^2_{k-1}}} \exp\left\{-\frac{1}{2}(x^2_k - x^2_{k-1})^2 / Q^2_{k-1}\right\} \tag{37a}$$

and $x^1_k \in S^1$ while

$$p(x_k \mid x_{k-1}) = \sum_{m=-\infty}^{\infty} \frac{1}{2\Pi\sqrt{Q^1_{k-1}Q^2_{k-1}}} \exp\left\{-\frac{1}{2}(x^1_k - x^1_{k-1} - x^2_{k-1}(T - c^{opt}_{k-1}) +\right.$$

$$\left. + \omega_0 c^{opt}_{k-1} + 2\Pi m)^2 / Q^1_{k-1}\right\} \times \exp\left\{-\frac{1}{2}(x^2_k - x^2_{k-1})^2 / Q^2_{k-1}\right\} \tag{37b}$$

$$x_k^1,\ x_{k-1}^1 \in S^1$$

$$p(y_k \mid x_k) = \sum_{n=-\infty}^{\infty} b_n(y_k)\, e^{jx_k^1 n} \tag{37c}$$

$$x_k^1 \in S^1$$

and let

$$p_{w_k^2}(x_k^2 - x_{k-1}^2) \stackrel{\text{df}}{=} \frac{1}{\sqrt{2\Pi\, Q_{k-1}^2}} \exp\left\{-\frac{1}{2}(x_k^2 - x_{k-1}^2)^2 / Q_{k-1}^2\right\}. \tag{37d}$$

Now the conditional density is represented as

$$p(x, t_k \mid Y_k) = \sum_{n=-\infty}^{\infty} c_n(k, x_k^2)\, e^{jx_k^1 n}, \tag{37e}$$

while

$$p(x, k_o \mid Y_0) = p(x_0), \quad k = 0, 1, \ldots \tag{37f}$$

Thus

$$p(x, t_k \mid Y_{k-1}) = \int p(x_k \mid x_{k-1})\, p(x_{k-1} \mid Y_k)\, \mathrm{d}x_{k-1} =$$

$$= \int_{-\Pi}^{\Pi} \int_{-\infty}^{\infty} \sum_{n_1=-\infty}^{\infty} a_n\, e^{jn_1(x_k^1 - x_{k-1}^1 - x_{k-1}^2(T - c_{k-1}^{opt} + \omega_0 c_{k-1}^{opt})} \times$$

$$\times\ p_{w_k^2}(x_k^2 - x_{k-1}^2) \sum_{n_2=-\infty}^{\infty} c_{n_2}(k-1, x_{k-1}^2)\, e^{jn_2 x_{k-1}^1}\, \mathrm{d}x_{k-1}^1\, \mathrm{d}x_{k-1}^2 =$$

$$= \int_{-\infty}^{\infty} \sum_{n_1=-\infty}^{\infty} \sum_{n_2=-\infty}^{\infty} a_{n_1}\, e^{jn_1(x_k^1 - x_{k-1}^1(T - c_{k-1}^{opt}) + \omega_0 c_{k-1}^{opt})} \times$$

$$\times\ p_{w_k^2}(x_k^2 - x_{k-1}^2) \times c_{n_2}(k-1, x_{k-1}^2) \int_{-\Pi}^{\Pi} e^{jx_{k-1}^1 (n_2 - n_1)}\, \mathrm{d}x_{k-1}^1\, \mathrm{d}x_{k-1}^2 =$$

$$= \sum_{n=-\infty}^{\infty} 2\Pi\, a_n \int_{-\infty}^{\infty} c_n(k-1, x_{k-1}^2)\, e^{jn(x_k^1 + \omega_0 c_{k-1}^{opt})}\, e^{-jnx_{k-1}^2(T - c_{k-1}^{opt})} \times$$

$$\times\ p_{w_k^2}(x_k^2 - x_{k-1}^2)\, \mathrm{d}x_{k-1}^2. \tag{38a}$$

Let

$$c_n'(k \mid k-1, x_k^2) \stackrel{\text{df}}{=} \int_{-\infty}^{\infty} c_n(k-1, x_{k-1}^2)\, e^{jnx_{k-1}^2(T - c_{k-1}^{opt})}\, p_{w_k^2}(x_k^2 - x_{k-1}^2)\, \mathrm{d}x_{k-1}^2 \tag{38b}$$

Then the density being looked for is

$$p(x, t_k | Y_{k-1}) = \sum_{n=-\infty}^{\infty} c_n(k|k-1, x_k^2)\, e^{jnx_k^1} \tag{38c}$$

and

$$c_n(k|k-1, x_k^2) \stackrel{\text{df}}{=} 2\Pi\, a_n\, c_n'(k|k-1, x_k^2)\, e^{jn\omega_0 c_{k-1}^{pn}} \tag{38d}$$

Secondly

$$p(y_k | Y_{k-1}) = \int p(y_k | x_K) p(x_K | Y_{k-1})\, \mathrm{d}x_k =$$

$$= \int_{-\infty}^{\infty} \int_{-\Pi}^{\Pi} \sum_{n_1=-\infty}^{\infty} b_{n_1}(y_k)\, e^{jx_k^1 n_1} \sum_{n_2=-\infty}^{\infty} c_{n_2}(k|k-1, x_k^2)\, e^{jn_2 x_k^1}\, \mathrm{d}x_k^1 \mathrm{d}x_k^2 =$$

$$= \int_{-\infty}^{\infty} \sum_{n_1=-\infty}^{\infty} \sum_{n_2=-\infty}^{\infty} b_{n_1}(y_k)\, c_{n_2}(k|k-1, x_k^2) \int_{-\Pi}^{\Pi} e^{jx_k^1(n_1+n_2)}\, \mathrm{d}x_k^1\, \mathrm{d}x_k^2 = \mathrm{d}x_k^1 \mathrm{d}x_k^2$$

$$= \sum_{n=-\infty}^{\infty} 2\Pi\, b_n(y_k) \int_{-\infty}^{\infty} c_{-n}(k|k-1, x_k^2)\, \mathrm{d}x_k^2 . \tag{39a}$$

Also

$$p(y_k | x_k)\, p(x_k | Y_{k-1}) = \sum_{n_1=-\infty}^{\infty} b_{n_1}(y_k)\, e^{jx_k^1 n_1} \times$$

$$\times \sum_{n_2=-\infty}^{\infty} c_{n_2}(k|k-1, x_k^2)\, e^{jn_2 x_k^1} = \sum_{n=-\infty}^{\infty} d_n(y_k, x_k^2)\, e^{jx_k^1 n} \tag{39b}$$

and

$$d_n(y_k, x_k^2) = \sum_{m=-\infty}^{\infty} c_{n-m}(k|k-1, x_k^2)\, b_m(y_k) . \tag{39c}$$

Thus

$$p(y_k | Y_{k-1}) = 2\Pi \int_{-\infty}^{\infty} d_0(y_k, x_k^2)\, \mathrm{d}x_k^2 . \tag{39d}$$

As a matter of fact the following algorithm is obtained for recursive calculation of required coefficients of the solution to the optimal control problem

$$c_n'(k|k-1, x_k^2) \stackrel{\text{df}}{=} \int_{-\infty}^{\infty} c_n(k-1, x_{k-1}^2)\, e^{-jnx_{k-1}^1 (T - c_{k-1}^{pn})}\, \mathrm{d}x_{k-1}^2 \tag{40a}$$

$$p_{w_k^2}(x_k^2 - x_{k-1}^2) c_n(k \mid k-1, x_k^2) = 2\Pi a_n c_n'(k \mid k-1, x_k^2) e^{jn\omega_0 c_{k-1}^{opt}}, \tag{40b}$$

$$c_n(k, x_k^2) = \frac{d_n(y_k, x_k^2)}{2\Pi \int_{-\infty}^{\infty} d_0(y_k, x_k^2)\, dx_k^2}, \tag{40c}$$

$$d_n(y_k, x_k^2) = \sum_{m=-\infty}^{\infty} c_{n-m}(k \mid k-1, x_k^2)\, b_m(y_k) \tag{40d}$$

$$k = 1, \ldots, \quad n = \ldots, -1, 0, 1, \ldots$$

The difference between this latter case and the former one is visible now form (31) and (40) in terms of Fourier coefficients of the evolving densities. However the integration of (40a) and (40c) should be facilitated to make this algorithm more feasible. This can be accomplished in terms of power series or Fourier integral. Namely the conditional density can be represented as

$$p(x, t_k \mid Y_k) = \sum_{n=-\infty}^{\infty} e^{jx_k^1 n} \frac{1}{2\Pi} \int_{-\infty}^{\infty} c_n(k, \omega) e^{j\omega x_k^2} d\omega. \tag{41}$$

Now it is given in terms of Fourier series and Fourier integral.

$$p(x, t_k \mid Y_{k-1}) = \int_{-\Pi}^{\Pi} \int_{-\infty}^{\infty} \sum_{n_1=-\infty}^{\infty} a_n e^{jn_1(x_k^1 - x_{k-1}^1 - x_{k-1}^2(T - c_{k-1}^{opt}) + \omega_0 c_{k-1}^{opt})} \times$$

$$\times\, p_{w_k^2}(x_k^2 - x_{k-1}^2) \sum_{n_2=-\infty}^{\infty} e^{jx_k^1 n_2} \frac{1}{2\Pi} \int_{-\infty}^{\infty} c_{n_2}(k-1, \omega) e^{j\omega x_{k-1}^2} d\omega\, dx_{k-1}^1\, dx_{k-1}^2 =$$

$$= \int_{-\infty}^{\infty} \sum_{n_1=-\infty}^{\infty} \sum_{n_2=-\infty}^{\infty} a_n e^{jn_1(x_k^1 - x_{k-1}^1(T - c_{k-1}^{opt}) + \omega_0 c_{k-1}^{opt})} \times p_{w_k^2}(x_k^2 - x_{k-1}^2) \times$$

$$\times \frac{1}{2\Pi} \int_{-\infty}^{\infty} c_{n_2}(k-1, \omega) e^{j\omega x_{k-1}^2} d\omega\, dx_{k-1}^2 =$$

$$= \sum_{n=-\infty}^{\infty} 2\Pi a_n e^{jnx_k^1} e^{jn\omega_0 c_{k-1}^{opt}} \times$$

$$\times \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{jnx_{k-1}^1(T - c_{k-1}^{opt})} e^{j\omega x_{k-1}^2} p_{w_{k-1}^2}(x_k^2 - x_{k-1}^2) dx_{k-1}^2 \times$$

$$\times \frac{c_n(k-1, \omega)}{2\Pi} d\omega. \tag{42a}$$

But

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{jx_{k-1}^2(\omega-n(T-c_{k-1}^{opt}))} p_{w_k^2}(x_k^2 - x_{k-1}^2)\,\mathrm{d}x_{k-1}^2,\ \frac{c_n(k-1,\omega)}{2\Pi}\mathrm{d}\omega =$$

$x_k^2 - x_{k-1}^2 = u$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{j(x_k^2-u\,(\omega-n(T-c_{k-1}^{opt}))} p_{w_k^2}(u)\,\mathrm{d}u \times \frac{c_n(k-1,\omega)\,\mathrm{d}\omega}{2\Pi} =$$

$$= \int_{-\infty}^{\infty} e^{jx_k^2(\omega-n(T-c_{k-1}^{opt}))} \int_{-\infty}^{\infty} e^{-ju(\omega-n(T-c_{k-1}^{opt}))} p_{w_k^2}(u)\,\mathrm{d}u \times$$

$$\times\ c_n\frac{(k-1,\omega)}{2\Pi}\mathrm{d}\omega =$$

$$= \int_{-\infty}^{\infty} e^{jx_k^2(\omega-n(T-c_{k-1}^{opt}))} P_{w_k^2}(\omega - n(T - c_{k-1}^{opt}))\frac{c_n(k-1,\omega)}{2\Pi}\mathrm{d}\omega =$$

$\omega' = \omega - n(T - c_{k-1}^{opt})$

$$= \int_{-\infty}^{\infty} e^{jx_{k-1}^2\omega'} P_{w_k^2}(\omega')\,\frac{c_n(k-1,\omega' + n(T - c_{k-1}^{opt}))}{2\Pi}\mathrm{d}\omega, \tag{42b}$$

while

$P_{w_k^2}(\omega)$ stands for the Fourier transform of $p_{w_k^2}(u)$;

$$P_{w_k^2}(\omega) = \mathscr{F}\{p_{w_k^2}(u)\},\ p_{w_k^2}(u) = \mathscr{F}^{-1}\{P_{w_k^2}(\omega)\}.$$

Let

$$p(x, t_k \mid Y_{k-1}) = \sum_{n=-\infty}^{\infty} e^{jx_k^1 n}\,\frac{1}{2\Pi}\int_{-\infty}^{\infty} c_n(k \mid k-1,\omega)\,e^{j\omega x_k^1}\mathrm{d}\omega. \tag{42c}$$

Then

$$c_n(k \mid k-1,\omega) = 2\Pi\,a_n\,e^{jn\omega_0 c_{k-1}^{opt}}\,c_n(k-1,\omega + n(T - c_{k-1}^{opt}))\,P_{w_k^2}(\omega). \tag{42d}$$

Secondly (see 39d)

$$p(y_k \mid Y_{k-1}) = \sum_{n=-\infty}^{\infty} 2\Pi\,b_n(y_k)\int_{-\infty}^{\infty} c_n(k \mid k-1, x_k^2)\,\mathrm{d}x_k^2 =$$

$$= \sum_{n=-\infty}^{\infty} 2\Pi\, b_n(y_k) \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{2\Pi}\, c_{-n}(k\,|\,k-1,\omega)\, e^{j\omega x_k^1 - 1}\, \mathrm{d}\omega\, \mathrm{d}x_k^2 =$$

$$= \sum_{n=-\infty}^{\infty} 2\Pi\, b_n(y_k) \int_{-\infty}^{\infty} \mathscr{F}^{-1}\{c_{-n}(k\,|\,k-1,\omega)\}\, \mathrm{d}x_k^2 =$$

$$= \sum_{n=-\infty}^{\infty} 2\Pi\, b_n(y_k) \int_{-\infty}^{\infty} \mathscr{F}^{-1}\{c_{-n}(k\,|\,k-1,\omega)\}\, \mathrm{e}^{j\omega x_k^1} \Big|_{\omega=0} \mathrm{d}x_k^2 =$$

$$= \sum_{n=-\infty}^{\infty} 2\Pi\, b_n(y_k)\, c_{-n}(k\,|\,k-1,\,0)\,. \tag{43a}$$

Also (see 39f–39h)

$$p(y_k\,|\,x_k)\, p(x_k\,|\,Y_{k-1}) = \sum_{n=-\infty}^{\infty} d_n(y_k, x_k^2)\, e^{jx_k^1 n} \tag{43b}$$

and

$$d_n(y_k, x_k^2) = \sum_{m=-\infty}^{\infty} c_{n-m}(k\,|\,k-1, x_k^2)\, b_m(y_k) \tag{43c}$$

Thus

$$d_n(y_k, x_k^2) = -\frac{1}{2\Pi} \int_{-\infty}^{\infty} \sum_{m=-\infty}^{\infty} c_{n-m}(k\,|\,k-1,\omega)\, b_m(y_k)\, e^{j\omega x_k^1}\, \mathrm{d}\omega\,. \tag{43d}$$

Let

$$d_n(y_k, \omega) \stackrel{\mathrm{df}}{=} \sum_{m=-\infty}^{\infty} c_{n-m}(k\,|\,k-1,\omega)\, b_m(y_k)\,. \tag{43e}$$

Then

$$p(y_k\,|\,x_k)\, p(x_k\,|\,Y_k) = \sum_{n=-\infty}^{\infty} e^{jx_k^1 n} \frac{1}{2\Pi} \int_{-\infty}^{\infty} d_n(y_k, \omega)\, e^{j\omega x_k^1}\, \mathrm{d}\omega\,. \tag{43f}$$

Thus

$$p(y_k\,|\,Y_{k-1}) = 2\Pi\, d_0(y_k, 0) \tag{43g}$$

Finally

$$c_n(k\,|\,k-1,\omega) = 2\Pi\, a_n\, e^{jn\omega_0 c_{k-1}^{opt}}\, c_n(k-1, \omega + n(T - c_{k-1}^{opt}))\, P_{w_k^2}(\omega) \tag{44a}$$

$$d_n(y_k, \omega) = \sum_{m=-\infty}^{\infty} c_{n-m}(k \mid k-1, \omega)\, b_m(y_k) \tag{44b}$$

$$c_n(k, \omega) = \frac{d_n(y_k, \omega)}{2\Pi\, d_0(y_k, 0)}. \tag{44c}$$

The marginal densities are readily obtained by using this form and the required statistics either. Further improvements involving the special functions are possible either [1].

## 7. THE SIMPLIFIED CASE OF OPTIMAL TRACKING OF PHASE OFFSET

1). Suppose the instabilities of phase are negligible. Then ($\delta$ stands for the function of Dirac)

$$p(x_k \mid x_{k-1}) = \delta(x_k - x_{k-1} + \omega_0\, c_{k-1}^{opt}) \tag{45}$$

and the integration of (27a) results in

$$p(x, t_k \mid Y_{k-1}) = \int \delta(x_k - x_{k-1} + \omega_0\, c_{k-1}^{opt}) \sum_{n=-\infty}^{\infty} c_n(k-1)\, e^{jnx_{k-1}}\, dx_{k-1} =$$

$$= \sum_{n=-\infty}^{\infty} c_n(k-1)\, e^{jn(x_k + \omega_0 c_{k-1}^{opt})}. \tag{46}$$

As a matter of fact the simplified form of (27b) is

$$c_n(k \mid k-1) = c_n(k-1)\, e^{jn\omega_0 c_{k-1}^{opt}} \tag{47}$$

$$n = \ldots, -1, 0, 1, \ldots$$

2). Suppose the instabilities of phase are negligible and observations are noiseless. Then

$$p(y_k \mid x_k) = \delta(y_k^1 - A_m \sin x_k)\, \delta(y_k^2 - A_m \cos x_k). \tag{48}$$

Now the integration of (28) is

$$p(y_k \mid Y_{k-1}) = \int p(y_k \mid x_k)\, p(x_k \mid Y_{k-1})\, dx_k =$$

$$= \int_{-\Pi}^{\Pi} \delta(y_k^1 - A_m \sin x_k)\, \delta(y_k^2 - A_m \cos x_k) \sum_{n=-\infty}^{\infty} c_n(k \mid k-1)\, e^{jx_k n}\, dx_k =$$

$$= \int_{-\Pi}^{\Pi} \delta\left(e^{jx_k} - \frac{y_k^2 + y_k^1}{A_m}\right) \sum_{n=-\infty}^{\infty} \frac{c_n(k \mid k-1)}{j}\, e^{jx_k(n-1)}\, d\, e^{jx_k} =$$

$$= \sum_{n=-\infty}^{\infty} \frac{c_n(k \mid k-1)}{j} \frac{(y_k^2 + jy_k^1)^{n-1}}{A_m}. \tag{49a}$$

$$\delta\left(e^{jx_k} - \frac{y_k^2 + jy_k^1}{A_m}\right) = \delta(y_k^1 - A_m \sin x_k)\,\delta(y_k^2 - A_m \cos x_k). \tag{49b}$$

Also

$$p(y_k|x_k)\,p(x_k|Y_{k-1}) = \delta(y_k^1 - A_m \sin x_k)\,\delta(y_k^2 - A_m \cos x_k) \sum_{n=-\infty}^{\infty} c_n(k|k-1)\,e^{jx_k n} =$$

$$= \delta\left(e^{jx_k} - \frac{y_k^2 + jy_k^1}{A_m}\right) \sum_{n=-\infty}^{\infty} c_n(k|k-1)\,e^{jx_k n}. \tag{50}$$

Thus

$$c_n(k) = \frac{c_n(k|k-1)\,\delta\left(e^{jx_k} - \dfrac{y_k^2 + jy_k^1}{A_m}\right)}{\dfrac{1}{2\Pi j} \displaystyle\sum_{n=-\infty}^{\infty} 2\Pi\, c_n(k|k-1) \left(\frac{y_k^2 + jy_k^1}{A_m}\right)^{n-1}}. \tag{51}$$

However

$$c_n(k|k-1) = c_n(k-1)\,e^{jn\omega_0 c_{k-1}^{opt}} \tag{52a}$$

$$c_n(0) = \frac{1}{2\Pi}\delta_n, \quad \delta_n = \begin{cases} 0 & n \neq 0 \\ 1 & n = 0 \end{cases} \tag{52b}$$

and

$$c_0^{opt} = 0 \tag{52c}$$

$$c_k^{opt} = \hat{x}_k^k / \omega_0. \tag{52d}$$

Thus

$$c_n(k) = \frac{c_n(k-1)\,e^{jn\omega_0 c_{k-1}^{opt}}\,\delta\left(e^{jx_k} - \dfrac{y_k^2 + jy_k^1}{A_m}\right)}{\dfrac{1}{2\Pi j} \displaystyle\sum_{n=-\infty}^{\infty} 2\Pi\, c_n(k-1)\,e^{jn\omega_0 c_{k-1}^{opt}} \left(\frac{y_k^2 + jy_k^1}{A_m}\right)^{(n-1)}} =$$

$$= \frac{c_0(k-1)\,\delta\left(e^{jx_k} - \dfrac{y_k^2 + jy_k^1}{A_m}\right)}{\dfrac{1}{2\Pi j}\, 2\Pi\, c_0(k-1) \left(\dfrac{y_k^2 + jy_k^1}{A_m}\right)^{-1}} =$$

$$= j\,\delta\left(e^{jx_k} - \frac{y_k^2 + jy_k^1}{A_m}\right)\left(\frac{y_k^2 + jy_k^1}{A_m}\right). \tag{53}$$

In view of (53)

$$\int_{-\Pi}^{\Pi} p(x_k, t_k|Y_k)\,dx = \int_{-\Pi}^{\Pi} j\,\delta\left(e^{jx_k} - \frac{y_k^2 + jy_k^1}{A_m}\right)\frac{(y_k^2 + jy_k^1)}{A_m}\,dx_k =$$

$$= \frac{(y_k^2 + jy_k^1)}{A_m} \int_{-\Pi}^{\Pi} \delta\left(e^{jx_k} - \frac{y_k^2 + jy_k^1}{A_m}\right) e^{-jx_k} \mathrm{d}e^{jx_k} = 1. \tag{54}$$

Moreover

$$\hat{x}_k^k = \int_{-\Pi}^{\Pi} x\, p(x, t_k \mid Y_k)\mathrm{d}x = \int_{-\Pi}^{\Pi} x_k j\, \delta\left(e^{jx_k} - \frac{y_k^2 + jy_k^1}{A_m}\right)\left(\frac{y_k^2 + jy_k^1}{A_m}\right)\mathrm{d}x_k =$$

$$= \int_{-\Pi}^{\Pi} x_k\, \delta\left(e^{jx_k} - \frac{y_k^2 + jy_k^1}{A_m}\right)\frac{(y_k^2 + jy_k^1)}{A_m} e^{-jx_k}\mathrm{d}e^{jx_k} =$$

$$= x_k : e^{jx_k} = \frac{y_k^2 + jy_k^1}{A_m} \text{ and } x_k < -\Pi, \Pi > . \tag{55}$$

Let define ($X$ stands for Cartesian product)

$$\arg: < -A_m, A_m > X < -A_m, A_m > \rightarrow < -\Pi, \Pi >,$$

$$(y_k^2, y_k^1) \rightarrow \arg \frac{y_k^1}{y_k^2} = x_k \quad \text{of (55).}$$

Now the optimal DPLL in noiseless case is shown in fig. 2. Note that it is also optimal in the case of noiseless observations only.

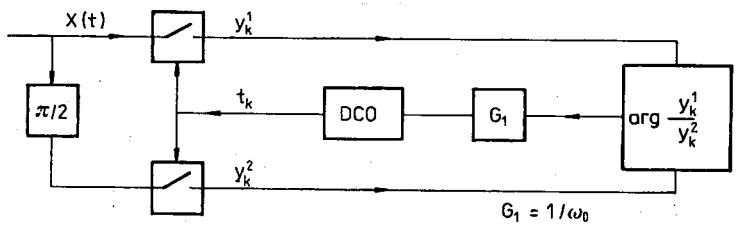



Fig 2. Optimal DPLL in noiseless case

## 8. THE SIMPLIFIED CASE OF OPTIMAL TRACKING OF PHASE AND FREQUENCY OFFSETS

1). Suppose the instabilities of frequency are negligible.
Then

$$p(x_k \mid x_{k-1}) = \sum_{m=\infty}^{\infty} \frac{1}{2\Pi\sqrt{Q_{k-1}^1 Q_{k-1}^2}} \times$$

$$\exp\left\{-\frac{1}{2}\frac{(x_k^1 - x_{k-1}^1 - x_{k-1}^2(T - c_{k-1}^{opt}) + \omega_0 c_{k-1}^{opt} + 2\Pi m)^2}{Q_{k-1}^1}\right\} \delta(x_k^2 - x_{k-1}^2). \quad (56)$$

Now the $p(x,t_k \mid Y_{k-1})$ of (38a) is

$$p(x,t_k \mid Y_{k-1}) = \sum_{n=-\infty}^{\infty} 2\Pi a_n c_n(k-1,x_k^2)\, e^{jn(x_k^1 + \omega_0 c_{k-1}^{opt})}\, e^{-jnx_k^2 (T - c_{k-1}^{opt})} =$$

$$= \sum_{n=-\infty}^{\infty} c_n(k \mid k-1, x_k^2)\, e^{jnx_k^1}. \quad (57)$$

Thus

$$c_n(k \mid k-1, x_k^2) = 2\Pi a_n\, c_n(k-1, x_k^2)\, e^{jn\omega_0 c_{k-1}^{opt}}\, e^{-jnx_k^2 (T - c_{k-1}^{opt})} \quad (58)$$

$$k = 1, \ldots \quad n = \ldots, -1, 0, 1, \ldots$$

If

$$p(x_0 \mid Y_0) = p(x_0) = p(x_0^1)\, p(x_0^2) \quad (59)$$

and

$$p(x_0^1) = 1/2\Pi, \quad x_0^1 \in <-\Pi, \Pi> \quad (60a)$$

$$p(x_0^2) = 1/2\Delta\omega_0, \quad x_0^2 \in <-\Delta\omega_0 + \omega_0, \omega_0 + \Delta\omega_0> \quad (60b)$$

then

$$c_n(0, x_0^2) = \frac{1}{2\Pi}\frac{1}{2\Delta\omega_0}\delta(n). \quad (61)$$

Also respectively

$$p(y_k \mid Y_{k-1}) = \int p(y_k \mid x_k)\, p(x_k \mid Y_{k-1})\, \mathrm{d}x_k =$$

$$= \int_{\omega_0 - \Delta\omega_0}^{\omega_0 + \Delta\omega_0} \int_{-\Pi}^{\Pi} \sum_{n_1=-\infty}^{\infty} b_{n_1}(y_k)\, e^{jx_k^1 n_1} \sum_{n_2=-\infty}^{\infty} c_{n_2}(k \mid k-1, x_k^2)\, e^{jn_2 x_k^1}\, \mathrm{d}x_k^1\, \mathrm{d}x_k^2 =$$

$$= \sum_{n=-\infty}^{\infty} 2\Pi b_n(y_k) \int_{\omega_0 - \Delta\omega_0}^{\omega_0 + \Delta\omega_0} c_{-n}(k \mid k-1, x_k^2)\, \mathrm{d}x_k^2 \quad (62)$$

$$p(y_k \mid x_k)\, p(x_k \mid Y_{k-1}) = \sum_{n_1=-\infty}^{\infty} b_{n_1}(y_k)\, e^{jx_k^1 n_1} \sum_{n_2=-\infty}^{\infty} c_{n_2}(k \mid k-1, x_k^2)\, e^{jn_2 x_k^1} =$$

$$= \sum_{n=-\infty}^{\infty} d_n(y_k, x_k^2)\, e^{jx_k^1 n} \quad (63)$$

$$d_n(y_k, x_k^2) = \sum_{m=-\infty}^{\infty} c_{n-m}(k \mid k-1, x_k^2)\, b_m(y_k) \quad (64)$$

$$p(y_k \mid Y_{k-1}) = 2\Pi \int_{-\Delta\omega_0+\omega_0}^{\omega_0+\Delta\omega_0} d_0(y_k, x_k^2)\,\mathrm{d}x_k^2 \tag{65}$$

$$n = \ldots, -1, 0, 1, \ldots \quad k = 1, \ldots$$

2). Suppose the instabilities of phase and frequency are negligible and the observations are noiseless.
Then

$$p(x_k \mid x_{k-1}) = \delta(x_k^1 - x_{k-1}^1 - x_{k-1}^2(T - c_{k-1}^{opt}) + \omega_0 c_{k-1}^{opt})\,\delta(x_k^2 - x_{k-1}^2) \tag{66}$$

and

$$p(y_k \mid x_k) = \delta(y_k^1 - A_m \sin x_k)\,\delta(y_k^2 - A_m \cos x_k). \tag{67}$$

Thus

$$p(x, t_k \mid Y_{k-1}) = \int p(x_k \mid x_{k-1})\,p(x_{k-1} \mid Y_{k-1})\,\mathrm{d}x_k =$$

$$\int \delta(x_k^1 - x_{k-1}^1 - x_{k-1}^2(T - c_{k-1}^{opt}) + \omega_0 c_{k-1}^{opt})\,\delta(x_k^2 - x_{k-1}^2 \sum_{n=-\infty}^{\infty} c_n(k-1, x_{k-1}^2) e^{jnx_{k-1}^1}\,\mathrm{d}x_{k-1}^2\,\mathrm{d}x_{k-1}^1 =$$

$$= \sum_{n=-\infty}^{\infty} c_n(k-1, x_k^2)\, ejn\,(x_n^1 - x_k^2(T - c_{k-1}^{opt}) + \omega_0 c_{k-1}^{opt}) =$$

$$= \sum_{n=-\infty}^{\infty} c_n(k \mid k-1, x_k^2)\, e^{jnx_k^1} \tag{68}$$

$$c_n(k \mid k-1, x_k^2) = c_n(k-1, x_k^2)\, e^{jn\omega_0 c_{k-1}^{opt}}\, e^{-jnx_k^2(T - c_{k-1}^{opt})} \tag{69}$$

$$n = \ldots, -1, 0, 1, \ldots, \quad k = 1, \ldots$$

As a matter of fact

$$c_n(k) = \frac{c_k(k \mid k-1, x_k^2)\,\delta\left(e^{jx_k^1} - \dfrac{y_k^2 + jy_k^1}{A_m}\right)}{\dfrac{1}{2\Pi j}\displaystyle\sum_{n=-\infty}^{\infty} 2\Pi c_n(k \mid k-1, x_k^2)\left(\dfrac{y_k^2 + jy_k^1}{A_m}\right)^{(n-1)}} \tag{70}$$

as in the phase offset only case.
Moreover

$$c_0^{opt} = 0 \tag{71a}$$

$$c_k^{opt} = (\hat{x}^{1k}_k + \hat{x}^{2k}_k T)/(\hat{x}^{2k}_k + \omega_0) \tag{71b}$$

$$c_n(0, x_0^2) = \frac{1}{2\Pi}\frac{1}{2\Delta\omega_0}\delta(n). \tag{71c}$$

$$k = 0, 1, \ldots$$

Next one obtains the result already known from (70) (see 53)

$$c_n(k) = j\delta\left(e^{jx_k^1} - \frac{y_k^2 + jy_k^1}{A_m}\right)\left(\frac{y_k^2 + jy_k^1}{A_m}\right). \tag{71d}$$

Thus

$$x_k^1 = \arg\frac{y_k^1}{y_k^2}, \quad k = 1, \ldots \tag{71e}$$

$$c_1^{opt} = x_1/\omega_0 \tag{71f}$$

$$\hat{x}_1^{21} = 0 \tag{71g}$$

But

$$x_2^1 = x_1^1 + (T - c_1^{opt})\, x_1^2 - \omega_0 c_1^{opt}. \tag{72a}$$

Thus

$$\hat{x}_2^{22} = x_1^2 = \frac{x_2^1 - x_1^1 + \omega_0 c_1^{opt}}{T - c_1^{opt}}. \tag{72b}$$

The phase error remains zero after two steps.

## 9. THE LINEAR APPROXIMATION OF CONTROL PROBLEM IN DPLL

Let the problem of optimal control is considered once again with simplified assumptions;

$$x_k^1 = x_{k-1}^1 + (T - c_k)\, x_{k-1}^2 - \omega_0 c_{k-1} + w_k^1 \tag{73a}$$

$$x_k^2 = x_{k-1}^2 + w_k^2 \tag{73b}$$

$$w_k^1 \sim N(0, Q_k^1), \quad w_k^2 \sim N(0, Q_k^2) \tag{73c}$$

$$y_k = A_m x_k^1 + v_k^1; \quad v_k^1 \sim N(0, R_k) \tag{73d}$$

$$k = 0, 1, \ldots$$

Now the equation of observation is linearized around the state of phase entrainment. The equations of evolution of densities are

$$p(x, t_0 \mid Y_0) = p(x_0) \tag{74a}$$

$$p(x, t_k \mid Y_k) = \frac{p(y_k \mid x_k)\, p(x_k \mid Y_{k-1})}{p(y_k \mid Y_{k-1})} \tag{74b}$$

$$k = 0, 1, \ldots$$

Before the evolving densities are given in hand let us state the problem in general way. The general equation are[1)]

$$x_k = A x_{k-1} + B c_{k-1} + w_k \tag{75a}$$

$$y_k = C x_k + v_k. \tag{75b}$$

[1)] Matrices **A**, **B**, **C**, are time variant generally. The dependence on time is not being marked to avoid proliferation.

A of (75) is the matrix of state, $B$ is the matrix of control, whereas $C$ is the matrix of observation. Also $w_k$ and $v_k$ stands for noise components in vector form.
The densities of (74) are gaussian. Thus

$$p(x, t_{k-1} | Y_{k-1}) \sim N(\hat{x}_{k-1}^{k-1}, P_{k-1}^{k-1}) \tag{76a}$$

$$p(y, t_k | x_k) \sim N(C x_k, R_k). \tag{76b}$$

Moreover

$$\hat{x}_k^{k-1} = E\{x_k | Y_{k-1}\} = E\{A x_{k-1} + B c_{k-1} + w_k | Y_{k-1}\} = A \hat{x}_{k-1}^{k-1} + B c_{k-1} \tag{76c}$$

and

$$x_k - \hat{x}_k^{k-1} = A(x_{k-1} - \hat{x}_{k-1}^{k-1}) + w_k. \tag{76d}$$

Also

$$P_k^{k-1} = E\{(x_k - \hat{x}_k^{k-1})(x_k - \hat{x}_k^{k-1})^T | Y_{k-1}\} = A P_{k-1}^{k-1} A^T + Q_k. \tag{76e}$$

That is why

$$p(x_k | Y_{k-1}) \sim N(\hat{x}_k^{k-1}, p_k^{k-1})$$

with $\hat{x}_k^{k-1}$ and $P_k^{k-1}$ of (76c) and (76e).

The same way

$$\hat{y}_k^{k-1} = E\{y_k | Y_{k-1}\} = C \hat{x}_k^{k-1} \tag{76f}$$

and

$$y_k - \hat{y}_k^{k-1} = C(x_k - \hat{x}_k^{k-1}) + v_k. \tag{76g}$$

Thus

$$E\{(y_k - \hat{y}_k^{k-1})(y_k - \hat{y}_k^{k-1})^T | Y_{k-1}\} = C P_k^{k-1} C^T + R_k. \tag{76h}$$

$$p(y, t_k | Y_{k-1}) \sim N(c\hat{x}_k^{k-1}, \mathrm{CP}_k^{k-1}\mathrm{C}^{\mathrm{T}} + \mathrm{R}_k)$$

Now in view of (74b)

$$p(x, t_k | Y_k) = \frac{1}{\sqrt{(2\Pi)^n | P_k^k|}} \cdot \exp\left\{-\frac{1}{2} \|x_k - \hat{x}_k^k\|^2_{P_k^{k-1}}\right\} =$$

$$= \frac{1}{\sqrt{(2\Pi)^m | R_k|}} \cdot \exp\left\{-\frac{1}{2} \|y_k - C x_k\|^2_{R_k^{-1}}\right\} \cdot$$

$$\cdot \frac{1}{\sqrt{(2\Pi)^n |P_k^{k-1}|}} \exp\left\{-\frac{1}{2} \|x_k - \hat{x}_k^{k-1}\|^2_{P_k^{k-1}-1}\right\} \cdot$$

$$\cdot \sqrt{(2\Pi)^m |\mathrm{CP}_k^k \mathrm{C}^{\mathrm{T}} + \mathrm{R}_k|} \exp\left\{\frac{1}{2} \|\mathrm{y} - \mathrm{C}\hat{\mathrm{x}}_k^k\|^2_{(\mathrm{CP}_k^k \mathrm{C}^{\mathrm{T}} + \mathrm{R}_k)^{-1}}\right\} \tag{77a}$$

$$\|\cdot\|^2_P = (\cdot)^T P(\cdot), \quad n = 2, m = 1.$$

That is why

$$\|x_k - \hat{x}_k^k\|^2_{P_k^{k-1}} = \|y_k - C x_k\|^2_{R_k^{-1}} + \|x_k - \hat{x}_k^{k-1}\|^2_{P_k^{k-1-1}} + \\ - \|y_k - C\hat{x}_k^k\|^2_{(CP_k^kC^T+R_k)^{-1}} \tag{77b}$$

Let $R_k^{-1} = A_1 - P_k^{k-1-1} = A_2$ and $(CP_k^kC^T + R_k)^{-1} = A_3$.
These matrices are positively definite and symmetric. Then the right hand side of (77b) is

$$(\cdot) = \|y_k\|^2_{A_1} - 2(y_k|Cx_k)_{A_1} + \|Cx_k\|^2_{A_1} + \\ + \|x_k\|^2_{A_2} - 2(x_k|\hat{x}_k^{k-1})_{A_2} + \|\hat{x}_k^{k-1}\|^2_{A_2} + \\ - \|y_k - Cx_k^k\|^2_{A_3}. \tag{77c}$$

$(|)$ stands for the scalar product of vectors: $(a|b)_A = a^TAb = (a|Ab)$.
Next

$$\|x_k\|^2_{A_2} + \|Cx_k\|^2_{A_1} - 2\big((y_k|Cx_k)_{A_1} + (x_k|\hat{x}_k^{k-1})_{A_2}\big) = \\ = \|x_k\|^2_{A_2} + \|x_k\|^2_{C^TA_1C} - 2\big((Cx_k|y_k)_{A_1} + (\hat{x}_k^{k-1}|x_k)_{A_2}\big) = \\ = \|x_k\|^2_{A_2+C^TA_1C} - 2\big((x_k|C^TA_1y_k) + (x_k|A_2\hat{x}_k^{k-1})\big) = \\ = \|x_k\|^2_{A_2+C^TA_1C} - 2\big((x_k|C^TA_1y_k + A_2\hat{x}_k^{k-1})\big) = \\ = \|x_k\|^2_{A_2+C^TA_1C} - 2(x_k|(A_2 + C^TA_1C)^{-1}(C^TA_1y_k + A_2\hat{x}_k^{k-1})^{-1})_{A_2+C^TA_1C} + \\ + \|(A_2 + C^TA_1C)^{-1}(C^TA_1y_k + A_2\hat{x}_k^{k-1})\|^2_{A_2+C^TA_1C} + \\ - \|(A_2 + C^TA_1C)^{-1}(C^TA_1y_k + A_2\hat{x}_k^{k-1})\|^2_{A_2+C^TA_1C} = \\ = \|x_k - (A_2 + C^TA_1C)^{-1}(C^TA_1y_k + A_2\hat{x}_k^{k-1})\|^2_{A_2+C^TA_1C} + \\ - \|(A_2 + C^TA_1C)^{-1}(C^TA_1y_k + A_2\hat{x}_k^{k-1})\|^2_{A_2+C^TA_1C} \tag{77d}$$

Now note then $x_k$ is not present in every component of (77c) but

$$\|x_k - (A_2 + C^TA_1C)^{-1}(C^TA_1y_k + A_2\hat{x}_k^{k-1})\|^2_{A_2+C^TA_1C} \tag{77e}$$

and the presented decomposition is unique. That is why the rest of components sum to zero while (77a) is gaussian distribution.
Thus

$$P_k^{k-1} = A_2 + C^TA_1C = P_k^{k-1-1} + C^TR_k^{-1}C \tag{77f}$$

and

$$\hat{x}_k^k = (A_2 + C^TA_1C)^{-1}(C^TA_1y_k + A_2\hat{x}_k^{k-1}) = \\ = P_k^k(C^TR_k^{-1}y_k + P_k^{k-1-1}\hat{x}_k^{k-1}). \tag{77g}$$

Summarizing one gets

$$x_0^0 = x_0; \quad P_0^0 = P_0 \tag{78a}$$
$$\hat{x}_k^{k-1} = A\hat{x}_{k-1}^{k-1} + Bc_{k-1} \tag{78b}$$
$$P_k^{k-1} = AP_{k-1}^{k-1}A^T + Q_k \tag{78c}$$
$$\hat{x}_k^k = P_k^k(C^TR_k^{-1}y_k + P_k^{k-1-1}\hat{x}_k^{k-1}) \tag{78d}$$

$$P_k^{k-1} = P_k^{k-1-1} + C^T R_k^{-1} C; \quad k = 1, \ldots \tag{78e}$$

The equations (78) are well known in filtering theory as Kalman's filter algorithm in the case of $c_k = 0$, $(k = 0,1,\ldots)$. The optimal controlling sequence $\{c_k^{opt};\ k = 0,1,\ldots\}$ minimizes the expected loss on each stage of control. Note that matrix $A$ is depedent on $c_{k-1}^{opt}$, $k = 1,\ldots$ (see 73).
Also

$$P_k^{k-1-1} + C^T R_k^{-1} C = P_k^{k-1-1}(I + P_k^{k-1} C^T R_k^{-1} C). \tag{79a}$$

Thus

$$P_k^k = (I + P_k^{k-1} C^T R_k^{-1} C)^{-1} P_k^{k-1}. \tag{79b}$$

Moreover by direct verification

$$(I + D)^{-1} = I - (I + D^{-1})^{-1} \tag{79c}$$

if well established. Inversion in generalized way $A^{-1}{}_{n\times m} A A^{-1} A = A$.
Respectively view of (79c)

$$D = P C^T R^{-1} C \tag{i}$$

$$D^{-1} = (R^{-1} C)^{-1} (P C^T)^{-1} \tag{ii}$$

$$I + D^{-1} = (P C^T + (R^{-1} C)^{-1})(P C^T)^{-1} \tag{iii}$$

$$(I + D^{-1})^{-1} = P C^T (P C^T + C^{-1} R)^{-1} = P C^T (c^{-1}(C P C^T + R))^{-1} = P C^T (C P C^T + R)^{-1} C. \tag{iiii}$$

Thus

$$P_k^k = P_k^{k-1} - P_k^{k-1} C^T (C P_k^{k-1} C^T + R^{-1}) C P_k^{k-1} = P_k^{k-1} - K_k C P_k^{k-1} \tag{80a}$$

and

$$K_k \stackrel{df}{=} P_k^{k-1} C^T (C P_k^{k-1} C^T + R)^{-1} \tag{80b}$$

Also by direct verification

$$(P^{-1} + C^T R^{-1} C)^{-1} C^T R^{-1} = P C^T (R + C P C^T)^{-1} = K. \tag{80c}$$

Now (78d) takes equivalent form from (80)

$$\hat{x}_k^k = K_k y_k + (P_k^{k-1} - K_k C P_k^{k-1-1}) P_k^{k-1} \hat{x}_k^{k-1} = (I - K_k C) \hat{x}_k^{k-1} + K_k y_k = \hat{x}_k^{k-1} + K_k (y_k - C \hat{x}_k^{k-1}). \tag{81}$$

The derived expressions are useful in optimizing the nonuniform sampling DPLL while its preformance is considered around the state of phase entrainment.

## 11. KALMAN FILTER ORIENTED APPROXIMATION TO THE PROBLEM OF OPTIMAL CONTROL IN DPLL

Kalman's filter equations are ($c_k^{opt}=0$)

$$\hat{x}_0^0 = x_0;\ P_0^0 = P_0 \tag{82a}$$

$$\hat{x}_k^{k-1} = A\hat{x}_{k-1}^{k-1} + B_0 \text{ (additional term } B_0) \tag{82b}$$

$$P_k^{k-1} = AP_{k-1}^{k-1}A^T + Q_k \tag{82c}$$

$$\hat{x}_k^k = P_k^k(C^T R_k^{-1} y_k + P_k^{k-1}\hat{x}_k^{k-1}) \tag{82d}$$

$$P_k^k = P_k^{k-1} - K_k C P_k^{k-1} \tag{82e}$$

$$K_k = P_k^{k-1}C^T(CP_k^{k-1}C^T + R)^{-1} = P_k^k C^T R_k^{-1} \tag{82f}$$

$$P_k^{k-1} = P_k^{k-1\,-1} + C^T R_k^{-1} C \tag{82g}$$

$$\hat{x}_k^k = \hat{x}_k^{k-1} + K_k(y_k - C\hat{x}_k^{k-1}) \tag{82h}$$

$$k = 0, 1, \ldots$$

Now let the model of state transitions refer to the zerocrossings of signal to the loop's input. First

$$T = \frac{2\Pi}{\omega_0 + \Omega_0} \underset{\Omega_0 \ll \omega_0}{\simeq} \frac{2\Pi}{\omega_0}\left(1 - \frac{\Omega_0}{\omega_0}\right) = T - T\frac{\Omega_0}{\omega_0} \tag{83a}$$

$$x_k^1 = t_k \tag{83b}$$

$$x_k^2 = -T\frac{\Omega_0}{\omega_0} \tag{83c}$$

$\omega_0$ stands for the radian frequency of input signal, $\Omega_0$ for frequency offsets and $t_k$ for sampling instant at k moment. In view of last relations the equation of state transitions in noiseless case is

$$\begin{bmatrix} x_k^1 \\ x_k^2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} x_{k-1}^1 \\ x_{k-1}^2 \end{bmatrix} + \begin{bmatrix} T \\ 0 \end{bmatrix}; \quad k = 1, \ldots \tag{83d}$$

On the other hand the equation of observation at zerocrossing instant is

$$y_k = Cx_k + v_k \tag{83e}$$

$$v_k \sim N(O, R_k); \quad \begin{array}{l} C=[1\ 0]\ A_m\omega_0\left(1 + \dfrac{\Omega_0}{\omega_0}\right) \\ \text{or} \\ C=[1\ 0]\ \omega_0\left(1 + \dfrac{\Omega_0}{\omega_0}\right). \end{array} \tag{83f}$$

Because that zerocrossing instants are observed

$$y_k = v_k. \tag{84a}$$

Secondly from (82h)

$$\hat{x}_k^k = \hat{x}_k^{k-1} - K_k(C\hat{x}_k^{k-1} - v_k). \tag{84b}$$

But $-v_k$ does not differ from $v_k$ due to gaussian distribution assumed. Thus

$$\hat{x}_k^k = \hat{x}_k^{k-1} - K_k Y_k. \tag{84c}$$

However $Y_k$ of (84c)

$$Y_k \stackrel{\text{df}}{=} C\hat{x}_k^{k-1} + v_k \tag{84d}$$

can be recognized as the sample of the input waveform at estimated zerocrossing instant. Moreover from (82b)

$$\hat{x}_{k+1}^k = A(x_k^{k-1} - K_k Y_k) + B_0 \tag{84e}$$

is the recurrence for the subsequent estimated zerocrossing instants while $B_0 = T$. Next by definition of information matrix

$$I_k^k \stackrel{\text{df}}{=} P_k^{k-1} \tag{85a}$$

one obtains from (82f–g)

$$K_k = I_k^{k-1} C^T R^{-1} \tag{85b}$$

$$I_k^k = A^{T^{-1}} I_{k-1}^{k-1} A^{-1} + C^T R^{-1} C \tag{85c}$$

$R_k = R$ while stationary gaussian white sequence is assumed.
Let

1). The initial information $I_0^0 = 0$

2). The state equation is noise free

Then the expression for information matrix $I_k^k$ is from (85c)

$$I_k^k = \sum_{i=1}^{k} (A^{T^{-1}})^{i-1} C^T R^{-1} C (A^{-1})^{i-1} + (A^{T^{-1}})^k I_0^0 (A^{-1})^k =$$

$$= \sum_{i=1}^{k} \begin{bmatrix} 1 & 0 \\ -(i-1) & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & -(i-1) \\ 0 & 1 \end{bmatrix} \omega'^2/R =$$

$$= \sum_{i=1}^{k} \begin{bmatrix} 1 & -(i-1) \\ -(i-1) & (i-1)^2 \end{bmatrix} \omega'^2/R = \begin{bmatrix} k & \dfrac{-k(k-1)}{2} \\ \dfrac{-k(k-1)}{2} & \dfrac{(k-1)k(2k-1)}{6} \end{bmatrix} \omega'^2/R. \tag{86a}$$

and

$$I_k^{k-1} = P_k^k = \begin{bmatrix} \dfrac{2(2k-1)}{k(k+1)} & \dfrac{6}{k(k+1)} \\ \dfrac{6}{k(k+1)} & \dfrac{12}{(k-1)k(k+1)} \end{bmatrix} R/\omega'^2 \tag{86b}$$

$$\omega' = \begin{matrix} A_m \omega_0 \left(1 + \dfrac{\Omega_0}{\omega_0}\right) \\ \text{or} \\ \omega_0 \left(1 + \dfrac{\Omega_0}{\omega_0}\right) \end{matrix} . \tag{86c}$$

Next

$$K_k = I_k^{k-1} C^T R^{-1} = \begin{bmatrix} \frac{2(2k-1)}{k(k+1)} & \frac{6}{k(k+1)} \\ \frac{6}{k(k+1)} & \frac{12}{(k-1)k(k+1)} \end{bmatrix} \frac{R}{\omega'^2} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \frac{\omega'}{R} =$$

$$= \begin{bmatrix} \frac{2(2k-1)}{k(k+1)} \\ \frac{6}{k(k+1)} \end{bmatrix} \frac{1}{\omega'} = \begin{bmatrix} k_1(k) \\ k_2(k) \end{bmatrix} \frac{1}{\omega'}, \quad k = 1, \ldots \tag{87a}$$

Now from (84e)

$$\begin{bmatrix} \hat{x}_{k+1}^{1k} \\ \hat{x}_{k+1}^{2k} \end{bmatrix} = \begin{bmatrix} \hat{x}_k^{1k-1} + \hat{x}_k^{2k-1} - (k_1(k) + k_2(k)) Y_k + T \\ \hat{x}_k^{2k-1} - k_2(k) Y_k \end{bmatrix}. \tag{87b}$$

The following expressions hold true in view of (87b)

$$\hat{t}_{k+1}^k = \hat{t}_k^{k-1} - \left( \sum_{i=1}^{k} k_2(i) + k_1(k) \right) Y_k + T \tag{87c}$$

$$\hat{x}_{k+1}^{2k} = - \sum_{i=1}^{k} k_2(i) Y_i; \quad k = 0, 1, \ldots \tag{87d}$$

The first equation refers to sampling instants while the second one to quantity proportional to frequency offset. Digital PLL which implements obtained algorithm is shown in fig. 3.

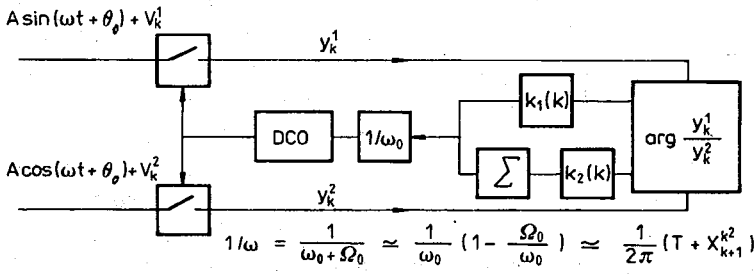



Fig 3. The nonuniform sampling DPLL with the filter of second order and certain loss of optimality due to the operation of argument

Because of instabilities the gain $K_k$ of (87a) cannot diminish to zero. Its final value can be design according to steady state criterion. If the jitter of phase and instabilities of frequency are known in terms of additional noise then the gain $K_k$ $k=1,\ldots$ can be calculated in advance within presented approach either. Subsequent coefficients should be stored in additional memory of the loop's filter. However the final gain should meet

the mentioned steady state criterion according to which the mean squared value of oscillation is established in steady state. Nevertheless on the first stage of control the calculated gain is dominant. Now let the loop of first order is considered. In this case

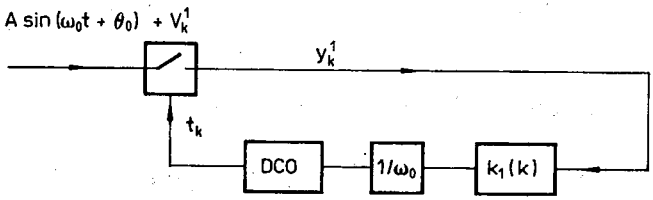
$$x_k = x_{k-1} + T \tag{88a}$$

$$x_{k+1}^k = x_k^{k-1} - K_k Y_k + T; \quad t_{k+1} = x_{k+1}^k \tag{88b}$$

$$I_k^k = \frac{\omega_0}{R} k; \quad Q_{k-1} = 0; \quad I_0^0 = 0, \quad k = 1, \ldots \tag{88c}$$

$$K_k = \frac{1}{\omega_0} k, \quad k = 1, \ldots \tag{88d}$$

The digital circuitry of the loop of first order is shown in fig. 4. The constraint on loop's gain from below should be posed in this case also.




Rys. 4. First-order DPLL with variable coefficient

## CONCLUSIONS

The approach to design of optimal nonuniform sampling digital phase-locked loops DPLL-s has been presented in systematic way. The optimal controlling algorithm on the circle has been considered first in the case of phase and phase and frequency offsets. The evolving densities have been given in terms of their coefficients of Fourier expansion. Similar case was considered by Willsky [1] while the problem of estimation was of concern only. Because the practical algorithm should take into good account the truncated expansion Cesaro – Féjer method of summation of Fourier series is of interest in order to avoid any spurious behaviour resulting from truncation. The limiting cases while the impact of noise is negligible have been considered either. Each of two proposed linear approaches is almost optimal while the variance of phase error process is small enough indicating the performance of DPLL in the linear region around the state of phase entrainment. Otherwise these linear algorithms are suboptimal only. As a matter of fact the presented paper can give rise to different controlling algorithms of practical importance in aim to improve the performance of nonuniform sampling DPLL-s.

Moreover the results presented in [4] can be used to establish convergence to the state of phase entrainment if necessary. The criterion according to which the mean squared value of oscillations is established in steady state may be chosen according to limit properties of Kalman's filter.

REFERENCES

1. Willsky A. S.: *Fourier Series and estimation on the circle with application to Synchronous Communication*, Part – I and Part – II, IEEE Trans. Information Theory, vol. IT – 20, no 5, Sept. 1974, pp. 577–590
2. Jaźwiński A. H.: *Stochastic processes and filtering theory*. Academic Press, New York and London, 1970
3. Lindsey W. C., Chie C. M.: *A survey of digital phase-locked loops*. Proceedings of the IEEE, vol. 69, no 4, April 1981
4. Żółtowski M.: *New stability criteria for the nonuniform sampling digital phase-locked loops*. International Symposium on Circuits and System, ISCAS, Sheraton Hotel, New Orleans, LA, May 1–3, 1990
5. Maurin K.: *Analiza*, cz. I, Elementy, Warszawa 1974

## APPENDIX 1

### THE PROOF OF THEOREM 1

$$\underset{x|Y}{E}\, L(x_t-\hat{x}_t|Y_\tau)=\int_{-\infty}^{\infty} L\,(\alpha+\hat{x}_t^\tau)\; p(\alpha,\, t|Y_\tau)d\alpha=$$

$$=\int_{-\infty}^{\infty} L(\alpha-\hat{x}_t^\tau+\hat{x}_t)\; p(\alpha,\, t|Y_\tau)d\alpha, \qquad (A1-1)$$

$$\alpha=x_t-\hat{x}_t^\tau,\; p(\alpha,\, t|Y_\tau)=p(-\alpha,\, t|Y_\tau),\; L(x)=L(-x).$$

while L is convex

$$\frac{1}{2}\,L(\alpha+\hat{x}_t^\tau-\hat{x}_t)+\frac{1}{2}\,L(\alpha-\hat{x}_t^\tau+\hat{x}_t)\geqslant L(\alpha) \qquad (A1-2)$$

and the theorem follows from;

$$\underset{x|Y}{E}\, L(x_t-\hat{x}_t|Y_\tau\geqslant \underset{x|Y}{E}\, L(x_t-\hat{x}_t^\tau|Y_\tau). \qquad (A1-3)$$

Suppose $\vee_{nxn}$ positively definite and

$$\underset{x|Y}{E}\, L(x_t-\hat{x}_t|Y_\tau)=\int_{-\infty}^{\infty}(x_t-\hat{x}_t)^T\vee(x_t-\hat{x}_t)\; p(x,\, t|Y_\tau)dx \qquad (A1-4)$$

Then

$$\nabla_{\hat{x}_t}\,\underset{x|Y}{E}\,(x_t-\hat{x}_t)^T\vee(x_t-\hat{x}_t)=-2\underset{x|Y}{E}\vee(x_t-\hat{x}_t) \qquad (A1-5)$$

and the minimum is gained by $\hat{x}_t=\underset{x|Y_\tau}{E}\, x_t=\hat{x}_t^\tau$.

## APPENDIX 2

The derivation of the equations of optimal nonlinear filtration. By the rule of Bayes

$$\begin{aligned} p(x, t_k | Y_k) &= p(x, t_k, y_k, Y_{k-1}) / p(y_k, Y_{k-1}) = \\ &= p(y_k | x, t_k, Y_{k-1})\, p(x, t_k, Y_{k-1}) / p(y_k, Y_{k-1}) = \\ &= p(y_k | x, t_k, Y_{k-1})\, p(x, t_k | Y_{k-1}) / p(y_k | Y_{k-1}) \end{aligned} \tag{A2-1}$$

But

$$p(y_k | x, t_k, Y_{k-1}) = p(y_k | x, t_k) \tag{A2-2}$$

in view of Markov property.
Thus

$$p(x, t_k | Y_k) = p(y_k | x, t_k)\, p(x, t_k | Y_{k-1}) / p(y_k | Y_{k-1}). \tag{A2-3}$$

Also

$$p(y_k | Y_{k-1}) = \int p(y_k | x, t_k)\, p(x, t_k | Y_{k-1})\, \mathrm{d}x \tag{A2-4}$$

and

$$p(x, t_k | Y_{k-1}) = \int p(x_k | x, t_{k-1})\, p(x, t_{k-1} | Y_{k-1})\, \mathrm{d}x. \tag{A2-5}$$

Finally

$$p(x, t_k | Y_k) = \frac{p(y_k | x_k) \int p(x_k | x_{k-1})\; p(x_{k-1} | Y_{k-1})\, \mathrm{d}x_{k-1}}{\iint p(y_k | x_k)\; p(x_k | x_{k-1})\; p(x_{k-1} | Y_{k-1})\, \mathrm{d}x_{k-1}\, \mathrm{d}x_k}. \tag{A2-6}$$

## APPENDIX 3

### The mathematical model of the nonuniform sampling digital phase locked loop DPLL while quantization is negligible

The tracked waveform to the loop's input is

$$X(t) = A_m \sin(\omega_0 t + \theta(t)) + v^1(t), \tag{A3-1}$$

$A_m$ is amplitude, $\theta(t)$ stands for the phase and $v^1(t)$ for disturbing noise. The sampling rate of DPLL is

$$t_{k+1} = t_k + T - c_k : k = 0, 1, \ldots \tag{A3-2}$$

$T$ is the nominal sampling period of the DCO of the loop while $c_k : k = 0,1,\ldots$ is the sequence of control. At sampling instants

$$X(t_k) = X_k = A_m \sin(\omega_0 t_k + \theta(t_k)) + v^1(t_k) = \\ = A_m \sin(\omega_0 t_k + \theta_k) + v_k^1. \tag{A3-3}$$

However in view of (A3–2)

$$t_k = t_0 + (k-1)T - \sum_{i=0}^{k-1} c_i. \tag{A3-4}$$

Thus

$$X_k \underset{t_0=0}{=} A_m \sin\left(\theta_k - \omega_0 \sum_{i=0}^{k-1} c_i\right) \tag{A3-5}$$

After defining the process of phase error

$$x_k \stackrel{\text{df}}{=} \begin{cases} \theta_0 & k = 0 \\ \theta_k - \omega_0 \sum_{i=0}^{k-1} c_i & k > 0 \end{cases} \tag{A3-6}$$

one gets the equation of DPLL

$$x_{k+1} = x_k + \theta_{k+1} - \theta_k - \omega_0 c_k \\ k = 0, 1, \ldots \tag{A3-7}$$

If $\theta(t)=\theta_0$ (the offset of phase) then

$$x_{k+1} = x_k - \omega_0 c_k + w_k^1 \tag{A3-8}$$

with noise component $w_k^1$, $(k=0,1,\ldots)$ added.
If $\theta(t)=\Omega_0 t+\theta_0$ (the offset of phase and frequency) then

$$x_{k+1} = x_k + \Omega_0 T - \Omega_0 c_k - \omega_0 c_k; \quad T = 2\Pi/\omega_0. \tag{A3-9}$$

By setting

$$x_k = x_k^1 \quad \text{and} \quad \Omega_0 = x_k^2 \tag{A3-10a}$$
$$\tag{A3-10b}$$

one gets

$$x_{k+1}^1 = x_k^1 + x_k^2 T - x_k^2 c_k - \omega_0 c_k + \mathrm{w}_k^1 \tag{A3-11a}$$
$$x_{k+1}^2 = x_k^2 + w_k^2. \tag{A3-11b}$$

The additional noise components $w_k^1$, $w_k^2 : k=0,1,\ldots$ are to model the instabilities of phase and frequency.

## APPENDIX 4

### The expansion of $p(y_k|x_k)$ into Fourier series

The conditional density $p(y_k|x_k)$ can be considered as smooth function of the form

$$p(y_k|x_k) = F_k(\sin x_k, \cos x_k) \tag{A4-1}$$

or

$$p(y_k|x_k) = F_k\left(\frac{e^{jx_k} - e^{-jx_k}}{2j}, \frac{e^{jx_k} + e^{-jx_k}}{2}\right) = \tag{A4-2}$$

$$= F_k\left(\frac{u_1 - u_2}{2j}, \frac{u_1 + u_2}{2}\right) = F_k(u_1, u_2) \tag{A4-3a}$$

and

$$u_1 = e^{jx_k}, \; u_2 = e^{-jx_k}. \tag{A4-3b}$$

According to Taylor series

$$F_k(u_1, u_2) = \sum_{n=0}^{\infty} \frac{d^n F_k(0, 0, u_1, u_2)}{n!} \tag{A4-4}$$

whereas

$$d^n F_k(0, 0, u_1, u_2) = \left(u_1 \frac{\partial}{\partial u_1} + u_2 \frac{\partial}{\partial u_2}\right)^n F_k(0, 0) =$$

$$= \sum_{r=0}^{n} \binom{n}{r} \frac{\partial^n F_k(0,0)}{\partial^r u_1 \, \partial^{n-r} u_2} u_1^r u_2^{n-r} = \sum_{r=0}^{n} c_r^n \cdot u_1^{2r-n}. \tag{A4-5a}$$

$$c_r^n = \binom{n}{r} \frac{\partial^n F_k(0,0)}{\partial^r u_1 \, \partial_{u_2}^{n-r}} \tag{A4-5b}$$

$$d^0 F_k(0, 0, u_1, u_2) = F_k(0, 0), \quad 0! = 1. \tag{A4-5c}$$

Thus

$$F_k(u_1, u_2) = \sum_{n=0}^{\infty} \sum_{r=0}^{n} \frac{C_r^n}{n!} u_1^{2r-n}. \tag{A4-6}$$

While coming back to required Fourier expansion

$$p(y_k|x_k) = \sum_{m=-\infty}^{\infty} b_m(y_k) e^{jmx_k} \tag{A4-7}$$

one obtains

$$b_m(y_k) = \sum_{n=0}^{\infty} \frac{c_r^n}{n!} \Bigg| \; 2r - n = m. \tag{A4-8}$$

M. ŻÓŁTOWSKI

# OPTYMALNE PĘTLE CYFROWE Z NIEJEDNOSTAJNYM PRÓBKOWANIEM ŚLEDZĄCE PRZEJŚCIA PRZEZ POZIOM ZEROWY PRZEBIEGU ODSTROJONEGO W FAZIE LUB W FAZIE I CZĘSTOTLIWOŚCI

Streszczenie

W pracy przedstawiono sposób projektowania cyfrowych pętli fazowych z niejednostajnym próbkowaniem poprzez optymalny algorytm sterowania na okręgu. Rozważono odstrojenia w fazie oraz w fazie i częstotliwości przebiegu wejściowego. Podobny przypadek był rozważany poprzednio przez Willsky'ego [1]. Niemniej jednak rozważany był problem estymacji. Ewoluujące rozkłady prawdopodobieństwa podano poprzez ich współczynniki Fouriera. Przedstawiono także przypadki graniczne i aproksymacje liniowe. Praca daje podstawy dla różnych algorytmów sterowania, których celem jest ulepszenie działania cyfrowych pętli fazowych z niejednostajnym próbkowaniem.

KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJI, 1992, 38, z. 1, ss. 121-152

# Cyfrowe pętle fazowe — wybrane problemy

MARIUSZ ŻÓŁTOWSKI

*Instytut Telekomunikacji, Politechnika Gdańska*

*Otrzymano 1991.02.10*

*Autoryzowano do druku 1991.03.20*

Podsumowano wyniki prac autora nad cyfrowymi pętlami fazowymi. Układ pętli fazowej jest układem o istotnym znaczeniu we współczesnej telekomunikacji, a jego zastosowanie stanowi również podstawę realizacji wielu technik pomiarowych. Po wstępie wprowadzającym w problematykę i podstawowe pojęcia krok po kroku udostępniane są czytelnikowi wyniki własnych badań, których zakres obejmuje najistotniejszą część problematyki dotyczącą cyfrowych pętli fazowych. Były one prezentowane na licznych seminariach, konferencjach krajowych i zagranicznych. Obok rezultatów podstawowych podano szereg wniosków o znaczeniu praktycznym. Przedyskutowano również obszary stosowalności rozważanych metod i dokonano porównań. W końcu przedstawiono perspektywy dalszej pracy nad cyfrowymi pętlami fazowymi oraz podano główne tezy związane z uzyskanymi wynikami.

## 1. WSTĘP

Pętla fazowa PLL (ang. — phase locked loop) datuje swoje istnienie od 1932 roku, kiedy to ukazała się praca H. de Bellescize'a pod tytułem „La reception synchrone" w czasopiśmie Onde Electrique, vol. 11, June 1932, pp. 230—240. Na tamte czasy była układem egzotycznym o niewielkim znaczeniu praktycznym. Obecnie natomiast znajdując szerokie zastosowanie we współczesnej telekomunikacji pętla fazowa jest nieodzowną częścią większości nowoczesnych urzędzeń telekomunikacyjnych. W telewizorach uklad ten służy do właściwego rozmieszczenia obrazu na ekranie kineskopu — synchronizacji linii, a także w systemie NTSC do właściwego odtworzenia sygnałów chrominancji decydujących o barwie odbieranych obrazów. W odbiornikach radiowych pętla fazowa wchodzi w skład dekodera sygnału stereofonicznego.

Ciągły postęp w zwiększaniu niezawodności, szybkości działania i jakości oraz jednocześnie zmniejszający się koszt i malejąca wielkość układów scalonych VLSI (ang. very large scale of integration) sprawiły, że istotnie wzrosło zainteresowanie zastosowaniem cyfrowych pętli fazowych DPLL (ang. digital phase locked loop), czyli pętli fazowych wykonanych w technice cyfrowej. Cyfrowość pętli pozwala bowiem ominąć szereg problemów dotyczących analogowej realizacji; między innymi czułość na wolnozmienne dryfy i nasycanie się elementów, potrzeby kalibracji i okresowego strojenia a również trudności z budowaniem pętli wyższego rzędu. Możliwość obróbki

sygnałów w czasie rzeczywistym czyni z pętli cyfrowej układ do wielu zastosowań natomiast cyfrowość pozwala na stosowanie różnorakiej adaptacji. Teoria cyfrowych pętli fazowych jest jeszcze w trakcie uzupełniania i z uwagi na działanie układu w czasie dyskretnym różni się od teorii pętli analogowych. Pojawiają się nowe zjawiska, które wynikają z nieliniowości pętli.

Celowość stosowania pętli fazowej, jako układu o wszechstronnym zastosowaniu we współczesnej telekomunikacji; znalazła potwierdzenie w ramach teorii optymalnej filtracji i detekcji sygnałów [R – 14,15].

Zastosowania pętli fazowych dotyczą;

1) Modulacji i demodulacji sygnałów AM, FM, PM lub mieszanych, cyfrowych lub analogowych o różnych specyfikacjach modulacji, w przypadku sygnałów analogowych i cyfrowych o różnych formatach.

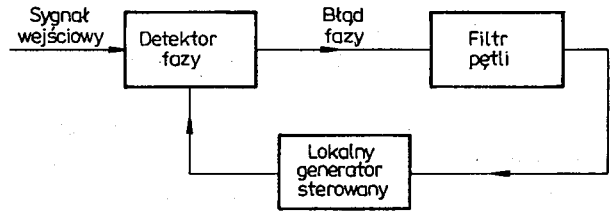
2) Rekonstrukcji sygnałów taktu w celu detekcji sygnałów.

3) Syntezy przebiegów o różnych częstotliwościach.

Do innych zastosowań należą filtry śledzące, filtry synchroniczne, przesuwniki fazy, sterowniki prędkości obrotowej silników, przetworniki analogowo cyfrowe. Układ pętli znajduje zastosowanie w systemach satelitarnych, w systemach sterowania na odległość, w tak zwanej telekomunikacji dalekiej przestrzeni, różnego rodzaju systemach telemetrycznych, pomiarowych i sterowania. Dlatego można uważać, że układ pętli fazowej ma podstawowe znaczenie we współczesnej telekomunikacji i stanowi również podstawę realizacji wielu technik pomiarowych. Zjawiska podobne do zjawisk w pętli fazowej, określane wspólnym mianem synchronizmu występują również w maszynach elektrycznych, układach biologicznych i wielu zdarzeniach przyrodniczych oraz innych dziedzinach techniki [R – 7]. Pętla fazowa jest układem nieliniowym ze sprzężeniem zwrotnym, w którym przebieg z generatora lokalnego, tak zwany przebieg odniesienia, nadąża za zmianami sygnału wejściowego. Proces przejściowy regulacji w pętli nazywamy synchronizacją lub procesem synchronizacji natomiast jego zakończenie określa się mianem stanu synchronizmu. Poszczególne układy pętli cyfrowej wykonane są w technice cyfrowej. Pamiętając o wymienionych zaletach cyfrowej realizacji istotną jakość w porównaniu z pętlą analogową wnosi generator pętli sterowany cyfrowo DCO (z ang. digitally controlled oscillator). Cyfrowa technika realizacji prowadzi w tym przypadku do liniowego i szerokiego zakresu przestrajania w funkcji częstotliwości generowanego przebiegu odniesienia o bardzo dużej stabilności (stabilizacja rezonatorem kwarcowym).

Wadą są ograniczenia technologiczne techniki cyfrowej i związana z nimi mniejsza w porównaniu z pętlą analogową osiągalna szybkość działania cyfrowej pętli fazowej. Tym niemniej w wielu zastosowaniach można stosować układy hybrydowe, w których obróbka szybkozmiennych sygnałów odbywa się w części analogowej natomiast przebiegi zmieniające się wolniej przetwarzane są w sposób cyfrowy. Rozwijana obecnie technologia oparta na arsenku galu może przesunąć obszar zastosowań cyfrowych układów synchronizacji o około dwa rzędy na skali częstotliwości.

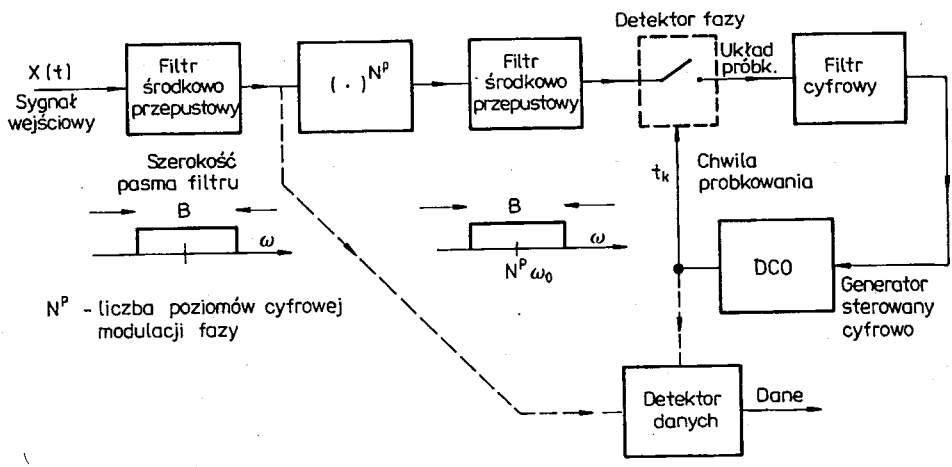
Teoria optymalnej filtracji potwierdzając zasadność struktury funkcjonalnej pętli fazowej (rys. 1) stwarza również podstawy do jej algorytmicznej realizacji w oparciu o procesor do przetwarzania sygnałów. Możliwość bezpośredniego programowego




Rys. 1. Schemat funkcjonalny pętli fazowej

sterowania układem synchronizacji jest nową cechą w porównaniu z pętlą analogową. Projektowanie cyfrowych pętli fazowych podobnie jak innych układów wymaga specyfikacji ich parametrów. Na tym etapie pomocna jest analiza teoretyczna, w której korzysta się z pojęcia modelu układu. W oparciu o analizę modelu można uzyskać wyrażenia określające parametry i zależności będące przedmiotem zainteresowania zamiast lub w uzupełnieniu do badań doświadczalnych. Modelem nazywamy matematyczny opis układu przeznaczony do przewidywania jego własności, określenia parametrów jego działania, wyjaśnienia obserwowanych w nim zjawisk.

Złożoność stosowanego modelu uwarunkowana jest jego przeznaczeniem. Odwołując się do fizyki wiemy, że nie potrzeba stosować elektrodynamiki kwantowej w sytuacjach, gdy wystarczy opis klasyczny, tak jak najistotniejszych cech ruchu najbliższych planet nie potrzeba wyjaśniać w oparciu o ogólną teorię grawitacji. Ponadto w porównaniu z przyrodnikiem inżynier ma możliwość konstruowania układów w zgodzie z przyjętym przez siebie modelem jeśli układem ogólnie nazwać źródło interesujących nas zjawisk. Streszczane wyniki uzyskano drogą analizy w ramach teorii układów dynamicznych, w tym teorii procesów Markowa. Zastosowano również modelowanie komputerowe.




Rys. 2. Schemat ideowy układu synchronizacji z operacją potęgowania w celu wydobycia informacji dla cyfrowej pętli fazowej [R—3,4]

## 2. TECHNIKI WYDOBYWANIA INFORMACJI DLA CELÓW SYNCHRONIZACJI

Informację potrzebną do synchronizacji otrzymuje się przez:

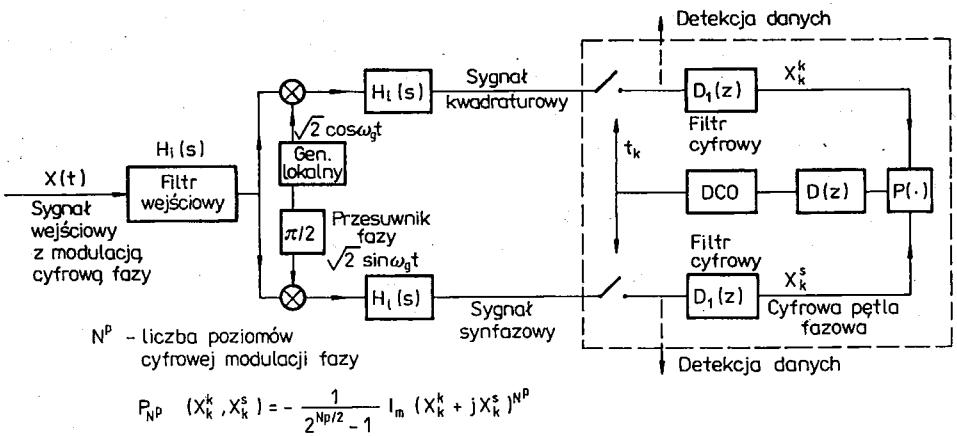
1) Odpowiednie zaprojektowanie filtru pętli w celu uzyskania wymaganych własności częstotliwościowych pętli w przypadku demodulacji sygnałów analogowych.

2) Odpowiednią konstrukcję detektora fazy (np. detektor w fazie i półfazie — ang. mid-in-phase detektor) w przypadku detekcji sygnałów cyfrowych.

3) Zastosowanie kodu synchronizującego (np. kodu Manchester).

4) Zastosowanie operacji potęgowania w celu usunięcia sygnału modulującego z sygnału w przypadku demodulacji i detekcji sygnałów cyfrowych lub innej operacji nieliniowej w celu wydzielania sygnału rytmu przesyłania danych.

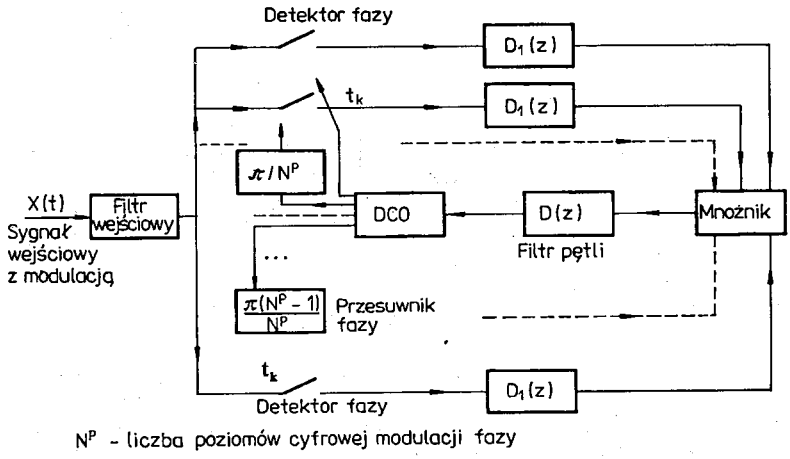
5) Zastosowanie pętli typu Costasa, lub S—K w celu osiągnięcia efektu podobnego do potęgowania [R—1].




Rys. 3. Wydobywanie informacji do synchronizacji w pętli S—K [R—1]

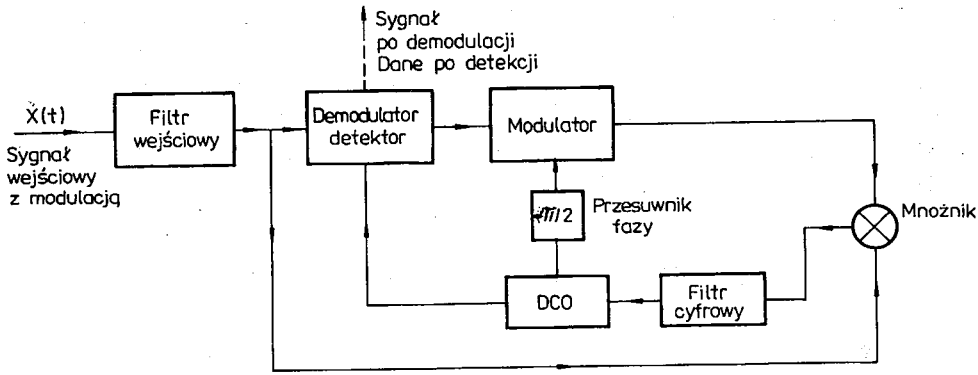
6) Zastosowanie pętli z remodulacją, w której stosuje się modulator w pętli sprzężenia zwrotnego w celu usunięcia wpływu modulacji na pomiar błędu nadążania przebiegu odniesienia z generatora lokalnego pętli za przebiegiem nośnym sygnału wejściowego.

Aczkolwiek każda z wymienionych technik może być przedmiotem osobnych dociekań i różnych porównań to efekt ich zastosowania pozwala na przyjęcie dla celów analizy uproszczonego modelu pętli, w którym przebieg wejściowy pozbawiony jest modulacji. Nie oznacza to oczywiście, że nie uzasadnione jest badanie konkretnych szczegółów realizacji układów synchronizacji. Na przykład przedmiotem zainteresowania autorów pracy [R—23] jest między innymi wpływ interferencji międzysymbolowych w pętlach wykorzystujących różne techniki wydobywania informacji dla celów synchronizacji. Tymniemniej niepożądany wpływ interferencji międzysym-

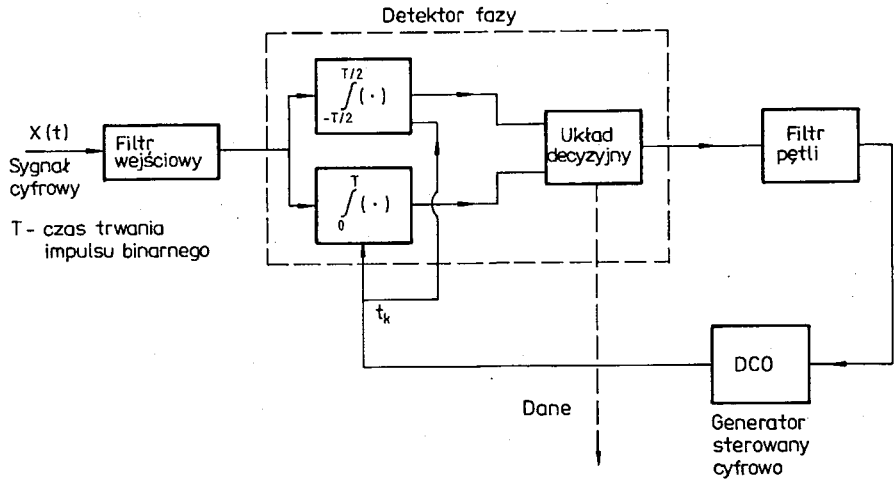



$N^P$ - liczba poziomów cyfrowej modulacji fazy

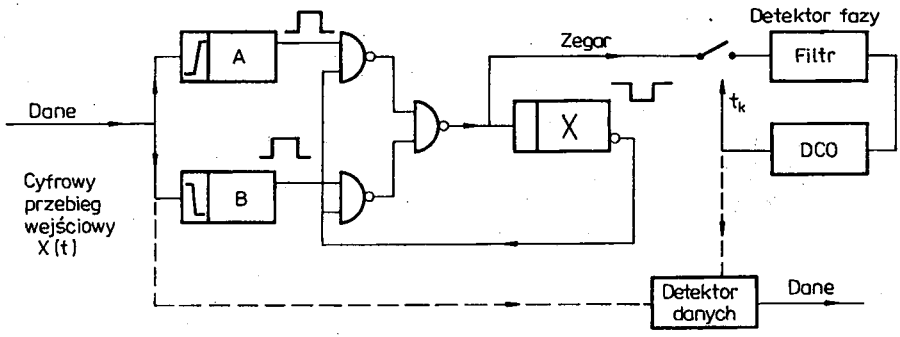
Rys. 4. Wydobywanie informacji do synchronizacji w pętli typu Costasa [R – 23]




Rys. 5. Wydobywanie informacji do synchronizacji w pętli z remodulacją [R – 23]




Rys. 6. Uzyskiwanie informacji do synchronizacji w układzie z detektorem fazy z całkowaniem w fazie i półfazie [R – 28]. Całkowanie może być wykonane w technice cyfrowej




Rys. 7. Wydobywanie informacji do synchronizacji w układzie z informacją cyfrową zakodowaną w kodzie Manchester. Zamiast przerzutników monostabilnych można użyć bardziej czuły detektor do wykrywania zboczy przebiegu wejściowego

bolowych, których nie można uniknąć powinien być znikomy jeśli pętla ma działać prawidłowo.
W celu zmniejszenia wpływu tych interferencji stosuje się adaptacyjne wyrównanie charakterystyki częstotliwościowej kanału telekomunikacyjnego. Istnieją również algorytmy realizujące łącznie wyrównywanie i synchronizację strumienia danych [R–27]. Przykłady metod wydobywania informacji przedstawiono na rys. 2–7.

## 3. PODZIAŁ CYFROWYCH PĘTLI FAZOWYCH

Podstawą do określenia parametrów ucyfrowienia sygnału wejściowego pętli są twierdzenia o próbkowaniu [R–25,26].

### TWIERDZENIE O PRÓBKOWANIU PRZEBIEGÓW DOLNOPASMOWYCH

Przebieg dolnopasmowy o rozpiętości częstotliwościowej $B$ może być odtworzony*

1) Na podstawie próbek pobieranych jednostajnie z częstotliwością próbkowania $f_s > 2B$

2) Na podstawie próbek pobieranych niejednostajnie w kolejnych jednakowych przedziałach czasu o czasie trwania $t_s = 1/f_s$ i częstotliwość próbkowania $f_s$ określona jest jak w p. 1.

### TWIERDZENIE O PRÓBKOWANIU SYGNAŁÓW ŚRODKOWO-PASMOWYCH

Przebieg środkowopasmowy o rozpiętości częstotliwościowej $B$ i skupiony wokół częstotliwości środkowej $f_0$ może być odtworzony*

* wierność odtworzenia zależy od zastosowanej interpolacji

1) Na podstawie próbek pobieranych jednostajnie z częstotliwością próbkowania $f_s$ spełniającą warunki

a) $f_s > 2B$

b) $\frac{2}{l+1}\left(f_0 + \frac{B}{2}\right) < f_s < \frac{2}{l}\left(f_0 - \frac{B}{2}\right)$

i $l$ jest liczbą całkowitą dodatnią.

2) Na podstawie próbek pobieranych niejednostajnie w jednakowych, kolejnych przedziałach czasu o czasie trwania $t_s = 1/f_s$ i częstotliwość próbkowania jest określona tak jak w punkcie 1).

Te znane twierdzenia dotyczące sygnałów dolnopasmowych i środkowopasmowych pozwalają określić częstotliwość próbkowania na podstawie rozpiętości częstotliwościowej próbkowanych sygnałów. W przypadku dużego oddziaływania zakłóceń (szumu) dobierając częstotliwość próbkowania należy uwzględnić zdolność śledzenia przez pętlę uskoków częstotliwości wynikających z nagłych odstrojeń przebiegu wejściowego od przebiegu odniesienia z generatora lokalnego pętli.

Poprzez zwiększenie częstotliwości próbkowania podnosi się stosunek mocy sygnału do mocy szumu. Ogólnie pętle cyfrowe można podzielić na dwie klasy; pętle z jednostajnym i niejednostajnym próbkowaniem. W pętlach z niejednostajnym próbkowaniem rolę detektora fazy pełni układ próbkujący. Proces synchronizacji w takiej pętli polega na regulacji odstępu czasu pomiędzy pobieraniem kolejnych próbek przebiegu synchronizowanego w taki sposób aby uzyskać zgodność chwil próbkowania z chwilami przejść przez poziom zerowy przebiegu synchronizowanego. Dlatego próbkowanie w takiej pętli jest niejednostajne co wyjaśnia jej nazwę.

Drugą klasę stanowią pętle z jednostajnym próbkowaniem. W tych pętlach próbkowanie odbywa się ze stałą częstotliwością. Natomiast poszczególne układy funkcjonalne pętli są wykonaną w technice cyfrowej cyfrową wersją ich analogowych odpowiedników. Często rolę detektora fazy spełnia cyfrowy układ mnożący. Obydwie klasy pętli można wstępnie scharakteryzować w sposób następujący. Pętla z niejednostajnym próbkowaniem jest układem bardziej nieliniowym w porównaniu z pętlą z jednostajnym próbkowaniem. Wzmocnienie takiej pętli zależy od częstotliwości przebiegu synchronizowanego. Dlatego w dłuższym przedziale czasu pętla ta nie nadąża, bez modyfikacji, za liniowo zmiennym uskokiem częstotliwościowym. Natomiast na pętle tą nie oddziaływują wyższe harmoniczne przebiegu wejściowego a oscylator lokalny pętli może być przestrajany w szerokim zakresie częstotliwości. W pętli z jednostajnym próbkowaniem i układem mnożącym jako detektorem fazy ograniczone jest wzmocnienie pętli wynikające z konieczności stłumienia wyższych harmonicznych. Detektor fazy takiej pętli jest oczywiście układem bardziej złożonym niż pętli z niejednostajnym próbkowaniem.

## 4. PRZEBIEG ZJAWISK W CYFROWYCH PĘTLACH — DYNAMIKA CYFROWYCH PĘTLI FAZOWYCH

Dynamikę pętli fazowych bada się głównie w celu określenia warunków istnienia synchronizmu, który powinien charakteryzować się wymaganą stabilnością i odpor-

nością na zakłócenia. Synchronizm $P:Q$ ($P, Q$ — liczby całkowite) to stabilne zjawisko, w przebiegu którego na $P$ okresów przebiegu śledzonego przypada $Q$ okresów z generatora lokalnego pętli. W modelu matematycznym synchronizm reprezentowany jest przez stabilny punkt okresowy mający własność przyciągania trajektorii względnie odpowiedni cykl graniczny, gdy istotne jest kwantowanie i jego efekt należy uwzględnić. W praktyce ma się do czynienia z zaburzonym synchronizmem.
Zaburzeniem jest odstrojenie przebiegu wejściowego w częstotliwości, modulacja, zakłócenie losowe itp. Gdy liczba poziomów kwantowania w układzie synchronizacji (liczba poziomów dotyczy głównie przetwornika analogowo-cyfrowego) jest duża to można kwantowanie pominąć. Ponieważ pętla cyfrowa jest układem, w którym regulacja odbywa się w czasie dyskretnym to badanie dynamiki pętli przy pomocy modelu matematycznego można sprowadzić do badania ciągu odwzorowań. W granicznym przypadku pominięcia efektów kwantowania można się posłużyć gładkim modelem topologicznym. Układ, którego model matematyczny wyznacza ciąg odwzorowań nazywamy układem z czasem dyskretnym. Pomijanie efektów kwantowania jest znanym zabiegiem teoretycznym, stosowanym na przykład przy określaniu charakterystyk czasowych i częstotliwościowych filtrów cyfrowych. Liczba poziomów kwantowania musi być odpowiednio duża, żeby szum kwantowania nie degradował własności projektowanych układów.

W przypadku modelu topologicznego istnienie synchronizmu i jego własności uzależnione są od położenia punktów równowagi, rodzaju ich stabilności, wielkości obszarów przyciągania itp. Kwantowanie można traktować również jako zaburzenie, które wprowadza cykle graniczne wokół punktów równowagi modelu topologicznego. W skrajnym przypadku, gdy kwantowanie jest bardzo zgrubne nie można je traktować jako zaburzenie. Stan synchronizmu w takim przypadku określony jest przez cykl graniczny. Związana z tym cyklem granicznym wielkość oscylacji powinna być odpowiednio mała w zależności od zastosowań.

Charakter wpływu zaburzenia w ramach gładkiego modelu topologicznego wynika z gładkiej ciągłości odwzorowań ze względu na parametr, gdy nie ma zmian bifurkacyjnych. Natomiast stan ustalony określony jest przez własności zbiorów granicznych. Stan synchronizmu można scharakteryzować również poprzez podanie miary na zbiorze punktów granicznych a ściślej puntów niebłądzących. Jest to rozkład punktowy w przypadku pojedynczego punktu równowagi stabilnej. Jeżeli w modelu uwzględnić dodatkowo zakłócenia o charakterze losowym to ich wpływ w stanie ustalonym można scharakteryzować przez podanie miary probabilistycznej, której nośnik zawiera punkt równowagi stabilnej modelu bez zakłóceń, określanym również jako przypadek deterministyczny.

Analogicznie do charakteru zmian argumentu czasowego, gdy mówimy o modelu z czasem dyskretnym i czasem ciągłym można mówić o ciągłej i dyskretnej przestrzeni stanu. W przypadku gładkiego modelu topologicznego mówimy o ciągłej przestrzeni stanu. Gdy nie zaniedbujemy zakłóceń losowych łańcuch Markowa modelujący dynamikę układu jest łańcuchem z ciągłą przestrzenią stanu. Natomiast, gdy uwzględnić kwantowanie przestrzeń stanu jest przestrzenią dyskretną a łańcuch Markowa łańcuchem z dyskretną przestrzenią stanu. Cykliczny charakter procesu

błędu fazy (błędu nadążania za przebiegiem synchronizowanym) stwarza możliwość rozważania modelu układu synchronizacji z bazową przestrzenią stanu wynikającą z operacji nakrycia [R—9]. W praktyce przestrzeń nakrywająca jest przestrzenią $R^n$, podczas gdy przestrzeń bazowa torusem $n$ wymiarowym $T^n$, gdy rząd pętli jest równy $n$. Zwartość przestrzeni bazowej jest jej istotną własnością, gdy wybrać ją na przestrzeń stanu. Spośród bardziej złożonych zjawisk dynamicznych autorowi udało się odkryć w pętli drugiego rzędu zjawisko chaosu przejściowego. Dotyczy ono parametrów pętli cyfrowej o znaczeniu użytecznym. W teorii układów synchronizacji znany jest inny klasyczny przykład skomplikowanej dynamiki spoza obszaru parametrów układu o znaczeniu użytecznym [R—20,21]. Charakteryzuje go kaskada bifurkacji prowadząca do chaosu. W streszczeniu omówione są następujące zagadnienia;

1) Synchronizm N:1 w cyfrowej pętli fazowej I-rzędu z dwupoziomowym kwantowaniem i idealnym układem kwantowania.

2) Synchronizm innych rzędów niż N:1 w pętli I-rzędu z niejednostajnym próbkowaniem i idealnym układem kwantującym.

3) Synchronizm N:1 w pętli z niejednostajnym próbkowaniem II-rzędu. Odkrycie zjawiska chaosu przejściowego w pętli.

4) Nowe kryteria stabilności dla cyfrowych pętli z niejednostajnym próbkowaniem dowolnego rzędu.

5) Stabilność lokalna pętli z jednostajnym próbkowaniem. Nowa metoda analizy.

6) Porównanie pętli z jednostajnym i niejednostajnym próbkowaniem.

7) Pętle z niejednostajnym próbkowaniem w obecności zakłóceń losowych. Wpływ szumu na zjawisko chaosu przejściowego w pętli II-rzędu. Pętla zmodyfikowana i jej własności.

8) Pętla z dwuwartościową korekcją błędu fazy. Zastosowanie teorii weryfikacji hipotez do optymalizacji pętli.

9) Optymalne pętle fazowe z niejednostajnym próbkowaniem do śledzenia przebiegu odstrojonego w fazie i częstotliwości względem przebiegu odniesienia.

Ponadto w streszczeniu porównano metody analizy i scharakteryzowano ich skuteczność. Streszczenie kończą tezy, dyskusja i szkic perspektyw dalszej pracy w dziedzinie cyfrowych układów synchronizacji.

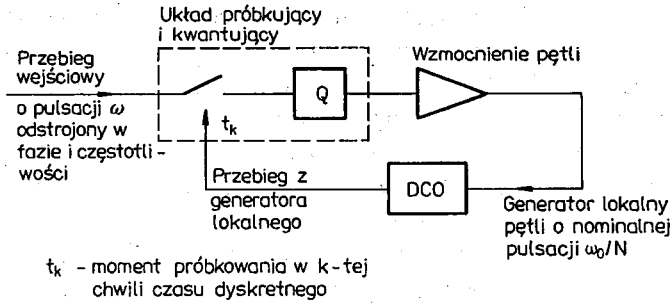
## 5. SYNCHRONIZM N:1 W CYFROWEJ PĘTLI FAZOWEJ I-RZĘDU Z DWUPOZIOMOWYM I IDEALNYM UKŁADEM KWANTUJĄCYM [S—2] (UWAGA W PRACY [S—2] $m=N$)

Pętlę cyfrową I-rzędu z niejednostajnym próbkowaniem, która wykorzystuje zjawisko synchronizmu $N$:1 cechuje przede wszystkim:

— Śledzenie chwil przejść przez poziom zerowy przebiegu wejściowego. W przypadku niezaburzonym na $N$ okresów przebiegu śledzonego przypada jeden okres przebiegu z generatora lokalnego pętli (rys. 8).

— Realizacja detektora fazy przy pomocy układu próbkującego.

— Filtr pętli pierwszego rzędu o działaniu proporcjonalnym.




Rys. 8. Schemat modelu pętli I-rzędu z niejednostajnym próbkowaniem. Odstrojenie względne $\omega_0/\omega$ (N = 1). Wyniki dla N = 1 łatwo wykorzystać, gdy N > 1

Własności takiej pętli zbadano przy pomocy własnej metody graficznej, gdy:

— śledzony przebieg harmoniczny jest odstrojony w fazie i częstotliwości względem przebiegu z generatora lokalnego pętli.

— nie ma zakłócającego szumu

— brane są pod uwagę dwa skrajne przypadki graniczne: idealny układ kwantujący — liczba poziomów kwantowania jest duża i dwupoziomowy układ kwantujący.

W rozważanej sytuacji można przy pomocy metody graficznej uzyskać informacje dotyczącą stabilności pętli odnośnie:

Pętli z idealnym układem kwantującym

Uzyskane informacje dotyczą:

a) sposobu osiągania stanu synchronizmu na podstawie badania warunków istnienia punktów równowagi stabilnej modelu matematycznego pętli w funkcji jego parametrów.

b) warunków braku synchronizmu, gdy punkty równowagi tracą własność przyciągania trajektorii.

c) warunków braku zjawiska poślizgu błędu fazy (ang. cycle slipping phenomenon).

Otrzymane wyniki można zilustrować graficznie w funkcji parametrów wzmocnienia pętli i odstrojenia względnego od częstotliwości nominalnej generatora lokalnego pętli.

Pętli z dwupoziomowym kwantowaniem

Uzyskane informacje dotyczą:

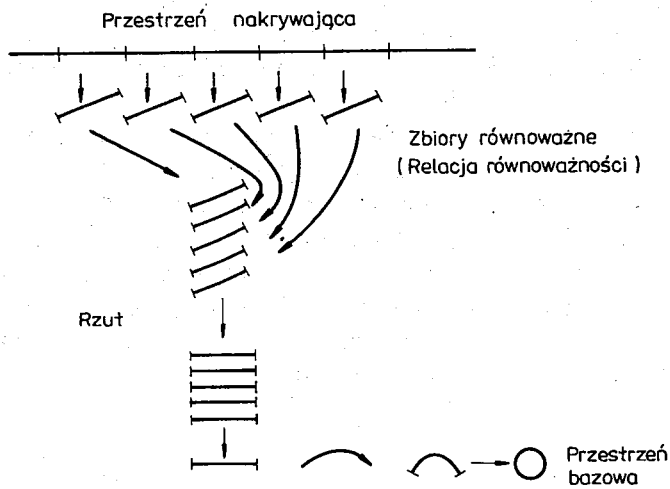
a) warunków braku zjawiska poślizgu błędu fazy w funkcji parametrów pętli; wzmocnienia pętli i odstrojenia względnego od częstotliwości nominalnej generatora pętli.

b) wielkości oscylacji w stanie ustalonym

c) ilustracji otrzymanych wyników

W obydwu rozważanych przypadkach pętli stan synchronizmu na płaszczyźnie wyznaczonej przez parametry, wzmocnienie pętli i odstrojenie względne, charakteryzuje strefa synchronizmu ograniczona od dołu przez znormalizowaną częstotliwość pracy układu, od góry przez warunki braku zjawiska poślizgu błędu fazy oraz

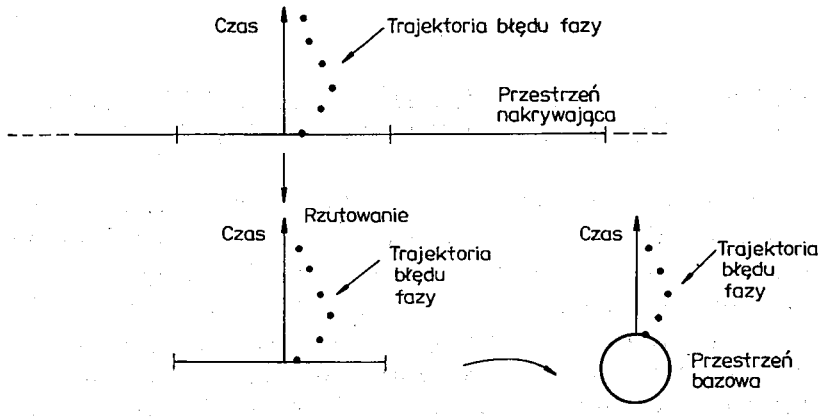
warunki istnienia punktów równowagi stabilnej a z boków przez warunki istnienia punktów równowagi stabilnej lub cyklu granicznego. Ponadto zastosowana metoda pozwala na ocenę czasu trwania synchronizacji. W pracy [S – 2] podano definicję braku zjawiska poślizgu (przeskoku błędu fazy) za Osborne [R – 3]. Definicję tę można uogólnić korzystając z pojęcia przestrzeni bazowej i przestrzeni nakrywającej (rys. 9) [R – 9]




Rys. 9. Ilustracja pojęcia przestrzeni nakrywającej i przestrzeni bazowej

## Definicja

Warunek braku poślizgu błędu fazy jest spełniony jeżeli trajektorie błędu fazy (zbiory kolejnych punktów reprezentujących chwilową wartość błędu fazy) w przestrzeni bazowej i pewnym zbiorze ze zbiorów równoważnych przestrzeni nakrywającej (rys. 10) są jednakowe dla każdego warunku początkowego błędu fazy.

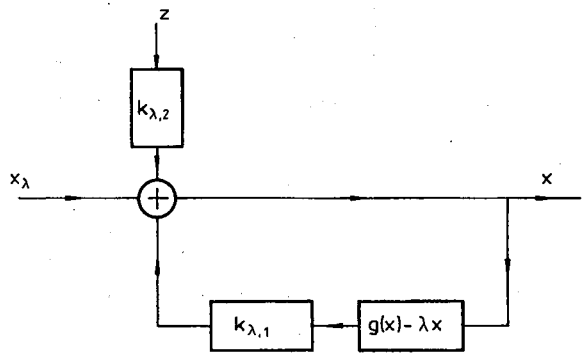



Rys. 10. Ilustracja warunku braku poślizgu błędu fazy

W przypadku pętli wyższego rzędu niż pierwszy poślizg błędu fazy może być nie do uniknięcia dla pewnych obszarów warunków początkowych [R—4].

Szerokość strefy synchronizmu mierzona wielkością przedziału częstotliwości, dla którego istnieje synchronizm niezależnie od warunków początkowych błędu fazy nazywamy zakresem wciągania lub chwytania (ang. pull in range). Często autorzy dodają jeszcze warunek braku poślizgu błędu fazy. Natomiast szerokość strefy wynikającą z warunku istnienia synchronizmu (punktu równowagi stabilnej równania pętli lub przyciągania cyklu granicznego) nazywamy zakresem trzymania (ang. hold in range). Zakres trzymania jest zwykle szerszy niż zakres chwytania i tak jest w przypadku pętli z niejednostajnym próbkowaniem.

Pojęcia zakresu trzymania i wciągania dotyczą modelu pętli bez zakłóceń losowych. Natomiast, gdy szum zakłócający istotnie wpływa na działanie pętli parametry charakteryzujące proces synchronizacji i synchronizmu mają sens statystyczny. Niektóre zjawiska w pętli z czasem dyskretnym nie mają bezpośrednich odpowiedników w pętlach analogowych z czasem ciągłym i ciągłą przestrzenią stanu. Wyniki pracy [S—2] dotyczące stabilności pętli I-rzędu są w części nowe. Dotyczy to na przykład dokładnego wyznaczenia obszaru nie występowania zjawiska poślizgu błędu fazy. Ich ścisłość ma sens w ramach badanych modeli.




Rys. 11. Ilustracja nowego modelu pętli z pracy [S—5]

## 6. SYNCHRONIZM INNYCH RZĘDÓW NIŻ *N*:1 W PĘTLI I-RZĘDU Z NIEJEDNOSTAJNYM PRÓBKOWANIEM

Synchronizm innych rzędów niż *N*:1 w pętli I-rzędu z niejednostajnym próbkowaniem jest przedmiotem rozważań w trzecim rozdziale pracy [S—7]. Omówiono również drogę do chaosu, gdy wzrasta wzmocnienie pętli w przypadku synchronizmu *N*:1. Pokazano diagramy bifurkacyjne obrazujące powstawanie oscylacji o różnych okresach. Zmiany bifurkacyjne pojawiają się w określonym porządku, opisanym w twierdzeniu Szarkowskiego. Zastosowano model topologiczny pomijając kwantowanie w amplitudzie przebiegu wejściowego. W przypadku małego wzmocnienia pętli dla synchronizmów $P:Q$ ($P$, $Q$ liczby całkowite) na płaszczyźnie wyznaczonej

przez parametry; wzmocnienie pętli i odstrojenie istnieją strefy synchronizmu nazwane językami Arnolda, który badał dyfeomorfizmy okręgu [R – 10]. Istnienie tych stref wiąże się z istnieniem stabilnego okresowego punktu równowagi na okręgu, odpowiedzialnego za stan synchronizmu. Stabilność oznacza przyciąganie trajektorii w sąsiedztwie punktu równowagi. W zasadzie w tym zakresie liczba obrotu zdefiniowana poraz pierwszy przez Poincaré w pełni charakteryzuje model pętli, gdy odwzorowanie opisujące dynamikę pętli jest dyfeomorfizmem okręgu. Podano twierdzenia dotyczące tego przypadku.

Gdy synchronizm jest innego rzędu niż $N$:1 to pętla nie jest układem śledzącym za przejściami przez poziom zerowy przebiegu synchronizowanego. Natomiast występują stabilne oscylacje błędu fazy, w których w niezburzonym przypadku na $P$ okresów przebiegu wejściowego przypada $Q$ okresów przebiegu z generatora lokalnego pętli i częstotliwości tych przebiegów są w stosunku $Q : P$. Wyjaśniono efekt nakładania się stref, gdy wzmocnienie pętli staje się duże, istnienia obszarów lokalnej stabilności oraz braku poślizgu błędu fazy. Wyjaśniono symetrię i regularność otrzymanych wykresów strefy synchronizmu.

Gdy model pętli opisuje dyfeomorfizm okręgu i parametry dotyczą strefy synchronizmu to występują wszystkie korzystne zjawiska z punktu widzenia przeznaczenia układu do synchronizacji. Mianowicie ma miejsce stabilność stanu synchronizmu dla dowolnego początkowego błędu fazy, brak poślizgu błędu fazy oraz stabilność strukturalna. Stabilność strukturalna oznacza zachowanie własności dotyczące synchronizmu przy małych zmianach dotyczących całego modelu. Szerokość strefy związana z synchronizmem N:1 jest największa, liniowo zależna od odstrojenia w częstotliwości przebiegu wejściowego. Dlatego ten synchronizm jest najbardziej interesujący z praktycznego punktu widzenia zważywszy dodatkowo większą liniowość – punkt stały zamiast punktu okresowego.

## 7. SYNCHRONIZM $N$:1 W PĘTLI II-RZĘDU Z NIEJEDNOSTAJNYM PRÓBKOWANIEM. ODKRYCIE ZJAWISKA CHAOSU PRZEJŚCIOWEGO [S – 3]

Szczegółowe rozważania dotyczące synchronizmu N:1 w pętli drugiego rzędu z niejednostajnym próbkowaniem przedstawiono w pracy [S – 3]. Zastosowano metody topologiczne. Dynamikę pętli II-rzędu z niejednostajnym próbkowaniem opisuje równanie nieliniowe, różnicowe drugiego rzędu. Wyznaczono model topologiczny z dwuwymiarową przestrzenią stanu. Rozwiązaniem równania jest przebieg błędu fazy charakteryzujący odstrojenie od stanu synchronizmu. Z uwagi na cykliczność błędu fazy jako przestrzeń stanu wykorzystano torus $T^2$. To samo równanie opisuje pętlę impulsową i pętlę cyfrową drugiego rzędu, gdy pominąć kwantowanie w amplitudzie przebiegu wejściowego. Rozważania przeprowadzono z punktu widzenia ogólnej teorii układów dynamicznych.

Na podstawie badania równania zlinearyzowanego można określić warunki istnienia stanu synchronizmu, gdy przebieg synchronizowany jest odstrojony w fazie

i częstotliwości względem przebiegu z generatora lokalnego pętli. W celu zbadania dynamiki na całym torusie dokonano jego podziału na obszary. Następnie zbadano obrazy w odwzorowaniu opisującym dynamikę pętli. Z uwagi na zagięcie przestrzeni torusa, ujawniono w otoczeniu punktu równowagi niestabilnej istnienie zbioru hiperbolicznego o złożonej strukturze topologicznej. Zbiór ten jest podzbiorem zbioru Cantora i można go scharakteryzować przy pomocy tzw. rozbicia Markowa i dynamiki symbolowej. Zbiór ten nie ma własności przyciągania trajektorii i nie jest atraktorem. Tym niemniej w sąsiedztwie tego zbioru trajektorie mają cechy trajektorii przebiegających w sposób przypadkowy. Dlatego zjawisko związane z tym modelem można nazwać chaosem przejściowym. Odkrycie tego zjawiska wskazuje na możliwość zawieszenia synchronizmu (ang. hang up phenomenon) o chaotycznej naturze i przedłużeniu procesu przejściowego pomimo, że lokalna szybkość osiągania stanu synchronizmu może być największa. Stwierdzono również możliwość istnienia synchronizmu innego rzędu, związanego z istnieniem trajektorii okresowej o okresie 4. Wykazano również przydatność obrazu rozmaitości stabilnych i niestabilnych punktów równowagi do scharakteryzowania obszarów przyciągania do punktu równowagi stabilnej reprezentującego stan synchronizmu w pętli. Całość rozważań osadzono w ramach ogólnej teorii układów dynamicznych podając liczne twierdzenia i definicje [S—3]. Okazjonalnie zjawisko chaosu przejściowego w pętli II-rzędu zostało potwierdzone w pracy [R—11]. Ustosunkowanie się do wyników w niej zawartych przedstawiono w pracy [S—5]. Ponadto wskazano na rozwiązania układowe, w których unika się tego zjawiska. Natomiast wpływ szumu na zjawisko chaosu przejściowego scharakteryzowano w pracy [S—9]. Powstanie podkowy charakterystycznej dla zjawisk chaotycznych w przedstawionym ujęciu wiąże się z zagięciem przestrzeni jaką jest torus a nie jak to się zwykle przyjmuje ze zginającymi własnościami samego odwzorowania.

## 8. NOWE KRYTERIA STABILNOŚCI DLA PĘTLI Z NIEJEDNOSTAJNYM PRÓBKOWANIEM

Autorowi tej pracy udało się również odkryć nowe kryteria stabilności dla pętli z niejednostajnym próbkowaniem [S—5]. Metoda, którą zastosowano przypomina dotyczącą twierdzenia Popowa dla układów z czasem ciągłym a jeszcze bardziej jego odpowiednika z czasem dyskretnym podanego przez Cypkina. Różnica wynika z rozważanej przestrzeni stanu. Zamiast przestrzeni euklidesowej rozważania dotyczą przestrzeni bazowej $S^1$ — okręgu. Do opisu procesu błędu fazy użyto funkcjonalne, nieliniowe równanie z czasem dyskretnym. Z uwagi na fakt, że pętla posiada jedno wejście i jedno wyjście równanie to jest równaniem skalarnym i może być rozważane na rozmaitości wymiaru 1, niezależnie od rzędu pętli. Równanie takie można otrzymać stosując technikę szeregów Laurenta znaną również jako przekształcenie $\mathscr{Z}$. Motywem poszukiwań nowych kryteriów stabilności była chęć określenia klasy pętli z niejednostajnym próbkowaniem, które osiągałyby zaburzony w częstotliwości stan synchronizmu $N$:1 możliwie szybko i niezależnie od początkowego błędu fazy.

Natomiast pierwszym impulsem w tym kierunku było odkrycie zjawiska chaosu przejściowego w pętli II-rzędu z sinusoidalną nieliniowością oraz współistnienie synchronizmu innego niż $N$:1 dla pewnego obszaru jedynie innych warunków początkowych błędu fazy. Dalsze szczegóły dotyczące tego zjawiska przedstawiono w poprzednim rozdziale tego streszczenia. Praca dotycząca nowych kryteriów stabilności składa się z trzech części. W pierwszej rozważono zaburzony synchronizm $N$:1, gdy zaburzenie dotyczy odstrojenia w częstotliwości, dodatkowego szumu zakłócającego i efektów kwantowania w amplitudzie przebiegu wejściowego. Rozważania oparto na pojęciu przestrzeni nakrywającej i przestrzeni bazowej. Równanie funkcjonalne pętli opisujące jej dynamikę rozważane jest na okręgu $S^1$ poprzez rzutowanie trajektorii punkt po punkcie.

Zgodnie z nowym modelem pętli z niejednostajnym próbkowaniem przebieg błędu fazy $x$ jest wypadkową znanego przebiegu $x_\lambda$ wyznaczonego na podstawie modelu zlinearyzowanego, przebiegu z pętli sprzężenia zwrotnego oraz przebiegu zakłócającego $z$. Funkcje z czasem dyskretnym $k_{\lambda,2}$ i $k_{\lambda,1}$ mają sens odpowiedzi impulsowych układów liniowych z czasem dyskretnym. Wyrażenie $g(x) - \lambda x$ ma sens odchylenia od modelu liniowego. Natomiast g(x) wyraża nieliniowość pętli (charakterystykę detektora fazy) zaś $\lambda > 0$ jest parametrem. Dzięki zdefiniowaniu funkcjonału $\rho$ o własnościach podobnych do normy można podać oszacowania dotyczące: liniowej części modelu pętli i nieliniowej części modelu pętli. Oszacowania te dotyczą każdej funkcji $\rho$. W ten sposób, gdy przestrzenią stanu jest okrąg $S^1$ można rozważyć różne przestrzenie sygnałów; przestrzeń sygnałów o skończonej pseudoenergii, przestrzeń sygnałów o skończonej amplitudzie w stanie ustalonym, gdy czas odpowiedzi zmierza do nieskończoności, przestrzeń przebiegów o skończonej wartości szczytowej, a w końcu przestrzeń przebiegów o skończonej mocy średniej. Można również w oparciu o wspomniane oszacowanie dotyczące nieliniowości i składowej liniowej modelu rozstrzygać przynależność procesu błędu fazy do wybranej przestrzeni.

Oszacowania dotyczące nieliniowości są ogólne. Wystarczy, gdy konkretna nieliniowość jest w obszarze ograniczonym przez dwie proste względnie przez większą ich ilość. W ten sposób można rozstrzygnąć istnienie stanu synchronizmu, szacować amplitudę oscylacji w stanie ustalonym w przypadku kwantowania w amplitudzie, szacować wartość średniokwadratową oscylacji w stanie ustalonym wynikającą z obecności szumu zakłócającego, wreszcie szacować maksymalne odchylenie od stanu synchronizmu. Można również podać prosty algorytm sterowania synchronizacją, gdy nieliniowość jest sinusoidalna. Należy zastosować kwantowanie dwupoziomowe w pierwszym etapie a potem wielopoziomowe. W ten sposób można uniknąć zjawiska chaotycznego zawieszenia stanu synchronizmu w pętli II-rzędu. Ponadto przy pomocy przedstawionej teorii można badać zakres chwytania i zakres trzymania synchronizmu w przypadku pętli dowolnego rzędu. W opraciu o uzyskane wyniki można określić klasę pętli, które osiągają stan synchronizmu N:1 niezależnie od początkowego błędu fazy. Do tej klasy należą pętle z nieciągłą charakterystyką detektora fazy. W pracy podano schematy takich pętli oraz ich modele matematyczne. Szereg wyników sformułowano również w postaci twierdzeń. Chociaż przeprowadzona analiza ma charakter jakościowy jako analiza najgorszego przypadku; jeżeli tylko nieliniowość

znajduje się w pewnych granicach to pewne własności mają miejsce, to jednak większość oszacowań tej pracy może być użyteczna również na etapie projektowania cyfrowych układów synchronizacji. Natomiast dokładniejsze studia dotyczące najdrobniejszych szczegółów mogą być przeprowadzone w oparciu o modelowanie komputerowe. W pracy przeglądowej Lindsey i Chie [R – 1, pp. 426 – 427] wskazali na niedogodność dotyczącą znanych im *„normowo"* zorientowanych metod w przypadku analizy pętli wyższych rzędów. Ponieważ opisana metoda jest „normowo" zorientowaną metodą, która została zastosowana do pętli wyższego rzędu również, to można stwierdzić, że niedogodność wskazana przez Lindsey'a i Chie została usunięta.

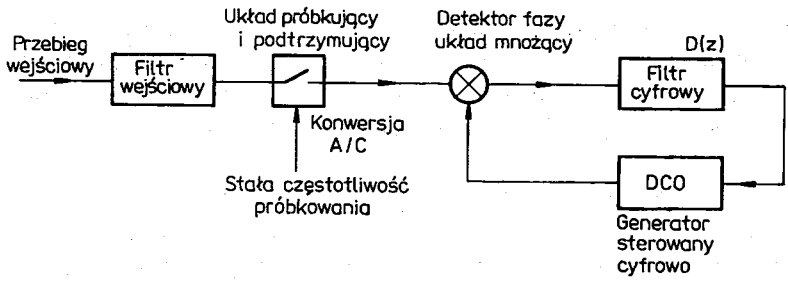
W części drugiej pracy metodę uogólniono na przypadek dodatkowej modulacji AM amplitudy i fazy PM i zaników [S – 6]. Natomiast w części trzeciej [S7] uwzględniono pętlę śledzącą uskok liniowo rosnącej częstotliwości odstrojenia występujący przy znanym efekcie Dopplera lub przeszukiwaniu zakresu częstotliwości oraz pętlę ze zmiennymi współczynnikami. Ponieważ częstotliwość chwilowa wejściowego przebiegu śledzonego wpływa na wzmocnienie pętli należy tym wzmocnieniem sterować w procesie synchronizacji by osiągnąć stabilny stan synchronizmu. Podano warunki wystarczające stabilności uzasadniając je w ramach przyjętego modelu pętli w postaci funkcjonalnego nieliniowego równania na okręgu bardziej złożonego niż w przypadku części I. Po raz pierwszy taki przypadek był badany przez Chie [R – 2].

Pętlę o zmiennych współczynnikach można stosować również w przypadku istotnego wpływu zakłóceń losowych. Cyfrowość układu stwarza możliwość łatwego przestrajania jej parametrów jeśli zastosować do tego celu na przykład mikroprocesor. Zmienność współczynników określa się drogą kompromisu między zapewnieniem szybkiej synchronizacji i jednoczesnej eliminacji wpływu zakłóceń. Najlepszy kompromis można osiągnąć stosując wyniki teorii nieliniowej filtracji. Jednak układ staje się bardzo złożony i problem spełnienia wymagania pracy w czasie rzeczywistym może być istotnym czynnikiem rozstrzygającym o zastosowaniu takiego układu. Pętla o zmiennych współczynnikach w obecności zakłóceń losowych była po raz pierwszy rozważana przez Rocha [R – 18]. Zastosowanie natomiast teorii filtracji do optymalnego przestrajania takiej pętli było przedmiotem zainteresowania autora tego streszczenia [I – 1]. Zastosowanie zmiennych współczynników pętli rzeczywiście przyspiesza proces synchronizacji i jednocześnie stopniowo eliminuje wpływ zakłóceń losowych (szum kanału telekomunikacyjnego) na działanie układu synchronizacji. Modelem matematycznym układu podobnie jak w częściach I i II jest nieliniowe równanie funkcjonalne na okręgu. Podane twierdzenia wynikają z nałożonych ograniczeń na liniową i nieliniową część modelu i z charakteru zakłóceń losowych.

## 9. STABILNOŚĆ LOKALNA PĘTLI Z JEDNOSTAJNYM PRÓBKOWANIEM NOWA METODA ANALIZY

W pracy [S – 4] podjęto problem stabilności lokalnej pętli z jednostajnym próbkowaniem. W przeciwieństwie do pętli z niejednostajnym próbkowaniem wzmoc-

nienie takiej pętli nie zależy od częstotliwości przebiegu wejściowego. Częstotliwość próbkowania pętli jest jednostajna i może być dobrana w oparciu o twierdzenia o próbkowaniu sygnałów dolnopasmowych lub środkowo-pasmowych (rys. 12).




Rys. 12. Model pętli z jednostajnym próbkowaniem

Podobnie, jak w przypadku pętli z niejednostajnym próbkowaniem jako model matematyczny użyto równanie funkcjonalne z czasem dyskretnym. Problem stabilności lokalnej rozważono w oparciu o zlinearyzowaną postać tego równania w sytuacji deterministycznej. Wpływ zakłóceń losowych uwzględniono w pracach [I – 3,4]. Problem stabilności lokalnej uznano za ważki z uwagi na występowanie w równaniu zlinearyzowanym składników oscylacyjnych, których stłumienie jest istotne do zapewnienia stabilności stanu synchronizmu. Badany synchronizm sklasyfikowano nie na podstawie częstotliwości próbkowania lecz na podstawie częstotliwości przebiegu z generatora lokalnego pętli, który podlega dyskretyzacji zgodnie ze stałą częstotliwością próbkowania i częstotliwości przebiegu wejściowego. Rozważono jako najbardziej istotny synchronizm 1:1.

Otrzymane zlinearyzowane równanie pętli można sklasyfikować jako odpowiednik z czasem dyskretnym równania Volterry drugiego rodzaju, o zmiennych w czasie parametrach. W przypadku pętli analogowej taki opis nie jest konieczny z uwagi na znaczne tłumienie wszystkich harmonicznych, gdyż pętla ma własności układu dolnoprzepustowego. Natomiast, gdy układ jest z czasem dyskretnym to jego charakterystyka częstotliwościowa jest okresowa i dlatego problem tłumienia wyższych harmonicznych zdaniem autora wymagał większej uwagi niż w przypadku pętli analogowych o modelu z czasem ciągłym. Zastosowano trzy metody badania otrzymanego równania stwierdzając zgodność otrzymanych przy ich pomocy wyników. Pierwsza metoda polegała na szacowaniu normy rozwiązania w oparciu o szacowanie górnej granicy normy opisującej wielkość oscylacji w stanie ustalonym. Podano warunki dostateczne istnienia i jednoznaczności rozwiązywania, istnienia stanu ustalonego, w oparciu o twierdzenie Banacha o punkcie stałym ciągu odwzorowań. Druga metoda polega na rozwiązaniu obciętego układu równań algebraicznych w celu dokładniejszego oszacowania rozwiązania w stanie ustalonym. Wreszcie trzecia metoda polegała na znalezieniu oszacowania wielkości oscylacji w stanie ustalonym przy pomocy symulacji komputerowej. Kolejno, metody te charakteryzują się rosnącą złożonością i skutecznością przy jednocześnie malejącej poglądowości. Z ich zastosowania wynika możliwość stłumienia oscylacji błędu fazy

o dowolną wielkość jeśli tylko częstotliwość próbkowania dobrana jest tak, że częstotliwość podstawowa oscylacji przebiegów harmonicznych z czasem dyskretnym jest różna od zera. Pozwalają ponadto na oszacowanie wielkości oscylacji.

Wyprowadzone wzory mogą być użyte na etapie projektowania. Tłumienie oscylacji pozwala na uniknięcie zjawiska poślizgu błędu fazy. Tym niemniej odbywa się to kosztem zmniejszenia wzmocnienia pętli. Porównanie obu klas pętli z jednostajnym i niejednostajnym próbkowaniem jest przedmiotem rozważań w następnym rozdziale.

Można zaproponować rozwiązanie pętli z jednostajnym próbkowaniem nie wymagające tłumienia w pętli wyższych harmonicznych. Jest ono w trakcie opracowania. Podobnie jak w przypadku pętli z niejednostajnym próbkowaniem zastosowana analiza jest słuszna dla pętli dowolnego rzędu, ponieważ równanie splotowe jest skalarne niezależnie od rzędu pętli.

## 10. PORÓWNANIE PĘTLI Z JEDNOSTAJNYM I NIEJEDNOSTAJNYM PRÓBKOWANIEM

W oparciu o wyniki opisane w poprzednim rozdziale, które dotyczą stabilności lokalnej pętli z jednostajnym próbkowaniem oraz podane wcześniej własności pętli z niejednostajnym próbkowaniem można dokonać porównania obu klas pętli. Pod względem konstrukcyjnym obie pętle różnią się zastosowanym detektorem fazy mierzącym odstrojenie od stanu synchronizmu. W pętli z niejednostajnym próbkowaniem śledzącej przebieg harmoniczny rolę detektora fazy pełni układ próbkujący, podczas gdy w pętli z jendostajnym próbkowaniem również o sinusoidalnym typie nieliniowości pełni tę funkcję układ mnożący. Jeśli porównać pętle z obu klas o zbliżonych częstotliwościach próbkowania to konieczność stłumienia wyższych harmonicznych w pętli z jednostajnym próbkowaniem ogranicza wzmocnienie tej pętli w porównaniu ze wzmocnieniem pętli z niejednostajnym próbkowaniem. Z opisanych w rozdziale V i VI własności wynika większa szerokość strefy synchronizacji, szerszy zakres chwytania i trzymania stanu synchronizmu pętli z niejednostajnym próbkowaniem oraz możliwość szybszego osiągania stanu synchronizmu. Dotyczy to pętli I-rzędu. Z uwagi na małe wzmocnienie własności pętli z jednostajnym próbkowaniem będą takie jak własności zaburzonego dyfeomorfizmu okręgu. Zalety obu pętli mogą być podobne jeśli duży poziom mocy zakłóceń ogranicza wzmocnienie obu pętli do tego stopnia, że możliwość wykorzystania większego wzmocnienia pętli z niejednostajnym próbkowaniem jest nierealna.

Pętla z niejednostajnym próbkowaniem jest układem bardziej nieliniowym. Wzmocnienie pętli zależy od częstotliwości przebiegu śledzonego przez pętlę i w pewnych zastosowaniach jak na przykład do śledzenia przebiegów wysyłanych przez przyspieszające obiekty należy stosować dodatkowy przebieg sterujący wzmocnienie pętli w celu uzyskania stanu synchronizmu. W przypadku pętli z jednostajnym próbkowaniem wzmocnienie to jest stałe.

Układ próbkujący jako detektor fazy wykorzystuje bezpośrednio kształt przebiegu śledzonego do mierzenia odstrojenia od chwil przejścia przez pozim zerowy próbkowanego przebiegu podlegającego synchronizacji. Dlatego częstotliwość próbkowania nie może być większa niż to wynika z częstotliwości przebiegu śledzonego [I — 8].

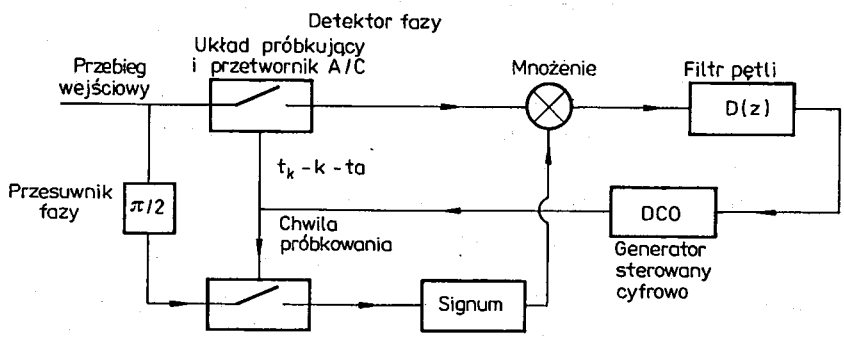
Natomiast w pętli z jednostajnym próbkowaniem przyspieszenie synchronizacji można w niektórych przypadkach osiągnąć przez zwiększenie częstotliwości próbkowania przebiegu podlegającego synchronizacji. Wprawdzie na zwiększenie częstotliwości próbkowania nie ma tego typu ograniczenia co w przypadku pętli z niejednostajnym próbkowaniem to jednak tej częstotliwości nie można zwiększać bez ograniczeń.

Częstotliwość podstawowa składowych niepożądanych, podlegających tłumieniu w pętli zależy odwrotnie proporcjonalnie od częstotliwości próbkowania i nie może być bliska zeru, gdyż wtedy trudno ulegają stłumieniu. Częstotliwość próbkowania w pętlach z jednostajnym próbkowaniem dobiera się w oparciu o twierdzenie o próbkowaniu sygnałów dolno i środkowopasmowych. Dokładniejsza analiza pokazuje, że warunki określające częstotliwość próbkowania wynikające z tych twierdzeń nie mogą być ominięte [S — 4]. Ponadto pętle z jednostajnym próbkowaniem mogą być przedmiotem łatwiejszej optymalizacji w oparciu o wyniki teorii nieliniowej filtracji. Wynika to stąd, że w modelach rozważanych w ramach teorii filtracji próbkowanie jest jednostajne.

## 11. PĘTLA Z NIEJEDNOSTAJNYM PRÓBKOWANIEM W OBECNOŚCI ZAKŁÓCEŃ LOSOWYCH. WPŁYW SZUMU NA ZJAWISKO CHAOSU PRZEJŚCIOWEGO W PĘTLI II-RZĘDU. PĘTLA ZMODYFIKOWANA I JEJ WŁASNOŚCI

Pętla w obecności zakłóceń losowych została poddana analizie w pracy [S — 8,9]. Praca ta składa się z dwóch części. W pierwszej części przedstawiono teorię, podczas gdy w drugiej przykłady. Obecność szumu zakłócającego jest własnością kanałów telekomunikacyjnych. Rozważono najważniejszy przypadek, gdy szum zakłócający jest addytywny. Wpływ szumu multiplikatywnego może być uwzględniony w podobny sposób. Dotychczas omawiane parametry pętli dotyczyły sytuacji deterministycznej. Można je traktować jako wielkości charakteryzujące graniczny przypadek, gdy wpływ szumu maleje. W przypadku niedużego stosunku mocy sygnału do mocy szumu, gdy układ charakteryzuje się pracą określoną w oparciu o jego charakterystyki jako pracę w pobliżu progu, można mówić jedynie o parametrach opisujących działanie pętli w sensie statystycznym. Do takich parametrów zalicza się średni czas do chwili osiągnięcia stanu synchronizmu, wartość średniokwadratowa oscylacji błędu fazy w stanie ustalonym, średni czas do poślizgu błędu fazy lub zerwania synchronizacji itp. Gdy pętlę zaprojektowaną w oparciu o model deterministyczny trzeba zastosować w sytuacji, gdy ma pracować w obecności istotnych zakłóceń losowych to jej parametry

należy poddać modyfikacji. Najczęściej wyraża się to zawężeniem charakterystyk częstotliwościowej charakteryzującej model liniowy pętli w celu zmniejszenia wpływu szumu zakłócającego. Odbywa się to jednak kosztem wydłużenia w czasie procesów przejściowych. Osiągnięcie w tym przypadku najlepszego kompromisu jest istotne z punktu widzenia efektywności działania układu synchronizacji. Dlatego można stosować pętlę z filtrem o zmiennych współczynnikach i inne rozwiązania wynikające z optymalizacji (rys. 13). Pętla o zmiennych parametrach była omawiana w jednym z poprzednich rozdziałów tego streszczenia.




Rys. 13. Schemat funkcjonalny pętli zmodyfikowanej. Mnożenie przez znak przebiegu przesuniętego o $\Pi/2^{[rad]}$ ma na celu ukształtowanie nieliniowości pętli. W efekcie nieliniowość jest przedziałami rosnąca z punktami nieciągłości, zbliżona kształtem do zębów piły [S—9]

W pierwszej części pracy [S—8,9] podano wyrażenia całkowe i macierzowe określające miarę stacjonarną i ewolucję tej miary opisującą proces synchronizacji w pętli, gdy jego modelem jest proces Markowa z ciągłą i dyskretną przestrzenią stanu. Podano twierdzenia zapewniające istnienie miary inwariantnej. Następnie w oparciu o wyniki teorii ergodycznej problem istnienia stanu ustalonego w obecności zakłóceń potraktowano jako zaburzenie miary inwariantnej z twierdzenia Bogolubowa i Kryłowa. Zastosowano operację nakrycia omawianą poprzednio w celu uzwarcenia przestrzeni stanu. Następnie w oparciu o znaną metodę funkcji tworzących lub przekształcenie $\mathscr{Z}$ — Laurenta podano wyrażenia określające pierwszy i drugi moment statystyczny czasu do pochłonięcia trajektorii układu o modelu Markowa przez zbiór pochłaniający.

W teorii układów synchronizacji model pochłaniania stosuje się do opisu procesu osiągania synchronizmu oraz procesu zrywania stanu synchronizmu, poślizgu błędu fazy. W oparciu o wyniki teorii ergodycznej; koncepcję układu spełniającego aksjomat A, warunek topologiczny stabilności w sensie Lapunowa, twierdzenie Poincaré' o powrocie i lemat Kaca przedyskutować można jakościowe zachowanie się procesu osiągania stanu synchronizmu. Jeżeli wpływ szumu na układ o zwartej przestrzeni stanu jest duży to taki układ charakteryzuje jednoznaczna miara probabilistyczna. Nośnikiem tej miary jest zwykle cała przestrzeń. Twierdzenie ergodyczne może być użyte do uzasadnienia stosowalności przedstawionych metod. Modelowanie komputerowe może być stosowane do numerycznego wyznaczania wybranych paramet-

rów. Własność mieszania gwarantuje zbieżność. Jednostajna ergodyczność definiowana istnieniem jednoznacznej miary probabilistycznej dotyczy pracy układu w pobliżu progu na charakterystykach jego działania.

Problem ewolucji miary probabilistycznej staje się interesujący w połączeniu z istnieniem zjawiska chaosu przejściowego w pętli II-rzędu. Ten problem i działanie pętli zmodyfikowanej poruszano w drugiej części pracy [S–9]. Część druga przedstawia zastosowanie teorii do analizy działania pętli z niejednostajnym próbkowaniem. Jak wiadomo w pętli drugiego rzędu może wystąpić zjawisko chaosu przejściowego, którego wpływ na szybkość synchronizacji jest niekorzystny. W pętlach fazowych zwykle stosuje się dodatkowy układ sterowania procesem synchronizacji. Układ ten zaczyna działać, gdy detektor stanu synchronizmu wskazuje brak synchronizmu. Działanie to polega na powolnej zmianie częstotliwości przebiegu z generatora lokalnego pętli w wybranym paśmie częstotliwości. Efekt takiego przestrajania podobny jest do efektu Dopplera, gdy obiekt emitujący przyspiesza lub hamuje. Takie przestrajanie zwiększa niezawodność osiągania stanu synchronizmu. Tym niemniej bardziej korzystne jest aby synchronizacja przebiegała możliwie szybko, bez nadrzędnego sterowania. Dlatego w jednym z poprzednich rozdziałów tej pracy rozważano klasę pętli z nieciągłą charakterystyką detektora fazy. Pętle z tej klasy charakteryzuje możliwość osiągania stanu synchronizmu możliwie szybko, niezależnie od początkowego błędu fazy.

Tym niemniej autora zainteresował wpływ modyfikacji układu na własności dotyczące synchronizmu w przypadku obecności zakłóceń losowych. Życzeniem było, żeby pętla osiągała stan synchronizmu możliwie szybko, w obecności zakłóceń losowych. Szczególnie interesujący z praktycznego punktu widzenia był wpływ modyfikacji na zjawisko poślizgu błędu fazy. W pętli zmodyfikowanej nie powinno wystąpić zjawisko chaotycznego zawieszenia stanu synchronizmu. W pracy pokazano jak na podstawie modelu układu można zbudować równanie całkowe z okręgiem jako przestrzenią stanu, opisujące ewolucję miary probabilistycznej. Pokazano również jak zastosować model Markowa ze stanem pochłaniającym do opisu zjawiska synchronizmu i poślizgu błędu fazy.

W oparciu o równanie całkowe można otrzymać wyrażenie określające średnią liczbę kroków do poślizgu błędu fazy oraz odchylenie standardowe tej liczby dla różnych stosunków mocy sygnału do mocy szumu. Pokazano jak błąd statyczny fazy $\varphi_s$ wpływa na degradację działania pętli zmniejszając liczbę kroków do poślizgu błędu fazy. Właśnie średnia liczba kroków do poślizgu fazy świetnie nadaje się do przedstawienia efektu progowego, lepiej niż wariancja opisująca oscylacje błędu fazy na okręgu w stanie ustalonym. Ten ostatni parametr porównano z przewidywanym w oparciu o teorię liniową. Pokazano również dla różnych stosunków mocy sygnału do mocy szumu miary stacjonarne opisujące stan ustalony układu. W oparciu o wyniki modelowania komputerowego podano średni czas do chwili osiągnięcia stanu synchronizmu oraz odchylenie standardowe tego parametru dla różnych stosunków mocy sygnału do mocy szumu. Poświęcono również uwagę zjawisku chaosu przejściowego w pętli II-rzędu z nieliniowością typu sinusoidalnego. Wnioski z komputerowego badania pętli w obecności szumu są następujące. Średni czas do chwili

osiągnięcia stanu synchronizmu jest prawie jednakowy dla trajektorii punktów z wybranego małego obszaru ze zbioru hiperbolicznego odpowiedzialnego za to zjawisko. Natomiast ujawniona w pracy [S – 3] czułość na niewielką zmienę warunku początkowego znajduje odzwierciedlenie w czułości wariancji tego czasu ze względu na warunek początkowy [S – 9].

Szum kwantowania ma podobny wpływ do wpływu szumu zakłócającego na działanie pętli. O tym wspomniano w pracy [S – 5] ustosunkowując się do pracy [R – 11]. W miarę malenia stosunku mocy sygnału do mocy szumu należy się spodziewać istnienia jednoznacznej miary probabilistycznej na okręgu, skoncentrowanej wokół stanu równowagi stabilnej związanego z istnieniem synchronizmu $N$:1, wypełniającej całą przestrzeń, która staje się jej nośnikiem. Wtedy różnica między działaniem układu zmodyfikowanego i niezmodyfikowanego maleje i dla małych stosunków mocy sygnału do mocy szumu korzyści z modyfikacji układu wydają się być znikome. Przeprowadzone obliczenia potwierdzają te spostrzeżenia. Odpowiednie wykresy i tabela ilustrują przedstawione wnioski. Jeśli chodzi o układ zmodyfikowany to przeprowadzone obliczenia nie wskazują na degradację jego własności pod wpływem modyfikacji a jedynie wskazują na poprawę działania dla większych stosunków mocy sygnału do mocy szumu. W dodatkach do pracy [S – 8,9] podano uzasadnienia stosowanych metod. Pokazano również jak wyznaczyć parametry łańcucha Markowa, gdy uwzględnia się kwantowanie sygnału wejściowego w amplitudzie. W pracy zbadano również wpływ kwantowania na działanie pętli i zamieszczono ilustrujące wykresy. Wykazano, że w miarę malenia stosunku mocy sygnału do mocy szumu stosowanie wielu poziomów kwantowania jest niecelowe a model z pominięciem efektów kwantowania nawet, gdy poziomów kwantowania jest stosunkowo niedużo nie prowadząc do istotnych różnic staje się słuszny.

## 12. PĘTLA Z DWUWARTOŚCIOWĄ KOREKCJĄ BŁĘDU FAZY ZASTOSOWANIE TEORII WERYFIKACJI HIPOTEZ DO OPTYMALIZACJI PĘTLI

W pracy [S – 1] pokazano przykład analizy pętli cyfrowej z dyskretną przestrzenią stanu. Analiza tej klasy pętli może być przeprowadzona w oparciu o model Markowa z przeliczalną przestrzenią stanu. Rozważono przebieg odstrojony w fazie od przebiegu z generatora lokalnego pętli oraz synchronizm $N$:1, przy czym w pracy [S – 1] użyto oznaczenia $L$ dla liczby okresów przebiegu wejściowego przypadającego na pobranie jednej próbki ($L \equiv N$).

Pętla należy do klasy pętli z niejednostajnym próbkowaniem. Z uwagi na dwuwartościową korekcję błędu fazy (dwupoziomowy układ kwantujący w amplitudzie przebieg wejściowy) problem sterowania pobieraniem próbek przez generator lokalny pętli potraktowano jako problem podejmowania decyzji w sensie statystycznym odnośnie znaku błędu fazy charakteryzującego odstrojenie pętli od stanu synchronizmu. W tym celu zastosowano zwykły test ilorazowy i ilorazowy test

sekwencyjny. Układy realizujące wymienione testy są stosunkowo proste. Są to sumatory, liczniki sumujące i liczniki rewersyjne. Filtry takie były stosowane w dotychczasowych rozwiązaniach układowych lecz uzasadnienie ich zastosowania nie było zorientowane pod kątem optymalnego podejmowania decyzji odnośnie procesu synchronizacji w pętli. Rozważane w pracy [S – 1] pętle mają cechy układów pierwszego rzędu w sensie pojęcia astatyzmu układu. Na pobudzenie uskokiem częstotliwości przebiegu wejściowego reagują stałym błędem fazy nadążania, jeśli odstrojenie jest na tyle małe, że nie prowadzi do utraty stabilności. W oparciu o omawiane w pracy filtry sekwencyjne można budować układy o własnościach układów drugiego rzędu. Dwupoziomowe kwantowanie w układzie pętli zapobiega niepożądanym zmianom amplitudy przebiegu wejściowego. Tym niemniej wzmocnienie pętli nie może być duże. Jego wielkość musi prowadzić do odpowiednio małych oscylacji w stanie synchronizmu. Patrz również wyniki pracy [S – 2] dotyczące pętli z dwupoziomowym kwantowaniem, które można traktować jako przypadek graniczny, gdy wpływ szumu maleje.

Parametrami charakteryzującymi pętlę w obecności zakłóceń losowych są: odchylenie standardowe oscylacji błędu fazy w stanie ustalonym, średni czas do osiągnięcia stanu synchronizmu, średni czas do zerwania synchronizmu (poślizgu błędu fazy) pod wpływem zakłóceń losowych, które można wyznaczyć w oparciu o zredukowane łańcuchy Markowa [S – 1]. Redukcja ta wynika z rozważenia dynamiki na okręgu $S^1$ oraz z symetrii stanów położonych na lewo i prawo od stanu synchronizmu.

Autorowi udało się znaleźć wyrażenia zamknięte określające te parametry, co wiąże się z uproszczeniem ich obliczeń. Własności rozważanych pętli zilustrowano pokazując zależność tych parametrów od prawdopodobieństwa błędnej decyzji układu decyzyjnego sterującego pobieraniem próbek oraz od stosunku mocy sygnału do mocy szumu. Parametr $N$ w tej pracy oznacza liczbę stanów łańcucha Markowa i zależy od wzmocnienia pętli. Im mniejsze to wzmocnienie to stanów tych jest więcej i wpływ szumu na pętle jest mniejszy kosztem wydłużenia procesów przejściowych i synchronizacji. Wyższość filtrów sekwencyjnych pokazano w oparciu o obliczone charakterystyki, na podstawie których można doszukać się również różnic ilościowych. Teoretycznym uzasadnieniem tego faktu jest twierdzenie Walda – Wolfowitza [S – 1]. Okazuje się, że ta przewaga nad filtrem niekonsekwencyjnym (sumatorem sumującym określoną z góry liczbę próbek) zależy od kształtu przebiegu śledzonego. Krótszy w statystycznym sensie czas podejmowania decyzji w przypadku filtru sekwencyjnego prowadzi do przyspieszenia procesu synchronizacji, gdy wartości wielkości charakteryzujących stan ustalony, synchronizm w pętli, są zbliżone [I – 1]. Pojęcie filtru rozumiane jest tu w szerszym sensie jako układu przekształcającego zadany przebieg w celu uzyskania określonych własności przebiegu wyjściowego z filtru. W szczególnym przypadku śledzenia przez pętlę przebiegu prostokątnego poprawa własności jaką się uzyskuje jest poprawą wynikającą z przewagi klasycznego testu sekwencyjnego nad testem niesekwencyjnym.

Stwierdzono rownież, że dwupoziomowe kwantowanie sygnału wejściowego przed układem decyzyjnym degraduje jego własności o 2 dB stosunku mocy sygnału do mocy

szumu na wejściu. Wprawdzie krótszemu średniemu czasowi do osiągnięcia stanu synchronizmu towarzyszy również krótszy średni czas do jego zerwania to jednak z uwagi na duże wartości tego ostatniego istotniejsze jest zmniejszenie pierwszego parametru. Pozostałe szczegóły można znaleźć w pracy [S—1, I—1].

## 13. OPTYMALNE PĘTLE FAZOWE Z NIEJEDNOSTAJNYM PRÓBKOWANIEM DO ŚLEDZENIA PRZEBIEGU ODSTROJONEGO W FAZIE I CZĘSTOTLIWOŚCI WZGLĘDEM PRZEBIEGU ODNIESIENIA Z GENERATORA LOKALNEGO PĘTLI

Wiadomo, że na zagadnienie projektowania cyfrowych pętli fazowych można spojrzeć z punktu widzenia algorytmicznego. Zadanie to polega na znalezieniu algorytmu, który zapewniłby osiąganie stanu synchronizmu w sposób optymalny. Niestety znane kryteria optymalności nie uwzględniają złożoności realizującego go układu. Często istotna jest bowiem praca w czasie rzeczywistym, która nakłada duże wymagania odnośnie szybkości działania i którym technika cyfrowa nie zawsze jest w stanie sprostać. Tym niemniej istotny postęp w dziedzinie technologii układów cyfrowych i architektury cyfrowego przetwarzania informacji pozwala w sposób optymistyczny patrzeć na problemy realizacji praktycznej coraz bardziej złożonych algorytmów. Dlatego rozważanie optymalnych pętli cyfrowych jest celowe.

Ponadto układ optymalny wyznacza granicę jakości działania osiągalną przez układ praktyczny. Znajomość tej granicy pozwala w sposób sensowny określić stopień złożoności układu, który warto realizować. Można przedstawić metodę projektowania cyfrowych pętli fazowych z niejednostajnym próbkowaniem poprzez optymalny algorytm sterowania na okręgu [S—10]. W pracy [S—10] rozważono przebieg wejściowy odstrojony w fazie i częstotliwości względem przebiegu z generatora lokalnego pętli. Podobny przypadek był rozważany poprzednio przez Willsky'ego. Ewolujące rozkłady prawdopodobieństwa określono poprzez ich współczynniki rozwinięcia w szereg Fouriera.

Willsky rozważał problem estymacji, podczas gdy w pracy [S—10] poruszono problem sterowania. W pracy [S—10] przedstawiono także przypadki graniczne dotyczące małego wpływu szumu na działanie pętli i aproksymacje liniowe. Można uważać, że praca ta daje podstawy do projektowania różnych algorytmów sterowania, których celem jest ulepszenie działania cyfrowych pętli z niejednostajnym próbkowaniem. W celu optymalizacji pętli przyjęto Bayesowski punkt widzenia. Zgodnie z nim gęstość prawdopodobieństwa wektora stanu uwarunkowana ze względu na zbiór dostępnych obserwacji tego wektora stanowi rozwiązanie problemu nieliniowej filtracji. Jest tak ponieważ gęstość ta zawiera całkowitą informację odnośnie wektora stanu układu na podstawie jego obserwacji i na podstawie warunku początkowego określonego przez rozkład prawdopodobieństwa. Gdy funkcja stanowiąca kryterium jakości jest wypukła to rozwiązaniem problemu nieliniowej filtracji jest średnia warunkowa. Jeżeli przyjąć Markowski model ewolucji wektora stanu i obserwacji tego wektora to na podstawie reguły Bayesa można otrzymać równanie całkowe określ-

lające ewolucję gęstości warunkowej prawdopodobieństwa koniecznej do rozwiązania problemu filtracji.

Po określeniu modelu pętli z niejednostajnym próbkowaniem, w którym dopuszcza się przebieg wejściowy odstrojony w fazie lub w fazie i częstotliwości jako cel optymalizacji autor przyjął minimalizację wartości odstrojenia od stanu synchronizmu na każdym etapie sterowania generatorem lokalnym pętli. Warto przypomnieć, że w stanie synchronizmu chwile próbkowania powinny możliwie wiernie śledzić chwile przejścia przez poziom zerowy przebiegu wejściowego. Takie kryterium jakości powinno zapewnić największą szybkość osiągania stanu synchronizmu zgodnie ze znaną zasadą optymalizacji sformułowaną przez Bellmana.

Przyjęto, że szum zakłócający oddziaływuje w sposób addytywny, a efekty kwantowania można pominąć. Traktując przypadek optymalnego śledzenia przebiegu odstrojonego w fazie ewolujące rozkłady prawdopodobieństwa rozważono na okręgu z uwagi na cykliczność procesu błędu fazy. Fakt ten uzasadnia reprezentację tych gęstości w postaci szeregów Fouriera. W efekcie zamiast równań całkowych otrzymuje się wyrażenia określające współczynniki rozwinięcia w szereg Fouriera w sposób rekurencyjny poszukiwanej gęstości warunkowej.

Operacjami jakie należy wykonać są operacje mnożenia, dzielenia i splotu. Optymalny ciąg sterowania ograniczony jest do okręgu. Następnie rozważono bardziej trudny przypadek śledzenia przez pętlę fazową przebiegu harminicznego odstrojonego w fazie i częstotliwości. Rozwiązanie podano w formie ewolujących rozkładów prawdopodobieństwa rozwiniętych w szereg Fouriera–Taylora oraz w postaci rozwinięcia w szereg Fouriera i reprezentacji w postaci całki Fouriera. Rozwiązaniu podobnie jak poprzednio nadano postać rekurencyjną.

W oparciu o uzyskane wyniki ogólne rozpatrzono przypadki szczególne dotyczące pomijalnego wpływu zakłóceń losowych na układ pętli w pierwszym i drugim przypadku. Otrzymano łatwo interpretowalne struktury optymalnych pętli, gdy wpływ szumu staje się pomijalny. Ciekawe z praktycznego punktu widzenia są dalsze rozważania dotyczące liniowych aproksymacji rozważanego zagadnienia. W tym przypadku w optymalnym algorytmie sterowania predykacja odbywa się w zasadzie w oparciu o filtr Kalmana, który jest rozwiązaniem probabilistycznego podejścia do problemu filtracji w przypadku liniowym i gaussowskim. Praktycznym aspektem linearyzacji jest uzyskanie ciągów wzmocnień filtru pętli zapewniających ich optymalne działanie.

Pokazano również jak wykorzystać jedynie algorytm filtracji do zaprojektowania suboptymalnej pętli i wyznaczenia ciągu sterowań. Z uwagi na własność Markowa rozwiązanie suboptymalne zbliża się do optymalnego w miarę upływu czasu, gdy wektor stanu znajduje się w liniowej strefie nieliniowości pętli wokół stanu reprezentującego synchronizm. Celem przeprowadzonych rozważań jest ulepszenie działania cyfrowych pętli fazowych z niejednostajnym próbkowaniem. Stosując optymalny algorytm należy dokonać aproksymacji zostawiając skończoną liczbę wyrazów szeregów Fouriera. Żeby uniknąć niepożądanych efektów wynikających z „obcięcia" szeregów można zastosować ich sumowanie według Cesaro–Féjera.

## 14. PORÓWNANIE METOD ANALIZY CYFROWYCH PĘTLI FAZOWYCH

Metody analizy cyfrowych pętli fazowych można podzielić na:
- metody topologiczne
- metody analityczne oparte na linearyzacji
- metody analityczne przybliżone
- metody teorii procesów Markowa
- metody oparte na komputerowym modelowaniu

Wymienione metody ułatwiają szczegółowe określenie poszukiwanych parametrów i przewidywanie własności układów w ramach przyjętych modeli. Jest to istotne w procesie projektowania. Oprócz metod analizy można wyróżnić metody, które można określić mianem metod syntezy. Do tej grupy należą metody wynikające z optymalizacji, gdy model układu pojawia się jako model matematyczny układu realizującego optymalny algorytm.

Do tej grupy można zaliczyć:

1) metody polegające na optymalnym rozwiązaniu problemu nieliniowego sterowania lub nieliniowej filtracji.

2) metody przybliżone.

Pomimo ciągle rosnących możliwości zastosowania optymalnych metod syntezy w praktycznym układzie ich stosowalność pozostaje nadal ograniczona. Dzieje się tak dlatego, że w kryterium optymalności nie uwzględnia się kosztów wynikających ze złożoności realizacji i innych technologicznych uwarunkowań. Metody optymalnej filtracji pozwalają na oszacowanie również interesujących parametrów działania układu a więc i analizę działania układów optymalnych. Z uwagi jednak na konieczność dokonywania uproszczeń istotna jest znajomość i metod analizy i syntezy w celu uzasadnienia stosowanych przybliżeń i określenia parametrów układu nie w pełni optymalnego, lecz cechującego się względną prostotą konstrukcyjną. Metody topologiczne stanowią ważną grupę metod. Dysponując najogólniejszym zbiorem pojęć można w oparciu o nie dokonać klasyfikacji badanej nieliniowej dynamiki przed zastosowaniem metod szczegółowych.

W pracy [S–2] (rozdział 5) zastosowano metodę graficzną analizy pętli I-rzędu. Poglądowość metody, skuteczność, łatwość stosowania to główne zalety, gdy wymiar przestrzeni stanu nie jest duży. Wygodnie jest posłużyć się komputerem gdy wymiar przestrzeni stanu jest większy od 1. W oparciu o metody topologiczne można wyznaczyć dokładne parametry układów synchronizacji na podstawie modelu deterministycznego bez zakłóceń losowych.

Metody analityczne polegające na linearyzacji mają uzasadnienie topologiczne. Model topologiczny układu zlinearyzowanego jest topologicznie sprzężony z badanym układem nieliniowym. W wyniku linearyzacji można określić rodzaj punktu równowagi, lokalny przebieg trajektorii a więc i lokalną szybkość osiągania stanu synchronizmu, zakres parametrów zapewniający lokalną stabilność itp. Gdy uwzględnić zakłócenia losowe model zlinearyzowany pozwala na względnie łatwe określenie wariancji fluktuacji błędu fazy w stanie ustalonym, a także trajektorii wartości średniej i wariancji, gdy pętla lokalnie osiąga stan synchronizmu to znaczy procesy dynamiczne

przebiegają wokół położenia punktu równowagi stabilnej. Model liniowy jest stosunkowo prosty do analizy i słuszny, gdy błąd fazy jest mniejszy od $\Pi/6$ [rad]. Inny przykład linearyzacji przedstawiono w pracy [S–4] (patrz rozdział 9). Zastosowanie modelu z operacją splotu prowadzi do funkcjonalnego równania, które po zlinearyzowaniu jest równaniem ze zmiennym parametrem. Odpowiednikiem tego równania z czasem ciągłym jest równanie Volterry II rodzaju. Stabilność lokalna pętli z jednostajnym próbkowaniem zależy od własności rozwiązań takiego równania. Przedstawione w pracy [S–4] metody rozwiązania stanowią „małą" teorię stabilności pętli z jednostajnym próbkowaniem.

Metody teorii procesów Markowa stanowią ważną klasę metod, w oparciu o które można obliczyć wybrane parametry łańcuchów Markowa modelujących dynamikę badanych układów. Jasność pojęciowa, ścisły sens statystyczny to główna zaleta tych metod, w przeciwieństwie do trudności obliczeniowych, które pojawiają się proporcjonalnie do złożoności dynamiki i wymiaru przestrzeni stanu. Traktowaniu numerycznemu podlegają równania całkowe, których rozwiązania dostarczają potrzebnej informacji lub ich macierzowe odpowiedniki w przypadku dyskretnej przestrzeni stanu. Metody rozwiązywania równań całkowych polegają na rozwinięciu w szeregi ortogonalne i potęgowe. Nakład obliczeniowy w przypadku stosowania metody rozwinięć ortogonalnych zależy od przyjętego systemu funkcji ortogonalnych. Również liczba funkcji wchodzących w skład rozwinięcia zależy od stosowanego systemu funkcji. Na ogół, gdy bardziej złożone i dopasowane do danego problemu funkcje, tym mniejsza wymiarowość problemu numerycznego. Bardziej jednak również złożone jest obliczanie współczynników rozwinięcia w szereg. W pracy [S–9] zastosowano funkcje ortogonalne o różnych szerokościach w celu osiągnięcia kompromisu między dokładnością a wymiarowością problemu numerycznego. Powodzenie zastosowania metody rozwinięcia w szeregi potęgowe zależy od stopnia gładkości poszukiwanych rozwiązań. Od czasów Poincaré wiadomo, że zjawiska znane obecnie jako chaotyczne charakteryzują się praktyczną nieanalitycznością i wyznaczanie rozwiązań w obszarach występowania tych zjawisk jest z tego powodu utrudnione. Dlatego metoda szeregów potęgowych może być stosowana z dużym powidzeniem do badania zjawisk regularnych.

Ostanią grupę metod stanowią metody oparte na komputerowym modelowaniu. Ich stosowanie można przyrównać do eksperymentu, w którym odpowiednie programy modelują układ, sygnały pobudzające i zakłócające oraz układy pomiarowe. Zakres stosowalności metod komputerowego modelowania jest największy. Ograniczeniem jest jedynie czas wymagany na przeprowadzenie obliczeń i mniejsza poglądowość w stosunku do metod teoretycznych. Dlatego, znając ogólny obraz zjawiska na podstawie rozważań teoretycznych lub doświadczenia, modelowanie komputerowe z powodzeniem warto stosować do rozważań bardziej szczegółowych, polegających na dokładnym obliczeniu interesujących parametrów i zależności. Modelowanie komputerowe jest w zasadzie zgodne z realizacją w sposób programowy badanych pętli fazowych (dotyczy układu cyfrowego).

Dysponując zbiorem metod należy posługiwać się nimi tak, aby wykorzystać ich główne zalety w obszarach ich stosowalności.

## 15. DYSKUSJA, PERSPEKTYWY DALSZEJ PRACY I GŁÓWNE TEZY

Motywem prac autora w dziedzinie cyfrowych pętli fazowych było podanie ich teorii i opis przewidywanych zjawisk dla celów poznawczych. W gruncie rzeczy chodziło przede wszystkim o solidne podwaliny teoretyczne dla projektowania systemów synchronizacji w oparciu o cyfrowe pętle fazowe. Poszukiwanie układu prawie optymalnego, o dużej niezawodności działania, do wszechstronnego zastosowania w telekomunikacji, nadającego się do scalenia w wielkiej skali integracji jest zgodne z tendencjami w dziedzinie produkcji układów mikroelektronicznych. Jeśli podane wyniki pozwolą zorientować się co do wyboru odpowiednich rozwiązań układowych to cel autora na pewno został osiągnięty.

Ponadto cały zbiór metod ogólnych może być łatwiejszy do zrozumienia i innych zastosowań jeśli prześledzić zastosowanie do cyfrowych pętli fazowych. W dziedzinie cyfrowych układów synchronizacji na pewno jest jeszcze miejsce dla dalszych prac. Na przykład celem może być szukanie nowych kompromisów między wymaganiami odnośnie praktycznej realizacji a spełnieniem wymogów optymalności. Wreszcie może to być optymalizacja wybranych układów. W końcu może to być poszukiwanie nowych rozwiązań, takich jak wspomniany nowy układ pętli z jednostajnym próbkowaniem w rozdziale 9.

Ogólnie można stwierdzić, że przedstawione w pracach autora układy spełniają zadanie synchronizacji. Rozważania dotyczą układów skonstruowanych tak, że modele są adekwatne. Dlatego układy próbkowania i podtrzymywania (ang. *sample and hold devices*) powinny posiadać stałą zaniku tak małą, by w czasie przetwarzania przez przetwornik analogowo-cyfrowy (ang. *analog-digital convertor*) zanik można było pominąć. Filtr i generator sterowany pętli mogą być realizowane przez program w oparciu o mikroprocesor lub w sposób sprzętowy. Natomiast generator sterowany cyfrowo pętli z niejednostajnym próbkowaniem na ogół realizowany jest sprzętowo w oparciu o programowany licznik.

Do najważniejszych wyników własnych w dziedzinie cyfrowych układów synchronizacji autor zalicza:

- Porównanie dla synchronizmu N:1 własności pętli z niejednostajnym próbkowaniem I-rzędu z idealnym i dwupoziomowym układem kwantującym. Uściślenie warunku braku zjawiska poślizgu błędu fazy w ramach przyjętego modelu w funkcji parametrów pętli. Podanie ogólnej definicji braku zjawiska poślizgu błędu fazy w oparciu o pojęcia przestrzeni bazowej i przestrzeni nakrywającej teorii gładkich układów dynamicznych.
- Rozważenie innych synchronizmów i pokazanie, że zjawisko poślizgu błędu fazy może dotyczyć każdej strefy synchronizacji.
- Odkrycie zjawiska chaosu przejściowego w pętli II-rzędu z niejednostajnym próbkowaniem i nieliniowością sinusoidalną. Opis tego zjawiska w przypadku deterministycznym, gdy brak zakłóceń losowych.
- Odkrycie nowych kryteriów stabilności dla pętli z niejednostajnym próbkowaniem dowolnego rzędu w oparciu o nowy model pętli w postaci równania funkcjonalnego z czasem dyskretnym na okręgu. Rozważanie w oparciu o nie szeregu

przypadków uzasadnionych względami praktyki i wyciągnięcie wniosków dotyczących między innymi warunku braku zjawiska chaosu przejściowego w pętli drugiego i dowolnego rzędu. Opracowanie tych kryteriów powoduje usunięcie niedogodności wskazanej przez Lindsey'a i Chie a dotyczącej stosowania kryteriów normopodobnych [R – 1, pp. 426 – 427] w przypadku pętli wyższego rzędu niż jeden.

- Podanie w miarę pełnej analizy pętli w obecności zakłóceń losowych. Rozwinięcie przez autora metody równań całkowych i macierzowych do wyznaczania podstawowych parametrów pętli oraz scharakteryzowanie wpływu zakłóceń losowych na zjawisko chaosu przejściowego w pętli II-rzędu. Dokonanie modyfikacji pętli II-rzędu i zbadanie jej własności.
- Rozwiązanie ścisłe zagadnienia stabilności lokalnej pętli z jednostajnym próbkowaniem. Zwiększenie przez to wagi wyników porównań obu klas pętli z jednostajnym i niejednostajnym próbkowaniem. Zaproponowanie nowego układu pętli z jednostajnym próbkowaniem o przewidywanych bardziej korzystnych własnościach.
- Pokazanie jak można wykorzystać teorię testowania hipotez statystycznych do optymalizacji pętli z dwuwartościową korekcją fazy i nadanie tej klasie pętli cech układów optymalnych.
- Rozwiązanie problemu optymalnego śledzenia przebiegów odstrojonych w fazie i częstotliwości w sposób ogólny w oparciu o teorię nieliniowej filtracji a także podanie różnych aproksymacji tego problemu pozwalających na ulepszanie układów praktycznych.

Ponadto dyskusję dotyczącą zastosowanych metod oraz perspektywą dalszej pracy w dziedzinie cyfrowych układów synchronizacji.

## BIBLIOGRAFIA

Cykl prac:

[S]

1. Żółtowski M.: *O śledzeniu zakłóconego przebiegu sinusoidalnego za pomocą cyfrowej pętli fazowej I-rzędu z dwuwartościową korekcją fazy*. Rozprawy Elektrotechniczne 1983, 29 z. 1, ss. 201 – 216
2. Żółtowski M.: *On the Acquisition Behaviour of a First-Order Digital Phase-Locked Loop with Negligible and Binary Quantization Effects*. Rozprawy Elektrotechniczne 1984, 30 z. 2, ss. 401 – 412
3. Żółtowski M.: *O globalnej stabilności synchronizmu pętli fazowej drugiego rzędu z niejednostajnym próbkowaniem*. Rozprawy Elektrotechniczne, 1988, 34, z. 4, ss. 37 – 58
4. Żółtowski M.: *On the local stability of the uniform sampling phase-locked loop*. Kwartalnik Elektroniki i Telekomunikacji – Electronics and Telecommunications Quarterly 1990, z. 3
5. Żółtowski M.: *New stability criteria for the nonuniform sampling zero-crossing digital phase-locked loops of any order*. Part I. Kwartalnik Elektroniki i Telekomunikacji, Electronics and Telecommunications Quarterly, 1991, z. 1 – 2
6. Żółtowski M.: *New stability criteria for the nonuniform sampling zero-crossing digital phase-locked loops of any order*. Part II. Kwartalnik Elektroniki i Telekomunikacji – Electronics and Telecommunications Quarterly, 1991, z. 1 – 2
7. Żółtowski M.: *New stability criteria for the nonuniform sampling zero-crossing digital phase-locked loops of any order*. Part III. Kwartalnik Elektroniki i Telekomunikacji, Electronics and Telecommunications Quarterly, 1991, z. 1 – 2

8. Żółtowski M.: *On the nonuniform sampling digital phase-locked loops in the presence of noise.* Part I. Kwartalnik Elektroniki i Telekomunikacji Quarterly, 1991, z.
9. Żółtowski M.: *On the nonuniform sampling digital phase-locked loops in the presence of noise.* Part II. Kwartalnik Elektroniki i Telekomunikacji — Electronics and Telecommunications Quarterly, 1991, z.
10. Żółtowski M.: *Optimal uniform sampling zero-crossing digital phase-locked loops for tracking the offsets in phase or in phase — and frequency.* Electronics and Telecommunications Quarterly, 1991, z.

## Inne prace autora dotyczące cyfrowych pętli fazowych

[I]

1. Żółtowski M.: *Cyfrowe pętle fazowe z niejednostajnym próbkowaniem.* Praca Doktorska, Politechnika Gdańska, Gdańsk 1983
2. Żółtowski M.: *O zastosowaniu filtru sekwencyjnego w układach cyfrowych pętli fazowych DPLL.* III Krajowe Sympozjum Nauk Radiowych Wrocław, 12—14, II, 1981
3. Żółtowski M.: *On the Comparison of the first-order digital phase-locked loops, DPLL-s with uniform and nonuniform sampling.* Seventh Colloqium on Microwave Communication $6^{th}$—$10^{th}$ September 1982, Budapest, Hungary
4. Żółtowski M.: *Niektóre własności procesu synchronizacji w cyfrowych pętlach fazowych pierwszego rzędu z jendostajnym i niejednostajnym próbkowaniem.* Zeszyty Naukowe Politechniki Gdańskiej, Nr 359, Elektronika LIV, 1983
5. Żółtowski M.: *O zjawisku poślizgu błędu fazy w cyfrowej pętli fazowej DPLL I-rzędu z idealnym i dwupoziomowym układem kwantującym.* V Krajowa Konferencja, Teoria Obwodów i Układy Elektroniczne, Łódź 1982
6. Żółtowski M.: *Cyfrowa pętla z niejednostajnym próbkowaniem o zmiennych w czasie współczynnikach.* V Krajowa Konferencja, Teoria Obwodów i Układy Elektroniczne, Łódź, 1982
7. Żółtowski M.: *Porównanie dwóch metod zmniejszenia dyspersji błędu fazy dla pętli z dwupoziomową korekcją fazy.* VI Krajowa Konferencja, Teoria Obwodów i Układy Elektroniczne, Gliwice 1983
8. Żółtowski M.: *Metody Zwiększenia Szybkości Synchronizacji w Cyfrowych pętlach Fazowych DPLL z Niejednostajnym Próbkowaniem — Przypadek deterministyczny.* VIII Krajowa Konferencja Teoria Obwodów i Układy Elektroniczne, Poznań, 15—18 X, 1985
9. Żółtowski M.: *O stabilności globalnej pętli fazowej II-rzędu o modelu z czasem dyskretnym i niejednostajnym próbkowaniem.* Teoria Obwodów i Układy Elektroniczne, IX Konferencja, Wrocław—Szklarska Poręba, 22—24 Października 1986
10. Żółtowski M.: *Chaotic transient hang-up phenomenon and N:1 phase entrainment in discrete time phase-locked loop of second order.* European Conference on Circuit Theory and Design, ECCTD, Paris, September, 1—4, 1987
11. Żółtowski M.: *On the local stability of the uniform sampling Phase-locked Loops.* European Conference on Circuit Theory and Design, 5—8 September 1989, ECCTD Brighton
12. Żółtowski M.: *Chaotic Transient Phenomenon in the Discrete-Time Phase-locked Loop of Second Order.* Electronics and Telecommunications Letters, vol. 2, No 3, 1987
13. Żółtowski M.: *New stability criteria for the nonuniform sampling digital phase-locked loops.* 1990 IEEE International Symposium on Circuits and Systems, Sheraton Hotel, New Orleans, LA, May 1—3, 1990

[R]

1. Lindsey W. C., Chie C. M.: *A survey of digital phase-locked loops.* Proc. IEEE, vol. 69, no 4, April, pp. 410—431, 1981
2. Chie C. M.: *Second-order Frequency aided digital-phase-locked loop for Doppler rate tracking.* IEEE Trans. Commun., vol. COM—28, no 8, pp. 1431—1436, August, 1980

3. Osborne H. C.: *Stability analysis of $N^{th}$ power digital phase-locked loop – Part I: First-order DPLL.* IEEE Trans. Commun., vol. COM–28, no 8, pp. 1342–1354, August 1980
4. Osborne H. C.: *Stability analysis of $N^{th}$ power digital phase-locked loop – Part II: Second and Third-Order DPLL,s.* IEEE Trans. Commun., vol. COM–28, no 8, August 1980
5. D'Anrea N. A., Russo F.: *Multilevel Quantized DPLL behavior with phase- and frequency-step plus noise input.* IEEE Trans. Commun., vol. COM–28, no 8, pp. 1373–1381, August 1980
6. Lindsey W. C., Chie C. M.: *Acquisition behavior of a first-order digital phase-locked loop.* IEEE Trans. on Commun. vol. COM–26, no 9, pp. 1364–1376, September 1978
7. Lindsey W. C.: *Synchronization Systems in Communication and Control.* Printice–Hall, 1972
8. Bitjuckij W. I., Sierdiukov P. N.: *Ocenka wriemieni do zrywa sinchronizma w impulsnoj sistemie FAPCZ.* Radiotechnika, ss. 95–97, 1973
9. Szlenk W.: *Wstęp do teorii gładkich układów dynamicznych,* Warszawa: PWN, 1982
10. Arnold W. I.: *Teoria równań różniczkowych.* Warszawa: PWN, 1983
11. Bernstein G. M., Lieberman M. A., Lichtenberg A. J.: *Nonlinear Dynamics of a Digital Phase Locked Loop.* IEEE Trans. Commun., vol. 37, no 10, October 1989
12. Jaźwiński A. H.: *Stochastic Processes and filtering theory.* Academic Press 1970, New York and London
13. Gill G. S., Gupta S. C.: *First-order discrete phase-locked loop with applications to demodulation of angle modulated carrier.* IEEE Trans. Commun., vol. COM–70, pp. 454–462, June 1972
14. Snyder D. L.: *The State–Variable Approach to Analog Communication Theory.* IEEE Trans. on Information Theory, vol. IT–14, no 1, pp. 94–104, January 1968
15. Polk D. R., Gupta S. C.: *Quasi-optimum digital phase-locked loops.* IEEE Trans. Commun., vol. COM–21, pp. 75–82, January 1973
16. Weinberg A., Liu B.: *Discrete time analysis of nonuniform sampling first- and second-order digital phase-locked loops.* IEEE Trans, Commun., vol. COM–22, no 2, pp. 123–137
17. Garodnic J., Greco J., Schiling D. L.: *Response of an all digital phase-locked loop.* IEEE Trans. Commun., vol. COM–22, no 6, pp. 751–764, June 1974
18. Rocha L. F.: *Simulation of a Discrete PLL with Variable Parameters.* Proc. of the IEEE, vol. 67, no 3, pp. 440–442, March 1979
19. Oppenheim A. V., Schafer R. W.: *Cyfrowe przetwarzanie sygnałów.* Warszawa: WKŁ, 1979, tłumaczenie z języka angielskiego
20. Kudrewicz J., Grudniewicz J., Swidzińska B.: *Chaos, phase slipping and Cantor-like sets in a discrete phase-locked loop.* European Conference on Circuit Theory and Design, ECCTD, September, 1–4, 1987, Paris
21. Belykh V. N., Lebedeva L. V.: *Investigation of a particular mapping of a circle,* PMM, U.S.S.R, vol. 46. no 5, pp. 616–620, 1983
22. Guckenheimer J., Holmes P.: *Nonlinear oscillations, Dynamical Systems and Bifurcations of Vector Fields.* Springer–Verlag, 1983
23. Beneviste A., Joindot M., Vandamme P.: *Analyse theoretique des boucles de phase numeriques en presence de canaux dispersifs.* Note technique NT/MER/TSF/1, Janvier 1980
24. Best A.: *Phase locked loops,* 1984
25. Szabatin J.: *Podstawy teorii sygnałów.* Warszawa: WKŁ, 1982
26. Kohlenberg A.: *Exact interpolation of band limited functions,* J. Appl. Phys. vol. 24, pp. 1432–1436, December 1953
27. Falconer D. D.: *Jointly adaptive equalization and carrier recovery in Two-dimensional Digital Communication Systems.* B.S.T.J., 55 no 3, 1976
28. Cessna J. R., Levy D. M.: *Phase noise and transient times for a binary quantized digital phase-locked loop in white gaussian noise.* IEEE Trans. on Commun., vol. COM–20, no 2, pp. 94–104, April 1972

M. ŻÓŁTOWSKI

# DIGITAL PHASE-LOCKED – SELECTED TOPICS


## Summary

The main results obtained by the author while the digital phase-locked loops have been of concern are summarized. Digital phase-locked loop is a circuit of noticeable importance in modern telecommunication and also being used in different measuring equipment. After the basic notions and problems have been introduced the step by step the original results by the author are presented while concerning the most important aspects of the digital phase-locked loop's performance. These were lectured during numerous colloquia and native and foreign conferences. Besides the results of fundamental and theoreticel meaning the conclusions of practical interest are given also. The comparisons of the methods being considered together with discussion concerning when they are applicable are included. Finally, the main topics gained by the author and the prospects of further work are pointed out.

KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJI, 1992, 38, z. 1, ss. 153-179

(621.382.066)

# Półprzewodnikowe układy dużej mocy — technologia, analiza, projektowanie

ANDRZEJ NAPIERALSKI

*Laboratoire d'Automatique et Analyse des Systemes du C.N.R.S.*
*Toulouse Cedex, France 7, av. du Colonel Roche, 31077*
*et*
*L'Institute des Sciences Appliquées de Toulouse*

*Otrzymano 1991.7.24*

*Autoryzowano do druku 1991.9.20*

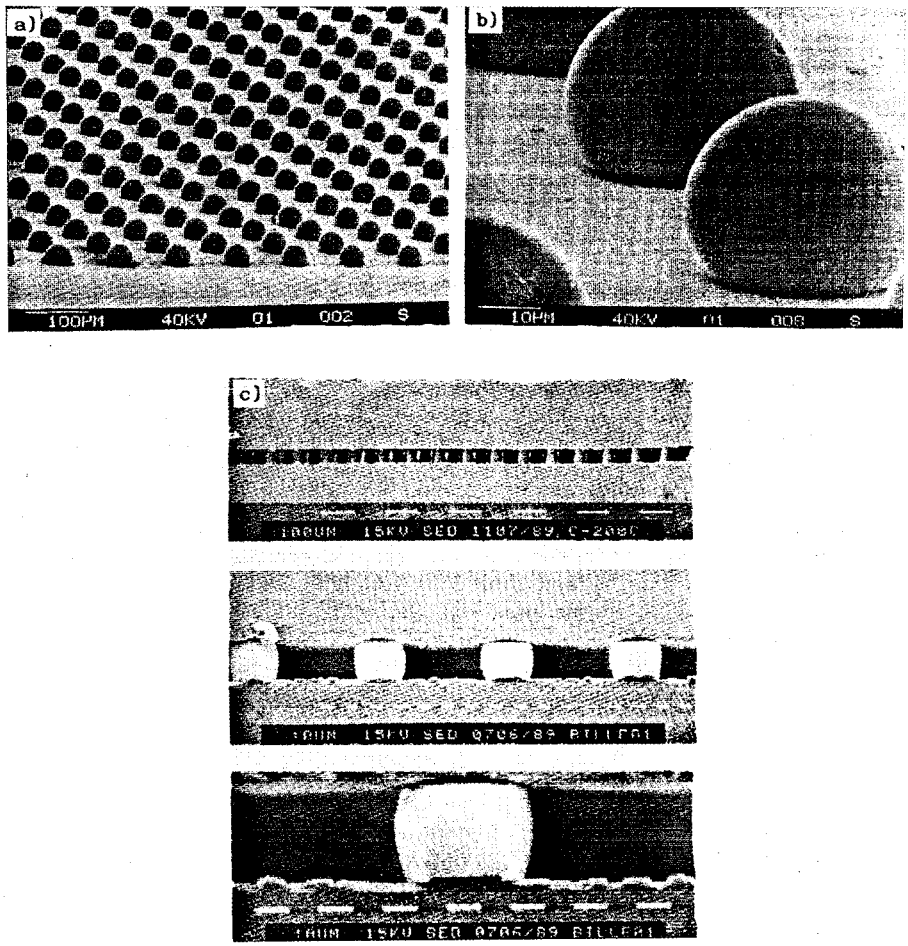
Olbrzymi postęp w technologii wytwarzania przyrządów półprzewodnikowych i stały wzrost gęstości wydzielanej w nich mocy doprowadził do wytwarzania układów G.S.I., W.S.I. i M.C.M's w przypadku układów monolitycznych, oraz układów hybrydowych, w których gęstości wydzielanej mocy przekraczają 100 W/cm². Modelowanie specyficznych problemów występujących w tych układach stało się aktualnie przedmiotem licznych programów badawczych. W artykule omówiono podstawowe problemy związane z technologią wytwarzania układów hybrydowych. Przedstawiono również kompletny projekt hybrydowego łącznika mocy, wykonany na zamówienie SSM (SAGEM – SAT – MICROELECTRONIQUE) i wdrożonego do seryjnej produkcji. Omówiono też perspektywy rozwoju i zastosowania współczesnych układów półprzewodnikowych typu „Smart Power", które zawierają jednocześnie elementy mocy i układy o dużej skali integracji.

## 1. WSTĘP

Wydzielanie mocy w przyrządach i układach półprzewodnikowych oraz, wraz z postępem technologii, zwiększanie jej gęstości, powoduje konieczność zapewnienia właściwego chłodzenia struktury półprzewodnikowej. W układach dyskretnych istnieje szereg znormalizowanych obudów (np. TO3 czy też press-pack) służących do odprowadzania ciepła. Nowe technologie i pojawienie się układów zawierających więcej niż jeden element wymagają ponownego rozważenia problemu odprowadzania ciepła i wyboru właściwego układu chłodzącego.

Stały postęp w technologii wytwarzania przyrządów półprzewodnikowych w ciągu ostatnich lat doprowadził do realizacji monolitycznych układów scalonych o coraz większej skali integracji (L.S.I., V.L.S.I., G.S.I. (Giga Scale Integration), W.S.I. (Wafer Scale Integration) i o coraz większej szybkości działania. W celu zwiększenia ilości funkcji realizowanych przez układy elektroniczne w ostatnich latach rozwinięto

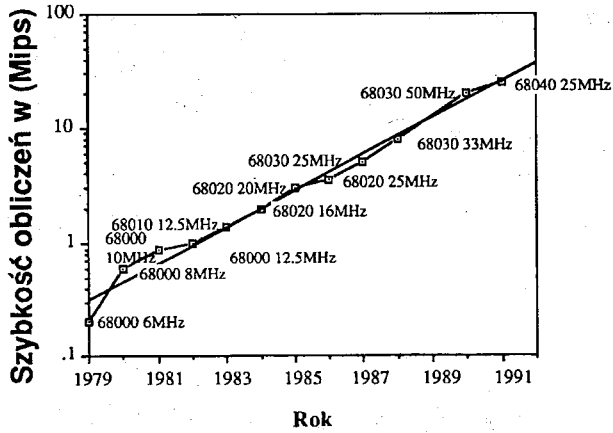
produkcję układów MCM (Multi Chip Modules), które są układami scalonymi o dużej gęstości upakowania i wysokiej częstotliwości granicznej, połączonymi pomiędzy sobą poprzez mikrokontakty (mikrokuleczki). Połączenia pomiędzy dwoma płytkami krzemu dokonywane są na całej powierzchni, dzięki czemu nie ma ograniczeń w ilości wejść i wyjść poszczególnych płytek krzemu. Na rysunku 1 przedstawiono przekrój struktury MCM wykonanej przez CEA – LETI [5,15]. Prace nad tymi nowymi strukturami oraz układami 3-wymiarowymi są przedmiotem europejskiego programu badawczego JESSI – BLR [15].




Rys. 1. Schematyczne przedstawienie połączeń pomiędzy dwoma płytkami krzemu
a) łatwotopliwe mikrokulki, b) powiększenie 10-krotne, c) przekrój struktury Si/Si połączonej przez mikrokulki

Układy typu MCM są aktualnie realizowane w technologii 0.5 $\mu$m (przewiduje się 0.25 $\mu$m w roku 1991 [15]) i ich duża gęstość upakowania powoduje wydzielanie mocy rzędu kilku watów (ponad 30 W w roku 2000). Problemy w ich projektowaniu stają się zbliżone do przypadku współczesnych układów hybrydowych.

Nawet klasyczna technologia CMOS, w jakiej wykonane są na przykład mikroprocesory Motoroli, znacznie zwiększyła w ciągu ostatnich dziesięciu lat gęstość upakowania elementów. Na rysunku 2 przedstawiono przykładowo wzrost szybkości obliczeń w milionach instrukcji na sekundę (Mips) mikroprocesorów Motoroli w latach 1979–1991. Procesor 68000 z 1979 roku zawierał zaledwie 68000 tranzystorów i pracował przy częstotliwości 6 MHz. Ostatni procesor Motoroli 68040, zawiera już 1 200 000 tranzystorów i pracuje przy częstotliwości 25 MHz. Jego moc obliczeniowa jest aż o dwa rzędy wielkości większa od tego pierwszego.



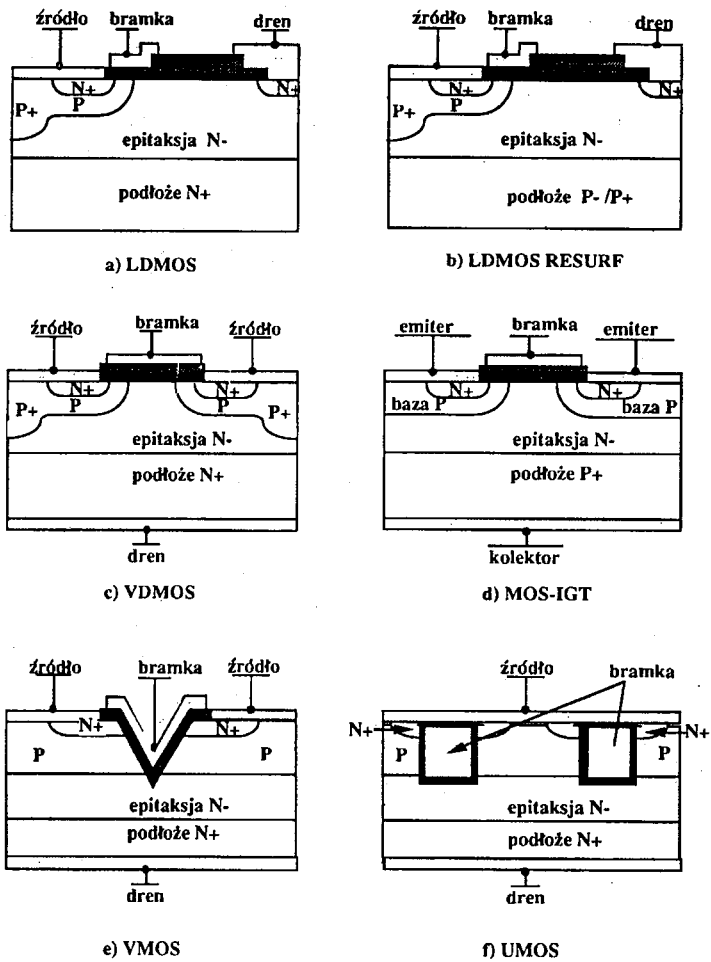

Rys. 2. Szybkość obliczeń w milionach instrukcji na sekundę (Mips) mikroprocesorów Motoroli w latach 1979–1991

W elementach dużej mocy od kilku lat obserwujemy podobne tendencje rozwoju. W 1970 roku granicznymi parametrami tyrystorów był prąd przewodzenia rzędu 1000 A i napięcie blokowania 3000 V [1]. W 1988 roku granica ta przesunęła się do 6500 V (10 000 V w laboratorium) [2].

Jeszcze większy postęp dokonał się w technologii tranzystorów MOS dużej mocy, których rozwój zaczął się dopiero w latach 70. Pierwszy tranzystor mocy MOS został wykonany w Uniwersytecie Stanford [6] w technologii LDMOS (**D** – double diffusion – podwójna dyfuzja, **L** – lateral – struktura pozioma). Aktualnie w elementach dyskretnych stosuje się technologię VDMOS (**V** – vertical – struktura pionowa) i w niej produkowane jest 95% tranzystorów MOS dużej mocy.

Na rys. 3 [4] przedstawiono przykładowo wybrane struktury tranzystorów MOS dużej mocy, a na rys. 4 wzrost mocy sterowanej (iloczyn prądu przewodzenia i napięcia blokowania) poprzez tranzystory MOS–FET i MOS–IGT (Insulated Gate Transistor) w latach 1980–1988 [2].

Przy zachowaniu tych samych parametrów, powierzchnia tranzystorów MOS dużej mocy jest już obecnie znacznie mniejsza niż kilka lat temu. Z drugiej strony zwiększyła się znacznie częstotliwość komutacji stosowanych w praktyce układów, od kilku kHz w początkowym okresie do kilkudziesięciu MHz obecnie. Powoduje to konieczność minimalizacji długości połączeń pomiędzy przyrządami mocy oraz ich

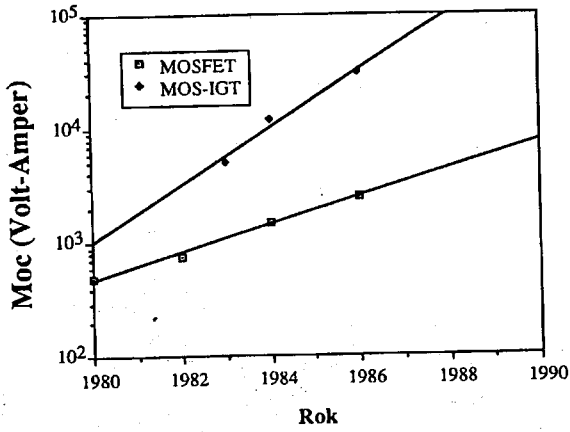



Rys. 3. Wybrane struktury tranzystorów MOS dużej mocy

układami sterowania i zabezpieczeń. Jednym z rozwiązań jest integracja monolityczna całego układu. Tego typu rozwiązania zwane są układami „Smart Power" („INTELIGENTNA MOC"). Jako przykład pierwszych tego typu układów można wymienić serię układów MPC 20... Motoroli, stosowanych do zabezpieczania układów zasilania 5 V, 12 V i 15 V.

W przypadku dużej mocy, do podwójnej tendencji: zwiększania upakowania i szybkości działania układów, dochodzi jeszcze zwiększenie ich całkowitej mocy. Duża gęstość mocy wydzielanej w układzie powoduje w konsekwencji znaczne zwiększenie jego maksymalnej temperatury pracy. Problemy termiczne stają się w tej chwili problemami najwyższej wagi i zostaną przedstawione dokładnie w dalszej części artykułu.

Po początkowych sukcesach układów „Smart Power", stwierdzono, że nie można w praktyce osiągnąć zbyt dużych wartości napięcia i sterowanej mocy. Dodatkowe trudności występują w praktycznej realizacji tych układów, w których trzeba jednocześnie wykonać część analogową, cyfrową oraz mocy, tak że technologia ta staje




Rys. 4. Wzrost mocy tranzystorów MOSFET i MOS–IGT w latach 1980–1988

się bardzo skomplikowana i kosztowna. Do realizacji funkcji, które nie mogą być scalone w krzemie, lub też których wykonanie jest zbyt kosztowne, używana jest technologia hybrydowa. Polega ona na umieszczeniu w tej samej obudowie przyrządów mocy, ich układów sterowania i zabezpieczeń. Tak zrealizowane układy nazywają się układami hybrydowymi mocy i można je opisać następującą definicją:

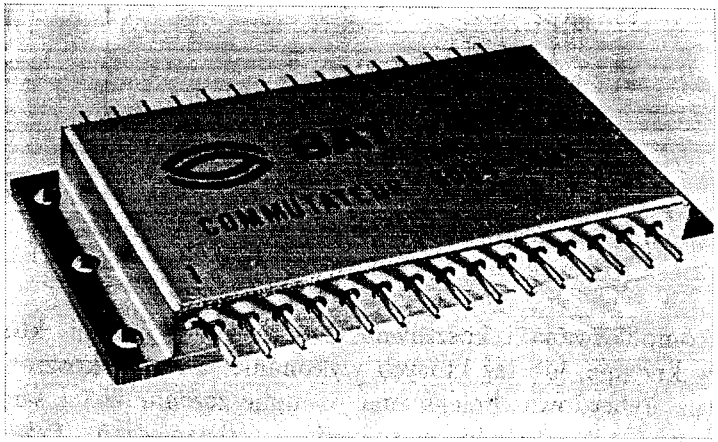
Def.: *Układem hybrydowym mocy nazywamy dowolny układ hybrydowy, zawierający minimum jeden przyrząd mocy i wydzielający moc powyżej 5W.*

Tego typu definicja dotyczy oczywiście tak samo wzmacniacza operacyjnego APEX w obudowie TO3, jak i modułu hybrydowego 600V/60A o częstotliwości pracy 100 kHz wykonanego przez SSM (SAGEM–SAT–MICROELECTRONIQUE) i pokazanego na rys. 5 [30], czy też układu POWER COMPACT [28] przedstawionego na rysunku 6.

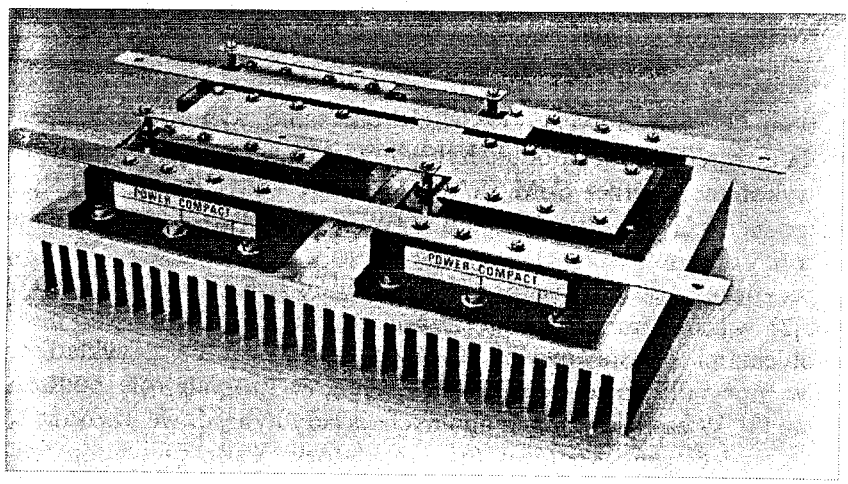
Inną specyficzną dziedzinę zastosowania układów hybrydowych stanowi technika kosmiczna [3], w której minimalizacja ciężaru układu i zwiększenie jego niezawodności są problemami podstawowymi. Crouzet opracował na przykład całą serię regulatorów mocy dla zastosowań kosmicznych, w których rolę podłoża spełnia tlenek berylu [3]. W zastosowaniach ogólnych układy hybrydowe używane są między innymi w technice samochodowej (np. regulatory prądnicy wykonywane przez Motorolę), w sterowaniu silników (np. gałęzie falowników, pełne mostki Gretz'a, itp., wykonywane przez POWER COMPACT [29]), w układach ogólnego zastosowania, oraz w produkcji wzmacniaczy RF (Motorola). W zależności od zastosowania, wydzielana w układzie hybrydowym moc zmienia się od 5 W do kilkuset watów i w miarę rosnących postępów w technologii górna granica wydzielanej mocy zbliża się powoli do 1 kW.

## 2. TECHNOLOGIA UKŁADÓW HYBRYDOWYCH

Realizacja praktyczna hybrydowych układów dużej mocy, jak na przykład układów przedstawionych na rys. 5 i 6, nie jest sprawą prostą. Wymaga ona rozwiązania dwóch podstawowych problemów jakimi są połączenia dla dużych prądów i ewakuacja energii cieplnej.



Rys. 5. Układ hybrydowy mocy zawierający 12 tranzystorów MOSFET




Rys. 6. Cztery moduły (500V, 180A, $t_F = 20$ nS) połączone razem

Konstruktor układu hybrydowego musi wybrać odpowiednie podłoże, system past, metodę łączenia płytek krzemu z podłożem oraz podłoża z obudową układu hybrydowego i opracować system połączeń całego modułu. Modelowanie układów

hybrydowych podzielić można na analizę technologiczną, oraz na analizę problemów elektrotermicznych. Oba te problemy są ze sobą ściśle związane. Jak zobaczymy w dalszej części artykułu, wybór technologii ma bezpośredni wpływ na wartość maksymalnej temperatury pracy układu hybrydowego.

## 2.1. ANALIZA TECHNOLOGICZNA

a) *Wybór podłoża*

Dla przyrządów półprzewodnikowych dużej mocy, tylna strona płytki krzemu jest aktywna, gdyż przez nią przepływa sterowany prąd (kolektor tranzystora bipolarnego, dren tranzystora VDMOS, itd.). W pewnych zastosowaniach, obudowa jest jedną z końcówek układu elektronicznego i płytki krzemu mogą być łączone bezpośrednio z podłożem, co pozwala na uproszczenie wielu problemów, jak na przykład problemu ewakuacji ciepła z przyrządu półprzewodnikowego.

W większości przypadków trzeba jednak zapewnić izolację galwaniczną pomiędzy układem elektrycznym i zewnętrzną stroną obudowy (2500V według normy VBE) Przyrządy półprzewodnikowe należy więc umieszczać na warstwie podłoża o grubości wystarczającej dla zapewnienia wymaganej izolacji. Do tego celu używane są zazwyczaj ceramiki i obecnie najczęściej używaną jest warstwa $Al_2O_3$ (ceramika alundowa) grubości 635 $\mu$m. Warstwa ceramiki alundowej tej grubości jest niestety dobrym izolatorem ciepła i w pewnych zastosowaniach jej rezystancja termiczna jest stanowczo zbyt duża (konduktywność termiczna $Al_2O_3$ wynosi 0.2 W/(cm·K)). Podłożem o bardzo dobrej konduktywności termicznej (2.5 W/(cm·K)) jest tlenek berylu BeO. Jest on jednak toksyczny i jego zastosowanie dla celów cywilnych nie będzie prawdopodobnie nigdy dozwolone.

Kilka lat temu pojawiła się na rynku nowa ceramika (AlN — azotek glinu), która ma bardzo dobre własności termiczne (konduktywność termiczna 0.7 – 1.7 W/(cm·K)) i mechaniczne (współczynnik rozszerzalności cieplnej zbliżony do krzemu) i jest obecnie uważana za idealny materiał dla produkcji układów hybrydowych. Podstawowym problemem jest uzyskanie powtarzalności parametrów termofizycznych tego materiału, co jest główną przyczyną dotychczasowego braku masowej produkcji układów hybrydowych w tej technologii.

b) *Wybór past*

Przy projektowaniu grubowarstwowego układu hybrydowego należy wybrać odpowiedni system past (muszą one być kompatybilne ze sobą), dla realizacji przewodników, rezystancji oraz dla zapewnienia izolacji pomiędzy przewodnikami w czasie ich nakładania. Dla zastosowań przemysłowych używa się na ogół zestawu „Ag – Pb", natomiast w przypadku układów specjalizowanych złota.

Przy projektowaniu modułu należy rozdzielić układy mocy oraz układy zabezpieczeń i sterowania, które realizowane są na osobnych podłożach. Dla układów

mocy, należy używać past przewodzących o bardzo małej rezystywności. Dobrze nadają się do tego celu pasty wykonane na bazie miedzi, których rezystywność jest tylko dwa razy większa od rezystywności litych przewodów miedzianych, a ponadto dają się one łatwo lutować, oraz łączyć techniką ultradźwiękową z przewodami doprowadzającymi [31]. Dla układów, w których prądy przekraczają 50 A, ścieżki przewodzące wykonane metodą sitodruku nie mogą już być stosowane i zamiast nich używa się przewodów z litej miedzi, nałożonych bezpośrednio na podłoże ceramiczne (D.B.C.), lub metodą termokompresji czy też lutowania aktywnego [27]. Tego typu podłoża są jednak bardzo drogie, i dlatego technika ich realizacji dostępna jest tylko nielicznym przedsiębiorstwom.

c) *Łączenie przyrządów półprzewodnikowych*

Łączenie przyrządów półprzewodnikowych dużej mocy, które mają tylną powierzchnię pokrytą warstwą złota lub srebra, z podłożem jest na ogół wykonywane poprzez przetapianie stopu zawierającego miękkie lutowie (SnPb, InPb, ...) w piecu przewłokowym lub w maszynie do przetapiania w fazie gazowej. Oprócz funkcji połączenia elektrycznego, lutowie, którego grubość jest rzędu 50 $\mu$m, pełni rolę amortyzatora sił wynikających z różnicy rozszerzalności cieplnej pomiędzy krzemem i metalizacją podłoża. Doświadczenie pokazuje, że bardzo trudno jest połączyć idealnie przyrząd półprzewodnikowy z podłożem i w związku z tym w lutowiu mogą wystąpić dziury, których następstwem jest występowanie ciepłych punktów [7,26]. W czasie pracy modułu przyrząd mocy podlega działaniu cykli cieplnych, które powodują stopniową degradację lutowia [16].

Inną metodą łączenia przyrządów mocy jest stosowanie klejów przewodzących opartych na bazie srebra. Jak do tej pory przewodność termiczna tych klejów jest jeszcze zbyt mała, aby zapewnić własności porównywalne z połączeniami poprzez lutowie. Coraz częstsze używanie klejów wynika głównie z ich małej wrażliwości na cykle cieplne i brak dziur trudnych do uniknięcia w przypadku lutowia.

d) *System połączeń wewnątrz modułu*

Metalizacja połączeń pomiędzy przyrządami mocy jest zazwyczaj wykonana z aluminium. Połączenia przyrząd-przewodnik, lub przyrząd-przyrząd są wykonane za pomocą przewodów aluminiowych o przekroju rzędu 500 $\mu$m techniką ultradźwiękową [31,32]. Możliwe jest również użycie do tego celu przewodów miedzianych, ale do tej pory technologia ta nie jest opracowana na skalę przemysłową. Połączenia pozostałych elementów wykonywane są metodami klasycznymi poprzez przetapianie, lub połączenie przewodami mocowanymi techniką ultradźwiękową.

e) *Obudowy*

Obudowy układów hybrydowych mogą być podzielone na dwa podstawowe rodzaje:
1) obudowy wykonane całkowicie z metalu (na ogół hermetyczne)

2) obudowy mające metalową podstawę i przykrycie wykonane z tworzyw sztucznych.

Do tej pory nie ma normalizacji dla obudów metalowych, które są wykonywane na zamówienie. W zależności od zastosowania podstawa obudowy wykonana jest z kowaru lub miedzi o grubości zoptymalizowanej dla zapewnienia jak najlepszych właściwości termomechanicznych. Istnieją też moduły, które nie mają podstawy metalowej, lecz podłoże ceramiczne jest w bezpośrednim kontakcie z radiatorem, na którym użytkownik umocował układ [29].

f) *Podłoże*

Przytwierdzenie podłoża wykonuje się różnymi metodami w zależności od typu układu. Układy sterowania i zabezpieczania wydzielają relatywnie małą moc i mogą być przyklejone do podstawy. Podłoże układów mocy jest najczęściej lutowane i dlatego jego dolna część musi być metalizowana.

Ze względu na bardzo dużą powierzchnię podlegającą lutowaniu problemy technologiczne są w tym wypadku jeszcze bardziej krytyczne niż w przypadku lutowania przyrządów półprzewodnikowych. Jedną z metod stosowanych do poprawy jakości lutowia, jest analiza wykonanych układów przy pomocy promieni X [7].

g) *Połączenia z otoczeniem zewnętrznym*

Układ hybrydowy łączy się z pozostałymi układami (zasilanie, obciążenie, sterowanie) poprzez przewody o możliwie małej rezystancji szeregowej. Ponadto powinny one być mało wrażliwe na proces starzenia. W przypadku układów wykonanych całkowicie z metalu, połączenia wykonywane są poprzez hermetyczne wyprowadzenia zatopione w szkle dla zapewnienia izolacji elektrycznej.

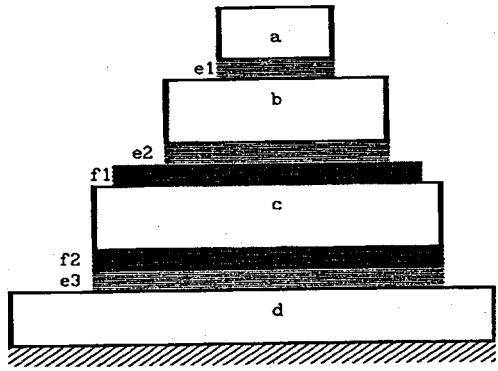
Dla układów seryjnych, połączenia wykonywane są przy pomocy prętów miedzianych i śrub. Ta tradycyjna metoda jest dobra zarówno pod względem elektrycznym, jak i termicznym.

## 2.2. PROBLEMY TERMOMECHANICZNE

Rozważmy teraz problemy, które są specyficzne dla hybrydowych układów mocy. Właściwości termiczne hybrydowego układu mocy mają bezpośredni wpływ na jego czas życia, gdyż niezawodność przyrządów półprzewodnikowych jest odwrotnie proporcjonalna (wykładniczo) do ich temperatury. Zmniejszenie temperatury pracy o 20°C powoduje czterokrotny wzrost ich czasu życia.

Poprawne zaprojektowanie układu hybrydowego ze względu na jego właściwości termiczne jest bardzo trudne, gdyż w miejscach gdzie przymocowane są przyrządy mocy trzeba odprowadzić moc o gęstości powyżej 100 W/cm$^2$, przy zachowaniu maksymalnej temperatury złącza poniżej 150°C. Strumień cieplny wytwarzany wewnątrz przyrządów mocy przekracza często 3000 W/cm$^3$. Główna część

generowanego ciepła jest odprowadzana przez radiator, na którym umocowany jest układ hybrydowy. Tak więc należy dążyć do minimalizacji rezystancji termicznej pomiędzy złączem a podstawą obudowy układu. Hybrydowy układ mocy można najprościej przedstawić w postaci pokazanej na rysunku 7.




Rys. 7. Schemat typowego układu hybrydowego dużej mocy
a — płytka krzemu, b — warstwa rozprzestrzeniająca ciepło, c — podłoże, d — obudowa, e — warstwy lutowia lub kleju, f — metalizacja podłoża

Współczynniki rozszerzalności cieplnej materiałów używanych do budowy układu są przedstawione w Tablicy 1:

Tablica 1

| MATERIAŁ | krzem | tlenek glinu ($Al_2O_3$) | miedź |
|---|---|---|---|
| WSP. ROZSZ. CIEPLNEJ $\alpha$ | $3 \cdot 10^{-6} K^{-1}$ | $6 \cdot 10^{-6} K^{-1}$ | $18 \cdot 10^{-6} K^{-1}$ |

Jak widać z powyższego porównania, można spodziewać się uszkodzeń wynikających z ich różnic. W praktyce, pomiędzy krzemem a $Al_2O_3$ umieszcza się warstwę metalu „b", która odgrywa rolę tłumika sił wynikających z różnicy rozszerzalności cieplnej tych dwóch materiałów. Oprócz tego ułatwia ona rozprzestrzenianie strumienia cieplnego, który wnika dzięki niej na znacznie większej powierzchni w podłoże.

Materiałem używanym najczęściej do tego celu jest molibden ($\alpha = 5 \cdot 10^{-6} K^{-1}$). Ma on gorszy współczynnik przewodności termicznej niż miedź, i w praktyce w tego typu układzie jako lutowie e1, e2 używa się past miedzianych. Grubość warstw lutowia e1, e2 musi być zoptymalizowana, w celu zapewnienia jak najlepszego przepływu strumienia cieplnego i maksymalnej odporności na cykle cieplne.

W czasie swojego życia, układ hybrydowy znajduje się w dwóch stanach funkcjonowania odpowiadających pracy i spoczynkowi. W czasie pracy podlega on impulsowym przegrzaniom cieplnym, odpowiadającym impulsowym pobudzeniom elektrycznym. W wyniku tego następuje stopniowa degradacja właściwości złącz lutowia na skutek zjawiska zwanego zmęczeniem cieplnym. Następuje stopniowe zwiększenie rezystancji termicznej i w konsekwencji układ ulega zniszczeniu.

## 3. METODY MODELOWANIA I ANALIZY TERMICZNEJ

Przed wykonaniem prototypu układu hybrydowego, co jest niezwykle kosztowne ze względu na konieczność wykonania masek, konstruktor powinien określić rozkład temperatury w układzie dla zadanej w specyfikacji maksymalnej wartości wydzielanej mocy. Metody obliczania temperatury w układach hybrydowych można podzielić na:

a) Metody przybliżone oparte na pojęciu rezystancji termicznej.

b) Metody analityczne pozwalające na otrzymanie rozwiązania w postaci rozwinięcia w szereg.

c) Metody numeryczne.

d) Metody analityczno-numeryczne.

ad. a) W przypadku jednowymiarowego przepływu ciepła można stosować analogię pomiędzy rezystancją termiczną i elektryczną, i metoda ta ze względu na swoją prostotę jest często stosowana przez praktyków. Niestety w rzeczywistości przepływ ciepła jest praktycznie zawsze trójwymiarowy i w przypadku źródeł mocy o bardzo małej powierzchni w porównaniu z grubością struktury może ona dawać wyniki obarczone dużym błędem (nawet kilka tysięcy procent! [19,26]). Jej zastosowanie ograniczone jest zatem do przypadków, w których wydzielanie mocy następuje na dużej powierzchni układu.

ad. b) Rozwiązanie analityczne równania przewodnictwa cieplnego jest możliwe w przypadku, gdy zastosowane założenia upraszczające pozwalają sprowadzić je do postaci równania Laplace'a. Rozwiązanie analityczne można otrzymać wówczas przy warunkach brzegowych typu Neumanna lub Dirichleta w postaci szeregu Fouriera, szeregu funkcji Bessela, czy też poprzez zastosowanie funkcji Greena [14]. Metody te są bardzo żmudne, gdyż dla każdego szczególnego przypadku warunków brzegowych istnieje inna postać rozwiązania, a co za tym idzie dowolna modyfikacja układu wymaga powtórzenia całego procesu obliczeniowego.

ad. c) Metody numeryczne, jak na przykład metoda różnic, czy też elementów skończonych, mogą być stosowane do układów o dowolnej geometrii i o dowolnym przestrzennym rozkładzie generacji ciepła pod warunkiem poprawnego zdefiniowania warunków brzegowych, a w przypadku symulacji dynamicznej, również i początkowych.

Podobnie jak w przypadku metod analitycznych każda zmiana geometrii, czy też gęstości mocy wydzielanej w układzie, powoduje konieczność ponownego rozpoczynania procesu obliczeń. Czas obliczeń jest bardzo długi i metody te stosowane są raczej do symulacji, a nie do projektowania układu[1].

ad. d) W celu uniknięcia wad wyżej wymienionych metod można zastosować metodę analityczno-numeryczną [17,26], która dla specyficznego przypadku struktury hybrydowej (patrz rys. 7) pozwala na szybką i dokładną analizę. Dzięki dużej szybkości obliczeń możliwą staje się jednoczesna analiza termiczna i elektryczna, a co za tym

[1] W przeciwieństwie do jednorazowej symulacji, proces projektowania polega na wielokrotnym powtórzeniu obliczeń w celu zoptymalizowania struktury układu w funkcji jego parametrów.

idzie, uwzględnienie zjawisk sprzężenia elektrotermicznego. Problem ten jest szczególnie krytyczny w przypadku tranzystorów bipolarnych. W pracach [20,22,26] problem ten został przedyskutowany szczegółowo. Tutaj ograniczymy się do przedstawienia pełnej analizy układu hybrydowego przeznaczonego do praktycznej realizacji.

## 4. ANALIZA CIEPLNA UKŁADU

### 4.1. WPROWADZENIE

Klasyczna metoda projektowania układu hybrydowego ze względu na jego właściwości cieplne, polegała do tej pory na uproszczonych obliczeniach, zakładających jednowymiarowy przepływ ciepła i na wstępnym obliczeniu jego rezystancji termicznej. Następnie wykonywano pomiary na pierwszym praktycznym modelu układu i model ten modyfikowano aż do osiągnięcia żądanych właściwości.
Jeżeli materiały i technika montażu jest już dobrze znana, można zastosować metody projektowania komputerowego (C.A.D.) w celu optymalizacji termicznej układu. Jej celem nie jest określenie temperatury w każdym punkcie układu z dokładnością do 0.01°C, lecz wypracowanie reguł projektowych, które pozwolą na skrócenie czasu i kosztu projektu poprzez znaczną redukcję ilości próbnych układów.

Problem modelowania właściwości termicznych modułu hybrydowego i jego wpływ na wybór ostatecznej wersji układu zostanie przedstawiony na przykładzie projektu łącznika 600V/60A wykonanego przez SSM (SAGEM – SAT – MICROELECTRONIQUE) (patrz rys. 5) [10,22].

### 4.2. HYBRYDOWY ŁĄCZNIK MOCY. ZAŁOŻENIA PROJEKTOWE

W przypadku układów elektronicznych dużej mocy, badania nad optymalną strukturą układu prowadzone są dopiero po pełnej analizie elektrycznej pracy przyszłego modułu. Pozwala ona na wybranie właściwych elementów mocy do montażu w układzie, oraz na określenie przewidywanych całkowitych strat mocy. Założone właściwości elektryczne projektowanego łącznika mocy są następujące [13,18,24]:

a) możliwość pracy jako przerywacz okresowy (chopper), ramię falownika, lub też falownik w układzie mostka,

b) w konfiguracji przerywacza dopuszczalny średni prąd przewodzenia wynosi 60 A,

c) maksymalne napięcie blokowania jest równe 600 V,

d) maksymalna częstotliwość komutacji wynosi 100 kHz.

### 4.3. WYBÓR TRANZYSTORÓW MOCY

Pierwotna koncepcja zakładała zastosowanie tranzystora bipolarnego (MEDL DT47) i osiągnięcie parametrów: 600 V, 60 A. Po wykonaniu analizy termicznej układu

i komputerowej symulacji sprzężeń elektrotermicznych [8,20,22,23] okazało się, że technologia bipolarna prowadzi do występowania zjawiska wtórnego przebicia termicznego, które doprowadzało do zniszczenia próbnych egzemplarzy układu.

Problemy wtórnego przebicia termicznego występujące w tranzystorach bipolarnych spowodowały zmianę pierwotnej koncepcji projektowej. Po wstępnej analizie, opartej na znajomości standardowej technologii firmy SSM (SAGEM – SAT – MICROELECTRONIQUE) (ceramika alundowa) [9], zaproponowano zastosowanie 12 tranzystorów MOSFET (MTC 8N60), połączonych w cztery grupy po trzy tranzystory. Po uwzględnieniu kształtu prądu przepływającego przez poszczególny tranzystor i jego charakterystyk, oszacowano moc strat w pojedynczym elemencie na 25 W, co daje łączne straty mocy rzędu 300 W w całym module.

### 4.4. WYBÓR OPTYMALNEJ STRUKTURY UKŁADU

W celu optymalizacji układu należy rozwiązać następujące problemy [10]:

a) wybór technologii ($Al_2O_3$, czy też AlN)

b) w przypadku technologii $Al_2O_3$, wybór optymalnej powierzchni i grubości warstwy miedzi służącej do rozprzestrzeniania ciepła (warstwa „b" rys. 7)

c) optymalizacja rozmieszczenia elementów mocy

d) optymalizacja struktury termicznej układu (dobranie odpowiednich grubości poszczególnych warstw i powierzchni całego układu)

e) wybór materiału i struktury radiatora

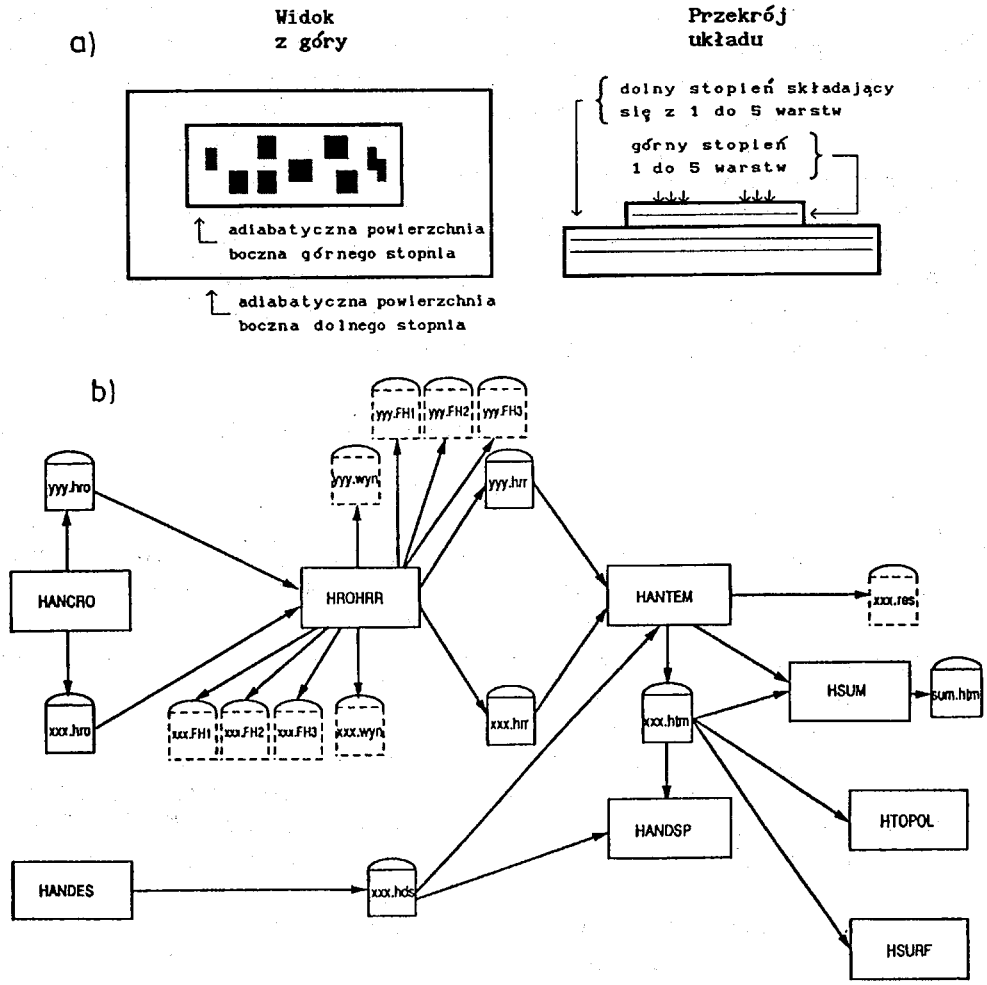
f) komputerowa symulacja różnych wersji kompletnego układu

W celu symulacji wyżej wymienionych problemów opracowano program PYRTHERM [25] (oparty na wcześniej opracowanym programie NAPOM [21]), który pozwala na analizę układów o strukturze przedstawionej na rysunky 8a. Jego schemat blokowy pokazano na rysunku 8b.

Jako przykład zastosowania programu rozważmy problem wyboru rozmiarów warstwy rozprzestrzeniającej ciepło w przypadku zastosowania technologii klasycznej. Grubość tej warstwy została ustalona na 0.5 mm lub 1.0 mm. Przykład pokazany na rys. 9 dotyczy poszukiwania optymalnej powierzchni warstwy miedzi (TAB). Na podstawie przedstawionego wykresu widać wyraźnie, że zwiększanie powierzchni tej warstwy powyżej pewnej granicy nie powoduje dalszego zmniejszania rezystancji termicznej układu. W praktyce konstruktor stara się zawsze określić, jaka najmniejsza powierzchnia miedzi może być jeszcze zastosowana w układzie. Wyniki symulacji z rys. 9 dają jednoznaczną odpowiedź na to pytanie.

Założenia projektowe z punktu 4.2 określają globalną rezystancję termiczną układu na 0.2 k/w. Odpowiada ona maksymalnej temperaturze złącza 150°C, przy wydzielaniu mocy 300 W i zewnętrznej temperaturze obudowy 90°C [13].

Przy projekcie technologicznym należy uwzględnić możliwości wykonawcze oraz koszty produkcji. Wybór powinien więc paść na możliwie standardową i sprawdzoną technologię. W naszym przypadku chcemy używać lutowia srebro – cyna – ołów, które jest najczęściej stosowane przez S.A.T., oraz stosować elementy (podłoże, obudowa, radiator) o standardowych wymiarach. Uwzględnienie tych ograniczeń




Rys. 8. a) Widok z góry oraz przekrój struktury, którą można symulować przy pomocy programu PYRTHERM, b) Ogólny schemat blokowy programu

prowadzi do wyboru szeregu wariantów układu kompatybilnych z możliwościami technologicznymi wykonawcy.

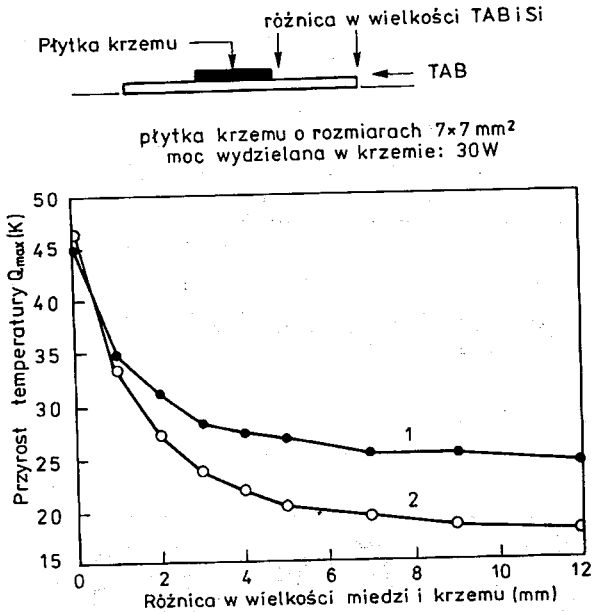
Projekt sprowadza się więc do:

a) Wyboru podstawy (obudowy) układu, która może być wykonana z miedzi lub stali.

b) Wyboru podłoża, które może być wykonane klasycznie w postaci dwóch warstw miedź/$Al_2O_3$ (odpowiednio grubości 0.5 mm i 0.64 mm) lub też z jednej metalizowanej warstwy AlN o grubości 0.7 mm lub 1 mm (pomijamy grubość metalizacji).

c) W przypadku wyboru technologii klasycznej trzeba określić wymiary podłużne warstwy miedzi pełniące rolę „rozprzestrzeniacza" ciepła.

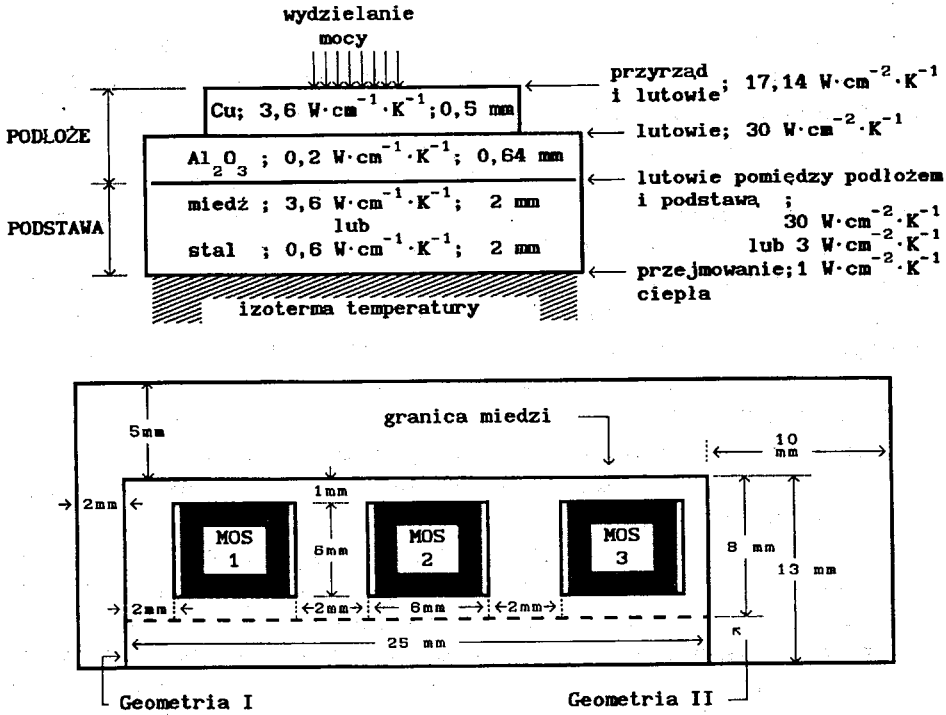
Układy przyjęte do symulacji przedstawione są na rysunku 10 (technologia klasyczna) i rysunku 11 (technologia AlN). Ze względu na symetrię ograniczono się do



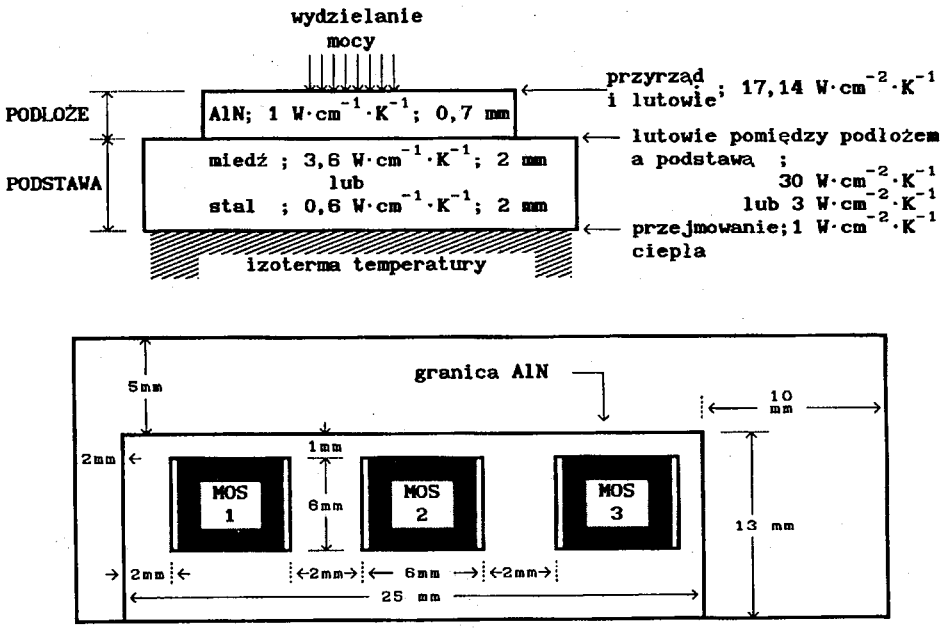

Rys. 9. Przykład obliczeń z wykorzystaniem programu PYRTHERM. Poszukiwanie optymalnej wielkości warstwy miedzi (TAB)

1 – grubość warstwy miedzi 0.5 mm

2 – grubość warstwy miedzi 1.0 mm




Rys. 10. Widok z góry oraz przekrój układu wykonanego w technologii klasycznej ($Cu/Al_2O_3$) przyjętego do wstępnej symulacji




Rys. 11. Widok z góry oraz przekrój układu wykonanego w technologii AlN przyjętego do wstępnej symulacji

pokazania (i do symulacji) 1/4 struktury układu. Pozwoliło to na znaczne skrócenie czasu obliczeń. Rezystancję termiczną płytek krzemu wraz z lutowiem uwzględniono jako rezystancję jednowymiarową o wartości 0.058 K·cm²/W (g = 17.14 W/(K·cm²). Konduktancje termiczne poszczególnych warstw lutowia (grubość lutowia srebro – cyna – ołów wynosi 80 $\mu$m) przyjęto dla dwóch skrajnych przypadków. Wartość teoretyczna konduktancji termicznej lutowia idealnego wynosi 30 W/(K·cm²), a dla lutowia zdegradowanego przyjęto 3 W/(K·cm²). We wszystkich przypadkach grubość podstawy niezależnie od wyboru miedzi, czy też stali jest 2 mm. We wszystkich symulacjach założono, że pomiędzy podstawą modułu, a układem chłodzenia znajdującym się w temperaturze odniesienia istnieje warstwa przejmowania ciepła o konduktancji termicznej 1 W/(K·cm²). Odpowiada ona przypadkowi przytwierdzenia modułu za pomocą tłuszczu termicznego.

W Tablicy 2 [13] przedstawiono wyniki 12 symulacji rezystancji termicznej układu (w K/W) wyrażonej w postaci stosunku przyrostu temperatury punktu najcieplejszego do całkowitej mocy wydzielanej w układzie.

Analiza Tablicy 2 prowadzi do wniosku, że osiągnięcie założonej wartości globalnej rezystancji termicznej układu poniżej 0.2 K/W jest możliwe w każdej z wybranych konfiguracji układu. Zastosowanie podstawy stalowej prowadzi jednak do znacznego zwiększenia rezystancji termicznej modułu i w przypadku degradacji właściwości termicznych lutowia może doprowadzić do przekroczenia założonej wartości Rth = 0.2 K/W. Należy również zauważyć, że w przypadku klasycznej technologii ($Cu/Al_2O_3$) wielkość powierzchni miedzi służącej do rozprzestrzeniania ciepła nie jest szczególnie krytyczna.

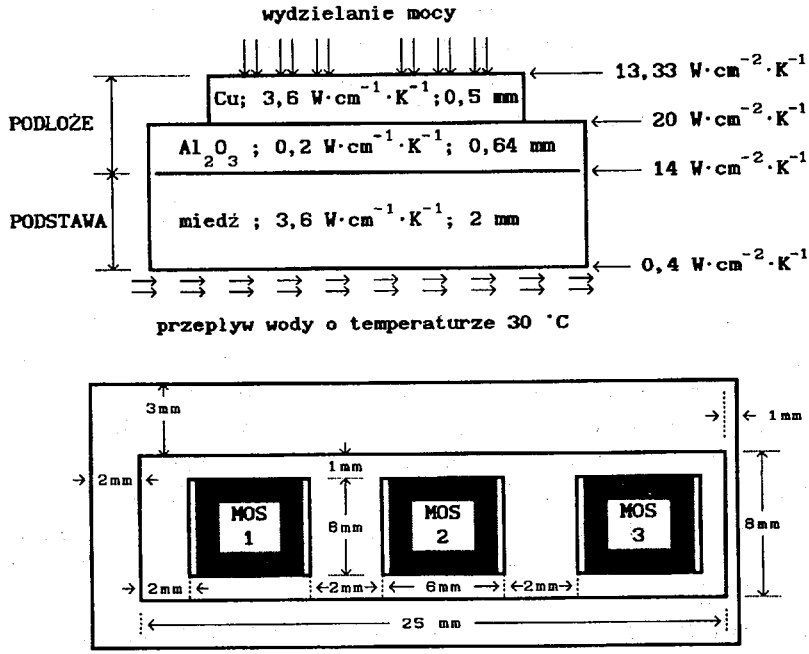
Tablica 2

**WSTĘPNY PROJEKT: Podstawowe wyniki. Rezystancje termiczne w (K/W) obliczone dla poszczególnych hipotez**

| WARUNKI SYMULACJI ↓ | Geometria → Struktura → Podstawa ↓ | geometria I „$Cu/Al_2O_3$" | geometria II „$Cu/Al_2O_3$" | geometria I „AlN" |
|---|---|---|---|---|
| Układ idealny pod względem termicznym | stalowa | 0,157 | 0,172 | 0,152 |
| | miedziana | 0,112 | 0,124 | 0,099 |
| Uwzględnienie Rth pomiędzy podłożem a podstawą | stalowa | 0,193 | 0,207 | 0,191 |
| | miedziana | 0,145 | 0,150 | 0,145 |

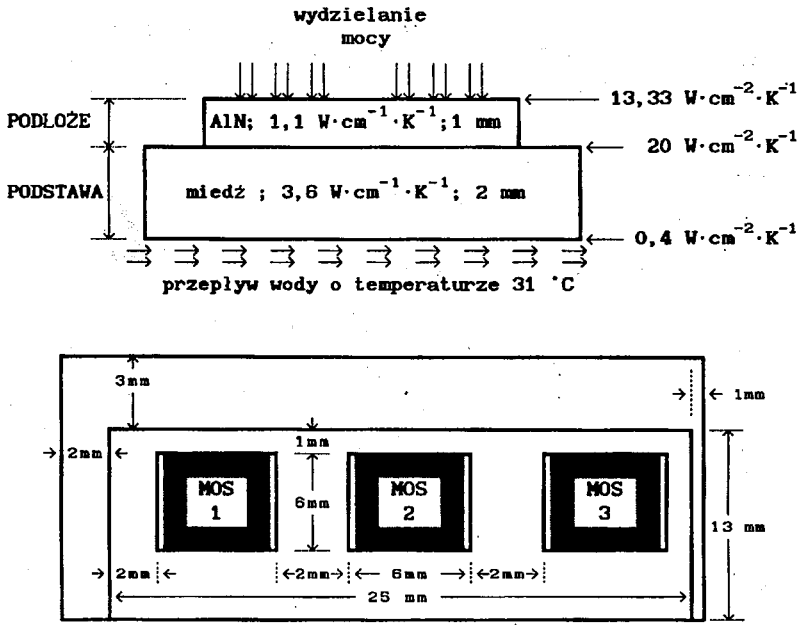
Najlepsze wyniki daje zastosowanie podłoża z AlN, pomimo faktu, że wrażliwość rezystancji termicznej na degradację lutowia jest w tym przypadku większa niż przy zastosowaniu technologii klasycznej. Na podstawie wstępnego projektu przedstawionego powyżej do praktycznej realizacji wybrano następujące rozwiązania:

a) Układ oparty na technologii $Cu/Al_2O_3$, z podstawą miedzianą, tranzystory MOS połączone są w grupach po trzy na płytce miedzi o szerokości 8 mm (geometria II na rysunku 12).




Rys. 12. Dane do symulacji modułu $Cu/Al_2O_3$. Struktura termiczna i chłodzenie

b) Układ oparty na technologii AlN z podstawą miedzianą. Położenie tranzystorów i wymiary takie, jak na rysunku 13.




Rys. 13. Dane do symulacji modułu AlN. Struktura termiczna i chłodzenie

## 4.5. KOŃCOWA OCENA WYBRANYCH ROZWIĄZAŃ

W celu właściwej oceny wybranych rozwiązań i wykonania końcowej symulacji komputerowej należy określić możliwie dokładnie parametry technologiczne układu. Największy problem sprawia zawsze ocena jakości i właściwości termicznych stosowanego lutowia. W przypadku współpracy z określonym producentem układów hybrydowych można dokonać charakteryzacji stosowanej technologii przy wykorzystaniu specjalnych próbek zawierających rozmaite kombinacje połączeń poszczególnych warstw. Metoda tej charakteryzacji była już przedmiotem publikacji [9] i nie będzie tu omawiana. Ograniczymy się tylko do przedstawienia wyników końcowych zawartych w Tablicy 3:

Przy uwzględnieniu rzeczywistych wartości parametrów zawartych w Tablicy 3, dokonano końcowej symulacji układu w założonym idealnym układzie chłodzenia (izotermiczny radiator, ze współczynnikiem przejmowania 1 W/(K·cm²). Otrzymano wartość rezystancji termicznej 0.13 K/W dla układu $Al_2O_3$ i 0.108 K/W dla układu AlN, które są bliskie wartościom z Tablicy 2 dla układu idealnego pod względem

Tablica 3

**Charakterystyka technologii modułu. Podstawowe wyniki.**

| | |
|---|---|
| lutowie pomiędzy przyrządem a podłożem | $g = 20\ W \cdot cm^{-2} \cdot K^{-1}$ |
| lutowie pomiędzy miedzią a $Al_2O_3$ podłoże „$Cu/Al_2O_3$" | $g = 20\ W \cdot cm^{-2} \cdot K^{-1}$ |
| lutowie pomiędzy podłożem a podstawą podłoże „$Cu/Al_2O_3$" | $g = 14\ W \cdot cm^{-2} \cdot K^{-1}$ |
| lutowie pomiędzy podłożem a podstawą podłoże „AlN" | $g = 20\ W \cdot cm^{-2} \cdot K^{-1}$ |
| przewodność cieplna $Al_2O_3$ | $k = 0{,}2\ W \cdot cm^{-1} \cdot K^{-1}$ |
| przewodność cieplna „AlN" | $k = 1{,}1\ W \cdot cm^{-1} \cdot K^{-1}$ |

termicznym. Następnym etapem jest symulacja modułów w realnych warunkach pracy, odpowiadających przyszłym warunkom pomiarowym układu.

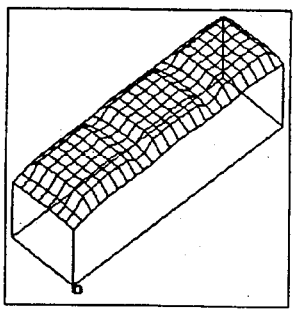
Na rysunku 12 i 13 przedstawiono dane przyjęte do symulacji układów wybranych do realizacji. Ze względu na symetrię ograniczono się do analizy 1/4 modułu. Wartość współczynnika przejmowania ciepła była w praktycznym układzie nieliniowa. W części lewej (gdzie wstrzykiwana była woda) wynosi on 0.8 W/(K·cm²), a w części prawej (odpływ wody) 0.4 W/(K·cm²). Do obliczeń przyjęto więc prawą stronę modułu (najgorszy przypadek), a więc wybrano współczynnik przejmowania ciepła jako równy 0.4 W/(K·cm²). Otrzymane mapy temperatury dla obu układów (przy założeniu wydzielania mocy 100 W) przedstawiono na rysunku 14. Obliczona rezystancja termiczna dla układu $Cu/Al_2O_3$ wynosi 0.216 K/W, a dla układu AlN 0.204 K/W. Wynik ten wskazuje na pozornie niewielką przewagę technologii AlN. Przekroczenie wartości maksymalnej rezystancji termicznej wynika z przyjęcia niezbyt dobrego układu chłodzącego, odpowiadającego ściśle warunkom doświadczalnym.

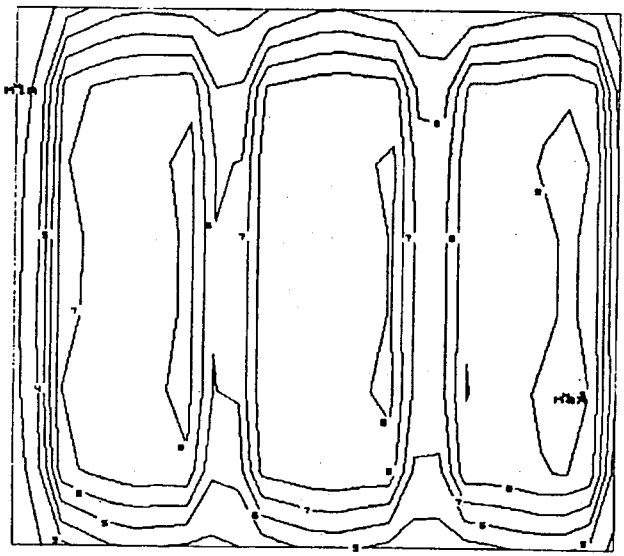
Po praktycznym wykonaniu modułów i zamontowaniu ich w będącym do dyspozycji układzie chłodzącym otrzymano mapy temperatury przedstawione na rysunku 15 dla modułu $Cu/Al_2O_3$, oraz na rysunku 16 dla modułu AlN. Pomiary wykonano przy pomocy termografu komputerowego HUGHES TVS 4100, po otwarciu układu i pomalowaniu go czarną farbą dla otrzymania jednorodnego współczynnika emisji w podczerwieni.

Asymetria otrzymanych map temperatury wynika z różnicy współczynnika przejmowania ciepła „h" dla lewej i prawej strony układu. Oprócz tego w przypadku technologii $Al_2O_3$, widać wyraźnie punkty ciepłe występujące po lewej stronie układu i nie dające się wytłumaczyć inaczej jak tylko występowaniem „dziur" w lutowiu w tych miejscach. Występowanie tych anomalii jest typowe dla układów opartych na technologii $Cu/Al_2O_3$, gdyż istnienie dodatkowej warstwy komplikuje proces technologiczny. Pomijając te defekty, pomierzone wartości rezystancji termicznych wynoszą 0.22 K/W dla struktury $Cu/Al_2O_3$ i 0.195 K/W dla struktury AlN. Są on więc bardzo bliskie rezultatom otrzymanym z sytuacji komputerowej

a)

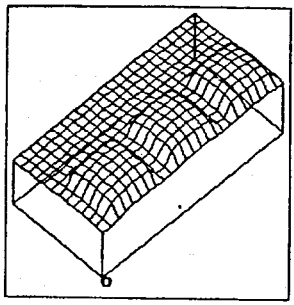
T = [ 15.612 ; 21.569 ]
X = [ 1.0 ; 25.0 ]
Y = [ 1.0 ; 8.0 ]
Isotherms values:
[1] 15.841 [6] 19.278
[2] 16.529 [7] 19.965
[3] 17.216 [8] 20.653
[4] 17.903 [9] 21.340
[5] 18.591

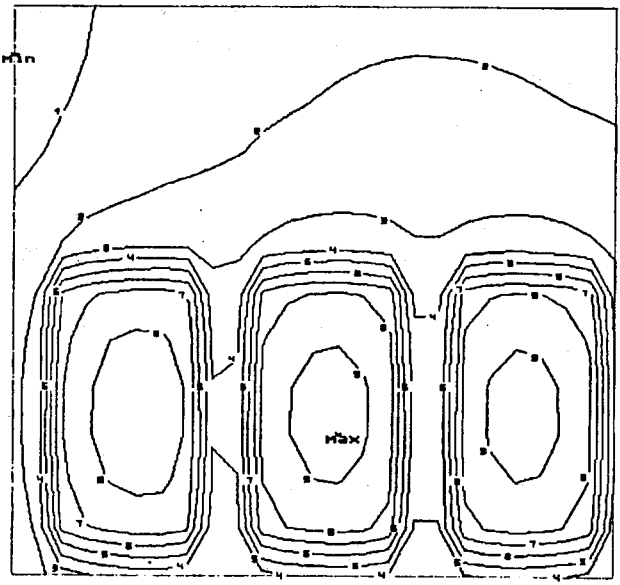




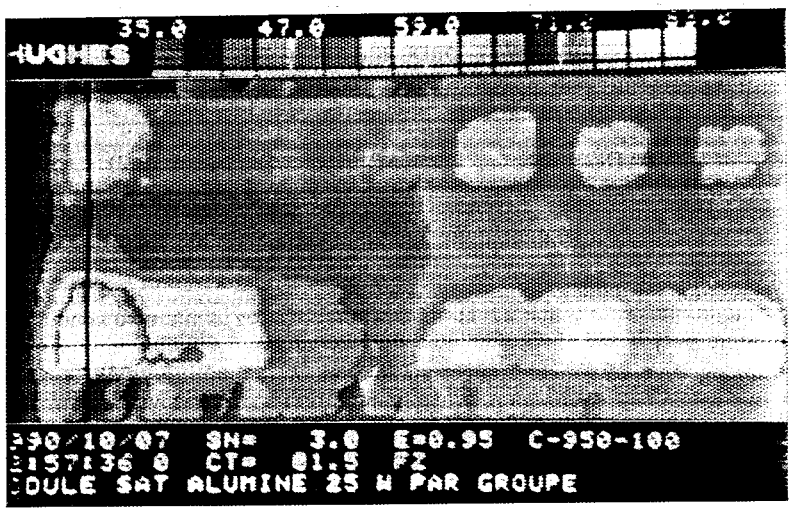

b)

T = [ 13.581 ; 20.442 ]
X = [ 1.0 ; 25.0 ]
Y = [ 1.0 ; 13.0 ]
Isotherms values:
[1] 13.845 [6] 17.803
[2] 14.637 [7] 18.595
[3] 15.428 [8] 19.387
[4] 16.220 [9] 20.178
[5] 17.012

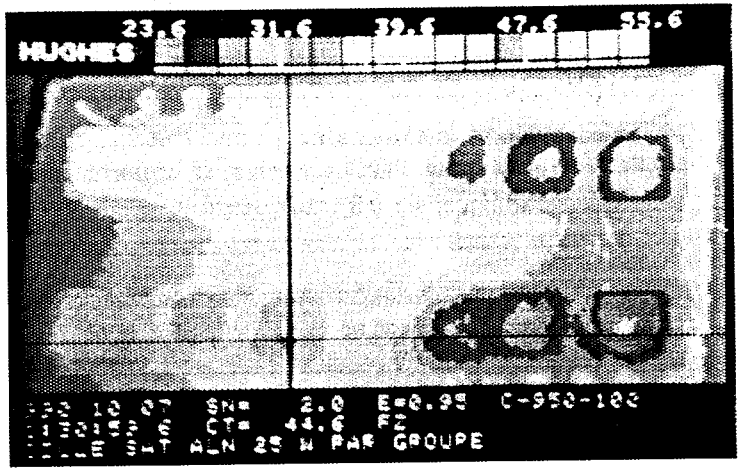





Rys. 14. Symulacja komputerowa układów przy założeniu wydzielania mocy 8.33W w każdym z tranzystorów (w sumie 100W)
a) Układ oparty na technologii $Cu/Al_2O_3$, b) Układ oparty na technologii AlN




Rys. 15. Pomierzona mapa temperatury modułu „$Cu/Al_2O_3$"
- kierunek przepływu wody: z lewej strony na prawą
- temperatura wody: 35°C
- moc wydzielana w pojedynczym tranzystorze 8.33W (moc całkowita 100W)




Rys. 16. Pomierzona mapa temperatury modułu „AlN"
- kierunek przepływu wody: z lewej strony na prawą
- temperatura wody: 32°C
- moc wydzielana w pojedynczym tranzystorze 8.33W (moc całkowita 100W)

przy uwzględnieniu rzeczywistych parametrów układu Tablicy 4 przedstawiono podsumowanie otrzymanych wyników:

Tablica 4

| UKŁAD ↓ | ZAŁOŻENIE[a)] | POMIAR[b)] (1) | POMIAR[b)] (2) | SYMULACJA[b)] (1) |
|---|---|---|---|---|
| | [K/W] | [K/W] | [K/W] | [K/W] |
| „AlN" | 0.108 | 0.195±0.01 | 0.176±0.01 | 0.204 |
| $Cu/Al_2O_3$ | 0.130 | 0.22±0.015 | 0.21±0.015 | 0.216 |

gdzie: a) zamocowanie układu na radiatorze idealnym poprzez tłuszcz termiczny o konduktancji termicznej $h=1$ W/(K·cm²).

b) chłodzenie wodne ze współczynnikiem przejmowania ciepła $h=0.4$ W/(K·cm²).

(1) stosunek temperatury najcieplejszego punktu z grupy sześciu tranzystorów znajdujących się po prawej stronie do całkowitej mocy wydzielanej w układzie.

(2) stosunek średniej z wartości maksymalnych temperatur każdego z sześciu tranzystorów znajdujących się po prawej stronie do całkowitej mocy wydzielanej w układzie.

Jak wynika z Tablicy 4, różnice w wartościach rezystancji termicznej wynikają głównie z różnic w sposobie chłodzenia układu.

Tablica 5

| MODUŁ | $\Delta\theta$ max max. pomierzona wartość | $\Delta\theta$ min min. pomierzona wartość | $\Delta\theta$ max max. obliczona wartość | $\Delta\theta$ min min. obliczona wartość |
|---|---|---|---|---|
| $Cu/Al_2O_3$ | 25 °C (±1) | 19 °C (±) | 21,6 °C | 20,6 °C |
| AlN | 20 °C (±1) | 16 °C (±1) | 20,4 °C | 19,5 °C |

W Tablicy 5 przedstawiono maksymalne pomierzone i obliczone przyrosty temperatury dla obu modułów, a w Tablicy 6 różnice w pomierzonej i obliczonej wartości $\Delta\theta_{max}$ w grupie składającej się z 3 tranzystorów.

Tablica 6

| MODUŁ | Zmienność pomierzonej wartości $\Delta\theta$ max w grupie składającej się z trzech tranzystorów | Zmienność obliczonej wartości $\Delta\theta$ max w grupie składającej się z trzech tranzystorów |
|---|---|---|
| $Cu/Al_2O_3$ | 6 °C / 27% | 1 °C / 4,8% |
| AlN | 4 °C / 22% | 0,9 °C / 4,5% |

Rezultatem powyższej symulacji jest wybór układu opartego na technologii AlN, który poza 30% zmniejszeniem wartości rezystancji termicznej jest mniej wrażliwy na błędy technologiczne, gdyż zawiera o jedną warstwę lutowia mniej.

### 4.6. ANALIZA STANÓW PRZEJŚCIOWYCH

Po optymalnym zaprojektowaniu układu, przy zastosowaniu metod symulacji termicznej w stanie ustalonym, w niektórych przypadkach interesujące jest określenie odpowiedzi układu na chwilowe przeciążenie [11,12,13]. Symulacja ta nie ma jednak

wpływu na wybór optymalnej struktury termicznej, gdyż zawsze stan ustalony jest warunkiem początkowym do analizy stanu przejściowego i jego minimalizacja odpowiada dobraniu odpowiednich warunków pracy.

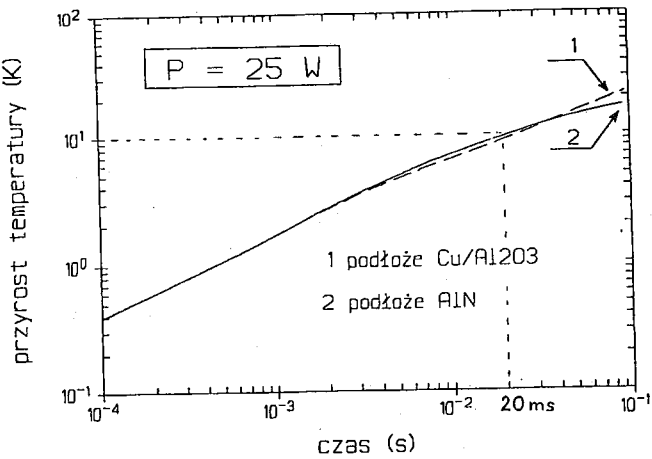
Rozpatrzmy w tym miejscu odpowiedź termiczną układu na krótkotrwałe przeciążenie. Jeżeli czas obserwacji jest bardzo krótki, to znaczy gdy strumień cieplny nie zdąży jeszcze rozprzestrzenić się pod górną warstwą lutowia poza obszar wstrzykiwania mocy, można wówczas obliczać przejściową odpowiedź termiczną układu na podstawie aproksymacji jednowymiarowej [12,22,26]. Parametry termiczne zastosowanych materiałów do obliczania odpowiedzi przejściowej układu przedstawiono w Tablicy 7.

Tablica 7

**Właściwości termiczne zastosowanych materiałów**

| MATERIAŁ ↓ | przewodność termiczna ($W \cdot cm^{-1} \cdot K^{-1}$) | ciepło właściwe izochoryczne ($J \cdot cm^{-3} \cdot K^{-1}$) |
|---|---|---|
| krzem | 1,2 | 1,7 |
| miedź | 3,6 | 3,4 |
| $Al_2O_3$ | 0,2 | 2,6 |
| AlN | 1,1 | 2,3 |

Krzywe odpowiedzi przejściowej na jednostkowy impuls mocy wydzielanej w pojedynczym tranzystorze równy 25 W (300 W w całym układzie) przedstawiono na rysunku 17. Jak widać, w czasie od 0 do 100 ms są one praktycznie identyczne dla obu przyjętych rozwiązań, tak więc wybór jednej lub drugiej technologii nie ma zasadniczego wpływu na zachowanie się układu w stanie przejściowym. Dokładne obliczenie odpowiedzi przejściowej dla czasów średnich i długich wymaga oczywiście przeprowadzenia analizy trójwymiarowej. Zauważmy jednak, że w przypadku zwarcia, czy




Rys. 17. Odpowiedź termiczna w stanie nieustalonym dla każdego z wybranych rozwiązań

też przeciążenia maksymalna temperatura złącza zostanie przekroczona w bardzo krótkim czasie. Na przykład, jak wynika z rysunku 17, dla impulsu mocy równego 250 W w jednym tranzystorze, przyrost temperatury o 100°C zostanie osiągnięty już po 20 ms. Tak więc analiza jednowymiarowa stanu nieustalonego wywołanego zwarciem, czy też uszkodzeniem modułu jest w pełni poprawna.

### 4.7. ANALIZA SPRZĘŻEŃ ELEKTROTERMICZNYCH

Problem ten został już częściowo poruszony w p. 2 i był dyskutowany szczegółowo w poprzednich pracach [20,22,26]. Jest on szczególnie krytyczny w przypadku stosowania technologii bipolarnej. W układach z tranzystorami MOSFET sprzężenia elektrotermiczne nie powodują zmian warunków pracy wiodących w kierunku zniszczenia układu, stąd obserwuje się ostatnio coraz częstsze stosowanie tej technologii w hybrydowych układach dużej mocy. Wybór tej technologii w przypadku zaprezentowanego projektu łącznika spowodowany był właśnie problemami wtórnego przebicia termicznego w bipolarnych tranzystorach dużej mocy.

## 5. PRÓBY PRZYSPIESZONEGO STARZENIA

Przed podjęciem produkcji seryjnej układ poddawany jest próbom przyspieszonego starzenia, składającym się z:

a) szoków termicznych
b) wielokrotnych cykli termicznych
c) pracy w podwyższonej temperaturze
d) wibracjom

W zależności od późniejszych zastosowań układu, próby te są mniej lub bardziej surowe.

ad.a) Szoki termiczne polegają na trzymaniu niezasilonego układu w niskiej temperaturze (na przykład $-55°C$ przez 5 minut), a następnie gwałtownemu przeniesieniu go do wysokiej temperatury ($+125°C$ na 5 minut). Próba ta wykazuje natychmiast zły projekt termomechaniczny układu, który ulega zniszczeniu już w ciągu pierwszych kilku cykli.

ad.b) Wielokrotne cykle termiczne polegają na trzymaniu niezasilanego układu w średniej temperaturze (na przykład $+55°C$ przez 10 minut), a następnie stopniowemu ogrzaniu go do wysokiej temperatury, utrzymaniu w niej przez pewien czas (na przkład w $+125°C$ przez 10 minut), ochłodzeniu do $-5°C$, itd.. Próba ta powoduje zmęczenie cieplne lutowia i wzrost rezystancji termicznej układu.

Cykle termiczne mogą być też wywoływane samonagrzewaniem się układu na skutek wydzielonej w nim mocy elektrycznej i chłodzenia go przez konwekcję wymuszoną. W czasie jednego cyklu temperatura złącza może wzrosnąć o $\Delta T = 150°C$. W tym przypadku, naprężenia mechaniczne powodują nie tylko zmęczenie cieplne lutowia, lecz mogą również spowodować wystąpienie pęknięć w krzemie. Próba ta może również wykazać degradację właściwości połączeń układu (przewody i ścieżki przewodzące).

ad. c) Próby pracy w podwyższonej temperaturze mogą spełniać funkcję wstępnego starzenia. Przyśpieszają one zniszczenie wadliwych przyrządów, lecz nie pozwalają na wyciąganie wniosków co do jakości układu.
ad. d) Wibracje mogą powodować wzrost naprężeń, a nawet pęknięcia w układzie. Powoduje to następnie wzrost lokalnej rezystancji termicznej i w czasie dłuższej pracy układ może ulec autodestrukcji.

## WNIOSKI

Rozwój układów elektronicznych dużej mocy, których nie można wykonać w postaci monolitycznego układu scalonego, doprowadził do realizacji układów hybrydowych zwanych „inteligentnymi" („smart"), z powodu realizacji przez nie bardzo skomplikowanych funkcji. Można przypuszczać, że najbliższe lata przyniosą dalszy gwałtowny rozwój produkcji tych układów, a ich stosowanie w przemyśle, a zwłaszcza w układach elektroniki samochodowej będzie coraz częstsze.

W celu umożliwienia wykonania poprawnego projektu układu, obserwuje się rozwój nowych, coraz szybszych i tańszych metod C.A.D. dla celów projektowania hybrydowych układów dużej mocy.

Wydzielanie dużej mocy powoduje wzrost temperatury pracy tych układów, i w związku z tym konieczność starannej optymalizacji ich struktury termicznej. Przedstawiony w artykule program PYRTHERM pozwala na wykonanie wstępnej analizy i wybranie właściwej technologii realizacji modułu. Jako przykład zastosowania programu przedstawiono realizację łącznika 600V/60A wykonanego przez SSM.

Pod względem technologicznym, duży postęp obserwuje się w stosowanych podłożach i ich metalizacji. Dla układów mocy podłożem coraz częściej używanym staje się azotek glinu (AlN).

Przedstawiona w artykule metoda projektowania układu hybrydowego jest w pełni ogólna. Zaprojektowany układ (patrz rysunek 5) jest już wdrożony do seryjnej produkcji przez SSM. Może on być z powodzeniem stosowany zarówno do celów cywilnych, jak i wojskowych.

Ostatnim problemem, który czeka jeszcze na rozwiązanie jest problem normalizacji obudów układów hybrydowych dużej mocy oraz układów „Smart Power".

## BIBLIOGRAFIA

1. B. J. B a l i g a: *Modern Power Devices.* Willey—Interscience, 1987 r.
2. B. J. B a l i g a: *Evolution of MOS—Bipolar Power Semiconductor Technology.* Proc. IEEE, 76, 4 (April 1988), pp. 409—418
3. C. C a z a o u l o u, R. B a r b o u: *Hybrides de puissance pour l'aérospatial.* Electronique de Puissance — 34, 08. 1989
4. G. C h a r i t a t: *Modélisation et réalisation de composants planar haute-tension.* Thése de Doctorat d' Etat (Sciences), 28.09.1990, UPS, Tolouse, France

5. *Connexions de circuits intégreés par microbilles fusibles* – notice interne, CEA–LETI, L/89
6. M. J. Declecq, D. Plummer: *Avalanche breakdown in high voltage DMOS devices.* IEEE Trans. Electron Devices, ED–23, 1976, pp. 1–6
7. Ph. Despagne: *Informacje własne*
8. J. M. Dorkel, E. Lachiver, A. Napieralski, Ph. Leturcq: *Modèles thermiques pour la conception de circuit hybrides de puissance.* RAPPORT FINAL LAAS N° 85331, Conratr MRT–SAT–LABINAL–LAAS, maj 1986
9. J. M. Dorkel, A. Napieralski, Ph. Leturcq, J. J. Picault: *La méthode des coefficients d'influence: Son application à l'identification des paramêters thermiques d'une technologie d'assemblage hybride.* Actes du 3ème Colloque sur la Thermique, l'Energie et l'Environnement, Perros–Guires Trégastel, France, 21–23 juin 1989
10. J. M. Dorkel, A. Napieralski, Ph. Leturcq, S. Tounsi: *Aide à la conception thermique d'un module interrupteur de puissance à transistor MOS.* RAPPORT FINAL LAAS N° 90975 dla SSM (SAGEM–SAT–MICROELECTRONIQUE), marzec 1990
11. J. M. Dorkel, A. Napieralski, Ph. Leturcq, S. Tounsi: *Outils et methodologie de la conception thermique des circuits de l'electronique automobile.* 4e Congrès International sur L'ELECTRONIQUE AUTOMOBILE, S.I.A., 4–6 Avril 1990, Toulouse
12. J. M. Dorkel, S. Tounsi, A. Napieralski, Ph. Leturcq: *Thermal C.A.D. of hybrid or integrated power circuits. DC state and T.R. analysis.* 3rd PRO–CHIP WORKSHOP, May 14, 1990, Paris
13. J. M. Dorkel, A. Napieralski, Ph. Letrcq, E. Mokrani, R. Picault, J. J. Simon: *Conception thermique d'un comutateur hybride de puissance.* Troisième Colloque National L'Electronique de Puissance du Futur EPF'90, Toulouse 10 au 12 Octobre 1990, pp. 109–114
14. M. Jacob: *Heat transfer.* vol. 1, New York, John Willey and Sons
15. Jessi BLR. – *Packaging Workshop.* LETI/Grenoble, France, 17–19. IV. 1990
16. J–C. Le Floch: *Etude de la fiabilité des composants électroniques. Le joint Pb–Sn en fatigue thermique.* ELECTRON BORDEAUX 1984
17. Ph. Leturcq, J. M. Dorkel, A. Napieralski, E. Lachiver: *A new approach to thermal analysis of power devices.* IEEE Trans. Electron Devices, vol. ED–34, No 5, May 1987, pp. 1147–1156
18. E. Mokrani, J. C. Houbart, R. Picault, J. J. Simon: *Comutateur hybride de puissance.* Troisième Colloque National L'Electronique de Puissance du Futur EPF'90, Toulouse 10 au 12 Octobre 1990, pp. 27–32
19. A. Napieralski, J. M. Dorkel, E. Lachiver, Ph. Leturcq: *Thermal characterization of cooling substrates for power semiconductors.* II nd International Conference on Power Electronics and Variable–speed Drives, 25–27 listopad 1986, Birmingham, Wielka Brytania, (str. 57–61)
20. A. Napieralski, J. M. Dorkel, Ph. Leturcq: *Electro–Thermal Coupling Simulation in High Power Bipolar Transistors.* NASECODE–87, 17–19 June 1987, Dublin, Ireland, (pp. 301–306)
21. A. Napieralski, F. Pompignac: *Thermal C.A.D. of power hybrid devices using light computer means.* IEE Proceedings on Electronic Circuits and Systems, Vol. 135, Pt. G, No. 1, Feb. 1988, (pp. 34–43)
22. A. Napieralski, Ph. Leturcq: *Simulation Electro–Thermique.* Journées GRECO CNRS Clamart, France, 21–22 Avril 1988, (8 pages)
23. A. Napieralski, J. M. Dorkel: *La thermique des composants et les couplages electro-thermiques.* Rapport LAAS N° 88372, dècembre 1988, 37 p.
24. A. Napieralski: *Współczesne układy hybrydowe dużej mocy. Modelowanie, perspektywy rozwoju i zastosowanie.* XIII Krajowa Konferencja „Teoria Obwodów i Układy Elektroniczne", Bielsko-Biała 16–18. X. 1990, str. 23–40
25. A. Napieralski, J. M. Dorkel, Ph. Leturcq: „PYRTHERM" Version 3.0, *Manuel de l'Utilisateur,* Novembre 1990, RAPPORT LAAS N° 90380
26. A. Napieralski: *Komputerowe projektowanie układów półprzewodnikowych dużej mocy ze szczególnym uwzględnieniem ich właściwości termicznych.* W. P. Ł., Łódź, 1988
27. D. E. Neil: *The direct bonding of copper to ceramic and resulting advantages and applications,* Colloque International sur les techniques de fabrication des circuits hybrides, Paris, Avril 1976

28. Power Compact: *Power hybrid technology,* internal notice
29. Power Compact: *Standard products catalog,* issued: dec. 1988
30. Sagem – Sat MICROELECTRONIQUE, *Commutateur hybride de puissance à MOSFET,* notice interne, 1990
31. C. Zardini, J. Pistre, F. Rodes, J–L, Aucoutourier, M. Monneraye, H. Baudry: *ISHM STRESA,* 1985
32. C. Zardini: *Informacje własne*

A. NAPIERALSKI

## HIGH POWER SEMICONDUCTOR CIRCUITS – TECHNOLOGY, ANALYSIS AND DESIGN

### Summary

Increasing power consumption and denser chip packaging of high complexity, high speed circuits give rise to thermal problems of the same nature as those encountered in the field of power electronics. Dissipated power density can exceed 100 W/cm$^2$. Thermal modeling of specific problems in such circuits is currently a subject of many research programs. In this paper basic problems of power hybrid circuits construction, modeling and designing are discussed. The all design process of an hybrid circuit containing 12 power MOS transistor is presented. Prospects of development and application of „Smart Power" circuit, which contains the power semiconductors and VLSI circuits in the same mount are discussed.

KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJI, 1992, 38, z. 1, ss. 181-195

621.314.57/.58

# Filtr aktywny wstrzykujący w zastosowaniu do wygładzania napięcia wyjściowego przekształtników tyrystorowych

PAWEŁ HEMPOWICZ

*Instytut Elektrotechniki Morskiej i Przemysłowej, Politechnika Gdańska*

*Otrzymano 1991.03.10*

*Autoryzowano do druku 1991.07.20*

Przedstawiono nową interpretację procesów fizycznych zachodzących w szeregowym filtrze wstrzykującym. Wyprowadzono bilans energetyczny filtru i wykazano, że filtr może mieć dużą sprawność energetyczną. Opisano przykład rozwiązania układu sterowania filtru. Przedstawiono wyniki praktycznej realizacji filtru zastosowanego do wygładzania napięcia za przekształtnikiem tyrystorowym.

## 1. WSTĘP

Do wygładzania napięcia wyjściowego tyrystorowych układów przekształtnikowych stosuje się filtry bierne. Filtry te powodują wydłużenie stałej czasowej układu, a jednocześnie nie zapewniają możliwości całkowitego wyeliminowania tej składowej napięcia wyjściowego, której przyczyną jest to, że w większości przekształtników napięcie wyjściowe, lub prąd wyjściowy mają postać impulsów prostokątnych.

Dużą wartość współczynnika filtracji bez jednoczesnego wydłużenia stałej czasowej układu można uzyskać stosując filtry aktywne wstrzykujące. Działanie filtrów wstrzykujących polega na wprowadzaniu do obwodu w którym filtr jest włączony porcji energii, przy czym ilość energii i czas jej wprowadzenia są zależne od sygnału sterującego. Energia ta może pochodzić z dodatkowego źródła, albo w części lub w całości może być uprzednio pobraną z obwodu do którego filtr jest włączony i zgromadzona w filtrze.

Rozróżnia się dwa typy aktywnych filtrów wstrzykujących: równoległe i szeregowe. Działanie filtru równoległego polega na sumowaniu prądów. Do prądu zawierającego składową zmienną którą należy wyeliminować, dodaje się prąd zmienny o przebiegu możliwie zbliżonym do niepożądanej składowej zmiennej prądu wygładzanego, lecz w przeciwnej fazie. Prąd sumaryczny może być pozbawiony niepożądanej składowej zmiennej. Filtry wstrzykujące równoległe są stosowane najczęściej do eliminowania wyższych harmonicznych prądu pobieranego przez przekształtniki z sieci zasilającej (1,2,3).

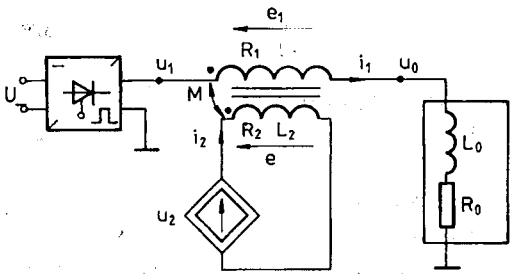
Działanie szeregowego filtru wstrzykującego polega na tym, że do napięcia zawierającego niepożądaną składową zmienną dodaje się SEM o przebiegu możliwie zbliżonym do tej składowej, lecz o przeciwnej fazie. W efekcie można spowodować, że napięcie za filtrem jest całkowicie wygładzone, lub zawiera składową zmienną o pożądanym przebiegu. Zmienną SEM wytwarza się w uzwojeniu, najczęściej nawiniętym na rdzeniu ferromagnetycznym.

Przyjmuje się, że wadą szeregowych filtrów wstrzykujących jest niska sprawność energetyczna i dlatego filtry te nie są stosowane w urządzeniach większej mocy, lub stosuje się je w takich przypadkach w których sprawność energetyczna nie ma istotnego znaczenia [4,5]. Dotychczas domeną stosowania szeregowych filtrów wstrzykujących są urządzenia niewielkiej mocy i w tych urządzeniach włącza się je najczęściej jako ostatni człon filtru złożonego, za filtrem biernym L lub LC.

W artykule przedstawiono nową interpretację zjawisk fizycznych zachodzących w szeregowym filtrze wstrzykującym. Celem artykułu jest wykazanie że szeregowy filtr wstrzykujący może mieć wysoką sprawność energetyczną i dzięki temu może być włączany za przekształtnikiem tyrystorowym jako jedyny filtr wyjściowy, bez poprzedzającego go filtru biernego. Dla zilustrowania możliwego do uzyskania stopnia filtracji napięcia wyjściowego przytoczono wyniki badań praktycznie zrealizowanych filtrów.

## 2. ZASADA DZIAŁANIA SZEREGOWEGO FILTRU WSTRZYKUJĄCEGO

Szeregowy filtr wstrzykujący zawiera nawinięte na rdzeniu ferromagnetycznym i dwa uzwojenia: główne o indukcyjności $L_1$ i dodatkowe o indukcyjności $L_2$ (rys. 1.). Uzwojenie główne włącza się między przekształtnik a odbiornik. Uzwojenie dodatkowe jest połączone do regulowanego źródła prądowego mającego dużą dynamiczną oporność wewnętrzną du/di dzięki której wartość prądu $i_2$ płynącego w uzwojeniu dodatkowym nie zależy od wartości SEM indukowanej w tym uzwojeniu, a jedynie od sygnałów sterujących doprowadzanych do źródła prądowego. Zmiany wartości prądu $i_2$ indukują w obu uzwojeniach SEM. Można wymusić taki przebieg zmian prądu $i_2$ przy którym SEM $e_1$ indukowana w uzwojeniu głównym dodając się do napięcia $u_1$ na wyjściu przekształtnika powoduje, że napięcie na odbiorniku $u_0$ ma wartość pożądaną, np. proporcjonalną do sygnału sterującego, a niezależną od chwilowej




Rys. 1. Schemat uproszczony aktywnego filtru wstrzykującego

wartości napięcia na wyjściu przekształtnika. Dotyczy to oczywiście tylko wartości chwilowych, średnia wartość napięcia na odbiorniku musi być równa średniej wartości napięcia na wyjściu przekształtnika, pomniejszonej o spadek napięcia na rezystancji głównego uzwojenia dławika.

W opisach filtru wstrzykującego (4,5,6,7,8,9) dławik o dwóch uzwojeniach jest traktowany jako transformator. Jest to założenie błędne. Traktując dławik filtru jako transformator i rozpatrując dla przykładu przypadek w którym filtr jest włączony bezpośrednio za przerywaczem prądu stałego, którego napięcie wyjściowe ma postać impulsów przedzielonych przerwami beznapięciowymi, należałoby przyjąć, że w czasie przerwy między impulsami odbiornik jest zasilany za pośrednictwem tego transformatora ze źródła prądowego zasilającego uzwojenia dodatkowe dławika. W efekcie moc chwilowa źródła prądowego musiałaby być zbliżona do mocy przekształtnika, a moc średnia — do połowy mocy przekształtnika. Takiemu założeniu przeczy bilans energetyczny filtru.

## 3. BILANS ENERGETYCZNY FILTRU WSTRZYKUJĄCEGO

### 3.1. BILANS ENERGETYCZNY OGÓLNY

Bilans energetyczny filtru wstrzykującego wyprowadza się wychodząc z równań zapisanych według II prawa Kirchoffa dla obwodów przedstawionych na rys. 1. Dla uproszczenia przyjęto, że rezystancje uzwojeń i rezystancja reprezentująca straty w rdzeniu dławika są pomijalnie małe, lub zostały przeniesione do odbiornika.

Dla obwodu obejmującego przekształtnik, uzwojenie główne dławika i odbiornik:

$$u_1 = e_1 + u_0 = L_1 \frac{di_1}{dt} + M \frac{di_2}{dt} + u_0 \tag{1}$$

$$W_{1t} = \int_0^t u_1 i_1 \, dt = \int_0^t L_1 i_1 \frac{di_1}{dt} dt + \int_0^t M i_1 \frac{di_2}{dt} dt + \int_0^t u_0 i_1 \, dt \tag{2}$$

$$W_{1t} = L_1 \int_{i_{10}}^{i_{1t}} i_1 \, di_1 + M \int_{i_{20}}^{i_{2t}} i_1 \, di_2 + \int_0^t u_0 i_1 \, dt; \tag{3}$$

gdzie: $W_{1it}$ — ilość energii wypływającej z przekształtnika w czasie od 0 do $t$
pozostałe oznaczenia — jak na rys. 1.

Ostatni człon równania (3) oznacza energię dostarczoną do odbiornika, pozostałe dwa człony — energię zgromadzoną w filtrze.

Jeśli prąd w uzwojeniu głównym dławika $i_1$ ma wartość stałą to pierwszy człon prawej strony równania (3) ma wartość zerową. Natomiast drugi człon może mieć wartość niezerową jeśli zmianie ulega wartość prądu $i_2$. Oznacza to że również przy stałej wartości prądu w uzwojeniu głównym, dławik może pobierać i gromadzić energię

z źródła zasilającego obwód w którym to uzwojenie jest włączone, lub oddawać zgromadzoną energię. Czynnikiem decydującym o przepływie energii są wtedy zmiany prądu w uzwojeniu dodatkowym $i_2$. Uzwojenie dodatkowe jest w tym dławiku elementem sterującym.

Dla obwodu obejmującego uzwojenie dodatkowe i źródło prądowe:

$$u_2 = L_2 \frac{di_2}{dt} + M \frac{di_1}{dt} \tag{4}$$

$$W_{2t} = \int_0^t u_2 i_2 \, dt = L_2 \int_0^t i_2 \frac{di_2}{dt} dt + M \int_0^t i_2 \frac{di_1}{dt} dt \tag{5}$$

$$W_{2t} = L_2 \int_{i_{20}}^{i_{2t}} i_2 \, di_2 + M \int_{i_{10}}^{i_{1t}} i_2 \, di_1, \tag{6}$$

gdzie: $W_{2t}$ – ilość energii wypływająca z źródła prądowego do dławika w czasie od 0 do $t$

Jeśli napięcie $u_1$ jest zmienne okresowo o okresie $T$, to

$$u_{10} = u_{1T} \quad i_{10} = i_{1T} \quad i_{20} = i_{2T}$$

i równania (3) i (6) dla całego okresu $T$ przyjmują postać:

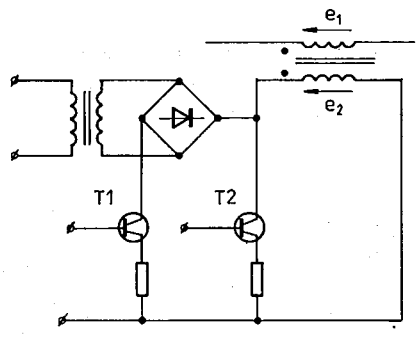
$$W_{1T} = \int_0^T u_0 i_1 \, dt \tag{7}$$

$$W_{2T} = 0 \tag{8}$$

gdzie: $W_{1T}$ i $W_{2T}$ – ilości energii wypływającej w ciągu całego okresu z przekształtnika z źródła prądowego.

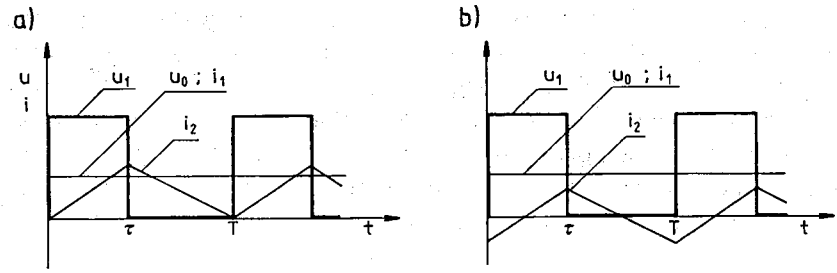
Z równania (7) wynika, że w obwodzie obejmującym głowne uzwojenie dławika cała energia w czasie pełnego okresu z przekształtnika dopływa do odbiornika, czyli ta część energii, która w części okresu jest gromadzona w filtrze, potem jest w całości oddawana do odbiornika. Z równania (8) wynika, że suma energii wypływającej z źródła prądowego w czasie całego okresu jest równa zeru, czyli że przez część okresu energia przepływa z źródła prądowego do uzwojenia dodatkowego, a przez pozostałą część okresu energia przepływa z uzwojenia dodatkowego do źródła prądowego. Gdyby źródło prądowe było zdolne gromadzić energię, która do niego wpłynęła w części okresu, a następnie zwracać ją do dławika w następnym okresie, to wygładzanie napięcia odbiornika odbywałoby się bez wydatkowania energii źródła prądowego, a sprawność filtru byłaby bliska 1. W rzeczywistości źródłem prądowym jest zazwyczaj układ tranzystorowy w którym nie można gromadzić energii. W efekcie energia dopływająca z uzwojenia dodatkowego do źródła prądowego zamienia się na ciepło w złączu kolektorowym tranzystora a później źródło prądowe musi

dostarczyć taką samą ilość energii do dławika. Równania(1—6) są ężnościami ogólnymi, a równania (7) i (8) są słuszne dla wszystkich przebiegów okresowych napięcia $u_1$ i prądu $i_1$. W czasie całego okresu żadna część energii wypływająca z źródła prądowego nie dopływa do odbiornika, jest ona jedynie gromadzona w dławiku a następnie przepływa do źródła prądowego. Potwierdza to konkluzję że uzwojenie dodatkowe spełnia rolę elementu sterującego przepływem energii w obwodzie w którym jest włączone uzwojenie główne.

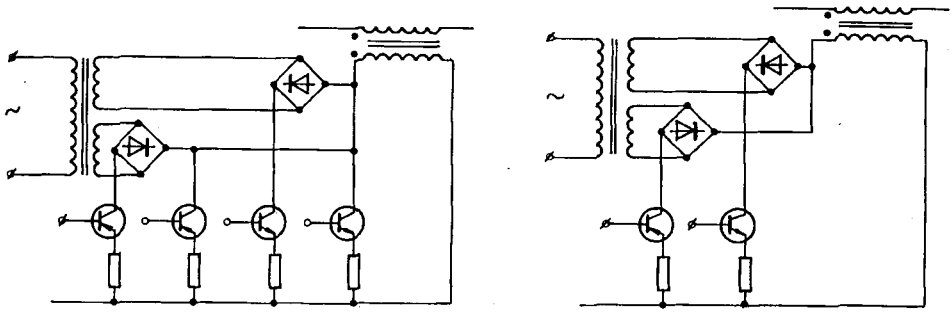



Rys. 2. Schemat przykładu rozwiązania źródła prądowego przy niesymetrycznym względem zera przebiegu prądu $i_2$

Rodzaj zastosowanego źródła prądowego decyduje o możliwym do uzyskania kształcie przebiegu prądu $i_2$. Najprostsze rozwiązanie przedstawiono na rys. 2. W czasie gdy SEM $e_1$ i $e_2$ są skierowane jak na rysunku, wysterowany jest tranzystor T1, a tranzystor T2 jest zablokowany. W czasie gdy SEM są skierowane odwrotnie, wysterowany jest tranzystor T2, a T1 jest zablokowany. W układach małej mocy można zrezygnować z zastosowania tranzystora T2. Wtedy w ciągu całego okresu przepuszczać prąd $i_2$ płynie przez tranzystor T1, lecz wtedy ilość energii traconej w źródle prądowym w ciągu całego okresu jest około dwukrotnie większa. Jeśli dla przykładu przyjmiemy że źródłem napięcia $u_1$ jest tyrystorowy przerywacz prądu stałego, to w rozwiązaniu przedstawionym na rys. 2. prąd $i_2$ zmienia się jak na rys. 3a. W przedziale czasu $(0,\tau)$ płynie w kierunku przeciwnym do kierunku SEM $e_2$ kosztem energii źródła prądowego. W przedziale czasu $(\tau,T)$ SEM $e_1$ i $e_2$ mają kierunek przeciwny do zaznaczanego na rys. 1. i rys. 2., tranzystor T1 jest zablokowany, wysterowany jest tranzystor T2 prąd $i_2$ płynie zgodnie z kierunkiem SEM $e_2$, energia jest czerpana z dławika, przepływa do źródła prądowego i jest w nim tracona. Jeśli prąd $i_2$ przy zachowaniu tych samych przyrostów zmienia się symetrycznie względem wartości $i_2=0$ tak jak przedstawiono to na rys. 3b., ilość energii przepływającej w czasie pełnego okresu z źródła prądowego do dławika może być 2-krotnie mniejsza. Przykłady rozwiązań źródła prądowego umożliwiających uzyskanie przebiegu prądu $i_2$ symetrycznego względem zera przedsatwiono na rys. 4a i 4b.




Rys. 3. Przebiegi napięć i prądów w aktywnym filtrze wstrzykującym: $u_1$ — napięcie wejściowe, $u_0$ — napięcie wyjściowe, $i_1$ — prąd wyjściowy, $i_2$ — prąd w dodatkowym uzwojeniu dławika, T — okres, $\tau$ — czas trwania impulsu napięcia $u_1$



Rys. 4. Schematy przykładów rozwiązania źródła prądowego przy symetrycznym względem zera przebiegu prądu $i_2$

### 3.2. PRZYKŁAD LICZBOWY BILANSU ENERGETYCZNEGO FILTRU

Charakter bilansu energetycznego dławika o dwóch uzwojeniach ujawnia się szczególnie wydatnie w konkretnym przykładzie liczbowym: Załóżmy że dławik o dwóch uzwojeniach został włączony jako element filtru aktywnego między odbiornikiem o rezystancji $R_0 = 1\,\Omega$ a przerywaczem tyrystorowym którego napięcie wyjściowe ma postać impulsów o amplitudzie $U_1 = 300$ V. Zakładamy że filtr zapewni pełne wygładzenie prądu $i_1$ więc indukcyjność odbiornika nie ma wpływu na działanie filtru. Przyjmijmy że główne uzwojenie dławika ma liczbę zwojów $z_1 = 1000$ i indukcyjność $L_1 = 100$ mH, a uzwojenie dodatkowe ma $z_2 = 100$ i $L_2 = 1$mH. Przyjmując dla uproszczenia idealne sprzężenie uzwojeń $k = 1$, otrzymamy indukcyjność wzajemną $M = k\sqrt{L_1 \cdot L_2} = 10$ mH. Przyjmijmy że częstotliwość pracy przerywacza $f = 333$ Hz, okres $T = 3$ ms, a obliczenia wykonamy dla współczynnika wypełnienia impulsów $\tau/T = 1/3$, czyli $\tau = 1$ ms. Przy takim wypełnieniu impulsów napięcie $U_0$ na odbiorniku oraz prąd $I_1$ płynący z przerywacza przez uzwojenie $z_1$ do odbiornika mają wartości:

$$U_0 = U_1 \frac{\tau}{T} = 100\,\text{V} \qquad I_1 = \cdot\frac{U_0}{R} = 100\,\text{A}$$

Przyjmijmy, że układ sterowania źródła prądowego wymusza przebieg prądu $i_2$ zgodny z rys. 3a. Wtedy również przebiegi napięcia $u_0$ oraz prądu $i_1$ są zgodne z przedstawionymi na rys. 3a. W czasie trwania impulsu napięcia $U_1$ w przedziale czasu $(0,\tau)$ SEM $e_1$ indukowana w uzwojeniu $z_1$ i pochodna prądu $i_2$ mają wartości:

$$e_1 = U_1 - U_0 = 200\,\mathrm{V} \qquad \mathrm{d}i_2/\mathrm{d}t = e_1/M = 200\,\mathrm{V}/10\,\mathrm{mH} = 20\,\mathrm{A/ms}$$

SEM $e_2$ indukowana przez zmiany prądu $i_2$ w uzwojeniu $z_2$

$$e_2 = e_1 \frac{z_2}{z_1} = 20\,\mathrm{V} \quad \text{lub} \quad e_2 = L_2 \frac{\mathrm{d}i_2}{\mathrm{d}t} = 20\,\mathrm{V}.$$

Wartość maksymalna prądu $i_2$ wynosi

$$i_2 = \tau \cdot \mathrm{d}i_2/\mathrm{d}t = 1\,\mathrm{ms} \cdot 20\,\mathrm{A/ms} = 20\,\mathrm{A}.$$

W przedziale czasu $(0,\tau)$ prąd $i_2$ płynie w kierunku przeciwnym do kierunku SEM $e_2$ pod wpływem napięcia źródła prądowego. Napięcie to musi mieć wartość $u_2 = e_2 + i_2 R_2 \cong e_2 = 20\,\mathrm{V}$. Ilość energii która przepływa z źródła prądowego do dławika w tym przedziale czasu wynosi:

$$\Delta W_{2\tau} = \int_0^{\tau} u_2 i_2 \,\mathrm{d}t = \frac{1}{2} u_2 i_{2max} \cdot \tau \cong \frac{1}{2} 20\,\mathrm{V} \cdot 20\,\mathrm{A} \cdot 1\,\mathrm{ms} = 0{,}2\,\mathrm{J}.$$

W tym samym przedziale czasu ze źródła napięcia $U_1$ (przerywacza) wypływa porcja energii o wartości

$$U_1 \cdot I_1 \cdot \tau = 300\,\mathrm{V} \cdot 100\,\mathrm{A} \cdot 1\,\mathrm{ms} = 30\,\mathrm{J}.$$

Część z niej dopływa do odbiornika, a część o wartości

$$\Delta W_{1\tau} = (U_1 - U_0) I_1 \cdot \tau = 200\,\mathrm{V} \cdot 100\,\mathrm{A} \cdot 1\,\mathrm{ms} = 20\,\mathrm{J}$$

zostaje zmagazynowana w polu magnetycznym dławika.

W czasie przerwy między impulsami napięcia $U_1$, w przedziale czasu $(\tau,T)$, odbiornik jest zasilany przez SEM $e_1$ indukowaną w uzwojeniu $z_1$ wskutek zmian prądu $i_2$, kosztem energii pola magnetycznego. SEM $e_1$ musi mieć wartość $-U_0$, a prąd $i_2$ indukujący SEM $e_1$ musi zmieniać się z pochodną o wartości:

$$\frac{\mathrm{d}i_2}{\mathrm{d}t} = \frac{-U_0}{M} = \frac{-100\,\mathrm{V}}{10\,\mathrm{mH}} = -10\,\mathrm{A/ms}.$$

Przyrost prądu $i_2$ w przedziale czasu $(\tau,T)$ wynosi

$$i_2 = (T - \tau) \frac{\mathrm{d}i_2}{\mathrm{d}t} = -2\,\mathrm{ms} \cdot 10\,\mathrm{A/ms} = -20\,\mathrm{A}$$

czyli ma taką samą wartość bezwzględną jak w przedziale czasu $(0,\tau)$ i w momencie $t = T$ prąd maleje do zera. W tym przedziale czasu kierunek SEM $e_2$ jest zgodny

z kierunkiem prądu $i_2$, prąd $i_2$ płynie przez tranzystor T2 (rys. 3), który działa jako rezystor o zmiennej rezystancji w którym wydziela się energia czerpana z pola magnetycznego dławika. Ilość tej energii w całym przedziale czasu $(\tau,T)$ jest równa energii która uprzednio, w przedziale czasu $(0,\tau)$ przepłynęła z źródła prądowego do dławika i wynosi 0,2 J.

Do odbiornika w tym samym przedziale czasu $(\tau,T)$ dopływa ilość energii

$$\Delta W_2(\tau,T) = U_0 \cdot I_1 \cdot (T - \tau) = 100\,\mathrm{V} \cdot 100\,\mathrm{A} \cdot 2\,\mathrm{ms} = 20\,\mathrm{J}.$$

Jest to ta sama ilość energii która w przedziale czasu $(0,\tau)$, pobrana z źródła napięcia $U_1$ (przerywacza), została zgromadzona w polu magnetycznym dławika. Ilość energii pobranej w ciągu całego okresu z źródła prądowego jest 100-krotnie mniejsza. Przykład ten potwierdza, że uzwojenie dodatkowe o mniejszej liczbie zwojów spełnia rolę elementu sterującego polegającą na tym, że kosztem niewielkiej ilości energii dostarczonej z źródła prądowego do tego uzwojenia można regulować wymianę znacznie większych ilości energii między dławikiem a obwodem połączonym z głównym uzwojeniem dławika. Wynika to z faktu, że wartość energii zgromadzonej w polu magnetycznym wytworzonym przez dwie cewki jest wyrażona równaniem:

$$W = \frac{1}{2}L_1 i_1^2 + Mi_1 i_2 + \frac{1}{2}L_2 i_2^2. \tag{I}$$

Jeśli prąd $i_1$ ma stałą wartość, a prąd $i_2$ wzrasta o wartość $\Delta i_2$ i powoduje przyrost strumienia, na przyrost energii pola magnetycznego składają się dwie porcje energii:

$$W_a = \frac{1}{2}L_2\left[(i_2 + \Delta i_2)^2 - i_2^2\right] = \frac{1}{2}L_2(\Delta i_2^2 - 2i_2\,\Delta i_2) \tag{II}$$

— jest to porcja dostarczona z źródła prądu $i_2$, oraz

$$\Delta W_b = Mi_1\,\Delta i_2 \tag{III}$$

— jest to porcja dopływająca z źródła prądu $i_1$. Jeśli liczba zwojów cewki $L_1$ jest znacznie większa od liczby zwojów cewki $L_2$, to $M > L_2$ i ilość energii określona wyrażeniem III może mieć znacznie większą wartość od ilości energii określonej wyrażeniem II, zwłaszcza gdy $i_2$ ma dużą wartość. Mechanizm przepływu energii $W_b$ polega na tym, że prąd $i_1$ płynie przez cewkę $L_1$ w czasie gdy istnieje w niej SEM wytworzona przez zmianę wartości prądu $i_2$. Energia potencjalna ładunków przepływających w obszarze działania tej SEM zostaje powiększona kosztem energii źródła prądu $i_1$ i gromadzi się w postaci energii pola magnetycznego.

### 3.3. BILANS ENERGETYCZNY DLA PRZEBIEGÓW IMPULSOWYCH

W szczególnym przypadku gdy źródłem napięcia $u_1$ jest falownik z jednobiegunową modulacją szerokości impulsów, lub przerywacz prądu stałego, napięcie ma postać impulsów o amplitudzie $U_1$, a przebiegi napięć i prądów w filtrze są podobne do przedstawionych na rys. 3a lub 3b. W typowej sytuacji gdy filtr jest

wykorzystywany do wygładzania napięcia $u_0$ i prądu $i_1$ zmieniających się powoli w porównaniu z częstotliwością napięcia $u_1$, można przyjąć, że:

$$\frac{di_1}{dt} = 0 \quad \text{oraz} \quad e_1 = L_1 \frac{di_1}{dt} + M \frac{di_2}{dt} \cong M \frac{di_2}{dt}.$$

Przy założeniu że napięcie $u_0$ zostanie w pełni wygładzone, jego wartość określa równanie:

$$u_0 = U_1 \frac{\tau}{T} \quad \text{gdzie } \tau - \text{czas trwania impulsu} \tag{9}$$

Dla uzyskania wygładzenia napięcia $u_0$, SEM $e_1$ oraz prąd $i_2$ muszą zmieniać się zgodnie z równaniami:

$$e_1 = u_1 - u_0 = u_1 - U_1 \frac{\tau}{T} \tag{10}$$

$$\frac{di_2}{dt} = \frac{e_1}{M} = \frac{1}{M}\left(u_1 - U_1 \frac{\tau}{T}\right) \tag{11}$$

gdzie: $u_1$ — wartość chwilowa napięcia wyjściowego przekształtnika, w czasie trwania impulsu równa $U_1$, w czasie przerwy między impulsami równa zeru.

Jeśli przebieg prądu $i_2$ jest zgodny z rys. 3a, to:

$$i_{2max} = \int_0^{\tau} \frac{di_2}{dt} dt = \frac{U_1}{M}\left(1 - \frac{\tau}{T}\right)\tau = \frac{1}{M}(U_1 - u_0)\tau. \tag{12}$$

Ilość energii którą źródło prądowe w czasie impulsu $(0;\tau)$ musi dostarczyć do dławika wynosi, zgodnie z r-mi (6), (12)

$$\Delta W_{2T} = L_2 \int_0^{i_{2max}} i_2 \, di_2 = \frac{1}{2} L_2 i_{2max}^2 = \frac{L_2 (U_1 - u_0)^2 \tau^2}{2M^2}. \tag{13}$$

W przedziale czasu $(\tau,T)$ taka sama ilość energii jest tracona.

W obwodzie obejmującym główne uzwojenia dławika część energii wypływająca z przekształtnika jest w czasie trwania impulsu napięcia $u_1$, czyli w przedziale czasu $(0;\tau)$ gromadzona w filtrze. Jej wartość oznaczoną $\Delta W_{1\tau}$ można określić równaniem (oznaczenia jak na rys. 3.):

$$\Delta W_{1\tau} = \int_0^{\tau} u_1 i_1 \, dt - \int_0^{\tau} u_0 i_1 \, dt = (U_1 - u_0) i_1 \tau. \tag{14}$$

Z zależności (7) wynika, że ta sama ilość energii jest następnie w czasie przerwy między impulsami oddawana z filtru do odbiornika.

Jako miarę sprawności energetycznej filtru można przyjąć stosunek energii pobranej w jednym okresie $T$ z źródła prądowego do energii zmagazynowanej w filtrze

w czasie trwania impulsu napięcia $u_1$, a następnie oddanej do odbiornika w czasie przerwy między impulsami. Stosunek ten wynosi:

$$\frac{\Delta W_{2T}}{\Delta W_{1\tau}} = \frac{L_2 (U_1 - u_0)^2 \tau^2}{2M^2 (U_1 - u_0) i_1 \tau}. \tag{15}$$

Uwzględniając że $M = k\sqrt{L_1 L_2}$ oraz r-e (9) otrzymamy:

$$\frac{\Delta W_{2T}}{\Delta W_{1\tau}} = \frac{u_0 (T - \tau)}{2k^2 L_1 i_1} = \frac{R_0}{2k^2 L_1}\left(1 - \frac{\tau}{T}\right) = \frac{1}{2k^2}\frac{T}{T_f}\left(1 - \frac{\tau}{T}\right). \tag{16}$$

gdzie: $T_f = \dfrac{L_1}{R_0}$.

Moc pobieraną z źródła prądowego można wyznaczyć mnożąc obie strony r-a (15 przez częstotliwość $f$. Po przekształceniach otrzymuje się zależność

$$P_2 = W_{2T} \cdot f = \frac{U_1^2}{2k^2 L_1 f}\left[\frac{\tau}{T} - \left(\frac{\tau}{T}\right)^2\right]^2. \tag{17}$$

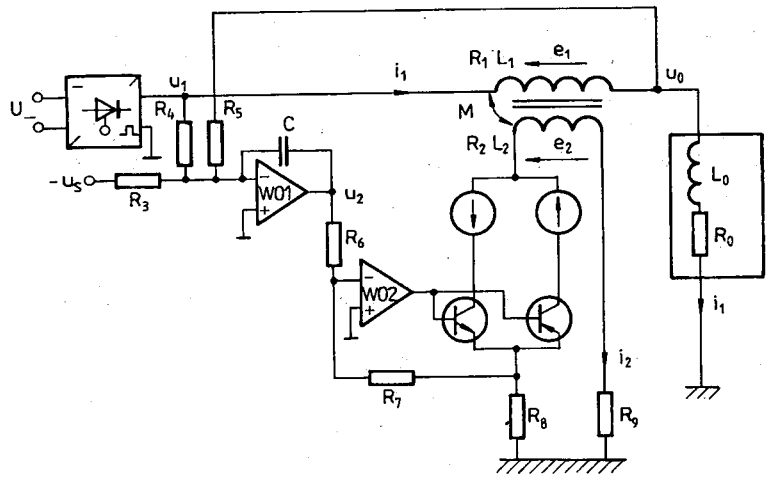
Funkcja $P(\tau/T)$ osiąga maksimum gdy $\tau/T = 0{,}5$ i ma wtedy wartość:

$$P_{2max} = \frac{U_1^2}{32 k^2 L_1 f} \tag{18}$$

## 4. STEROWANIE FILTRU

Założenie niezmienności, lub małej prędkości zmian prądu w uzwojeniu głównym $i_1$ przyjęto w rozdziale 3. tylko w celu uproszczenia bilansu energetycznego filtru. Wstrzykujący filtr aktywny może być również wykorzystywany do wygładzania przebiegu napięcia za przekształtnikiem tyrystorowym o zmiennym współczynniku wypełnienia impulsów, zasilającym odbiornik prądem o zmiennej wartości (w tym również np. prądem sinusoidalnym).

Jeśli filtr aktywny jest włączony za przekształtnikiem którego współczynnik wypełnienia impulsów napięcia $u_1$, a wraz z nim wartość średnia tego napięcia zależy liniowo od sygnału sterującego $u_s$ doprowadzonego do układu sterowania przekształtnika, to zmieniając w odpowiedni sposób prąd $i_2$ w dodatkowym uzwojeniu filtru, można spowodować, że wartość chwilowa napięcia za filtrem $u_0$ będzie zależna od wartości sygnału sterującego $u_s$, a całkowicie niezależna od chwilowych wartości napięcia $u_1$. Dla prawidłowej realizacji takiej funkcji filtru prąd $i_2$ w dodatkowym uzwojeniu dławika musi zmianiać się w sposób zależny jednocześnie od napięcia za przekształtnikiem $u_1$ i od sygnału sterującego $u_s$. Schemat układu elektrycznego realizującego sterowanie prądu $i_2$ przedstawiono na rys. 5. Uzwojenie główne dławika jest włączone między przekształtnikiem a odbiornikiem. Uzwojenie dodatkowe jest




Rys. 5. Schemat przykładu rozwiązania układu sterowania filtru wstrzykującego

połączone z układem sterującym składającym się z integratora sumującego (zbudowanego z wykorzystaniem wzmacniacza operacyjnego WO1) i źródła prądowego sterowanego napięciowo (zbudowanego z wykorzystaniem wzmacniacza WO2 i tranzystorów). Do wejścia integratora podawany jest przez $R_3$ odwrócony sygnał sterujący $-u_s$, oraz przez $R_4$ sygnał kompensacji i przez $R_5$ sygnał sprzężenia zwrotnego proporcjonalny do napięcia na odbiorniku. Zależności między wartościami prądów i napięć w filtrze można przedstawić w postaci operatorowej (oznaczenia w równaniach są zgodne z oznaczeniami na rys. 5., dla uproszczenia przyjęto zerowe warunki początkowe), $R9$ jest bocznikiem pomiarowym o małej rezystancji.

$$u_2(s) = \frac{-1}{sC}\left[\frac{u_1(s)}{R_4} + \frac{u_0(s)}{R_5} - \frac{u_s(s)}{R_3}\right] \tag{19}$$

$$i_2(s) = -u_2(s)\frac{R_7}{R_6 R_8} \tag{20}$$

$$e_1(s) = i_2(s) \cdot s \cdot M \tag{21}$$

$$u_1(s) = e_1(s) + i_1(s)[R_1 + R_0 + s(L_1 + L_0)] \tag{22}$$

$$u_0(s) = u_1(s) - e_1(s) - i_1(s)(R_1 + sL_1) = i_1(s)(R_0 + sL_0). \tag{23}$$

Po przekształceniach i wprowadzeniu następujących oznaczeń:

$$k = \frac{M R_7}{C R_3 R_6 R_8} \qquad \beta_1 = \frac{R_3}{R_5} \qquad \beta_2 = \frac{R_3}{R_4} \tag{24}$$

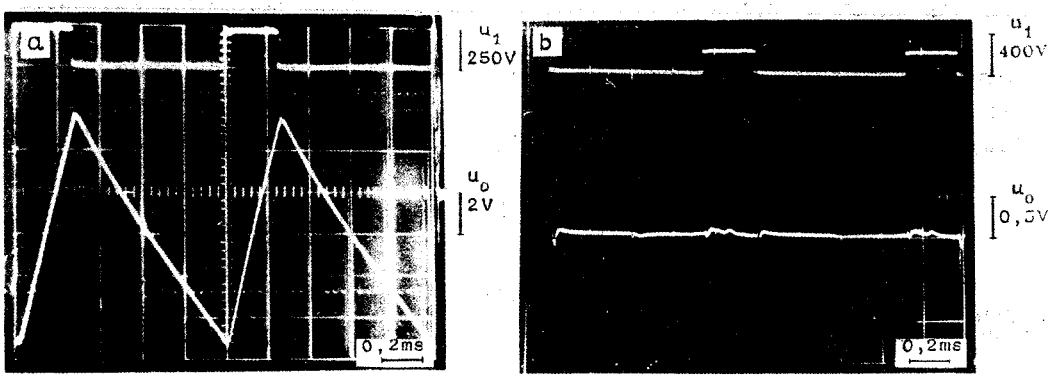
otrzymamy zależność (pełne wyprowadzenie w dodatku):

$$u_0(s) = u_s(s)\frac{k}{1 + k\beta + \dfrac{R_1 + sL_1}{R_0 + sL_0}} + u_1(s)\frac{1 - k\beta_2}{1 + k\beta_1 + \dfrac{R_1 + sL_1}{R_0 + sL_1}}. \tag{25}$$

Jeśli elementy układu zostaną dobrane w taki sposób, że iloczyn $k\beta_2 = 1$ to licznik drugiego członu równania stanie się równy zeru, co oznacza że napięcie na odbiorniku $u_0$ zależy tylko od sygnału sterującego, przestaje zależeć od chwilowej wartości napięcia $u_1$, czyli nie zawiera składowych zmiennych wynikających z impulsowego charakteru działania przekształtnika. Oznacza to całkowite wygładzenie napięcia za przekształtnikiem.

## 5. WYNIKI PRAKTYCZNEJ REALIZACJI FILTRU

Sposób włączenia aktywnego filtru wstrzykującego bezpośrednio za przekształtnikiem bez poprzedzającego go filtru biernego, oraz sposób sterowania filtru przedstawiony w rozdz. 4. zostały przez autora zrealizowane praktycznie w wzmacniaczach tyrystorowych o mocach 8 kVA i 20 kVA zbydowanych w Instytucie Elektroniki Morskiej i Przemysłowej Politechniki Gdańskiej. Wysoką sprawność energetyczną filtru potwierdza fakt, że w większym wzmacniaczu składającym się z tyrystorowego falownika z jednobiegunową modulacją szerokości impulsów napięcia wyjściowego o apmlitudzie $U_1 = 230$ V i maksymalnym prądzie wyjściowym $i_{1max} = 100$ A oraz z filtru aktywnego, zastosowane w filtrze źródło prądowe miało moc łączną 400 W.



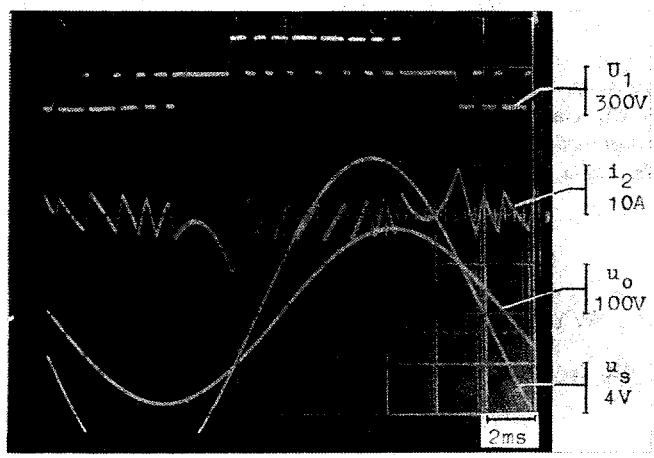

Rys. 6. Oscylogramy napięć wejściowego i wyjściowego: a) w filtrze biernym indukcyjnym, b) w filtrze aktywnym o tej samej indukcyjności głównego uzwojenia dławika

Zdolność wygładzania napięcia przez filtr ilustrują oscylogramy rys. 6a. i 6b. Oscylogramy te wykonano badając wzmacniacz (mniejszy, o mocy nominalnej 8 kVA) był obciążony rezystorem 4Ω. Falownik wytwarzał impulsy napięcia o amplitudzie 220 V, częstotliwości 1 kHz i współczynniku wypełnienia około 0,27. Do odbiornika płynął prąd o wartości średniej 15 A. Oscylogramy przedstawiają przebieg napięcia na wyjściu falownika i składowej zmiennej napięcia na odbiorniku. Oscylogram rys. 6a wykonano przy wyłączonym układzie sterowania filtru i braku prądu $i_2$, dławik

spełniał rolę filtru biernego. Oscylogram rys. 6b wykonano w tym samym stanie pracy falownika, lez z włączonym układem sterowania filtru i w innej skali napięć.

Na oscylogramie rys. 6a zapis składowej zmiennej napięcia na odbiorniku $u_0$ ma amplitudę $12\,V_{pp}$ czyli tylko 18 razy mniejszą od amplitudy impulsów napięcia $u_1$. Na oscylogramie rys. 6b zapis składowej zmiennej napięcia $u_0$ ma amplitudę około $0{,}2\,V_{pp}$ czyli 1000–krotnie mniejszą od amplitudy impulsów napięcia na wyjściu falownika. Tę przeszło 50-krotnie lepszą filtrację napięcia uzyskano bez dodania do obwodu głównego nowych indukcyjności lub pojemności, czyli bez wydłużenia stałej czasowej układu w porównaniu ze stanem gdy dławik działał jako prosty filtr bierny. Rzeczywisty przebieg napięcia $u_0$ obserwowany na ekranie oscyloskopu zawierał w momentach skoku napięcia $u_1$ impulsy o bardzo krótkim czasie trwania widoczne w postaci „szpilek" tak cienkich, że taśma fotograficzna nie zarejestrowała ich. Impulsy te były przekazywane poprzez filtr przez pojemności międzyzwojowe uzwojenia dławika, pojemności montarzowe i sprzężenia indukcyjne. Amplituda ich wynosiła około 4 V, a częstotliwość podstawowej harmonicznej tych impulsów była większa od 150 kHz, więc można było je odfiltrować stosując niewielkie proste filtry bierne małej mocy.

Oscylogramy rys. 6a i 6b wykonano przy stałym współczynniku wypełnienia impulsów w celu dokładnego zademonstrowania zdolności filtru do wygładzania napięcia. Jako przykład działania filtru w innych warunkach przedstawiono na oscylogramie rys. 7. przebiegi napięć i prądów przy pracy falownika z nadążną modulacją szerokości impulsów jednobiegunowych przy sinusoidalnym przebiegu napięcia sterującego o częstotliwości 55 Hz. Przy takiej pracy zapisu przebiegu napięcia $u_0$ nie można było wykonać w równie wielkiej skali jak na rys. 6b.




Rys. 7. Przebiegi napięć i prądów w aktywnym filtrze wstrzykującym przy sinusoidalnym przebiegu sygnału sterującego o częstotliwości 55 Hz: $u_1$ napięcie wejściowe, $u_s$ – napięcie sygnału sterującego, $i_2$ – prąd w uzwojeniu dodatkowym dławika, $u_0$ – napięcie wyjściowe

## WNIOSKI

1. W aktywnym szeregowym filtrze wstrzykującym procesem gromadzenia w polu magnetycznym dławika energii czerpanej z źródła zasilającego obwód główny (np. z przekształtnika) można sterować za pomocą małych ilości energii dostarczanej do uzwojenia dodatkowego. Dzięki temu sprawność energetyczna poprawnie zaprojektowanego filtru jest wysoka.

2. Aktywny szeregowy filtr wstrzykujący może być stosowany jako samodzielny filtr wyjściowy przekształtników tyrystorowych dużej mocy bez konieczności włączania filtrów biernych między przekształtnik a filtr aktywny.

3. Aktywny szeregowy filtr wstrzykujący umożliwia uzyskanie bardzo dokładnego wygładzenia napięcia wejściowego przekształtnika. Wydłużenie stałej czasowej układu jest niewielkie w porównaniu z wydłużeniem spowodowanym przez zastosowanie filtrów biernych o podobnym stopniu wygładzenia napięcia wyjściowego.

### BIBLIOGRAFIA

1. Gyn – Ha Choe, Min – Ho Park: *Analysis and Control of Active Power Filter with Optimized Injection.* IEEE Transactions on Power Elektronics Vol. 4 No 4 October 1989
2. Nakajima T., Masada E.: *An active filter with monitoring of harmonic spectrum.* EPE'89 Aachen 1989 Vol. III
3. Swan J. P., Bausiere R., Seynier G.: *Eliminating 5th and 7th Line Current Harmonics in AC/DC Converters by Harmonic Current Iniection.* EPE Grenoble 1987 Vol. II
4. M. Kosicki: *Filtry aktywne w układach zasilających prądu stałego.* Przegląd Elektrotechniczny 1974/11
5. R. Prajoux, G. Juanel: *Ètude d'un filtre actif á transistor associé á un redresseur polphasé de grand puissance fonnant en régime d'impulsions.* Rewvue de Physique Appliquée t. 5: 1070 nr 4
6. J. N. Szuwajew: *Tranzistornyje sgłaživajuszczyje ustrojstwa s poslledowatelnym i paralelnym wkluczeniem kompensirujuszczego transformatora.* Elektronnaja technika w awtomatike 1980/11
7. Pat. USA 3551780 Off. Gazette 1970/5
8. *Ustrojstwo dla kompensacji pulsacji naprażenia.* Patent ZSRR 411586 Bul. Izobr. 1974 nr 2
9. *Ustrojstwo dla sgłaživania pulsacji naprażenia.* Patent ZSRR 547940 Bul. Izobr. 1977 nr 2

### DODATEK

## WYPROWADZENIE ZALEŻNOŚCI (25)

Po podstawieniu równań 19 do 20 a następnie 20 do 21 otrzymamy:

$$e_1(s) = \frac{MR_7}{CR_3 R_6 R_8}\left[u_1(s)\frac{R_3}{R_4} + u_0(s)\frac{R_3}{R_5} - u_s(s)\frac{R_3}{R_3}\right] \tag{A1}$$

po wprowadzeniu oznaczeń (24)

$$e_1(s) = k[u_1(s)\beta_2 + u_0(s)\beta_1 - u_s(s)], \tag{A2}$$

do r-a (22) podstawiamy (A2)

$$u_1(s) = u_1(s)k\beta_2 + u_0(s)k\beta_1 - u_s(s)k + i_1(s)(R_0 + sL_0)\left(1 + \frac{R_1 + sL_1}{R_0 + sL_0}\right) \tag{A3}$$

po podstawieniu (23)

$$u_1(s)(1 - k\beta_2) = -u_s(s)k + u_0(s)\left(k\beta_1 + 1 + \frac{R_1 + sL_1}{R_0 + sL_0}\right) \tag{A4}$$

po uporządkowaniu (A4) otrzymamy (25).

P. HEMPOWICZ

## ACTIVE INJECTING FILTER APPLIED IN SMOOTHING OF THE OUTPUT VOLTAGE OF THYRISTOR CONVERTER

Summary

A new interpretation of physical proceses occuring in serial injecting filter is presented. Energetic equation of the filter was derived to prove its high efficiency. An example of the control unit is described. Results of a practical realization of such a filter in application in smothing of the output voltage of thyristor converter is presented.

KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJI, 1992, 38, z. 1, ss. 197-209

(621.375.026)

# Tyrystorowy wzmacniacz dużej mocy o małych zniekształceniach

PAWEŁ HEMPOWICZ

*Instytut Elektrotechniki Morskiej i Przemysłowej, Politechnika Gdańska*

*Otrzymano 1991.03.10*

*Autoryzowano do druku 1991.07.20*

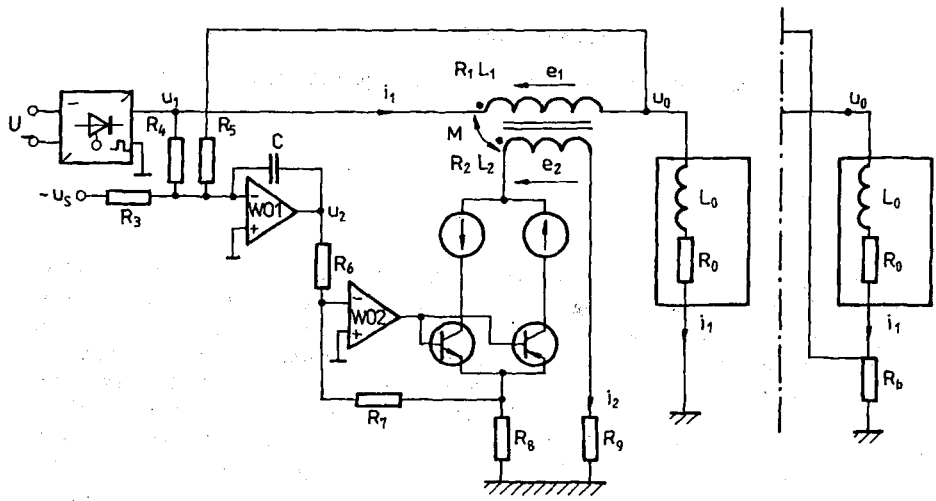
Przedstawiono nowe, oryginalne rozwiązanie układu tyrystorowego w którym szczególnie dobre właściwości uzyskano dzięki ingerencji przekształtnika tyrystorowego z filtrem aktywnym o wysokiej sprawności energetycznej. Układ charakteryzuje się znikomą zawartością zniekształceń w napięciu wyjściowym i jednocześnie krótką stałą czasową. Napięcie wyjściowe praktycznie nie zawiera zniekształceń spowodowanych impulsowym charakterem przewodzenia tyrystorów. Układ może być realizowany jako wzmacniacz napięciowy lub prądowy, albo jako źródło regulowanego napięcia lub prądu stałego lub przemiennego, w tym również sinusoidalnego o regulowanej częstotliwości i amplitudzie. Przedstawiono wyniki praktycznej realizacji układu.

## 1. WSTĘP

Jednym z rodzajów filtrów nadających się do wygładzania napięcia wyjściowego przekształtników tyrystorowych jest aktywny szeregowy filtr wstrzykujący. Jego podstawową zaletą jest zdolność do bardzo dokładnego wygładzania napięcia wyjściowego bez znaczącego wydłużania stałej czasowej układu, do którego filtr jest włączony. Filtr ten nie znajduje dotychczas szerszego zastosowania z dwóch przyczyn. Pierwszą jest przekonanie, że sprawność energetyczna tego filtru jest mała [1,2,3]. W pracach [4,5] wykazano, że jest to przekonanie błędne. Drugą przyczyną jest brak możliwości praktycznego wykorzystania potencjalnie bardzo dużej zdolności filtru do wygładzania napięcia, spowodowany trudnościami w uzyskaniu poprawnej współpracy filtru z źródłem napięcia wygładzanego. Celem artykułu jest przedstawienie nowego, oryginalnego rozwiązania układu, w którym szeregowy filtr wstrzykujący jest wykorzystywany do wygładzania napięcia za falownikiem tyrystorowym. Dzięki zintegrowaniu układów sterowania filtru i falownika uzyskano optymalną ich współpracę, a w efekcie możliwość uzyskania na wyjściu napięć niedokształconych, m.in. sinusoidalnych o regulowanej częstotliwości i amplitudzie. W artykule przedstawiono koncepcję tego rozwiązania, niektóre szczegóły jego technicznej realizacji oaz wyniki uzyskane w praktycznie zrealizowanych układach.

## 2. SZEREGOWY FILTR WSTRZYKUJĄCY

Elementem wykonawczym tego filtru jest rdzeń ferromagnetyczny i dwa nawinięte na nim uzwojenia: główne i dodatkowe (rys. 1). Uzwojenie główne włącza się między przekształtnik a odbiornik. Uzwojenie dodatkowe jest podłączone do regulowania źródła prądowego mającego dużą dynamiczną oporność wewnętrzną. Zmieniając wartość prądu $i_2$ w dodatkowym uzwojeniu dławika można indukować w obu uzwojeniach SEM. SEM $e_1$ indukowana w uzwojeniu głównym dodaje się do napięcia na wyjściu przekształtnika $u_1$ i łącznie z nim decyduje o przebiegu chwilowych wartości napięcia na odbiorniku $u_0$.




Rys. 1. Schemat ideowy aktywnego filtru wstrzykującego: a. we wzmacniaczu napięcia, b. we wzmacniaczu prądowym

$R_1 L_1$ – rezystancja i indukcyjność głównego uzwojenia filtru
$R_2 L_2$ – rezystancja i indukcyjność uzwojenia dodatkowego
$R_0 L_0$ – rezystancja i indukcyjność odbiornika

Jeśli filtr działa poprawnie, to wartość napięcia za filtrem jest bliska średniej wartości napięcia $u_1$, lecz nie zależy od wartości chwilowych tego napięcia. Jeśli współczynnik wypełnienia impulsów napięcia przed filtrem $u_1$, a wraz z nim wartość średnia tego napięcia, zależą liniowo od sygnału sterującego $u_s$ doprowadzonego do układu sterowania przekształtnika, to zmieniając w odpowiedni sposób prąd $i_2$ w dodatkowym uzwojeniu dławika, można spowodować, że nie tylko wartość średnia, lecz również wartość chwilowa napięcia za filtrem będzie w każdym momencie proporcjonalna do sygnału sterującego. W takiej sytuacji przekształtnik wraz z filtrem staje się wzmacniaczem napięciowym dużej mocy. Można również zbudować układ, w którym współczynnik wypełnienia impulsów nie jest bezpośrednio proporcjonalny do wartości chwilowej sygnału sterującego, lecz zmienia się w taki sposób, że prąd płynący do odbiornika włączonego za filtrem będzie proporcjonalny do sygnału sterującego. Układ taki jest źródłem prądowym sterowanym napięciowo, lub wzmacniaczem prądowym.

Przykład rozwiązania filtru wraz z układem sterowania przedstawiono na rys. 1a. W tak zbudowanym układzie zależność napięcia na wyjściu filtru $u_0$ od sygnału

sterującego $u_s$ i napięcia przed filtrem $u_1$ jest opisana równaniem (1) [1] (oznaczenia w równaniu są zgodne z oznaczeniami na rys. 1; $R_9$ jest bocznikiem o małej rezystancji):

$$u_0(s) = u_s(s) \frac{k}{1 + k\beta_1 + \frac{R_1 + sL_1}{R_0 + sL_0}} - u_1(s) \frac{1 - k\beta_2}{1 + k\beta_1 + \frac{R_1 + sL_1}{R_0 + sL_0}} \quad (1)$$

gdzie: $k = \frac{MR_7}{CR_3R_6R_8} \quad \beta_1 = \frac{R_3}{R_5} \quad \beta_2 = \frac{R_3}{R_4}$

Jeśli elementy układu są dobrane w taki sposób że $k\beta_2 = 1$ to napięcie na odbiorniku $u_0$ zależy tylko od sygnału sterującego $u_s$, czyli układ jest wzmacniaczem napięciowym o małych zniekształceniach. Typową dla opisu wzmacniaczy postać zależności między sygnałem wejściowym (sterującym) a napięciem na wyjściu wzmacniacza uzyskamy po przekształeniu $r-a$ (1) i wprowadzeniu następujących oznaczeń:

$$K = \frac{kR_0}{R_0(1 + k\beta_1) + R_1} \quad T' = \frac{L_0}{R_0} \quad T'' = \frac{L_1 + L_0(1 + k\beta_1)}{R_1 + R_0(1 + k\beta_1)}$$

jeśli $k\beta_2 = 1$, to $$u_0(s) = u_s(s) \cdot K \cdot \frac{1 + sT'}{1 + sT''} \quad (2)$$

Oznacza to, że przekształtnik łącznie z filtrem jest wzmacniaczem którego charakter zależy od stosunku stałych czasowych $T'$ i $T''$. Jeśli dławik filtru jest poprawnie skonstruowany, to spełniona jest zależność

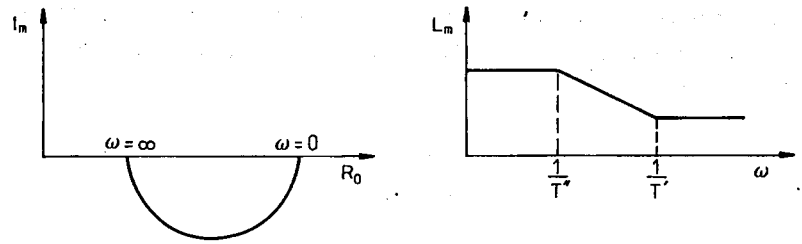
$$\frac{L_1}{R_1} \gg \frac{L_0}{R_0}$$

a wtedy również $T'' > T'$ i układ jest wzmacniaczem proporcjonalno inercyjnym. Jego charakterystyki amplitudowo-fazowa i asymptotyczna logarytmiczna amplitudową mają postać przedstawioną na rys. 2.

Wzmacniacz prądowy różni się od napięciowego tym, że sygnał sprzężenia zwrotnego dostarczony przez $R_5$ do wejścia integratora jest brany nie z zacisku odbiornika, jak na rys. 1a., lecz z bocznika jak na rys. 1b. Funkcja przenoszenia tego wzmacniacza opisana jest równaniem (3) (wprowadzenie w aneksie), w którym oznaczenia są takie same jak w r-u (1) i (2) z następującymi wyjątkami:

$$\beta_1 = \frac{R_b R_3}{R_5} \quad T = \frac{L_0 + L_1}{k\beta_1 + R_0 + R_1} \quad K = \frac{k}{k\beta_1 + R_0 + R_1} \quad (3)$$

jeśli $k\beta_2 = 1$, to $$i_1(s) = u_s(s) \frac{K}{1 + sT} \quad (4)$$




Rys. 2. Charakterystyki wzmacniacza napięcia złożonego z przekształtnika i filtru wstrzykującego w/g rys. 1 a. charakterystyka amplitudowo-fazowa, b. charakterystyka asymptotyczna logarytmiczna amplitudowa

Oznacza to, że w tej konfiguracji układu sterowania, przekształtnik łącznie z filtrem jest inercyjnym źródłem prądowym, sterowanym napięciowo.

## 3. OGRANICZENIA WSPÓŁPRACY PRZEKSZTAŁTNIKA Z FILTREM AKTYWNYM WSTRZYKUJĄCYM

Układ sterowania filtru może różnić się od przedstawionego w schemacie rys. 1, ale każdy układ sterowania filtru wstrzykującego musi zawierać integrator. Wynika to z faktu, że SEM wytwarzana w uzwojeniu głównym dławika jest proporcjonalna do pochodnej prądu w uzwojeniu dodatkowym $i_2$. Dlatego pochodna $di_2/dt$ powinna być proporcjonalna do różnicy sygnałów: proporcjonalnego do chwilowej wartości napięcia przed filtrem $u_1$ i sygnału sterującego $u_s$, a wartość chwilowa prądu $i_2$ powinna być proporcjonalna do całki z różnicy tych sygnałów. Całkowanie różnicy sygnałów dokonuje się w integratorze sumującym. Wartość składowej stałej prądu $i_2$ nie ma bezpośredniego wpływu na efekt działania filtru, a układ sterowania filtru nie może o wartości składowej stałej prądu $i_2$ decydować.

Uzyskanie potencjalnie możliwej bardzo dobrej filtracji napięcia w aktywnym filtrze wstrzykującym włączonym za przekształtnikiem napotyka na dwa ograniczenia: Pierwsze polega na tym, że obecność integratora w układzie sterowania filtru powoduje, że prąd $i_2$ może zawierać składową stałą o nieokreślonej wartości. Jest to stan niekorzystny i aby go uniknąć należy w układzie sterowania wprowadzić ujemne sprzężenie zwrotne od składowej stałej prądu $i_2$ lub słabe ujemne sprzężenie zwrotne od całkowitej chwilowej wartości prądu $i_2$. Najprostszym sposobem jest włączenie rezystora równolegle do kondensatora w integratorze. Mimo istnienia tego sprzężenia zwrotnego, przy szybkiej zmianie wartości sygnału sterującego pojawia się w prądzie $i_2$ składowa przejściowa o wartości początkowej proporcjonalnej do przyrostu sygnału sterującego, malejąca wykładniczo z długą stałą czasową zależną od wzmocnienia obwodu sprzężenia zwrotnego. Jednocześnie istnienie sprzężenia zwrotnego od prądu $i_2$ znacznie pogarsza filtrację.

Pogorszenie filtracji i składową przejściową prądu $i_2$ widać na oscylogramach rys. 3a., 3b. Na oscylogramach przedstawiono przebiegi: sygnału sterującego $u_s$, napięcia na odbiorniku $u_0$, prądu w uzwojeniu dodatkowym filtru wstrzykującego $i_2$ oraz napięcia przed filtrem $u_1$ — we wzmacniaczu tyrystorowym w którym układ