POLSKA AKADEMIA NAUK
KOMITET ELEKTRONIKI I TELEKOMUNIKACJI

PLISSN 0035-9386

ROZPRAWY ELEKTROTECHNICZNE

TOM XXXV — ZESZYT 1

KWARTALNIK

PAŃSTWOWE WYDAWNICTWO NAUKOWE
WARSZAWA 1989

POLSKA AKADEMIA NAUK
KOMITET ELEKTRONIKI I TELEKOMUNIKACJI

ROZPRAWY ELEKTROTECHNICZNE

TOM XXXV — ZESZYT 1

KWARTALNIK

PAŃSTWOWE WYDAWNICTWO NAUKOWE
WARSZAWA 1989

ADRES REDAKCJI

00-661 Warszawa, ul. Nowowiejska 15/19
Politechnika, Instytut Telekomunikacji, Gmach im. JANUSZA GROSZKOWSKIEGO

Dyżury Redakcji: środy i piątki, godz. 14—16
tel. 28 89 81 lub 210 07-737

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

PAŃSTWOWE WYDAWNICTWO NAUKOWE — WARSZAWA

Nakład 490 (400+90)	Oddano do składania 14.IX.1988 r.
Ark. druk. 14,0	Podpis. do druku w lutym 1990 r.
Ark. wyd. 19,0	Druk ukończono w marcu 1990 r.
Papier druk. kl. III, 70 g.	Zam. 6912/12/88. Cena zł 350,—

DRUKARNIA IM. REWOLUCJI PAŹDZIERNIKOWEJ, WARSZAWA

KOMUNIKAT

Jak już informowaliśmy naszych AUTORÓW i CZYTELNIKÓW

ROZPRAWY ELEKTROTECHNICZNE

decyzją Wydziału IV Nauk Technicznych Polskiej Akademii Nauk od 1988 r. stały się czasopismem Komitetu Elektroniki i Telekomunikacji PAN, natomiast ARCHIWUM ELEKTROTECHNIKI jest czasopismem Komitetu Elektrotechniki PAN. W związku z powyższym uległy zmianie profile tematyczne obu tych kwartalników.

ROZPRAWY ELEKTROTECHNICZNE publikują prace naukowe związane z podstawami elektroniki i telekomunikacji oraz prace z zakresu mikroelektroniki i zastosowań elektroniki, radiotechniki, telekomunikacji i elektroniki medycznej.

Zamieszczane są prace oryginalne, w tym przyczynkowe oraz zawierające analizy porównawcze metod lub systemów, systematyczne ujęcie określonej dziedziny wiedzy, omówienie aktualnego stanu lub postępu danej gałęzi techniki oraz omówienie perspektyw rozwojowych. Czasopismo jest przeznaczone dla specjalistów oraz studentów zajmujących się powyższą tematyką.

ARCHIWUM ELEKTROTECHNIKI natomiast publikuje prace naukowe dotyczące podstaw elektrotechniki oraz prace z zakresu energoelektroniki, elektroenergetyki, konstrukcji maszyn i urządzeń elektrycznych itp.

Redakcja ROZPRAW ELEKTROTECHNICZNYCH poczynając od początku roku 1989 przyjmuje prace odpowiadające nowemu profilowi tematycznemu czasopisma.

W związku z powyższymi zmianami, za zgodą władz Polskiej Akademii Nauk, od pierwszego zeszytu 1990 roku

ROZPRAWY ELEKTROTECHNICZNE

będą wydawane pod nowym tytułem

KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJI

(ELECTRONICS AND TELECOMMUNICATIONS QUARTERLY)

Redakcja
ROZPRAW ELEKTROTECHNICZNYCH

ROZPRAWY ELEKTROTECHNICZNE, 1989, 35, z. 1, ss. 5-11

621.372—517

Rozwiązanie zagadnienia własnego $[K]\{\Phi\} = \lambda[M]\{\Phi\}$ metodą iteracji podprzestrzennych przy nieokreślonej macierzy $[K]$

KAROL ANISEROWICZ

Wydział Elektryczny, Politechnika Białostocka

Otrzymano 1988.01.20

Autoryzowano do druku 1988.03.30

Przedstawiono uogólnienie metody iteracji podprzestrzennych, oparte na analizie Galerkina. Pozwala to na rozszerzenie zastosowań obliczeniowych metody na przypadek nieokreślonej symetrycznej macierzy $[K]$ w zagadnieniu $[K]\{\Phi\} = \lambda[M]\{\Phi\}$. Omówiono wykorzystanie przesunięcia widma do poprawy zbieżności algorytmu oraz do wyznaczania wartości i wektorów własnych z dowolnie obranego pasma widma.

1. WSTĘP

Zastosowanie metody elementów skończonych m.in. do analizy rozkładów pól elektromagnetycznych w falowodach [1, 2, 3, 8] prowadzi do algebraicznego zagadnienia własnego

$$[K]\{\Phi\} = \lambda[M]\{\Phi\}, \tag{1}$$

przy czym macierze $[K]$ i $[M]$ są stosunkowo duże. Macierze te są symetryczne i mają budowę pasmową.

Kilka najmniejszych wartości własnych problemu (1) i odpowiadające im wektory własne dla przypadku falowodów jednorodnych mogą być efektywnie obliczone przy użyciu metody iteracji podprzestrzennych [4, 5, 6].

Dotychczasowy algorytm tej metody był oparty na analizie Rayleigha- Ritza, użytecznej dla dodatnio określonych macierzy $[K]$ i $[M]$.

Przy badaniu właściwości dyspersyjnych falowodów z wypełnieniem niejednorodnym macierz $[K]$ może być nieokreślona, zatem możliwości zastosowania iteracji podprzestrzennych były ograniczone [3].

Niniejszy artykuł jest kontynuacją prac [2, 3], rozszerzając stosowalność iteracji podprzestrzennych na przypadek nieokreślonej, symetrycznej macierzy $[K]$.

Uogólnienie to jest możliwe dzięki oparciu algorytmu iteracyjnego na postępowaniu Galerkina. Przedstawiona analiza umożliwia również obliczanie kilku wartości i wektorów własnych zagadnienia (1) z dowolnego pasma widma.

2. POSTĘPOWANIE GALERKINA

Załóżmy, że macierze $[K]$ i $[M]$ są symetryczne i określone w n-wymiarowej przestrzeni rzeczywistej R^n, przy czym macierz $[M]$ jest dodatnio określona.

Rozważmy wektor $\{\bar{\Phi}\}$, który jest liniową kombinacją wektorów bazowych $\{\varphi_j\}$, $j = 1, 2, \ldots, p$:

$$\{\bar{\Phi}\} = \sum_{j=1}^{p} x_j \{\varphi_j\}. \tag{2}$$

Wektor $\{\bar{\Phi}\}$ należy do p-wymiarowej podprzestrzeni $R^p \subset R^n$, rozpiętej nad wektorami bazowymi. Wektory $\{\varphi_j\}$ są liniowo niezależne i tworzą układ zupełny w R^n.

Dążymy do wyznaczenia takich wektorów $\{\bar{\Phi}_i\}$, które są najlepszymi przybliżeniami poszukiwanych wektorów własnych $\{\Phi_i\}$, $i = 1, 2, \ldots, p$. Zbiór błędów przybliżenia reprezentowany jest przez wektor resztowy $\{R\}$

$$[K]\{\bar{\Phi}\} - \bar{\lambda}[M]\{\bar{\Phi}\} = \{R\}. \tag{3}$$

Minimalizujemy resztę $\{R\}$ poprzez ortogonalizację jej w stosunku do układu wektorów wagowych $\{w_i\}$

$$\{w_i\}^T \cdot \{R\} = 0, \quad i = 1, 2, \ldots, p. \tag{4}$$

W metodzie Galerkina przyjmuje się $\{w_i\} = \{\varphi_i\}$. Otrzymujemy stąd układ równań:

$$\sum_{j=1}^{p} x_j (\{\varphi_i\}^T [K]\{\varphi_j\} - \bar{\lambda}\{\varphi_i\}^T [M]\{\varphi_j\}) = 0, \quad i = 1, 2, \ldots, p. \tag{5}$$

Zależność (5) stanowi zredukowane zagadnienie własne, które w zapisie macierzowym ma postać

$$[\tilde{K}]\{x\} = \bar{\lambda}[\tilde{M}]\{x\}, \tag{6}$$

przy czym $[\tilde{K}]$ i $[\tilde{M}]$ są macierzami o wymiarach $p \times p$, a ich elementy

$$\tilde{k}_{ij} = \{\varphi_i\}^T [K]\{\varphi_j\}, \tag{7}$$

$$\tilde{m}_{ij} = \{\varphi_i\}^T [M]\{\varphi_j\}, \tag{8}$$

zaś $\{x\}$ jest wektorem, którego współrzędne są poszukiwanymi współczynnikami we wzorze (2):

$$\{x\}^T = [x_1 \quad x_2 \ldots x_p]. \tag{9}$$

Po rozwiązaniu (6) otrzymuje się p wartości własnych $\bar{\lambda}_1, \bar{\lambda}_2, \ldots, \bar{\lambda}_p$ oraz p wektorów

$$\begin{aligned} \{x^{(1)}\}^T &= [x_1^{(1)} \quad x_2^{(1)} \ldots x_p^{(1)}] \\ &\vdots \qquad\qquad \vdots \\ \{x^{(p)}\}^T &= [x_1^{(p)} \quad x_2^{(p)} \ldots x_p^{(p)}] \end{aligned} \tag{10}$$

Obliczone wartości $\overline{\lambda}_1, \overline{\lambda}_2, \ldots, \overline{\lambda}_p$ stanowią przybliżenia odpowiednich wartości własnych problemu (1). Oczywiście możliwe jest wyznaczenie ewentualnych wielokrotnych wartości własnych, tzn. $\overline{\lambda}_j = \overline{\lambda}_{j+1} = \ldots = \overline{\lambda}_{j+k}$. Układ przybliżonych wektorów własnych uzyskuje się z (2) i (10):

$$\{\overline{\Phi}_i\} = \sum_{j=1}^{p} x_j^{(i)} \{\varphi_j\}, \quad i = 1, 2, \ldots, p. \tag{11}$$

Porównując zależności (6)—(11) z wynikami analizy Rayleigha-Ritza [6] widać, że postępowanie Galerkina prowadzi do takich samych rezultatów w przypadku, gdy macierz $[K]$ jest dodatnio określona. W szczególności, przy dodatnio określonej macierzy $[K]$, przybliżenia wartości własnych są obliczone z nadmiarem, tzn. $\lambda_1 \leqslant \overline{\lambda}_1, \lambda_2 \leqslant \overline{\lambda}_2, \ldots, \lambda_p \leqslant \overline{\lambda}_p \leqslant \lambda_n$.

3. METODA ITERACJI PODPRZESTRZENNYCH

Błędy wyznaczenia wartości i wektorów własnych przy stosowaniu zarówno analizy Rayleigha-Ritza, jak i Galerkina silnie zależą od wyboru wektorów $\{\varphi_j\}$. W celu zapewnienia zadowalającej dokładności wyników praktycznych obliczeń komputerowych konieczne jest zastosowanie rozszerzonej procedury, umożliwiającej iteracyjne doskonalenie wektorów bazowych. Jest nią metoda iteracji podprzestrzennych [4, 5, 6], której podstawowe kroki pozostają bez zmian, dzięki analogii wzorów uzyskanych z postępowania Rayleigha-Ritza i Galerkina.

Metoda ta służy do wyznaczania p najmniejszych dodatnich wartości własnych λ_i oraz odpowiadających im wektorów własnych zagadnienia (1), przy czym pierwsza dodatnia wartość własna ma indeks 1. Należy nadmienić, że w analizie falowodów ujemnym wartościom własnym odpowiadają uboczne, niefizyczne rozkłady pól [8].

Dzięki odpowiedniemu doborowi początkowej podprzestrzeni bazowej R^q (przy czym $R^p \subset R^q \subset R^n$) oraz właściwościom wektorowej iteracji odwrotnej [6] zastosowanej jako część składowa algorytmu, uzyskuje się zbieżność obliczeń do wartości własnych o najmniejszym module. Poniżej przedstawiono algorytm metody iteracji podprzestrzennych, bardziej efektywny, niż stosowany w pracach [2, 3]:

1. Wyznaczenie zbioru q wektorów początkowych (stosowane dalej indeksy w nawiasach mają znaczenie wskaźników iteracji)

$$[\Psi_0] = [\{\psi_1^{(0)}\}, \{\psi_2^{(0)}\}, \ldots, \{\psi_p^{(0)}\}, \ldots, \{\psi_q^{(0)}\}], \tag{12}$$

przyjmując wskaźnik iteracji $k = 0$ oraz $\lambda_1^{(0)} = \lambda_2^{(0)} = \ldots = \lambda_p^{(0)} = 1$, przy czym $q > p$ w celu polepszenia zbieżności obliczeń. Regułę optymalnego wyboru wektorów początkowych opisano w pracy [6], a jej rozszerzenie w [2].

2. Zastosowanie równoczesnej wektorowej iteracji odwrotnej dla q wektorów oraz analizy Galerkina, w celu uzyskania aproksymacji pierwszych p wartości i wektorów własnych. Składają się na to następujące działania:

— rozkład macierzy $[K]$ na czynniki

$$[K] = [L][D][L]^T; \tag{13}$$

— obliczenie $[\bar{\Phi}_{k+1}]$, wykorzystując (13), z zależności:

$$[K] = [\bar{\Phi}_{k+1}][\Psi_k]; \tag{14}$$

— obliczenie kolejno:

$$[\tilde{K}_{k+1}] = [\bar{\Phi}_{k+1}]^T[\Psi_k], \tag{15}$$

$$[\bar{\Psi}_{k+1}] = [M][\bar{\Phi}_{k+1}], \tag{16}$$

$$[\tilde{M}_{k+1}] = [\bar{\Phi}_{k+1}]^T[\bar{\Psi}_{k+1}]; \tag{17}$$

— rozwiązanie zredukowanego zagadnienia własnego w q-wymiarowej podprzestrzeni

$$[\tilde{K}_{k+1}][Q_{k+1}] = [\tilde{M}_{k+1}][Q_{k+1}][\Lambda_{k+1}]; \tag{18}$$

— obliczenie aproksymacji wektorów własnych

$$[\Psi_{k+1}] = [\bar{\Psi}_{k+1}][Q_{k+1}]; \tag{19}$$

— sprawdzenie zbieżności obliczeń według kryterium

$$\left|\frac{\lambda_i^{(k+1)} - \lambda_i^{(k)}}{\lambda_i^{(k+1)}}\right| \leqslant \varepsilon, \quad i = 1, 2, \ldots, p, \tag{20}$$

gdzie ε jest zakładaną dokładnością wyników. Jeśli nie osiągnięto założonej dokładności, to powtarza się obliczenia (14)—(20);

— obliczenie ostatecznej postaci wektorów własnych $[\bar{\Phi}_{k+1}]$ z zależności (14).

3. Po stwierdzeniu zbieżności według kryterium (20) należy określić błąd wyznaczenia wektorów własnych. Sposób obliczania błędów jest opisany w pracach [2, 6]. Należy również sprawdzić, które pary własne $(\lambda_i, \{\Phi_i\})$ zostały obliczone. Kontrola zbieżności wyników do odpowiednich par własnych jest oparta na właściwościach rozkładu macierzy $[K]-\mu[M]$ na czynniki $[L][D][L]^T$. Wykorzystuje się fakt, że liczba ujemnych elementów macierzy $[D]$ jest równa liczbie wartości własnych mniejszych od μ. Wykonując dwukrotnie wspomniany rozkład, dla $\mu = 0$ oraz dla $\mu = 1{,}01\lambda_p^{(l)}$ (gdzie l — indeks ostatniej iteracji; wartość współczynnika 1,01 została sprawdzona praktycznie [2, 6]), oblicza się odpowiednie liczby ujemnych elementów w macierzy $[D]$. Ich różnica powinna wynosić p. Jeśli różnica ta jest większa, to nie osiągnięto zbieżności do p pierwszych dodatnich wartości własnych. Należy wówczas powiększyć zbiór wektorów startowych $[\Psi_0]$ lub zmienić je i powtórzyć obliczenia.

4. PEWNE ASPEKTY OBLICZENIOWE

W przypadku analizy pól w falowodach jednorodnych lub w pobliżu częstotliwości granicznych w falowodach niejednorodnych (dopóki macierz $[K]$ jest dodatnio określona) obliczenia można przeprowadzać tak, jak opisano w pracach [2, 3]. Jednakże uogólnienie algorytmu iteracji podprzestrzennych na przypadek nieokreślonej macierzy $[K]$ pozwala

poprawić zbieżność metody przez zastosowanie przesunięcia widma problemu (1) w kierunku mniejszych wartości własnych:

$$([K-\mu[M]]\{\Phi\} = (\lambda-\mu)[M]\{\Phi\}. \tag{21}$$

Zabieg ten poprawia asymptotyczny wskaźnik zbieżności iteracji λ_i/λ_{p+1} (im mniejszy wskaźnik, tym lepsza zbieżność) [6]. Optymalna wartość μ znajduje się w pobliżu λ_1. Zmniejszenie liczby kroków iteracyjnych (14)—(20) widać szczególnie wyraźnie podczas obliczania pól przy jednorodnych warunkach berzegowych Neumanna (pola typu H), dla których macierz $[K]$ jest dodatnio półokreślona ($\lambda_1 = 0$). W tym przypadku zalecano dotychczas przesunięcie widma zagadnienia (1) w kierunku większych wartości własnych [6, 18, 20]. Pogarszało to zbieżność metody. Po zastosowaniu zabiegu (21) w obliczeniach testowych dla fal typu E (warunki brzegowe Dirichleta) osiągano zmniejszenie liczby iteracji o jedną-dwie, zaś dla fal typu H — o dwie-trzy. Jest to zauważalna poprawa, gdyż ogólna liczba iteracji zwykle jest mniejsza od dziesięciu.

Podobną operację można wykonać w przypadku, gdy macierz $[K]$ jest nieokreślona, szczególnie gdy przewidujemy wystąpienie zerowej wartości własnej. Przesunięcie widma pozwala wówczas uniknąć trudności numerycznych przy rozkładzie (13). Problemy, które mogą pojawić się przy dekompozycji nieokreślonej macierzy $[K]$ na czynniki $[L][D][L]^T$ oraz sposoby ich unikania są omówione np. w pracy [6].

Wykorzystanie wektorowej iteracji odwrotnej w algorytmie iteracji podprzestrzennych pozwala rozszerzyć jej zastosowania na obliczanie kilku par własnych z dowolnego fragmentu widma. Dotychczas były one ograniczone do wyznaczania par własnych z krańców widma problemu (1). Wystarczy wykonać przesunięcie (21) o odpowiednią dużą wartość μ. Zostaną wówczas obliczone wartości własne $\eta_i = \lambda_i - \mu$ o najmniejszym module i odpowiadające im wektory własne $\{\Phi_i\}$. Oznacza to wyznaczenie wartości λ_i z pobliża μ.

5. WNIOSKI

Przedstawione uogólnienie algorytmu iteracji podprzestrzennych umożliwia rozwiązywanie zagadnienia własnego (1) przy nieokreślonej, symetrycznej macierzy $[K]$. Dzięki temu można efektywnie analizować np. właściwości dyspersyjne falowodów niejednorodnych dla kilku pierwszych rodzajów fal w pełnym zakresie częstotliwości.

Właściwy dobór wektorów startowych $\{\Psi_0\}$ ma istotny wpływ na zbieżność obliczeń do odpowiednich par własnych.

Przesunięcie widma (21) poprawia zbieżność algorytmu. Dzięki niemu można uzyskać podobne wartości wskaźników zbieżności przy analizie zagadnień brzegowych Neumanna i Dirichleta.

Uogólniona metoda iteracji podprzestrzennych może być stosowana do wyznaczania kilku par własnych z dowolnego pasma widma, przy czym algorytm jej jest prostszy, niż proponowany w pracy [16].

BIBLIOGRAFIA

1. S. Ahmed, P. Daly: *Finite-element methods for inhomogeneous waveguides.* Proc. IEE, 1969, Vol. 116, No. 10, pp. 1661—1664
2. K. Aniserowicz: *Numeryczna analiza pól elektromagnetycznych w bezstratnych falowodach jednorodnych.* Rozprawy Elektrotechniczne, 1988, (w druku)
3. K. Aniserowicz: *Numeryczna analiza pól elektromagnetycznych w bezstratnych falowodach z niejednorodnym wypełnieniem dielektrycznym.* Rozprawy Eektrotechniczne, 1988 (w druku)
4. K.J. Bathe, E.L. Wilson: *Large eigenvalue problems in dynamic analysis.* ASCE, Journal of Engineering Mechanics Division, 1972, Vol. 98, pp. 1471—1485
5. K.J. Bathe, E.L. Wilson: *Solution methods for eigenvalue problems in structural mechanics.* Int. J. Num. Meth. Eng., 1973, Vol. 6, pp. 213—226
6. K.J. Bathe, E.L. Wilson: *Numerical methods in finite element analysis.* Prentice-Hall, Englewood Cliffs, New Jersey 1976
7. R.B. Corr, A. Jennings: *A simultaneous iteration algorithm for eigenvalue problems.* Int. J. Num. Meth. Eng., 1976, Vol. 10, pp. 647—663
8. Z.J. Csendes, P. Silvester: *Numerical solution of dielectric loaded waveguides: I — Finite-element analysis.* IEEE Trans. MTT-18, 1970, No. 12, pp. 1124—1131
9. P. Daly: *Finite-element coupling matrices.* Electronics Letters, November 1969, Vol. 5, No. 24, pp. 613—615
10. M. Dryja, J. i M. Jankowscy: *Przegląd metod i algorytmów numerycznych.* Część 2, Warszawa: WNT, 1982
11. B.A. Finlayson: *The method of weighted residuals and variational principles.* Academic Press, New York 1972
12. K.K. Gupta: *Solution of eigenvalue problems by the Sturm sequence method.* Int. J. Num. Meth. Eng., 1972, Vol. 4, pp. 379—404
13. K.K. Gupta: *Eigenproblem solution by a combined Sturm sequence and inverse iteration technique.* Int. J. Num. Meth. Eng., 1973, Vol. 7, pp. 14—42
14. A. Jennings, D.R.L. Orr: *Application of the simultaneous iteration method to undamped vibration problems,* Int. J. Num. Meth. Eng., 1971, Vol. 3, pp. 13—24
15. S.G. Michlin, C.L. Smolicki: *Metody przybliżone rozwiązywania równań różniczkowych i całkowych.* Warszawa: PWN, 1970
16. J. Orkisz, B. Wrana: *Metoda obliczania wartości i wektorów własnych problemu* $[A]\{u\} = \lambda[B]\{u\}$ *w dowolnym pasmie widma,* Mechanika i Komputer, t. 5, IPPT PAN, Warszawa 1983, pp. 97—106
17. S.S. Rao: *The finite element method in engineering.* Pergamon Press 1982
18. H.R. Schwarz: *Methode der finiten Elemente,* Teubner, Stuttgart 1980
19. J.H. Wilkinson: *The algebraic eigenvalue problem.* Clarendon Press, Oxford 1965
20. O.C. Zienkiewicz: *Metoda elementów skończonych,* Arkady, Warszawa 1972

K. ANISEROWICZ

SOLUTION BY SUBSPACE ITERATION OF THE EIGENPROBLEM $[K]\{\Phi\} = \lambda[M]\{\Phi\}$ WITH AN INDEFINITE MATRIX $[K]$

Summary

A generalization of the subspace iteration method based on Galerkin analysis is described. This allows to extend the application of the method to an indefinite symmetric matrix $[K]$ case of the eigenproblem $[K]\{\Phi\} = \lambda[M]\{\Phi\}$. The use of spectrum shift for the improvement of the algorithm convergence and for the calculation of eigenvalues and eigenvectors in an arbitrary chosen band of spectrum, is considered.

K. ANISEROWICZ

RESOLUTION DU PROBLÈME PROPRE $[K]\{\Phi\} = \lambda[M]\{\Phi\}$ PAR MÉTHODE D'ITÉRATION SOUS-ESPACES POUR LA MATRICE INDÉFINIE $[K]$

Résumé

On a présenté la généralisation de la méthode d'itérations sous-espaces basée sur l'analyse de Galerkin. Ce la permet d'étendre les applications de calcul de la méthode sur le cas de la matrice indéfinie symétrique $[K]$, dans le problème de $[K]\{\Phi\} = \lambda[M]\{\Phi\}$. On a discuté la possibilité de la mise à profit du déplacement du spectre pour l'amélioration de la convergence de l'algorithme et pour la détermination de la valeur et des vecteurs propres à partir de la bande du spectre arbitrairement choisie.

K. ANISEROWICZ

LÖSUNG DES EIGENWERTPROBLEMS $[K]\{\Phi\} = \lambda[M]\{\Phi\}$ MITTELS SIMULTANER VEKTORITERATION BEI INDEFINITIVER MATRIX $[K]$

Zusammenfassung

Es wurde die auf der Galerkin-Analyse gestützte Verallgemeinerung der simultanen Vektoriteration geschildert. Dies gestattet, die Berechnungsanwendungen dieser Methode auf den Fall der indefinitiven symmetrischen Matrix $[K]$ für das Problem $[K]\{\Phi\} = \lambda[M]\{\Phi\}$ auszubreiten. Erwogen wurde die Ausnutzung der Spektrumverschiebung zwecks Konvergenzverbesserung des Algorithmus sowie zur Bestimmung des Eigenwertes und der Eigenwerte vom beliebig gewählten Spektrumband.

К. АНИСЕРОВИЧ

РЕШЕНИЕ ПРОБЛЕМЫ СОБСТВЕННЫХ ЗНАЧЕНИЙ $[K]\{\Phi\} = \lambda[M]\{\Phi\}$ МЕТОДОМ ПОДПРОСТРАНСТВЕННЫХ ИТЕРАЦИЙ В СЛУЧАЕ НЕОПРЕДЕЛЁННОЙ МАТРИЦЫ $[K]$

Резюме

Представленое обобщение метода подпространственных итераций, основано на анализе Галеркина. Позволяет оно расширить вычислительные применения метода на случай неопределённой, симметрической матрицы $[K]$ в задаче $[K]\{\Phi\} = \lambda[M]\{\Phi\}$. Описано использование сдвига спектра для улучшения сходимости алгоритма а также для вычисления собственных значений и собственных векторов из любого диапазона спектра.

ROZPRAWY ELEKTROTECHNICZNE, 1989, 35, z. 1, ss. 13-29

621.317.332.3—517

Piecewise-linear analysis of nonlinear resistive circuits using the block Gauss-Seidel and Jacobi methods

MICHAŁ TADEUSIEWICZ, MAREK OSSOWSKI

Instytut Podstaw Elektrotechniki, Politechnika Łódzka

Received 1987.09.18

Authorized 1988.02.17

The paper deals with the, application of the block Gauss-Seidel and Jacobi methods to piecewise-linear analysis of nonlinear resistive circuits. The problem of convergence of the methods is considered in detail. Sufficient conditions guaranteeing the existence of a unique solution and convergence of the iterative sequences to the solution for each initial guess, are formulated. Some weaker conditions related to the solvability and computation of the circuits are developed. A computer algorithm is described and numerical examples are demonstrated. The block Gauss-Seidel and Jacobi methods are compared with other well-known methods of piecewise-linear analysis. Special attention is paid to D.C. diode-transistor networks.

1. INTRODUCTION

The D. C. analysis of nonlinear networks is a very important problem in computer-aided design of electronic devices. Piecewise-linear methods and computational techniques are very efficient for this purpose. They are widely spread and often used both for analysis and synthesis [3], [4], [6], [14], [15] of nonlinear circuits. Katzenelson's method [8] extended and generalized in many papers, e.g. [2], [6], the Newton-Raphson method [5] and the algorithm developed by Chua and Ying [4] are the most important tools of piecewise-linear analysis of resistive circuits. The growth in the size of networks demands very efficient methods and algorithms.

Recently, the point and block Gauss-Seidel methods have been applied for D. C. piecewise-linear analysis of nonlinear electronic circuits [20], [21], [11], [23]. The point Gauss-Seidel iteration does not ask for a matrix inversion, which can be of great advantage. Both point and block Gauss-Seidel methods allow to replace the process of solving a system of equations by succesive analysis of smaller systems. Very similar to the Gauss-Seidel is the Jacobi iterative method. Both methods are classified as relaxation methods and they are discussed in many papers e.g. [13], [16], [19].

In papers [9] and [7] special attention is paid to the block relaxation methods. Ilic-Spong, Katz, Dai, Zaborszky [7] generalized some of More's [12] results concerning the point Gauss-Seidel and Jacobi methods to the block methods using the idea of weakly

block Ω-diagonally dominant matrices. Krätschmer [9] extended the Rheinboldt [16] results and defined block-off-diagonally antitone and block-diagonally-inverse-isotone functions.

A different approach to the investigation of convergence of the block relaxation (Gauss-Seidel and Jacobi) methods is presented in the paper. A generalization of the idea developed in [20], [21] is achieved and new results for the block methods are obtained.

In Section 2 the authors formulate the basic assumption, write down the equations of networks and introduce the idea of the block Gauss-Seidel and Jacobi methods. The main achievements of the paper are Theorems 1, 2 and 3 in Section 3 concerning the solvability and the convergence of the block relaxation algorithms. In Section 4 some aspects of computer analysis of diode-transistor networks using the block methods are discussed. Section 5 contains several numerical examples, discussion and concluding remarks. In the Appendix several definitions are presented.

2. PRELIMINARY DISCUSSION

We consider D. C. resistive circuits described by equation

$$\boldsymbol{A}\boldsymbol{f}(\boldsymbol{N}\boldsymbol{x}+\boldsymbol{z})+\boldsymbol{C}\boldsymbol{x}=\boldsymbol{y}, \tag{1}$$

where $\boldsymbol{x}=[x_1 \dots x_n]^T \in \boldsymbol{R}^n$ is a response vector of the circuit driven by $\boldsymbol{y}=[y_1 \dots y_n]^T \in \boldsymbol{R}^n$ and $\boldsymbol{z}=[z_1 \dots z_m]^T \in \boldsymbol{R}^m$; $\boldsymbol{C}=[c_{ij}]_{n\times n}$, $\boldsymbol{A}=[a_{ij}]_{n\times m}$ and $\boldsymbol{N}=[n_{ij}]_{m\times n}$ are real matrices; $\boldsymbol{f}(\boldsymbol{u})=[f_1(u_1) \dots f_m(u_m)]^T$ is a function: $\boldsymbol{R}^n \to \boldsymbol{R}^n$ composed of piecewise-linear, continuous and monotonically increasing functions $f_i(u_i)(i=1, \dots, m)$. We assume, that the function

$$\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{z})=[g_1(\boldsymbol{x}, \boldsymbol{z}) \dots g_n(\boldsymbol{x}, \boldsymbol{z})]^T=\boldsymbol{A}\boldsymbol{f}(\boldsymbol{N}\boldsymbol{x}+\boldsymbol{z}) \tag{2}$$

satisfies the following conditions:

$$\frac{\partial g_i}{\partial x_i} \geqslant 0, \quad \frac{\partial g_i}{\partial x_j} \leqslant 0 \quad i \neq j \quad (i, j=1, \dots, n) \tag{3}$$

(at the break point the left hand side derivative is taken) and that diagonal elements of the matrix $\boldsymbol{C}$ are positive.

If $\boldsymbol{N}=\boldsymbol{A}^T$, $\boldsymbol{A}=\delta_N$, then equation (1) describes the networks discussed in [20] and [21] consisting of linear and piecewise-linear resistors, linear voltage-controlled current sources and independent D. C. voltage and current sources. If $m=n$, $N=1$, $z=0$ and

$$\boldsymbol{A}=\boldsymbol{T}=\left[\begin{array}{cccc|c} \boldsymbol{T}_1 & 0 & \cdot\ \cdot & 0 & \\ 0 & \boldsymbol{T}_2 & \cdot\ \cdot & 0 & \\ \cdot & \cdot & & \cdot & 0 \\ 0 & 0 & \cdot\ 0 & \boldsymbol{T}_r & \\ \hline & & 0 & & 1 \end{array}\right], \quad \boldsymbol{T}_k=\begin{bmatrix} 1 & -\alpha_R^{(k)} \\ -\alpha_F^{(k)} & 1 \end{bmatrix}, \quad k=1, \dots, r$$

then equation (1) constitutes a well-known description of diode-transistor circuits [18], [22]:

$$\boldsymbol{T}\boldsymbol{f}(\boldsymbol{x})+\boldsymbol{C}\boldsymbol{x}=\boldsymbol{y}, \tag{4}$$

where $f_i(x_i)$ is a piecewise-linear approximation of the function $K_i(e^{\lambda_i x_i}-1)$ ($K_i > 0$, $\lambda_i > 0$). Each of the above classes of networks satisfies conditions (3). For practical purposes we wish only to find a solution in a finite interval. Therefore, we may assume that for sufficient large $|u_k|$ characteristics $f_k(u_k)$ ($k = 1, \ldots, m$) are saturated (they have zero slope).

The matrix $\boldsymbol{B}$ obtained from $\boldsymbol{C}$ after replacing all positive off-diagonal elements of $\boldsymbol{C}$ by zeros is introduced and the following decomposition is made

$$\boldsymbol{C} = \boldsymbol{B}+\boldsymbol{Q}. \tag{5}$$

Next we define the matrix

$$\boldsymbol{D} = \boldsymbol{B}-\boldsymbol{Q} = [d_{ij}]_{n\times n} \tag{6}$$

having elements: $d_{ii} = c_{ii}$, $d_{ij} = -|c_{ij}|$ $i \neq j$ $(i, j = 1, \ldots, n)$.

To perform the block Gauss-Seidel or Jacobi iterative algorithm we split equation (1) into ν blocks

$$\begin{aligned} p_1 &= g_1(x, z)+C_1 x-y_1 = 0 \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ p_\nu &= g_\nu(x, z)+C_\nu x-y_\nu = 0, \end{aligned} \tag{7}$$

where $x = [x_1^T \ldots x_\nu^T]^T$, $x_1 = [x_1 \ldots x_{\gamma_1}]^T$, $\ldots$, $x_\nu = [x_{(\gamma_{\nu-1}+1)} \ldots x_{\gamma_\nu}]^T$;
$y = [y_1^T \ldots y_\nu^T]^T$, $y_1 = [y_1 \ldots y_{\gamma_1}]^T$, $\ldots$, $y_\nu = [y_{(\gamma_{\nu-1}+1)} \ldots y_{\gamma_\nu}]^T$;
$g_i(x, z) = [g_{(\gamma_{i-1}+1)}(x, z) \ldots g_{\gamma_i}(x, z)]^T$;

$$C_i = \left[\begin{array}{ccc|c|ccc} c_{(\gamma_{i-1}+1,1)} & \cdots & c_{(\gamma_{i-1}+1,\gamma_1)} & \cdots & c_{(\gamma_{i-1}+1,\gamma_{\nu-1}+1)} & \cdots & c_{(\gamma_{i-1}+1,\gamma_\nu)} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ c_{(\gamma_i,1)} & \cdots & c_{(\gamma_i,\gamma_1)} & \cdots & c_{(\gamma_i,\gamma_{\nu-1}+1)} & \cdots & c_{(\gamma_i,\gamma_\nu)} \end{array}\right] = [C_{i1} \ldots C_{i\nu}]$$

where $\gamma_0 = 0$, $\gamma_\nu = n$. Hence, the division of $\boldsymbol{B}$ and $\boldsymbol{Q}$ into submatrices is of the form:

$$B = \begin{bmatrix} B_{11} B_{12} \ldots B_{1\nu} \\ B_{21} B_{22} \ldots B_{2\nu} \\ \cdots\cdots\cdots \\ B_{\nu 1} B_{\nu 2} \ldots B_{\nu\nu} \end{bmatrix}, \quad Q = \begin{bmatrix} Q_{11} Q_{12} \ldots Q_{1\nu} \\ Q_{21} Q_{22} \ldots Q_{2\nu} \\ \cdots\cdots\cdots \\ Q_{\nu 1} Q_{\nu 2} \ldots Q_{\nu\nu} \end{bmatrix}.$$

The block Gauss-Seidel iterative method requires solving succesively the blocks p_i with respect to vectors x_i according to the formula:

$$p_i([(x_1^{k+1})^T \ldots (x_{i-1}^{k+1})^T x_i^T (x_{i+1}^k)^T \ldots (x_\nu^k)^T]^T) = 0 \tag{8}$$

and setting $x_i^{k+1} = x_i$ $(i = 1, \ldots, \nu, k = 0, 1, 2, \ldots)$.

The block Jacobi iterative method is as follows: solve for x_i $(i = 1, \ldots, \nu)$ individual blocks

$$p_i([(x_1^k)^T \ldots (x_{i-1}^k)^T x_i^T (x_{i+1}^k)^T \ldots (x_\nu^k)^T]^T) = 0 \tag{9}$$

and take $x_i^{k+1} = x_i$. Set $k := k+1$ and repeat the procedure.

3. CONVERGENCE OF THE BLOCK GAUSS-SEIDEL AND JACOBI METHODS

The following theorem relates of the problem of solvability and computation of the solution of piecewise-linear networks described by equation (1) using the block Gauss-Seidel and Jacobi methods.

Theorem 1

If $(A\Delta N+D) \in M$ for each diagonal matrix $\Delta \geqslant 0$ then there exists a unique solution x^* of equation (1) for each $y \in R^n$ and $z \in R^m$. The solution can be found using the block Gauss-Seidel as well as Jacobi method and the iterative sequences are convergent to that solution for each initial guess $x^0 \in R^n$.

Proof

The proof will be performed only for the Gauss-Seidel method. Let us consider the function

$$p(s, r) = \begin{bmatrix} p_1(s, r) \\ \vdots \\ p_\nu(s, r) \end{bmatrix} = g(s, z)+Bs+Qr-y. \tag{10}$$

In [21] it has been shown that $\forall x^0 = [(x_1^0)^T \dots (x_\nu^0)^T]^T \in R^n \; \exists : v^0 = [(v_1^0)^T \dots (v_\nu^0)^T]^T \in R^n$ and $w^0 = [(w_1^0)^T \dots (w_\nu^0)^T]^T \in R^n$ such that

$$v^0 \leqslant x^0 \leqslant w^0 \tag{11}$$

$$p(v^0, w^0) \leqslant 0 \quad p(w^0, v^0) \geqslant 0. \tag{12}$$

The following suquence is defined:

$$\begin{aligned} \{x^0 &= [(x_1^0)^T(x_2^0)^T(x_3^0)^T \dots (x_\nu^0)^T]^T, \\ x^1 &= [(x_1^1)^T(x_2^0)^T(x_3^0)^T \dots (x_\nu^0)^T]^T, \\ x^2 &= [(x_1^1)^T(x_2^1)^T(x_3^0)^T \dots (x_\nu^0)^T]^T, \dots\}. \end{aligned} \tag{13}$$

Generally, the k-th element of the above sequence can be written in the form

$$x^k = [(x_1^s)^T(x_2^s)^T \dots (x_{i-1}^s)^T(x_i^s)^T(x_{i+1}^{s-1})^T \dots (x_\nu^{s-1})^T]^T, \tag{14}$$

where $s = 1+\mathrm{INT}\left(\dfrac{k-1}{\nu}\right)$, $i = k-(s-1)\nu$.

Let us define

$$x^{1l} = [(x_1^{1l})^T(x_2^0)^T \dots (x_\nu^0)^T]^T \tag{15}$$

where

$$x_1^{1l} = [x_1^j x_2^j \dots x_{b-1}^j x_b^j x_{b+1}^{j-1} \dots x_{\gamma_1}^{j-1}]^T \tag{16}$$

$$l = 1, 2, \dots, \quad j = 1+\mathrm{INT}\left(\frac{l-1}{\gamma_1}\right), \quad b = l-(j-1)\gamma_1.$$

v^{1l} and w^{1l} are defined similarly. The components x_b^j, v_b^j, w_b^j are found by solving the equations:

$$p_b(x^{1l}, x^{1,l-1}) = 0$$
$$p_b(v^{1l}, w^{1,l-1}) = 0 \tag{17}$$
$$p_b(w^{1l}, v^{1,l-1}) = 0$$

where the superscript 1, 0 is equivalent to 0.

Let $l = 1$, then x_1^1, v_1^1, w_1^1 are the solutions of the equations

$$p_1(x^{11}, x^0) = 0$$
$$p_1(v^{11}, w^0) = 0 \tag{18}$$
$$p_1(w^{11}, v^0) = 0$$

Hence, the following relations hold:

$$p_1(v^{11}, w^{11}) = 0 \tag{19}$$

$$p_1(w^{11}, v^{11}) = 0 \tag{20}$$

$$v_1^1 \geqslant v_1^0 \qquad w_1^1 \leqslant w_1^0. \tag{21}$$

Thus

$$v^{11} \geqslant v^0 \qquad w^{11} \leqslant w^0. \tag{22}$$

Moreover, taking into account (18) and (11), we find

$$p_1([(v_1^1 x_2^0 \ldots x_{\gamma_1}^0)(x_2^0)^T \ldots (x_\nu^0)^T]^T, x^0) \leqslant p_1(v^{11}, w^0) = 0,$$

which implies that

$$x_1^1 \geqslant v^1. \tag{23}$$

Similarly we obtain

$$x_1^1 \leqslant w_1^1. \tag{24}$$

The relations (22), (23) and (24) lead to the following inequalities:

$$v^0 \leqslant v^{11} \leqslant x^{11} \leqslant w^{11} \leqslant w^0. \tag{25}$$

On the basis of (12) and properties of the function p we find

$$p_1(v^{11}, w^{11}) \leqslant 0 \tag{26}$$
$$p_1(w^{11}, v^{11}) \geqslant 0.$$

When the procedure is continued, further elements of the sequences $\{v^{1l}\}$, $\{w^{1l}\}$, $\{x^{1l}\}$ are determined, and following the previous reasoning it is established that

$$v^0 \leqslant v^{11} \leqslant \ldots \leqslant v^{1l} \leqslant x^{1l} \leqslant w^{1l} \leqslant \ldots \leqslant w^{11} \leqslant w^0 \tag{27}$$

$$p_b(v^{1l}, w^{1l}) = 0 \qquad p_b(w^{1l}, v^{1l}) = 0 \tag{28}$$

$$p_1(v^{1l}, w^{1l}) \leqslant 0 \qquad p_1(w^{1l}, v^{1l}) \geqslant 0. \tag{29}$$

After performing γ_1 steps, the first computation cycle is finished and the process is repeated. The sequence $\{v^{1l}\}$ is nondecreasing and bounded above and the sequence $\{w^{1l}\}$ is nonincreasing and bounded below. Thus, the sequences are convergent, or $\lim_{l\to\infty} v^{1l} = v^1$, $\lim_{l\to\infty} w^{1l} = w^1$. Let us choose arbitrarily the index b and keep it fixed. The subsequence $\left\{\begin{bmatrix} v^{1l_i} \\ w^{1l_i} \end{bmatrix}\right\}_{i=1,2,\ldots}$ consisting of elements of the sequence $\left\{\begin{bmatrix} v^{1l} \\ w^{1l} \end{bmatrix}\right\}$ with superscripts l_i

satisfying relation $l_i = b + \left(\mathrm{INT}\left(\frac{l_i - 1}{\gamma_1}\right)\right) \cdot \gamma_1$ is convergent to the limit $\begin{bmatrix} v^1 \\ w_1 \end{bmatrix}$ and its elements fulfil equations:

$$p_b(v^{1l_i}, w^{1l_i} = 0 \quad p_b(w^{1l_i}, v^{1l_i}) = 0$$

Hence, we find $p_b(v^1, w^1) = 0$, $p_b(w^1, v^1) = 0$, $b = 1, \ldots, \gamma_1$, or

$$p_1(v^1, w^1) = 0, \quad p_1(w^1, v^1) = 0. \tag{30}$$

It can be proved [21] that sequence $\{x^{1l}\}$ is convergent to limit x^1 and $p_1(x^1, x^1) = 0$.

In fact $\{x_1^{1l}\}$ constitutes the point Gauss-Seidel iterative sequence. Inequalities (27) show that

$$v^1 \leqslant x^1 \leqslant w^1, \quad p(v^1, w^1) \leqslant 0, \quad p(w^1, v^1) \geqslant 0.$$

For next blocks we obtain similarly

$$p_i(x^k, x^k) = 0, \quad p_i(v^k, w^k) = 0, \quad p_i(w^k, v^k) = 0$$

$$k = 2, 3, \ldots, \quad i = k - \mathrm{INT}\left(\frac{k-1}{\nu}\right)\nu \text{ and}$$

$$v^0 \leqslant v^1 \leqslant \ldots \leqslant v^{k-1} \leqslant v^k \leqslant x^k \leqslant w^k \leqslant w^{k-1} \leqslant \ldots \leqslant w^1 \leqslant w^0. \tag{31}$$

Being bounded, sequences $\{v^k\}$ and $\{w^k\}$ (nondecreasing and nonincreasing, respectively) are convergent, or $\lim_{k\to\infty} v^k = v^*$, $\lim_{k\to\infty} w^k = w^*$. Fixing an arbitrary index i and taking into account the subsequence $\left\{\begin{bmatrix} v^{k_j} \\ w^{k_j} \end{bmatrix}\right\} \left(k_j = i + \mathrm{INT}\left(\frac{k_j - 1}{\nu}\right)\nu\right) (j = 1, 2, \ldots)$ of the sequence $\left\{\begin{bmatrix} v^k \\ w^k \end{bmatrix}\right\} \to \begin{bmatrix} v^* \\ w^* \end{bmatrix}$ we find

$$p_i(v^{k_j}, w^{k_j}) = 0 \quad p_i(w^{k_j}, v^{k_j}) = 0.$$

Hence, it follows

$$p_i(v^*, w^*) = 0 \quad p_i(w^*, v^*) = 0 \quad i = 1, \ldots, \nu,$$

or

$$\begin{aligned} p(v^*, w^*) &= Af(Nv^* + z) + Bv^* + Qw^* - y = 0 \\ p(w^*, v^*) &= Af(Nw^* + z) + Bw^* + Qv^* - y = 0. \end{aligned} \tag{32}$$

Subtracting the sides of the above equations gives

$$(A\Delta N + D)(v^* - w^*) = 0, \tag{33}$$

where $\Delta = \mathrm{diag}(\Delta_1, \ldots, \Delta_m) \geqslant 0$. Because $(A\Delta N + D) \in M$ (see assumption of theorem 1) then (33) implies: $v^* = w^*$. Taking into account (31) we obtain

$$\lim_{k\to\infty} x_i^k = x^* = v^* = w^*, \quad p(x^*, x^*) = 0.$$

The proof of uniqueness of the solution x^* is presented in paper [21].

The proof for the block Jacobi iterative method is similar and will be omitted.

Note

Since each of the equations $p_i(x, x) = 0$ has a unique solution for $x_i (i = 1, \ldots, v)$ then it can be solved using an arbitrary method, not necessarily the point Gauss-Seidel method as it was done in the proof.

Applying the concept of Ω-diagonally dominant matrix the following corollary can be formulated.

Corollary 1

If the matrix C^T is Ω-diagonally dominant and the matrix $A\Delta N = [k_{ij}]_{n\times n}$ satisfies the conditions: $k_{ii} \geqslant 0$, $k_{ij} \leqslant 0$, $\sum_{i=1}^{n} k_{ij} \geqslant 0$ $(j = 1, \ldots, n)$

$\forall \Delta = \operatorname{diag}(\Delta_1, \ldots, \Delta_m) \geqslant \mathbf{0}$, then conclusions of Theorem 1 hold.

Proof

Since C^T is Ω-diagonally dominant matrix then $D^T \in M$ [12]. Hence, using Lemma 2 [21] we find $(A\Delta N + D) \in M \quad \forall \Delta \geqslant \mathbf{0}$.

It is worth noting that conditions concerning the matrix $A\Delta N$ formulated in Corollary 1 fulfil diode-transistor circuits described by (4) as well as the circuits considered in [21].

The hypotheses of Theorem 1 can be relaxed and very useful results concerning solvability and computation of the networks using the block Gauss-Seidel and Jacobi methods can be obtained.

Theorem 2

If $D = [d_{ij}]_{n\times n} \in M$ then

(i) for each $y \in R^n$ and $z \in R^m$ there exists at least one solution of equation (1);

(ii) if equation (1) is split in such a way that each block has a unique solution and the point Gauss-Seidel (Jacobi) method is convergent to that solution for each initial guess, then the elements x^k of sequence generated by the block Gauss-Seidel (Jacobi) method belong to the respective sets $[v^k, w^k]$ converging to set $[v^*, w^*]$ in which there is at least one solution of equation (1);

(iii) if the sequence $\{x^k\}$ is convergent, then $x^* = \lim_{k\to\infty} x^k$ is a solution of equation (1).

Proof

To prove part (i), the point Jacobi iterative process will be considered. Let $p(x, x) := p(x)$ or $p_i(x, x) := p_i(x)$ $(i = 1, \ldots, n)$ and according to the Jacobi procedure we define the function

$$h(x) = [h_1(x_{(1)}) \ldots h_n(x_{(n)})]^T$$

where $x_{(i)} = [x_1 \ldots x_{i-1} x_{i+1} \ldots x_n]^T$ and $h_i(x_{(i)}) = \hat{x}_i$ $(i = 1, \ldots, n)$ is a unique solution of equation

$$p_i(\hat{x}_i, x_{(i)}) = 0.$$

Since $p_i(x_i, x_{(i)})$ is a strictly monotonically increasing function of x_i then for arbitrary $\varepsilon > 0$: $p_i(\hat{x}_i - \varepsilon, x_{(i)}) < 0$ and $p_i(\hat{x}_i + \varepsilon, x_{(i)}) > 0$. Because p_i is a continuous function of $x_{(i)}$ there exists such $\eta > 0$ that $p_i(\hat{x}_i - \varepsilon, x'_{(i)}) < 0$, $p_i(\hat{x}_i + \varepsilon, x'_{(i)}) > 0$ for each $x'_{(i)}$ fulfilling the condition: $\|x_{(i)} - x'_{(i)}\| < \eta$. Hence, for each $\varepsilon > 0$ there exists $\eta > 0$ such that

$||x_{(i)}-x'_{(i)}|| < \eta$ implies $|\hat{x}_i-\hat{x}'_i| < \varepsilon$, where $\hat{x}'_i = h_i(x'_{(i)})$. Thus, the function $h_i(x_{(i)})$ is continuous for $i = 1, \ldots, n$ and, therefore, also $\boldsymbol{h}(x)$ is a continuous function of x.

Since $\boldsymbol{D} \in M$, then for each $x^\circ \in \boldsymbol{R}^n$ there exist such vectors v^0 and w^0 that relations (11) and (12) hold (see Lemma 1 in [21]). Similarly as in the proof of Theorem 1, it can be shown that sequences $\{v^k\}$ and $\{w^k\}$ associated with the point Jacobi method are convergent to v^* and w^*, respectively and $v^* \leqslant w^*, v^k \leqslant x^k \leqslant w^k, \boldsymbol{p}(v^*, w^*) = \boldsymbol{0}, \boldsymbol{p}(w^*, v^*) = \boldsymbol{0}$. Hence, for each $x^0 \in [v^*, w^*]$ it holds $v^* \leqslant x^k \leqslant w^*$ $(k = 1, 2, \ldots)$. Thus, the continuous function $\boldsymbol{h}(\cdot)$ maps the bounded, closed and convex set $\Gamma = [v^*, w^*]$ into itself. From the Brouwer theorem [10] it follows that in the set Γ there exists at least one fixed point. Therefore, the set contains an element x such that $x = \boldsymbol{h}(x)$, which means that $\boldsymbol{p}(x) = \boldsymbol{0}$.

Part (ii) of Theorem 2 is proved similarly as Theorem 1. Part (iii) follows straightforwardly from the way in which the Gauss-Seidel (Jacobi) iterative process is performed.

Theorem 2 requires weaker assumptions than Theorem 1. The condition $\boldsymbol{D} \in M$ does not guarantee the global convergence of the block Gauss-Seidel and Jacobi methods. However, the fact that almost all elements of $\{x^k\}$ remain in some bounded set usually ensures the convergence to the solution.

For diode-transistor circuits described by (4) Theorem 2 holds under much weaker assumptions. To formulate a respective theorem we give up the requirements that characteristics $f_i(u_i)$ have zero slopes for large positive u_i. Instead, we assume that the last right segment of $f_i(u_i)$ has a sufficiently steep slope.

Theorem 3

If the matrix $\boldsymbol{B} = [b_{ij}]_{n\times n}$ (see formula (5)) belongs to the class M, then parts (i), (ii) and (iii) of Theorem 2 hold.

Proof

Using the idea as in Lemma 7 [22], it is easy to show that for each $x^0 \in \boldsymbol{R}^n$ there exist $v^0 \in \boldsymbol{R}^n$ and $w^0 \in \boldsymbol{R}^n$ fulfiling (11) and (12). The remaining part of the proof is similar as in Theorem 2.

Note

Let us consider a network consisting of transistors, positive resistors and independent voltage and current sources described by equation (4). If equation (4) is split into blocks in such a way that every block is associated only with one transistor and $\boldsymbol{B} \in M$, then it can be proved that each block has a unique solution and the point Gauss-Seidel (Jacobi) sequence is convergent to this solution.

In Theorems 1, 2 and 3 the key role is played by the class M of matrices. Definition and some properties of M-matrices enabling their identification are presented in Appendix.

4. APPLICATION OF THE BLOCK GAUSS-SEIDEL AND JACOBI METHODS TO ANALYSIS OF PIECEWISE-LINEAR DIODE-TRANSISTOR CIRCUITS

Let us consider a network consisting of transistors, diodes, positive linear resistors and independent voltage and current sources. By extracting transistors and diodes, a li-

near resistive n-port is obtained (as e.g. in [22]). To find an admittance representation of the n-port, the metod of systematic elimination [5] has been used. A temporary tree has been generated with the following preference: diode-controlled source combinations (of transistors), diodes and voltage sources. A set of preferred elements which does not belong to the tree indicates the points where some small resistors have to be included. Finally the tree is formed containing all preferred elements.

The efficiency of block relaxation method (Gauss-Seidel and Jacobi) depends on satisfying convergence conditions and on the structure of equations (quasi-diagonal). To check convergence conditions, a procedure of identification of M-matrices (see Appendix) has been developed.

To estimate the equation structure, the following parameter has been introduced defining some properties of the admittance matrix $\boldsymbol{C} = [c_{ij}]_{n\times n}$:

$$d_m(w_p) = \max_k d_k(w_p) \quad (k = 1, \ldots, v), \tag{34}$$

where $w_p = [w_{p1} \ldots w_{pv}]^T$ and w_{pi} $(i = 1, \ldots, v)$ denotes the number of equations forming i-th block,

$$d_k(w_p) = \frac{\sum_{j=1, j\neq k}^{v} ||C_{kj}||}{||C_{kk}||} \quad (k = 1, \ldots, v).$$

If the assumptions of Theorem 1, 2 or 3 are satisfied, if $d_m < 1$ and there exists an effective method for solving individual blocks, then usually the block Gauss-Seidel algorithm is faster than the well-known Katzenelson's and Newton-Raphson methods.

The simplified description of a computer algorithm of piecewise-linear diode-transistor circuits using the block Gauss-Seidel method is shown in Fig. 1. An important part

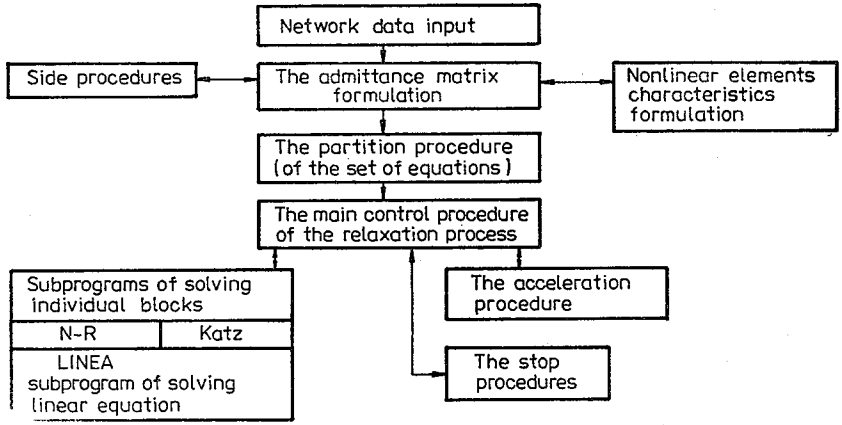


Fig. 1. The simplified scheme of a computer algorithm of solving piecewise-linear diode-transistor circuits using the block Gauss-Seidel method

of the program is the creation of the vector w_p. If the assumption of Theorem 1 is satisfied, then equation (4) is split in such a way that each block is associated with one transistor or diode. If weaker conditions hold (Theorem 2 or 3), then we form blocks for which d_m

is as small as possible and each block has a unique solution and satisfies requirements of convergence of the point Gauss-Seidel method. In the case when no convergence conditions are fulfilled or if $d_m > 1$, the blocks are successively enlarged until $d_k(w_p) < d_e$ ($d_e \cong 1$). An additional limitation is the dimension of an individual block which has to be smaller than some computed $w_{\max}$ depending on the number of network equations and the number of break-points of piecewise-linear functions.

To accelerate the convergence of the algorithm (close to the solution), a special procedure has been used. The idea of the procedure is similar as in the case of the linear iterative process [1]:

$$\begin{aligned} x^{k+1} &= \boldsymbol{H}_I z^k + \boldsymbol{e}_I \quad x^0 = z^0 \\ z^{k+1} &= x^{k+1} \quad \text{for} \quad k < k_a \\ z^{k+1} &= \mu z_i^k + (1-\mu) x^{k+1} \quad \text{for} \quad k = k_a. \end{aligned} \tag{35}$$

Under some assumptions the vector $\Delta x^k = x^k - x^{k-1}$ can be expressed as follows:

$$\Delta x^k = \sum_{i=1}^{n} d_i u_i (\lambda_i)^k,$$

where u_i denotes an eigenvector relating to the eigenvalue λ_i. For sufficiently large k it holds:

$$\begin{aligned} \Delta x_j^k &\cong d_m (u_m)_j (\lambda_m)^k \\ \Delta x_j^{k+1} &\cong d_m (u_m)_j (\lambda_m)^{k+1} \end{aligned} \qquad j = 1, \ldots, n$$

where index m relates to the eigenvalue of the maximum absolute value. Hence, we find

$$\beta_j^{k+1} = \frac{\Delta x_j^{k+1}}{\Delta x_j^k} \cong \lambda_m \quad j = 1, \ldots, n$$

Thus, for sufficiently large k the sequence $\{\Delta x_j^k\}$ ($j = 1, \ldots, n$) constitutes a geometric series. Setting $\mu = -\dfrac{\lambda_m}{1-\lambda_m}$ in formula (35) we obtain

$$z_j^{k+1} = x_j^k + \frac{\Delta x_j^{k+1}}{1-\lambda_m}. \tag{36}$$

For the nonlinear relaxation process a similar procedure can be formulated:

$$\begin{aligned} x^{k+1} &= H(z^k) \quad x^0 = z^0 \\ z^{k+1} &= x^{k+1} \quad \text{for } k < k_a \\ z_j^{k+1} &= x_j^k + \frac{\Delta x^{k+1}}{1-\beta_j^{k+1}} \quad \text{for} \quad k \geqslant k_a. \end{aligned} \tag{37}$$

To find k_a we calculate $|\Delta\beta_j^k| = |\beta_j^k - \beta_j^{k-1}|$ ($k = 2, 3, \ldots; j = 1, \ldots, n$). If $\max\limits_j |\beta_j^k| < 1$ and $|\Delta\beta_j^k| < \varepsilon_{AC}$ then $k_a = k$.

5. DISCUSSION AND CONCLUDING REMARKS

The following numerical examples are presented for illustration of the problems discussed in the paper.

Example 1

The equation (4) of a two-transistor network (Fig. 2) satisfies the conditions of The-

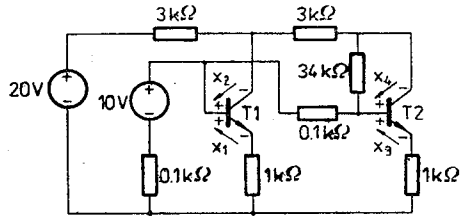


Fig. 2. A two-transistor network

orem 1. Matrix C and vector y are of the form:

$$C = \begin{bmatrix} 0.9183 & -0.0296 & -0.0745 & 0.0024 \\ -0.0296 & 0.6461 & 0.0024 & -0.3227 \\ -0.0745 & 0.0024 & 0.8439 & -0.0272 \\ 0.0024 & -0.3227 & -0.0272 & 0.3529 \end{bmatrix}, \quad y = \begin{bmatrix} -8.7110 \\ 3.5068 \\ -7.9424 \\ 0.2562 \end{bmatrix}$$

Table 1

Numerical solution of the circuit of Fig. 2.

Step	x_1 [V]	x_2 [V]	x_3 [V]	x_4 [V]	
0	0	0	0	0	
1	0.60126	0.58085	0.60189	0.59580	G-S/ /N-R/2
2	0.60134	0.58204	0.60189	0.59580	
1	0.60006	0.56261	0.59760	0.57772	G-S/Point
2	0.60007	0.57618	0.60089	0.58686	
3	0.60101	0.57914	0.60140	0.59138	
.	.	.	.	.	
.	.	.	.	.	
.	.	.	.	.	
13	0.60134	0.58204	0.60189	0.59580	
1	0.60006	0.56261	0.59760	0.57772	G-S/Point/ACC
.	.	.	.	.	
.	.	.	.	.	
.	.	.	.	.	
ACC 5	0.60126	0.58132	0.60177	0.59472	
6	0.60134	0.58204	0.60189	0.59580	
Solution	0.60134	0.58204	0.60189	0.59580	

Each transistor is described by the *DC* Ebers-Moll model containing diodes whose characteristics are approximated by piecewise-linear seven-segments function (see the legend the Table 2). The zero-valued starting vector has been used and the results of three types of iteration procedures are presented in Table 1 (the partition vector $w_p = [2\ 2]^T$).

Example 2

The transistor circuit of Fig. 3 has description (4) and the matrix $\boldsymbol{B}$ (5) belongs to

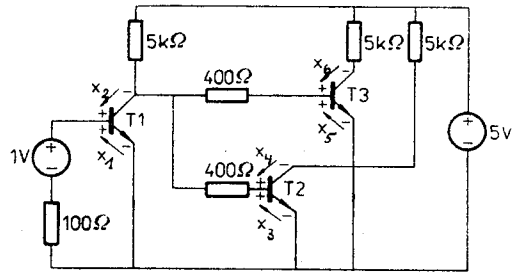


Fig. 3. A three-transistor network (NOR element)

the class M. Using the same diode characteristics as in Example 1, starting vector $x_0 = [0.41\ 0.41\ 0.51\ 0.51\ 0.51\ 0.51]^T$ and the partition vector $w_p = [2\ 2\ 2]^T$, the solution

$x_1 = 0.5772$ V, $x_2 = 0.5754$ V, $x_3 = 0.0018$ V, $x_4 = -4.9982$ V, $x_5 = 0.0018$ V, $x_6 = -4.9982$ V

was found after

20 steps of Katzenelson's method (number of long operations NLO = 4077),
4 steps of Newton-Raphson method (NLO = 712),
3 steps of block Gauss-Seidel procedure (NLO = 462),
5 steps of block Jacobi process (NLO = 865).

(*Note*: each block of the relaxation methods was solved by N-R method)

Table 2 contains a comparison between basic piecewise-linear methods. The columns a, b, c, d correspond with the following piecewise-linear approximations of diode characteristics:

(a) three segments: $f(u_d) = -36.72+67.85u_d+5.97|u_d-0.45|+61.88|u_d-0.55|$

(b) five segments: $f(u_d) = -1041.99+1743.32u_d+1.52|u_d-0.45|+8.63|u_d-0.5|+$
$+57.70|u_d-0.55|+1675.47|u_d-0.6|$

(c) seven segments: $f(u_d) = -1041.99+1743.32u_d+1.52|u_d-0.45|+4.14|u_d-0.5|+$
$+8.97|u_d-0.525|+23.21|u_d-0.55|+60.01|u_d-0.575|+1645.47|u_d-0.6|$

(d) nine segments: $f(u_d) = -398.58+654.26u_d+1.52|u_d-0.45|+8.63|u_d-0.5|+$
$+18.87|u_d-0.55|+17.64|u_d-0.5625|+28.37|u_d-0.575|+45.63|u_d-0.5875|+$
$+132.36|u_d-0.6|+401.23|u_d-0.625|$.

(The zero valued starting vector was used in Example 1 and the vector $x_0 = [0.41\ 0.41\ 0.41\ 0.51\ 0.51]^T$ in Example 2).

In both above examples (see Table 2), the block Gauss-Seidel method with "*natural*" partition (each nonlinear element corresponds to one simple block of the equation) is

Table 2

Comparison of some algorithms on the basis of analysis of circuits shown in Fig. 2 and 3

Ex	No	Method	W	Number of steps				NLO				Time $t_w = \frac{t}{t_k}$				ACC?
				a	b	c	d	a	b	c	d	a	b	c	d	*/–
Example 1	1	Katzenelson		4	14	14	17	414	611	1321	1603	1	1	1	1	
	2	Newt.-Raph.		2	3	3	5	104	204	252	500	0.45	0.27	0.29	0.40	
	3	G-S/Point	1	13	13	13	13	428	664	902	1184	1.48	1.24	0.95	0.99	—
	4	G-S/Katz	2	2	2	2	2	202	294	541	659	0.69	0.33	0.49	0.49	—
	5	G-S/N-R	2	2	2	2	2	100	192	256	434	0.52	0.28	0.29	0.38	—
	6	G-S/Point	1	6	6	6	6	172	328	454	624	0.58	0.32	0.38	0.43	*
Example 2	1	Katzenelson		6	11	17	16	1321	2311	3484	3315	1	1	1	1	
	2	Newt.-Raph.		3	4	6	6	390	616	712	1212	0.49	0.41	0.31	0.51	
	3	G-S/Point	1	12	13	13	13	732	1137	1488	1851	0.92	0.82	0.79	0.97	—
	4	G-S/Point	1	10	9	9	10	612	801	1056	1455	0.75	0.53	0.53	0.72	*
	5	G-S/Katz	3	4	4	4	4	905	1264	1564	1716	0.95	0.73	0.59	0.67	—
	6	G-S/Katz	2	4	3	3	3	493	619	798	901	0.67	0.46	0.41	0.45	—
	7	G-S/N-R	3	4	4	4	4	385	622	825	934	0.62	0.50	0.41	0.46	—
	8	G-S/N-R	2	4	3	3	3	306	358	462	642	0.54	0.35	0.28	0.38	—

W — block dimention (all blocks have the same dimention)
NLO — number of long operations
t_w — relative time of the procedure with respect to the Katzenelson method.

much better than Katzenelson's algorithm and often better than the Newton-Raphson piecewise-linear procedure.

Example 3

The equation of an operational amplifier stage (Fig. 4) does not fulfil any convergence conditions but the parameter $d_m(34)$ is close to one ($w_p = [4\ 4\ 4]^T$).

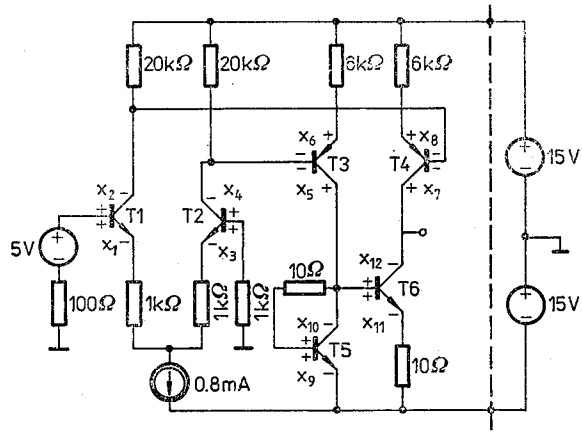


Fig. 4. An operation amplifier stage

The solution:

$$\begin{aligned} x_1 &= 0.517\ \text{V} & x_5 &= 0.001\ \text{V} & x_9 &= 0.450\ \text{V} \\ x_2 &= -5.922\ \text{V} & x_6 &= -29.549\ \text{V} & x_{10} &= 0\ \text{V} \\ x_3 &= -3.681\ \text{V} & x_7 &= 0.548\ \text{V} & x_{11} &= -0.450\ \text{V} \\ x_4 &= -14.997\ \text{V} & x_8 &= 0.498\ \text{V} & x_{12} &= -25.970\ \text{V} \end{aligned}$$

was obtained after

19 steps of Katzenelson's algorithm (NLO = 19 207)

12 steps of block Gauss-Seidel method (NLO = 5 745)

(starting vector x^0:$x_i^0 = 0.51$ $i = 1, \ldots, 12$) using the Newton-Raphson algorithm for solving individual blocks.

On the basis of the above discussion the following conclusions can be formulated.

1) Block relaxation methods lead to an effective algorithm of D. C. piecewise-linear analysis of nonlinear electronic circuits. The algorithm is simple and easy for computer implementation.

2) Theorems 1, 2 and 3 contain some useful results concerning solvability and computation of nonlinear resistive networks using the block Gauss-Seidel and Jacobi methods.

3) Efficiency of the block Gauss-Seidel and Jacobi methods depends on the splitting of equations into blocks. The quasi-diagonal structure is desirable. In the case of transistor circuits both equations associated with a transistor should be included into the same block. The splitting using structural priority is suitable only if $d_m \gg 1$.

4) The block Gauss-Seidel method is usually more effective then the block Jacobi method. Both methods are similar for equations having a good diagonal structure.

5) If the conditions formulated in the paper are satisfied, the block Gauss-Seidel iterative process is usually faster than the Katzenelson or Newton-Raphson methods. The block Gauss-Seidel method together with the Newton-Raphson method used for solving individual blocks is recommended. The acceleration procedure, described in Section 4, is very useful and considerably improves the efficiency of the algorithm.

APPENDIX

We consider a real square matrix $A = [a_{ij}]_{n\times n}$.

Definition 1 [12]

A matrix A is Ω-diagonally dominant if there is a network $\Omega = (N, \wedge)$ so that:
(i) A is diagonally dominant with respect to the network Ω.
(ii) there is a nonempty subset J of N such that $\forall i \in J: |a_{ij}| > \sum_{j=1, j\neq i}^{n} |a_{ij}|$ and $\forall i \notin J$ there is a path in $\wedge$ from i to some $j = j(i) \in J$.

Definition 2

A matrix A where $\forall i \neq j$ $(i, j = 1, \ldots, n)$: $a_{ij} \leqslant 0$ is said to be M-matrix (belongs to the class M) if it is nonsingular and all elements of A^{-1} are nonnegative.
Some properties of M-matrix

a) All diagonal elements of a M-matrix are positive [24].
b) If a matrix A satisfying condition: $a_{ij} < 0$ for $i \neq j$ $(i, j = 1, \ldots, n)$ is strictly diagonally dominant then it is a M-matrix [24].
c) If a matrix A satisfying conditions: $a_{ii} > 0$, $a_{ij} \leqslant 0$ for $i \neq j$ $(i, j = 1, \ldots, n)$ is Ω-diagonally dominant then it is an M-matrix [12].
d) The Stieltjes matrix (real, symmetric, positive definite, having nonpositive off-diagonal elements) is an M-matrix [24].
e) If A is an M-matrix and A_0 is a matrix obtained after replacing some (arbitrary) off-diagonal elements of A with zeros, then A_0 is also an M-matrix [24].
f) A matrix A having all off-diagonal elements nonpositive is an M-matrix if its all principal angle minors are positive. Generally, the best way of testing if a matrix belongs to the class M is determining all principal angle minors performing the forward course of the Gaussion elimination process.

REFERENCES

1. A. Bjorck, G. Dahlquist: *Metody numeryczne.* Warszawa: PWN, 1983
2. M. J. Chien, E. S. Kuh: *Solving Piecewise-Linear Equations for Resistive Networks,* Circuit Theory and Applications, 1976, vol. 4, pp. 3—24
3. L. O. Chua, R. L. P. Ying: *Finding all Solutions of Piecewise-Linear Circuits.* Circuit Theory and Applications, 1982, vol. 10, pp. 201—229
4. L. O. Chua, R. L. P. Ying: *Canonical Piecewise-Linear Analysis,* IEEE Transactions on Circuits and Systems, 1983, vol. CAS-30, No. 3, pp. 125—140

5. L. O. Chua, P. M. Lin: *Komputerowa analiza układów elektronicznych.* Warszawa: WNT, 1981
6. T. Fujisawa, E. S. Kuh, T. Ohtsuki: *A Sparse Matrix Method for Analysis of Piecewise-Linear Resistive Networks.* IEEE Transactions on Circuit Theory, 1972, vol. CT-19, No. 6, pp. 571—584
7. M. Ilic-Spong, I. N. Katz, H. Dai, J. Zaborszky: *Block Diagonal Dominance for Systems of Nonlinear Equations with Application to Load Flow Calculations in Power Systems.* Mathematical Modelling, 1984, vol. 5, pp. 275—297
8. J. Katzenelson: *An Algorithm for Solving Nonlinear Resistor Networks.* The Bell System Techn. Journal, 1965, vol. 48, pp. 1605—1620
9. M. Krätzschmer: *Inverse-isotone Mappings, M-matrices, M-functions and their application to nonlinear block-Gauss-Seidel iterations.* Math. Res., Berlin: Akademie-Verlag, 1984, vol. 21 pp, 9—18
10. J. Kudrewicz: *Analiza funkcjonalna dla automatyków i elektroników.* Warszawa: PWN, 1976
11. L. G. Mao, I. N. Hajj: *A Piecewise-Linear Approach to DC Analysis of Large-Scale Integrated Circuits,* Proc. Int. Symp. Net. Theor., Sarajevo, 1984
12. J. J. More: *Nonlinear Generalizations of Matrix Diagonal Dominance with Application to Gauss-Seidel Iterations.* SIAM J. Numer. Anal., 1972, vol. 9, No. 2
13. J. M. Ortega, W. C. Rheinboldt: *Monotone Iterations for Nonlinear Equations with Application to Gauss-Seidel Methods.* SIAM J. Numer. Anal., 1967, vol. 4, No. 2, pp. 171—190
14. S. Osowski, A. Tobola: *Projektowanie obwodów nieliniowych aproksymowanych wieloodcinkowo przy użyciu metody optymalizacji.* Mat. IX SPETO, Gliwice, 1986
15. S. Osowski, A. Cichocki, S. Filipowicz: *Some aspects of nonlinear function simulation.* Rozprawy Elektrotechniczne, 1986, z. 2
16. W. C. Rheinboldt: *On M-functions and Their Application to Nonlinear Gauss-Seidel Iterations and to Network Flows.* J. of Math. Anal. and Appl., 1970, vol. 32, pp. 274—307
17. W. C. Rheinboldt, J. S. Vandergraft: *On Piecewise Affine Mappings in Rn.* SIAM. J. Appl. Math., 1975, vol. 29, No. 4, pp. 680—688
18. I. W. Sandberg, A. N. Willson: *Some Network-Theoretic Properties of Nonlinear DC Transistor Networks.* The Bell S. Techn. J., 1969, vol. 48, No. 5, pp. 1293—1311
19. S. Schechter: *Iterations methods for nonlinear problems.* Trans. Am. Math. Soc., 1962
20. M. Tadeusiewicz: *Analysis of Piecewise-Linear Resistive Networks Using the Gauss-Seidel Method.* Proc. ECCTD'83, Stuttgart, 1983
21. M. Tadeusiewicz: *D. C. Analysis of Piecewise-Linear Networks Using the Gauss-Seidel Method.* Int. J. Comp. Math. El. Electron. Eng. (COMPEL), 1986, vol. 5, No. 4, pp. 195—205
22. M. Tadeusiewicz: *On the Solvability and Computation of D. C. Transistor Networks,* Circuit Theory and Applications. 1981, vol. 9, pp. 251—267
23. M. Tadeusiewicz, M. Ossowski: *Application of the Block Gauss-Seidel Method to DC Piecewise-Linear Analysis.* Proceedings of the AMSE Conference, Sorrento, 1986
24. R. S. Varga: *Matrix Iterative Analysis.* New Jersey: Englewood Cliffs, 1962

M. TADEUSIEWICZ, M. OSSOWSKI

ANALIZA ODCINKOWO-LINIOWA REZYSTANCYJNYCH OBWODÓW NIELINIOWYCH ZA POMOCĄ BLOKOWYCH METOD GAUSSA-SEIDELA I JACOBIEGO

Streszczenie

Tematem pracy jest odcinkowo-liniowa analiza nieliniowych obwodów rezystancyjnych przy wykorzystaniu blokowych metod relaksacyjnych Gaussa-Seidela i Jacobiego. Sformułowano warunki wystarczające istnienia jednego rozwiązania i zbieżności ciągów iteracyjnych dla dowolnego przybliżenia początkowego. Rozpatrzono problem rozwiązalności i zbieżności przy różnych założeniach. Omówiono pewne aspekty algorytmu i podano schemat programu komputerowego. Przytoczono kilka przykładów i dokonano wnikliwej oceny blokowych metod Gaussa-Seidela i Jacobiego na tle znanych metod analizy odcinkowo-liniowej. Szczególną uwagę poświęcono obwodom diodowo-tranzystorowym.

M. TADEUSIEWICZ, M. OSSOWSKI

ANALYSE LINÉAIRE PAR MORCEAUX DES CIRCUITS RÉSISTANTS NON-LINÉAIRES OÙ L'ON A APPLIQUÉ LES MÉTHODES DE BLOC DE GAUSS-SEIDEL ET DE JACOBI

Résumé

L'analyse linéaire par morceaux dans laquelle on a appliqué des relaxantes méthodes de bloc de Gauss-Seidel et de Jacobi fait le sujet de l'article. On a formulé les conditions suffisantes de l'existence d'une solution unique et de la convergence des suites itératives pour une approximation initiale arbitraire. On a considéré le problème de l'existence des solutions et de la convergence sous diverses hypotheses. On a traité certains aspects de l'algorithme et on a donné le schéma d'un programme pour l'ordinateur. On a cité quelques exemples et on a fait une analyse approfondie des méthodes de bloc de Gauss-Seidel et de Jacobi tout en la comparant avec les méthodes connues de l'analyse linéaire par morceaux. On a porté particulierement attention aux circuits de diodes et de transistors.

M. TADEUSIEWICZ, M. OSSOWSKI

STÜCKWEISE LINEARE ANALYSE DER NICHTLINEAREN RESISTANZKREISE MITTELS GAUSS-SEIDELSCHEN UND JACOBI-BLOCKVERFAHREN

Zusammenfassung

In dem Beitrag wird die stückweise lineare Analyse der nichtlinearen Resistanzkreise mittels Gauss-Seidelschen- und Jacobi-Blockverfahren erörtert. Es wurden ausreichende Bedingungen für das Bestehen einer einzigen Lösung und Konvergenz iterativer Folgen bei beliebiger Anfangsapproximation angegeben. Das Problem der Lösbarkeit und Konvergenz unter verschiedenen Voraussetzungen wurde diskutiert. Einige Aspekte des Algorithmus wurden erwogen, und ein Computerprogrammschema erstellt. Einige Beispiele hierfür wurden genannt, und eine eingehende Auswertung der Gauss-Seidelschen- und Jacobi-Blockverfahren im Vergleich mit bekannten Methoden der stückweisen linearen Analyse wurde durchgeführt. Besondere Aufmerksamkeit wurde den Dioden-Transisorenkreisen geschenkt.

М. ТАДЕУСЕВИЧ, М. ОССОВСКИ

КУСОЧНО-ЛИНЕЙНЫЙ АНАЛИЗ РЕЗИСТИВНЫХ НЕЛИНЕЙНЫХ ЦЕПЕЙ ПРИ ПОМОЩИ БЛОЧНЫХ МЕТОДОВ ГАУССА-ЗЕЙДЕЛЯ И ЯКОБИ

Резюме

Обсужден кусочно-линейный анализ нелинейных резистивных цепей с использованием блочных релаксационных методов Гаусса-Зейделя и Якоби. Сформулированы достаточные условия существования единственного решения и сходимости итерационных последовательностей при любой начальной точке. Исследован вопрос существования решения и сходимости при различных условиях. Обсуждены некоторые аспекты алгоритма и представлена схема программы для ЕВМ. Приведено несколько примеров и оценка блочных методов Гаусса-Зейделя и Якоби на фоне известных методов кусочно-линейного анализа. Особенное внимание посвящено цепям с диодами и транзисторами.

ROZPRAWY ELEKTROTECHNICZNE, 1989, 35, z. 1, ss. 31-44

621.3.016.2

Definicje pewnych mocy dla układów wielozaciskowych o przebiegach odkształconych

MAREK BRODZKI, MARIAN PASKO

Instytut Podstawowych Problemów Elektrotechniki i Energoelektroniki, Politechnika Śląska

Otrzymano 1987.10.20

Autoryzowano do druku 1988.05.27

W niniejszej pracy przedstawiono uogólnienie pojęcia rozkładu ortogonalnego prądu odkształconego elementu dwuzaciskowego, który to rozkład określony został w pracy [4], na przypadek przebiegów odkształconych na wejściu elementu $(n+1)$ — zaciskowego $(n \in N)$. W tym celu zdefiniowana została odpowiednia przestrzeń Hilberta wraz z pewnym zamkniętym układem ortonormalnym. Następnie w ślad za tym rozkładem prądów, zachodzącym przy spełnieniu przez ww. element pewnych warunków, wprowadzono moce z nim związane. Podano motywację fizykalną powyższego postępowania. Przeprowadzono rozkład omawianej przestrzeni Hilberta na sumę prostą trzech parami ortogonalnych podprzestrzeni.

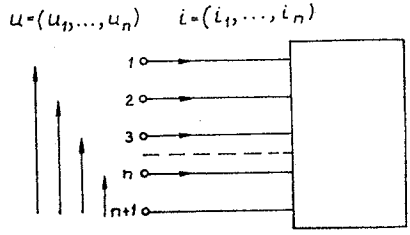
1. WSTĘP

Podstawowym pojęciem mocy w teorii obwodów elektrycznych jest pojęcie mocy chwilowej. Potrzeba wprowadzania definicji dowolnej wielkości w fizyce powinna być umotywowana. Motywację taką dla mocy chwilowej stanowi odpowiednia zasada zachowania, będąca konsekwencją obu praw Kirchhoffa ([3], str. 42). Zasada ta ma swój odpowiednik w teorii pola elektromagnetycznego oraz szerzej w teorii pól sprzężonych. Jednakże na zdefiniowaniu mocy chwilowej nie poprzestajemy, określając dla przebiegów okresowych moc czynną jako wartość średnią z mocy chwilowej (zakładając, że funkcja tej mocy jest całkowalna w przedziale $\langle 0; T \rangle$). Motywacją dla tego pojęcia może być dobrze znany fakt, że w układach o dostatecznie dużej bezwładności termicznej lub mechanicznej obserwując temperaturę lub np. prędkość, czynimy niewielki błąd operując zamiast pojęciem mocy chwilowej pojęciem mocy czynnej. Użycie zaś tej ostatniej upraszcza wyznaczanie przebiegów czasowych temperatury lub prędkości. Z kolei dla przebiegów sinusoidalnych i elementów dwuzaciskowych powstaje zagadnienie strat (oraz ich minimalizacji) w przewodach doprowadzających do nich moc czynną. Wielkością o charakterze mocy, służącą do oceny tych strat jest moc bierna, rozumiana jako część urojona mocy symbolicznej (zespolonej). Moc tę można kompensować, a prąd doprowadzający określoną moc czynną do danego elementu dwuzaciskowego minimalizować, przez równoległe dołączenie doń kondensatora (lub cewki zależnie od znaku tej mocy).

W tym mniej więcej momencie kończy się powszechna zgoda w zakresie definicji różnych mocy. W przypadku przebiegów odkształconych występują już kontrowersje. Obszerne ich omówienie, jak też literaturę dotyczącą teorii mocy, znajdzie czytelnik w pracy [4]. Jej autor, L. Czarnecki, proponuje tam (str. (94—97)), w przypadku przebiegów odkształconych, rozkład funkcji prądu pobieranego przez dwuzaciskowy element (opisanej jako punkt przestrzeni $L^2(\langle 0; T\rangle)$) na trzy wzajemnie ortogonalne składniki. Jeden równoległy do funkcji napięcia zasilającego ten element, drugi — kompensowalny z dowolną dokładnością, w sensie przyjętej normy, z pomocą elementu dwuzaciskowego złożonego ze skończonej liczby elementów L, C oraz „*resztę*". Z tym rozkładem prądu związana jest definicja trzech mocy: pierwsza jest mocą czynną, druga zdaje sprawę z możliwości kompensacji wspomnianym układem L, C, trzecia reprezentuje część niekompensowalną prądu z pomocą takiego układu. Opisaną procedurę można spróbować uogólnić w dwóch kierunkach: jeden dotyczy przebiegów czasowych prądów i napięć, które mogą być nieokresowe, drugi — elementów wielozaciskowych. W niniejszej pracy zajmiemy się tym drugim problemem, pozostając przy przebiegach odkształconych.

2. SFORMUŁOWANIE I ROZWIĄZANIE PROBLEMU

Niech więc danym będzie element $(n+1)$–zaciskowy $(n \in N)$ zilustrowany rys. 1.



Rys. 1. Element $(n+1)$-zaciskowy

Wnętrze tego elementu może być nieznane („czarna skrzynka"). Zakładamy, że jego napięcia oraz prądy opisane są funkcjami rzeczywistymi zmiennej rzeczywistej (czasu) u_α, i_α, $\alpha \in \{1, \ldots, n\}$, okresowymi o okresie T, mierzalnymi (w sensie Lebesgue'a) na przedziale domkniętym $\langle 0; T\rangle$. (Gdy będzie rzeczą obojętną, czy chodzi o napięcie, czy o prąd będziemy stosować oznaczenie f_α.) Funkcje te mają poza tym całkowalny (w sensie Lebesgue'a) kwadrat, tzn.:

$$\int_0^T f_{\alpha_j}^2(t)\,\mathrm{d}t < \infty. \tag{1}$$

W dalszym ciągu będziemy zakładać, że nasze funkcje f_α mają dziedzinę obciętą do zbioru $\langle 0; T\rangle$. Jeśli operujemy jednym stanem napięciowo-prądowym na zaciskach tego elementu, jest rzeczą obojętną jaka zależność wiąże funkcje prądu $i = (i_1, \ldots, i_n):\langle 0; T\rangle \to$ $\to R^n$ z analogicznymi funkcjami napięcia $u = (u_1, \ldots, u_n):\langle 0; T\rangle \to R^n$ (gdy zmienna u

przebiega jakiś zbiór zawierający więcej niż jeden element). Może to być wówczas nawet element nieliniowy (por.: [4], str. (100—102)).

Chcąc teraz dla takich elementów rozwinąć pożądaną teorię mocy, trzeba rozpocząć od konstrukcji odpowiedniej przestrzeni Hilberta. Najpierw więc, jak wygląda zbiór tej przestrzeni? Rzecz jasna, powinien on składać się z opisanych punktów $f = (f_1, \ldots, f_n)$, a dokładniej, nasz ciąg powinien mieć jako wyrazy klasy opisanych funkcji różniących się na zbiorach miary zero, tzn. takich, dla których całka (1) jest identyczna ([9], str. 297). Dla uproszczenia, będziemy oznaczać zarówno funkcję f_α jak i jej klasę $[f_\alpha]$ tym samym symbolem f_α oraz zamiast klasą $[f_\alpha]$ operować jej reprezentantem f_α jak to często praktykuje się w tym przypadku w literaturze matematycznej (np. w pozycji [6]). Powód, dla którego trzeba operować klasami funkcji poznamy za chwilę. Nasz zbiór oznaczymy $L_n^2(\langle 0; T\rangle)$.

Następnie w zbiorze tym wprowadzamy strukturę przestrzeni liniowej definiując działania dodawania elementów i mnożenia przez liczby rzeczywiste z pomocą wzorów

$$f+g = (f_1+g_1, \ldots, f_n+g_n), \tag{2}$$

$$c \cdot f = (cf_1, \ldots, cf_n), \tag{3}$$

dla dowolnych punktów $f, g \in L_n^2(\langle 0; T\rangle)$, $c \in R$.

Dodawanie dwóch klas funkcji rozumiemy jako zdefiniowane przez dodawanie dowolnych ich elementów. Mnożenie klasy przez liczbę rzeczywistą — jako mnożenie przez tę liczbę dowolnego jej elementu. Są to działania poprawnie określone. Na mocy nierówności $(f_\alpha(t)+g_\alpha(t))^2 \leqslant 2(f_\alpha^2(t)+g_\alpha^2(t))$ $t \in R$ stwierdzamy, że z przynależności funkcji (ściślej ciągów klas funkcji) f i g do zbioru $L_n^2(\langle 0; T\rangle)$ wynika również przynależność funkcji $f+g$ do niego. Przynależność funkcji $c \cdot f$ jest wówczas oczywista. Uporządkowana czwórka $(L_n^2(\langle 0; T\rangle), R_c, +, \cdot)$ jest więc przestrzenią liniową nad ciałem liczb rzeczywistych R_c.

Podkreślamy, że argumenty i wartości naszych funkcji prądu i napięcia pozbawione są wymiarów fizykalnych i dlatego możemy traktować je jako elementy tej samej przestrzeni liniowej. Przy uwzględnieniu tych wymiarów trzeba operować osobno odpowiednią przestrzenią prądową i napięciową. Można też wówczas zbudować odwzorowanie (bijekcję) z jednej do drugiej przestrzeni, polegającą na wzajemnie jednoznacznym przyporządkowaniu sobie tych elementów przestrzeni, które stają się identyczne po „zapomnieniu" wymiarów i następnie operować jedną z tych przestrzeni. Powód, dla którego wolimy operować wielkościami bezwymiarowymi, jak też porównanie rachunków „bezwymiarowych" i „wymiarowych" opisane są w artykule [2].

Pozostaje nam wyprowadzić iloczyn skalarny $(|)_n : L_n^2(\langle 0; T\rangle) \times L_n^2(\langle 0; T\rangle) \to R$. Podamy go z pomocą wzoru

$$(f|g)_n = \sum_{\alpha=1}^{n} \frac{1}{T} \int_0^T f_\alpha(t) g_\alpha(t) \mathrm{d}t. \tag{4}$$

Całkowalność iloczynu funkcji $f_\alpha g_\alpha$ wynika ze spełnienia nierówności $f_\alpha(t)g_\alpha(t) \leqslant \frac{1}{2}(f_\alpha^2(t)+$ $+g_\alpha^2(t))$, $t \in R$. Sprawdzenie spełnienia czterech aksjomatów iloczynu skalarnego, przez odwzorowanie (4), które dla wygody czytelnika cytujemy (dla naszego modelu) za książką [6], str. 62

$$\Lambda_{f,g\in L_n^2(\langle 0;T\rangle)}((f|g)_n = (g|f)_n), \tag{5}$$

$$\Lambda_{f,g,h\in L_n^2(\langle 0;T\rangle)}((f+g|h)_n = (f|h)_n + (g|h)_n), \tag{6}$$

$$\Lambda_{(c\in R)\wedge(f,g\in L_n^2(\langle 0;T\rangle))}((c\cdot f|g)_n = c(f|g)_n), \tag{7}$$

$$\Lambda_{f\in L_n^2(\langle 0;T\rangle)}((f\neq 0)\Rightarrow((f|f)_n > 0))\wedge((0|0)_n = 0)), \tag{8}$$

jest proste. Trzeba tu wyjaśnić, że w aksjomacie (8) symbol 0 występuje w dwóch znaczeniach: oprócz normalnego oznaczenia liczby rzeczywistej zero oznacza element zerowy naszej przestrzeni liniowej. Gdyby nie przyjąć umowy, że elementami naszej przestrzeni liniowej są ciągi złożone z klas funkcji różniących się na zbiorach miary zero, aksjomat (8), jak łatwo spostrzec, nie byłby spełniony. Mamy więc przestrzeń unitarną $((L_n^2(\langle 0; T\rangle), R_c, +, \cdot), (|)_n)$.

Aby stwierdzić, czy jest ona przestrzenią Hilberta, trzeba sprawdzić, czy jest zupełną. Wprowadzimy normę $\|\ \|_n$ indukowaną przez iloczyn skalarny $(|)_n$

$$\|f\|_n = \sqrt{(f|f)_n}, \quad f\in L_n^2(\langle 0; T\rangle). \tag{9}$$

Wprowadźmy również w podobny sposób normę $\|\ \|$ w przestrzeni $((L^2(\langle 0; T\rangle), R_c, +, \cdot), (|))$. (Iloczyn skalarny podany jest wówczas wzorem (4), gdzie: $n = 1$.) Teraz wystarczy zauważyć, że jeśli mamy ciąg $\overset{(f)}{k}$, $k\in N_0 = N\cup\{0\}$ spełniający warunek Cauchy'ego ([6], str. 25) w naszej pierwszej przestrzeni, to każdy ciąg $\overset{(f_\alpha)}{k}$, dla ustalonego wskaźnika $\alpha\in\{1, \ldots, n\}$, spełnia też warunek Cauchy'ego w drugiej przestrzeni. Na mocy zupełności tej drugiej ([6], str. 66) implikuje to, że ciągi $\overset{(f_\alpha)}{k}$ (dla ustalonych „α") zbieżne są do pewnych punktów zbioru $L^2(\langle 0; T\rangle)$. To zaś implikuje, że tematowy ciąg $\overset{(f)}{k}$ zbieżny jest do pewnego punktu zbioru $L_n^2(\langle 0; T\rangle)$. Czyli przestrzeń unitarna $((L_n^2(\langle 0; T\rangle), R_c, +, \cdot), (|)_n)$ jest zupełna i staje się przestrzenią Hilberta.

Ponieważ rozkładu prądu naszego elementu dokonywać będziemy w oparciu o pewien wyróżniony układ ortonormalny w skonstruowanej przestrzeni Hilberta, więc teraz podamy go. Stanowi on następujący zbiór:

$$\begin{aligned}&\{(1, 0, \ldots, 0), \ldots, (0, \ldots, 0, 1), \ldots,\\ &(\sqrt{2}\cosh\omega(\cdot), 0, \ldots, 0), \ldots, (0, \ldots, 0, \sqrt{2}\cosh\omega(\cdot)),\\ &(\sqrt{2}\sinh\omega(\cdot), 0, \ldots, 0), \ldots, (0, \ldots, 0, \sqrt{2}\sinh\omega(\cdot)), \ldots\},\end{aligned} \tag{10}$$

gdzie $\omega = \dfrac{2\pi}{T}$, $h\in N$.

Jest to oczywiście zbiór przeliczalny. Jeśli jego elementy ponumerujemy, począwszy od liczby zero, w kolejności napisania ich we wzorze (10) — otrzymamy ciąg $\overset{(e)}{k}$, $k\in N_0$. Łatwo wykazać, że:

$$(\overset{(e)}{k}|\overset{(e)}{l}) = \delta_{kl} = \begin{cases}0, & k\neq l,\\ 1, & k = l,\end{cases} \quad k, l\in N_0, \tag{11}$$

czyli, że powyższy układ jest ortonormalny. W przestrzeni Hilberta $((L^2(\langle 0; T\rangle), R, +, \cdot)$ $(|))$ mamy podobny układ ortonormalny $\{1, \ldots, \sqrt{2}\cosh\omega(\cdot), \sqrt{2}\sinh\omega(\cdot), \ldots\}$. Układ ten jest zamknięty, ponieważ zbiór kombinacji liniowych tych funkcji jest gęsty w tej przestrzeni ([6], str. 76, 78, 79). Fakt ten w prosty sposób decyduje o tym, że zbiór kombinacji liniowych funkcji należących do układu ortonormalnego (10) jest również gęsty w przestrzeni Hilberta $((L_n^2(\langle 0; T\rangle), R_c, +, \cdot), (|)_n)$., czyli że układ (10) jest zamknięty.

Wiadomo ([6], str. 75, 76), że względem takiego układu dowolny element $f \in L_n^2(\langle 0; T\rangle)$ ma następujący szereg Fouriera (zbieżny do „f"):

$$f = \sum_{k=0}^{\infty} \underset{k}{(f|e)_n} \cdot \underset{k}{e}. \tag{12}$$

Ze względu na wygodę przeprowadzanych rachunków, wprowadzając metodę symboliczną, zamiast wzoru (12) stosować będziemy wzór

$$f_\alpha = F_{\alpha 0} + \sqrt{2}Re \sum_{h=1}^{\infty} F_{\alpha h}\exp\left(jh\omega(\cdot)\right), \alpha \in \{1, \ldots, n\}. \tag{13}$$

Związki pomiędzy współczynnikami $\underset{k}{(f|e)_n}$ szeregu Fouriera (12) a odpowiednimi współczynnikami $F_{\alpha 0}$, $F_{\alpha h}$ są, na podstawie znanych z metody symbolicznej rozważań, łatwe do ustalenia i nie będziemy ich tu wypisywać. We wzorze (13) operujemy celowo funkcjami, a nie ich wartościami, aby przypomnieć w ten sposób, że szereg po prawej stronie wcale nie musi być w każdym punkcie $t \in \langle 0; T\rangle$ zbieżny do „$f_\alpha(t)$". Zachodzi tu zbieżność, w sensie przyjętej normy, do elementu f, a ściślej, dotyczy to klas funkcji różniących się na zbiorze miary zero. Zgodnie ze zwyczajami przyjmowanymi często przez matematyków, funkcję posiadającą wartość $f(t)$ oznaczamy $f(\cdot)$.

Zamkniętość układu ortonormalnego (10) oznacza, dla dowolnego elementu $f \in L_n^2(\langle 0; T\rangle)$, spełnienie wzoru

$$\|f\|_n^2 = \sum_{k=0}^{\infty} \underset{k}{(f|e)_n^2}. \tag{14}$$

W oparciu o ten fakt, metodą podobną do zastosowanej w pracy [8], str. 5, 6, wykazujemy dla dowolnych elementów $f, g \in L_n^2(\langle 0; T\rangle)$ wzór

$$(f|g)_n = Re \sum_{h=0}^{\infty} F_{\alpha h} G_{\alpha h}^*, \tag{15}$$

gdzie dla uproszczenia zapisu stosujemy od tego momentu konwencję sumacyjną Einsteina odnośnie wskaźnika α (oraz innych wskaźników np. β, γ numerujących zaciski opisywanego elementu) powtarzającego się w iloczynie w zakresie: $\alpha \in \{1, \ldots, n\}$. Dla wskaźnika h nie przyjmujemy tej konwencji.

Po tych przygotowaniach możemy przystąpić do zapowiedzianego rozkładu prądu odbiornika. Niech zatem u, $i \in L_n^2(\langle 0; T\rangle)$ oraz

$$u_\alpha = U_{\alpha 0} + \sqrt{2}Re \sum_{h=0}^{\infty} U_{\alpha h}\exp\left(jh\omega(\cdot)\right), \tag{16}$$

$$i_\alpha = I_{\alpha 0} + \sqrt{2}Re \sum_{h=0}^{\infty} I_{\alpha h} \exp(jh\omega(\cdot)), \tag{17}$$

gdzie

$$I_{\alpha 0} = G_{\alpha\beta 0} U_{\beta 0}, \quad I_{\alpha h} = Y_{\alpha\beta h} U_{\beta h}, \quad Y_{\alpha\beta h} = G_{\alpha\beta h} + jB_{\alpha\beta h}, \; h \in N. \tag{18}$$

Jeśli nasz element, umownie zwany odbiornikiem, jest opisany (np. przy znanym wnętrzu) układem równań różniczkowych liniowych o stałych parametrach i o postaci normalnej, lub też z punktu widzenia zacisków wejściowych splotem, będąc w takim sensie liniowym, to wzory (18) spełnione są dla ustalonych parametrów $G_{\alpha\beta 0}$, $Y_{\alpha\beta h}$ przy dowolnych wartościach $U_{\beta 0}$, $U_{\beta h}$, przy których tylko mamy wzór $u \in L_n^2(\langle 0; T\rangle)$ oraz prąd i, określony wzorem (17) ($i \in L_n^2(\langle 0; T\rangle)$ i wyliczony na podstawie analizowanych równań różniczkowych lub też splotu, stanowi odpowiedź na napięcie wymuszające u. Dla wygody dalszych rozważań przyjmujemy wzór (18) niezależnie od sposobu opisu odbiornika, byle tylko był spełniony warunek $u, i \in L_n^2(\langle 0; T\rangle)$. W przypadku odbiornika nieliniowego, opisanego jednym tylko napięciem u i prądem i wybór współczynników $G_{\alpha\beta 0}$, $Y_{\alpha\beta h}$ we wzorze (18), który teraz chcielibyśmy traktować jako definicyjny dla nich (por.: [4], str. 101), byłby niejednoznaczny. Można wówczas przyjąć, że:

$$I_{\alpha 0} = G_{|\alpha|0} U_{\alpha 0}, \; I_{\alpha h} = Y_{|\alpha|h} U_{\alpha h}, \text{ jeśli tylko } U_{\alpha 0}, U_{\alpha h} \neq 0, \tag{19}$$

gdzie ujęcie wskaźnika α w dwie pionowe kreski oznacza zakaz sumowania podług niego. Oto nasz rozkład:

$${}_a i_\alpha = GU_{\alpha 0} + \sqrt{2}Re \sum_{h=1}^{\infty} GU_{\alpha h} \exp(jh\omega(\cdot)), \tag{20}$$

$${}_s i_\alpha = (G_{\alpha\beta 0} - G\delta_{\alpha\beta}) U_{\beta 0} + \sqrt{2}Re \sum_{h=1}^{\infty} (G_{\alpha\beta h} - G\delta_{\alpha\beta}) U_{\beta h} \exp(jh\omega(\cdot)), \tag{21}$$

$${}_r i_\alpha = \sqrt{2}Re \sum_{h=1}^{\infty} jB_{\alpha\beta h} U_{\beta h} \exp(jh\omega(\cdot)), \tag{22}$$

gdzie

$$G = \frac{P}{||u||_n^2}, \quad ||u||_n \neq 0, \; P = (u|i)_n. \tag{23}$$

Oczywiście „P" oznacza, w myśl wzoru (4) moc czynną pobieraną przez nasz odbiornik. Z widocznych powodów przyjmujemy założenie $||u||_n \neq 0$. Łatwo spostrzec, że z warunku $u, i \in L_n^2(\langle 0; T\rangle)$ wynika warunek ${}_a i, {}_s i, {}_r i \in L_n^2(\langle 0; T\rangle)$.

Stosując wzór (15), skontrolujemy wartości iloczynów skalarnych wprowadzonych składników rozkładu

$$({}_s i|_r i)_n = Re \sum_{h=1}^{\infty} -j(G_{\alpha\beta h} - G\delta_{\alpha\beta}) U_{\beta h} B_{\alpha\gamma h} U_{\gamma h}^*, \tag{24}$$

$$({}_r i|_a i)_n = Re \sum_{h=1}^{\infty} jB_{\alpha\beta h} U_{\beta h} G U^*_{\alpha h}, \tag{25}$$

$$({}_a i|_s i)_n = Re \sum_{h=0}^{\infty} G U_{\alpha h} (G_{\alpha\beta h} - G\delta_{\alpha\beta}) U^*_{\beta h}. \tag{26}$$

Jeśli założyć:

$$B_{\alpha\beta h} = B_{\beta\alpha h}, \quad \alpha, \beta \in \{1, \ldots, n\}, h \in N, \tag{27}$$

czyli w zapisie macierzowym

$$B_h = \underset{t}{B_h}, \tag{28}$$

to wyrażenia $B_{\alpha\beta} U_{\beta h} U^*_{\alpha h}$ (dla poszczególnych wskaźników h) są rzeczywiste i stąd

$$({}_r i|_a i)_n = 0. \tag{29}$$

Przy spełnieniu założenia (27) mamy na mocy wzoru (15)

$$P = Re \sum_{h=0}^{\infty} U_{\alpha h} G_{\alpha\beta h} U^*_{\beta h} + Re \sum_{h=1}^{\infty} U_{\alpha h} (-jB_{\alpha\beta h}) U^*_{\beta h} =$$

$$= Re \sum_{h=0}^{\infty} U_{\alpha h} G_{\alpha\beta h} U^*_{\beta h}. \tag{30}$$

Stąd w oparciu o wzory (15), (23), (30) mamy:

$$({}_a i|_s i)_n = G\left(P - \frac{P}{||u||_n^2} ||u||_n^2\right) = 0. \tag{31}$$

Zakładamy teraz dodatkowo (oprócz spełnienia wzoru (27)):

$$G_{\alpha\beta h} B_{\alpha\gamma h} = G_{\alpha\gamma h} B_{\alpha\beta h}, \ \alpha, \beta \in \{1, \ldots, n\}, h \in N, \tag{32}$$

czyli macierzowo

$$\underset{t}{G_h B_h} = \underset{t}{B_h G_h} = B_h G_h. \tag{33}$$

Wówczas, stosując podobne rozumowanie jak przy dowodzie wzoru (29), otrzymujemy:

$$({}_s i|_r i)_n = 0. \tag{34}$$

Warunki (27), (32) są więc wystarczającymi do tego, by parami zachodziła ortogonalność wprowadzonych składników prądu (wzory (29), (31), (34)). Nie są one warunkami koniecznymi. O ile warunek (27) jest dobrze umotywowany, gdyż jak wiadomo, jest spełniony w przypadku, gdy odbiornik złożony jest ze skończonej liczby elementów R, L, C, to warunek (32) nie ma tak jasnej motywacji. Zyskuje ją, gdy go jeszcze wzmocnimy zakładając:

$$G_{\alpha\beta h} = \delta_{\alpha\beta} G_h, \ \alpha, \beta \in \{1, \ldots, n\}, h \in N, \tag{35}$$

(tu „G_h" jest zmienną rzeczywistą, a nie macierzową jak we wzorze (33)). Oczywiście spełnienie warunku (35) implikuje, przy założeniu warunku (27), spełnienie warunku (32). Sens elektrotechniczny warunku (35) jest dobrze widoczny. Zauważmy, że dla ($n = 1$)

warunki (27), (32) nie krępują doboru parametrów odbiornika — są wówczas spełnione zawsze.

Zastanówmy się teraz nad motywacją przyjęcia takiego rozkładu. Wyobraźmy sobie, że przewody doprowadzające prądy do naszego odbiornika (rys. 1) posiadają rezystancje ΔR. Załóżmy, że są one takie same dla wszystkich n przewodów a $(n+1)$-szy ma rezystancję zerową. Wówczas przy skończonej liczbie harmonicznych prądu, stosując np. metodę nieoznaczonego czynnika Lagrange'a ([7], str. 151, 152), można wykazać, przy założeniu, że zadana jest moc czynna pobierana przez odbiornik oraz zadane są jego napięcia, że minimum strat w przewodach występuje właśnie, gdy moc ta dostarczana jest przez prąd ${}_a i$. Oczywiście zachodzą wzory

$$(u|_a i)_n = P, \tag{36}$$

$$(u|_r i)_n = 0, \tag{37}$$

$$(u|_s i)_n = 0, \tag{38}$$

czyli jedynym nośnikiem mocy czynnej jest prąd ${}_a i$. To stanowi oczywiście przyczynę zdefiniowania prądu ${}_a i$. Lecz przypominamy, że obowiązuje ona przy założeniu symetrii rezystancyjnej przewodów doprowadzających prądy. Zakładamy również milcząco, że odbiornikowi jest „wszystko jedno" w jaki sposób zadana moc czynna została mu dostarczona pod względem jej rozkładu na poszczególne harmoniczne i przewody. Jeśli obowiązuje twierdzenie (analogiczne jak w przypadku ($n = 1$)), mówiące że dla dowolnej skończonej liczby harmonicznych prąd ${}_r i$ można skompensować dołączając równolegle do odbiornika element kompensujący złożony ze skończonej liczby elementów L, C (wymaga to dowodu), to sens wyprowadzonego prądu ${}_r i$ staje się jasny. Jest to część prądu i kompensowalna, dla nieskończonej liczby harmonicznych — z dowolną dokładnością w sensie przyjętej normy $\| \|_n$, z pomocą w/w układów L, C. Prąd ${}_s i$ stanowi resztę całkowitego prądu w tym rozkładzie. Przyczyną jej wystąpienia jest, dla dowolnej harmonicznej „rozrzut" przewodności $G_{\alpha\beta h}$ wokół przewodności $G\delta_{\alpha\beta}$, $(h \in N_0)$.

Zamiast prądami ${}_a i$, ${}_s i$, ${}_r i$ możemy operować odpowiednimi mocami. Mamy, w oparciu o wzory (29), (31), (34), prostą do wykazania zależność

$$\|i\|_n^2 = \|{}_a i\|_n^2 + \|{}_s i\|_n^2 + \|{}_r i\|_n^2. \tag{39}$$

Wprowadzając definicje mocy

$$Q_s = \|u\|_n \|{}_s i\|_n, \tag{40}$$

$$Q_r = \|u\|_n \|{}_r i\|_n, \tag{41}$$

$$P_m = \|u\|_n \|i\|_n \tag{42}$$

oraz dowodząc wzór

$$|P| = \|u\|_n \|{}_a i\|_n, \tag{43}$$

mamy

$$P_m^2 = P^2 + Q_{s_i}^2 + Q_r^2. \tag{44}$$

Wzór ten można, w znany sposób, interpretować geometrycznie z pomocą prostopadłościanu. Uwspółcześniając starożytnych ([1], str. 427) i abstrahując od jego elektrotechnicz-

nej interpretacji można go nazwać twierdzeniem Pitagorasa. Oczywiście opisana motywacja wprowadzenia prądów ${}_a i$, ${}_s i$, ${}_r i$ przenosi się na moce.

Łącząc równolegle m odbiorników mamy dla prądu całkowitego i tych odbiorników równość (I prawo Kirchhoffa)

$$i = \underset{1}{i} + \ldots + \underset{m}{i}. \tag{45}$$

Stąd też mnożąc skalarnie przez „u" otrzymujemy

$$P = \underset{r}{P} + \ldots + \underset{m}{P}. \tag{46}$$

Jest to znana zasada zachowania mocy czynnej. Pozostałe moce Q_s, Q_r, P_m nie są zachowawcze, ponieważ w przestrzeni unitarnej mamy związek (który zapisujemy w naszej interpretacji równoległego połączenia dwóch odbiorników)

$$||i||_n \leqslant ||\underset{1}{i}||_n + ||\underset{2}{i}||_n. \tag{47}$$

Nierówność ta przechodzi w równość (przy założeniu $\underset{1}{i}$, $\underset{2}{i} \neq 0$) wtedy i tylko wtedy, gdy ([6], str. 64): $\underset{2}{i} = c\underset{1}{i}$, $c \in R$, $c > 0$, co na ogół nie zachodzi dla prądów ${}_s i$, ${}_r i$, i.

Do tej pory znając zadane napięcie u rozkładaliśmy indywidualnie prądy i na trzy składniki określone wzorami (20), (21), (22). Na zagadnienie to można spojrzeć „kolektywnie". Przy zadanym napięciu $u \in L_n^2(\langle 0; T\rangle)$ oraz przy spełnionych warunkach $n = 1$, U_0, $U_h \neq 0$ rozpatrujemy nie jeden odbiornik, a zbiór odbiorników realizujących wszystkie przewodności G_0, G_h (dające wszystkie moce czynne P) oraz wszystkie elementy B_h, przy których wzór (22) określa prąd przestrzeni $L^2(\langle 0; T\rangle)$. Czyli dowolny prąd tej przestrzeni jest osiągalny przez odpowiedni dobór admitancji. Nietrudno wykazać, że operując przewodnościami G uzyskujemy jednowymiarową podprzestrzeń liniową domkniętą rozpatrywanej przestrzeni Hilberta, a operując takimi parametrami B_h — nieskończenie wymiarową podprzestrzeń liniową domkniętą tejże. Część komplementarna w stosunku do sumy prostej tych dwóch jest też podprzestrzenią liniową domkniętą omawianej przestrzeni Hilberta ([1], str. 421). Mamy więc rozkład naszej przestrzeni Hilberta na sumę prostą trzech podprzestrzeni liniowych domkniętych parami ortogonalnych. (Są one oczywiście też przestrzeniami Hilberta.) Oznaczając ją literą H oraz oznaczając pozostałe ww. podprzestrzenie odpowiednio H_a, H_s, H_r mamy:

$$H = H_a \oplus H_s \oplus H_r \tag{48}$$

oraz

$$H_a \perp H_s, \quad H_s \perp H_r, \quad H_r \perp H_a. \tag{49}$$

Podobny rezultat można uzyskać w oparciu o wzory (19), dla dowolnego $n \in N$ i $U_{\alpha 0}$, $U_{\alpha h} \neq 0$. Spełnienie tych wzorów dla zadanego (jedynego) stanu napięciowo-prądowego implikuje spełnienie warunków (27), (32). Po przyjęciu poprzedniej interpretacji indywidualnych rozkładów prądów i, interpretacja powyższego rozkładu przestrzeni Hilberta H nie nastręcza trudności.

3. ZAKOŃCZENIE

Składnik ${}_si$ określony wzorem (21) można rozłożyć na dwa dalsze ${}_{as}i$, ${}_di$ (por.: [5])

$$ {}_{as}i = (G_{\alpha\beta 0} - G_0\,\delta_{\alpha\beta})\,U_{\beta 0} + \sqrt{2}\,Re\sum_{h=1}^{\infty}(G_{\alpha\beta h} - G_h\,\delta_{\alpha\beta})\,U_{\beta h}\exp(jh\omega(\cdot)), \tag{50} $$

$$ {}_di = (G_0 - G)\,\delta_{\alpha\beta}\,U_{\beta 0} + \sqrt{2}\,Re\sum_{h=1}^{\infty}(G_h - G)\,\delta_{\alpha\beta}\,U_{\beta h}\exp(jh\omega(\cdot)), \tag{51} $$

gdzie

$$ G_h = \frac{P_h}{U_h^2}, \qquad U_h^2 = U_{\alpha h}U_{\alpha h}^*, \qquad P_h = Re(U_{\alpha h}I_{\alpha h}^*) \tag{52} $$

Przy spełnieniu warunków (27), (32), z pomocą podobnych obliczeń można stwierdzić ortogonalność wszystkich parami różnych składników prądu i: ${}_ai$, ${}_ri$, ${}_{as}i$, ${}_di$. Składnik ${}_{as}i$ związany jest z asymetrią fazową przewodności odbiornika dla danej harmonicznej o numerze h; składnik ${}_di$ — z dyspersją częstotliwościową przewodności przy zachowaniu symetrii fazowej.

Ilustrację przedstawionej teorii, wprowadzonych pojęć i definicji przedstawiamy na prostych przykładach.

Przykład 1

Podajemy przykład dotyczący porównania mocy biernej Budeanu Q_B z wprowadzoną mocą reaktancyjną Q_r.

Czynimy założenia, że $G_{\alpha\beta h} = G_{\beta\alpha h}$, $B_{\alpha\beta h} = B_{\beta\alpha h}$. Kwadrat mocy reaktancyjnej Q_r zgodnie z wzorem (41) wynosi

$$ \begin{aligned} Q_r^2 &= ||u||_n^2||i||_n^2 = \sum_{h=0}^{\infty}\sum_{\alpha=1}^{n}|U_{\alpha h}|^2\sum_{k=0}^{\infty}\sum_{\beta=1}^{n}|B_{\beta\gamma k}U_{\gamma k}|^2 = \\ &= \sum_{h=0}^{\infty}\sum_{\alpha=1}^{n}|U_{\alpha h}|^2\sum_{k=0}^{\infty}\sum_{\beta=1}^{n}|{}_bI_{\beta k}|^2 = \sum_{l=0}^{\infty}|U_l|^2\sum_{m=0}^{\infty}|{}_bI_m|^2 = \\ &= \sum_{l=0}^{\infty}|U_l|^2|{}_bI_l|^2 + \sum_{\substack{l,m=0\\ l>m}}^{\infty}(|U_l|^2|{}_bI_m|^2 + |U_m|^2|{}_bI_l|^2), \end{aligned} $$

gdzie ${}_bI_{\beta h} = B_{\beta\gamma h}U_{\gamma h}$.

W powyższym wzorze wykorzystano możliwość ustawienia ciągów podwójnych $(U_{\alpha h})$. $({}_bI_{\beta h})$ w ciągi pojedyncze (U_l), $({}_bI_m)$. Sposób tego ustawienia jest dowolny na skutek bezwzględnej zbieżności szeregów wartości skutecznych odpowiednich napięć i prądów. Moc bierna Budeanu Q_B

$$ \begin{aligned} Q_B &= Im\sum_{h=0}^{\infty}U_{\alpha h}(G_{\alpha\beta h} - jB_{\alpha\beta h})\,U_{\beta h}^* = Im\Big(-j\sum_{h=0}^{\infty}U_{\alpha h}B_{\alpha\beta h}U_{\beta h}^*\Big) = \\ &= -\sum_{h=0}^{\infty}U_{\alpha h}\,{}_bI_{\alpha h}^* = -\sum_{l=0}^{\infty}U_{lb}I_l^*. \end{aligned} $$

$$Q_B^2 = \sum_{l=0}^{\infty} |U_l|^2 |_b I_l|^2 + 2Re \sum_{\substack{l,m=0 \\ l>m}}^{\infty} U_{l\,b} I_l^* U_{m\,b}^* I_m.$$

Wykorzystując nierówność $2Reab^* \leqslant |a|^2 + |b|^2$, gdzie $a = U_{lb} I_m$, $b = U_{mb} I_l$ otrzymujemy

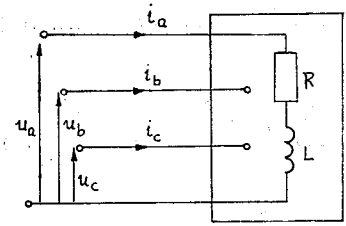
$$Q_B^2 \leqslant Q_r^2, \quad |Q_B| \leqslant Q_r.$$

Z ostatniej nierówności wynika, że dla skończonej liczby harmonicznych prądu $_r i$ można za pomocą skończonej liczby elementów LC skompensować większą moc Q_r niż moc bierną Budeanu Q_B.

Przykład 2

Dla układu przedstawionego na rys. 2 dokonać rozkładu prądów oraz podać odpowiednie moce zakładając, że

$$U_{ah} = E_h, \; U_{bh} = E_h e^{-jh\frac{2\pi}{3}}, \; U_{ch} = E_h e^{jh\frac{2\pi}{3}}, \; \sum_{h=0}^{\infty} |E_h|^2 < \infty.$$



Rys. 2. Obwód 4 — zaciskowy do przykładu 2

Prądy odbiornika dla kolejnych harmonicznych przyjmują wartości

$$I_{ah} = \frac{R - jh\omega L}{R^2 + (h\omega L)^2} E_h, \quad I_{bh} = 0, \quad I_{ch} = 0.$$

Na skutek poczynionych założeń mamy:

$$u, i \in L_3^2(<0; T>).$$

Moc czynna P_h, przewodność G_h określone są wyrażeniami

$$P_h = \frac{R}{R^2 + (h\omega L)^2} |E_h|^2, \quad G_h = \frac{P_h}{|U_{ah}|^2 + |U_{bh}|^2 + |U_{ch}|^2} = \frac{1}{3} \frac{R}{R^2 + (h\omega L)^2}.$$

Natomiast całkowita moc czynna P i przewodność równoważna G

$$P = \sum_{h=0}^{\infty} \frac{R}{R^2 + (h\omega L)^2} |E_h|^2, \quad G = \frac{P}{3 \sum_{h=0}^{\infty} |E_h|^2}.$$

Składowe prądów zgodnie z przyjętymi definicjami (20), (22), (50), (51) dla kolejnych harmonicznych przyjmują wartości

$$ {}_aI_h = GE_h,\ {}_aI_{bh} = GE_h e^{-jh\frac{2\pi}{3}},\ {}_aI_{ch} = GE_h e^{jh\frac{2\pi}{3}}. $$

$$ {}_rI_{ah} = -\frac{jh\omega L}{R^2+(h\omega L)^2}E_h, \quad {}_rI_{bh} = 0, \quad {}_rI_{ch} = 0. $$

$$ {}_{as}I_{ah} = \frac{2}{3}\frac{R}{R^2+(h\omega L)^2}E_h, \quad {}_{as}I_{bh} = -\frac{1}{3}\frac{R}{R^2+(h\omega L)^2}E_h e^{-jh\frac{2\pi}{3}}, $$

$$ {}_{as}I_{ch} = -\frac{1}{3}\frac{R}{R^2+(h\omega L)^2}E_h e^{jh\frac{2\pi}{3}}. $$

$$ {}_dI_{ah} = (G_h-G)E_h, \quad {}_dI_{bh} = (G_h-G)E_h e^{-jh\frac{2\pi}{3}}, \quad {}_dI_{ch} = (G_h-G)E_h e^{jh\frac{2\pi}{3}}. $$

Odpowiednie moce

$$ Q_r^2 = 3\sum_{h=0}^{\infty}|E_h|^2\sum_{k=0}^{\infty}\frac{(k\omega L)^2}{(R^2+(k\omega L)^2)^2}|E_k|^2, $$

$$ Q_{as} = 2\sum_{h=0}^{\infty}|E_h|^2\sum_{k=0}^{\infty}\frac{R^2}{(R^2+(k\omega L)^2)^2}|E_k|^2, $$

$$ Q_d = 9\sum_{h=0}^{\infty}|E_h|^2\sum_{k=0}^{\infty}(G_k-G)^2|E_k|^2. $$

Rozważania nasze skażone są jednak pewną wadą. Prezentowany rozkład prądów oparty jest o pewien wyróżniany układ ortonormalny, mianowicie o układ (10). Wydaje się, że albo powinno prowadzić się podobne rozważania dla dowolnego układu ortonormalnego — analogia do formułowania w geometrii różniczkowej współzmienniczych wzorów względem pewnego atlasu danej rozmaitości różniczkowej, albo w ogóle nie używać pojęcia układu ortonormalnego. Lecz jak to zrobić?

BIBLIOGRAFIA

1. A. Alexiewicz: *Analiza funkcjonalna*, Warszawa: PWN, 1969
2. M. Brodzki: *Kilka uwag o matematycznej naturze wielkości fizykalnych.* Zeszyty Naukowe Politechniki Śląskiej, Elektryka z. 100, Gliwice 1985
3. M. Brodzki: *Wstęp do teorii liniowych obwodów elektrycznych w ujęciu geometrycznym.* Skrypty Uczelniane Politechniki Śląskiej nr 847, Gliwice 1979
4. L. Czarnecki: *Interpretacja, identyfikacja i modyfikacja właściwości energetycznych obwodów jednofazowych z przebiegami odkształconymi*, Zeszyty Naukowe Politechniki Śląskiej, Elektryka z. 91, Gliwice 1984
5. L. Czarnecki: *Ortogonalny rozkład prądu źródła napięcia odkształconego zasilającego asymetryczny,*

nieliniowy odbiornik trójfazowy. Prace Seminarium z Podstaw Elektrotechniki i Teorii Obwodów. X — SPETO, Gliwice — Wisła 1987
6. W. Kołodziej: *Wybrane rozdziały analizy matematycznej*, Warszawa: PWN, 1970
7. F. Leja: *Rachunek różniczkowy i całkowy*, Warszawa: PWN, 1969
8. Z. Nowomiejski: *Moc i energia elektryczna w układach elektrycznych o dowolnych ustalonych przebiegach*, Zeszyty Naukowe Politechniki Śląskiej, Elektryka z. 15, Gliwice 1963
9. R. Sikorski: *Rachunek różniczkowy i całkowy*. Warszawa: PWN, 1977

M. BRODZKI, M. PASKO

DEFINITIONS OF CERTAIN POWERS OF MULTITERMINAL SYSTEMS WITH DISTORTED SIGNALS

Summary

The authors deal with the generalized notion of orthogonal distribution of the distorted current of a two-terminal element, defined in paper [4] for distorted signals on the input of the $(n+1)$ terminal element $(n \in N)$. In this case, the adequate Hilbert space together with a certain closed orthonormal system are defined. The current distribution (appearing when the conditions for a given element are met) is followed by introduction of respective powers. The physical motivation of the proposed procedure is given. Finally, the distribution of the discussed Hilbert space into a simple sum of 3 pairs of orthogonal sub-spaces is made.

M. BRODZKI, M. PASKO

DÉFINITIONS DE CERTAINES PUISSANCES POUR LES SYSTÈMES MULTIPÔLES À PHÉNOMÈNES PÉRIODIQUES

Résumé

Dans le travail on a présenté i généralisation de la notion de la décomposition orthonormale du courant périodique d'un élément dipôle, défini dans l'ouvrage [4], pour en embrasser aussi le cas de l'élément multipôle à phénomènes périodiques. Dans ce but on a déterminé l'espace hilbertien convenable et un certain système orthonormal fermé. Ensuite, à la base de cette décomposition des courants (ayant lieu quand le dit élément accomplit certaines conditions), on a introduit des puissances adéquates. On a proposé la motivation physique du procédé mentionné. Enfin, on décompose l'espace hilbertien ci-dessus cité en somme directe de trois sous-espaces réciproquement orthogonaux par paires.

M. BRODZKI, M. PASKO

DEFINITIONEN FÜR GEWISSE LEISTUNGEN BEI MEHRKLEMMEN-SYSTEMEN MIT DEFORMIERTEM VERLAUF

Zusammenfassung

In diesem Artikel befassen wir uns mit der Verallgemeinerung des Begriffs des orthogonalen Strome beim deformierten Zweiklemmenelement, das in dieser Bearbeitung mit [4] bezeichnet worden ist, für den Fall eines deformierten Verlaufes am Eingang des Elementes $(n+1)$ — Klemmenelement $(n \in N)$. Zu diesem Zweck wurde der entsprechende Hilbert-Raum mit einem geschlossenen orthogonalen System definiert. Weiterhin, nach dieser Stromverteilung, die bei Erfüllung durch das obengenannte Element gewisser Bedingungen stattfindet, führen wir die damit verbundenen Leistungen ein. Wir geben auch die physikalische Begründung des vorliegenden Verfahrens an. Zum Schluß führen wir die Zerlegung des besprochenen Hilbert-Raumes in eine gerade Summe von drei paarweise orthogonalen Unterräumen durch.

М. БРОЦКИ, М. ПАСКО

ДЕФИНИЦИИ НАДЁЖНЫХ МОЩНОСТЕЙ ДЛЯ МНОГОЗАЖИМНЫХ СИСТЕМ С ДЕФОРМИРОВАННЫМИ ПРОБЕГАМИ

Р е з ю м е

Обсуждено обощение понятия ортогонального распределения тока деформированного двухзажимного элемента, который определён в работе [4], в случае деформированных пробегов у входа элемента $(n+1)$ — зажимного $(n \in N)$. Определено соответствующее пространство Гильберта вместе с надёжной замкнутой ортонормальной системой. На основании распределения токов (происходящим при осуществлении вышеупомянутым элементом надёжных условий), вводятся связанные с ним мощности. Представлена физическая мотивировка вышеуказанного процесса. В заключение проведено перераспределение обсуждаемого пространства Гильберта на прямую тремя парами ортогональных подпространств.

Errata:

W artykule Pana Marka Brodzkiego w zeszycie nr 2/1987 r. Autor dostrzegł błąd niezauważony przez Redakcję. Prawidłowe brzmienie tytułu: „Certaine interprétation de la théorie des réseaux électriques dans la théorie des fibrations". Autora i Czytelników przepraszamy.

ROZPRAWY ELEKTROTECHNICZNE, 1989, 35, z. 1, ss. 45-59

53.082.62:53.087.92

Środki i sposoby ujednolicenia miary napięcia przemiennego

ANDRZEJ BARAŃSKI

Polski Komitet Normalizacji, Miar i Jakości w Warszawie
Zakład Metrologii Elektrycznej

Otrzymano 1987.11.25

Autoryzowano do druku 1988.04.08

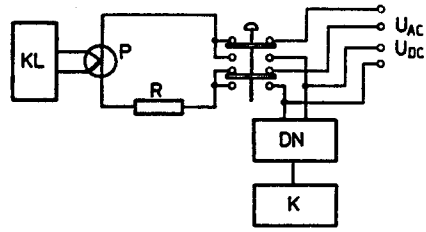
Przedstawiono etapy rozwoju i doskonalenia konstrukcji przetworników termoelektrycznych jedno- i wielozłączowych od zastosowań w pomiarach precyzyjnych napięcia przemiennego aż po wykorzystanie do odtwarzania jednostki napięcia przemiennego. Dokonano przeglądu układów pomiarowych do wyznaczania charakterystyk wspomnianych przetworników. Powtarzalność uzyskiwanych wyników dała podstawę podjęcia unifikacji miary w skali międzynarodowej. W ramach współpracy państwowych biur miar i instytutów metrologicznych prowadzone były komparacje etalonów specjalnych napięcia przemineego. Na tle osiągnięć światowych przedstawiono wkład krajowy w tym zakresie.

Do chwili obecnej nie jest znany sposób odtwarzania jednostki napięcia przemiennego za pośrednictwem wzorca parametrycznego odpowiednika np. ogniwa Westona, a podstawowa jednostka elektryczna — amper — zdefiniowana została wyłącznie przy prądzie stałym. Wynika stąd, że brak jest technicznych możliwości bezpośredniego odtwarzania i definiowania wspomnianych jednostek przy prądzie przemiennym. Problem rozwiązano pośrednio, przez zastosowanie tzw. etalonów specjalnych, za pośrednictwem których odbywa się dystrybucja jednostki napięcia na całą dziedzinę pomiarową. Jako podstawa etalonów specjalnych w biurach miar i państwowych laboratoriach metrologicznych, stosowane bywają przetworniki termoelektryczne różnych konstrukcji. Nie są one jedynymi środkami pomiarowymi, służącymi powiązaniu jednostki napięcia przemiennego z jednostką napięcia stałego, lecz nie mają obecnie równorzędnych sobie w zakresie dokładności komparacji i pasma częstotliwości przenoszenia. Mimo pewnych wad, takich jak np. stosunkowo duża inercja, dryft czasowy, uciążliwość odtwarzania charakterystyk metrologicznych — brak jest zdecydowanych przesłanek do rezygnacji z ich stosowania w charakterze urządzeń transferowych — obecnie i w najbliższej przyszłości.

Celem opracowania jest dokonanie chronologicznego przeglądu doskonalenia konstrukcji i rozwoju zastosowań przetworników jedno — i wielozłączowych do pomiarów precyzyjnych napięcia przemiennego, aż po wykorzystanie ich jako elementy etalonów specjalnych. Na tle prac prowadzonych przez znaczące w świecie ośrodki metrologiczne przedstawiono wkład krajowy.

ZASADA OGÓLNA POMIARU

Zastosowanie przetwornika termoelektrycznego jako urządzenia transferowego do pomiarów napięcia przemiennego przedstawiono na rys. 1. Równoważenie układu odbywa się



Rys. 1. Zasada pomiaru napięcia przemiennego przez porównanie z napięciem stałym
KL-kompensator Lindecka, DN- oporowy dzielnik napięcia, K-kompensator napięcia stałego

dwuetapowo — przy napięciu przemiennym a potem stałym w chwili, gdy ciepło Joule'a wytworzone w grzejniku przetwornika przy obydwu rodzajach zasilania jest takie samo. Wówczas to wartość skuteczna napięcia przemiennego i wartość napięcia stałego są sobie równe. W praktyce jednak zrównoważenie układu nie oznacza równości napięć wejściowych, pomiar bowiem obarczony jest błędem komparacji:

$$\delta = \frac{U_{AC} - U_{DC}}{U_{DC}} \tag{1}$$

gdzie:

U_{AC} — wartość skuteczna napięcia przemiennego

U_{DC} — wartość napięcia stałego, obydwie określone gdy $E_{AC} = E_{DC}$, w którym:

E_{AC} — siła termoelektryczna wyjściowa przy napięciu wejściowym przemiennym,

E_{DC} — siła termoelektryczna wyjściowa przy napięciu stałym.

Błąd komparacji zależny jest od częstotliwości i rośnie z jej kwadratem [1], dlatego bywa często definiowany jako błąd częstotliwościowy

$$\delta_f = \frac{U_{AC} - U_{AC(0)}}{U_{AC(0)}} \tag{2}$$

gdzie:

$U_{AC(0)}$ — wartość skuteczna napięcia przemiennego przy częstotliwości odniesienia (np. $f_0 = 1$ kHz).

Źródłem niejednoznaczności wartości U_{DC} występującej w wyrażeniu (1) jest błąd asymetrii definiowany jako:

$$\delta_a = \frac{U_{DC(+)} - U_{DC(-)}}{U_{DC}} \tag{3}$$

gdzie: $U_{DC(+)}$ — wartość napięcia stałego określona przy polaryzacji zgodnej, $U_{DC(-)}$ wartość napięcia stałego określona przy polaryzacji odwrotnej.

Błąd ten powstaje na skutek występowania zjawisk termoelektrycznych Thomsona

i Peltiera, głównie na złączach grzejnika z podporami. Dryft błędu komparacji (dryft charakterystyki) określa zależność:

$$v_t = \delta_t - \delta_0 \tag{4}$$

gdzie: δ_t — błąd komparacji wyznaczony po upływie czasu t od chwili, w której wyznaczono δ_0; δ_0 — błąd komparacji wyznaczony w chwili początkowej t_0.

PRZETWORNIKI JEDNOZŁĄCZOWE

Termoelement jest najprostszym przetwornikiem termoelektrycznym. Rozwój i doskonalenie konstrukcji termoelementów doprowadziły do wyraźnego ich podziału na dwa główne wykonania: do pracy przy małych częstotliwościach (m.cz.) i wielkich częstotliwościach (w.cz.). Stąd powstały odmiany zróżnicowane kształtem i wymiarami, a połączone wspólnymi cechami konstrukcyjnymi takimi jak np. [2]:

— grzejnik wykonany z prostego odcinka drutu oporowego,
— złącze termoelektryczne odizolowane lub nieodizolowane od grzejnika,
— podpory i złącza wykonane z materiałów antymagnetycznych,
— ampułka szklana — próżniowa lub wypełniona gazem albo powietrzem,
— perła szklana, w której zatopione jest złącze termoelektryczne.

Termoelement umieszczony w metalowej obudowie i wyposażony w typowe złącza elektryczne: wejściowe i wyjściowe ma zastosowanie jako przetwornik termoelektryczny w pomiarach napięcia przemiennego.

Uważa się, że wykorzystanie przetworników termoelektrycznych do pomiarów komparacyjnych napięcia przemiennego z napięciem stałym podjął L. F. Hermach, który stworzył podwaliny teoretyczne dziedziny, a w r. 1952 opublikował po raz pierwszy fundamentalną zależność wiążącą błąd komparacji przetwornika ze stałymi fizycznymi [3]:

$$\delta = -25H\Theta_{DC}q^2 \tag{5}$$

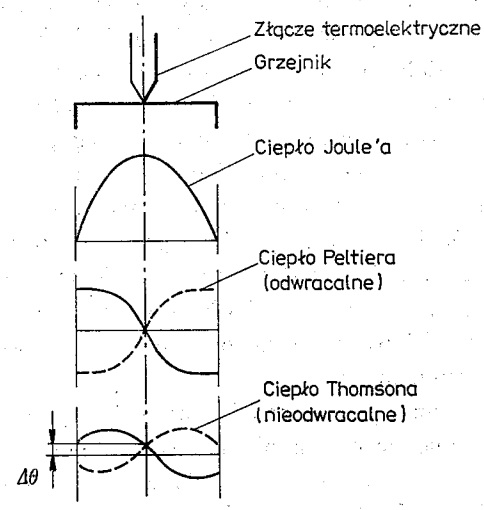
gdzie: $H = h + 0{,}5$ AB, w którym: h — zmienna zależna od współczynnika temperaturowego rezystancji grzejnika, jego długości, pola przekroju, konduktywności, emisyjności całkowitej, temperatury otoczenia i stałej Stefana — Boltzmana; A i B — stałe w zależności siły elektromotorycznej w funkcji przyrostu temperatury $\left(E = A\Theta + \frac{B}{2}\Theta^2\right)$, Θ_{DC} — przyrost temperatury w punkcie środkowym grzejnika przy prądzie stałym, $q = \frac{D}{\omega l^2}$, w którym: D — cieplna dyfuzyjność materiału grzejnika, $\omega = 2\pi f$ (f — częstotliwość prądu płynącego przez grzejnik, dla którego określany jest błąd komparacji), l — połowa długości grzejnika.

Podał ponadto przybliżone rozwiązania równań nieliniowych, różniczkowych, rządzących przemianą prądu elektrycznego w ciepło, z uwzględnieniem ciepła Thomsona i Peltiera. Opisał konstrukcję opracowanych w NBS [1] przetworników termoelektrycznych

[1] National Bureau of Standards — USA

z ich aplikacją jako wzorów odniesienia przy częstotliwościach sieciowych i akustycznych. Prowadził prace nad rozszerzeniem zastosowań opracowanych przetworników do pomiarów napięć w paśmie częstotliwości: 3 Hz ... 30 MHz [4]. Połączony w szereg z termoelementem rezystor potraktował w rozważaniach jako linię długą o długości $2\,l$ z równomiernie rozłożoną impedancją szeregową i admitancją bocznikującą. Błąd komparacji przy tych założeniach uzależnił od parametrów konstrukcyjnych scharakteryzowanych rezystancją, pojemnością, indukcyjnością i wymiarami geometrycznymi.

Na początku lat 60-tych F. J. Wilkins i F. C. Widdis z NPL [2] wyprowadzili równania wiążące asymetrię geometryczną grzejnika z ciepłem Thomsona i Peltiera, wskazując na istnienie nieodwracalnego ciepła Thomsona — por. rys. 2 [5, 6]. Ustalili ponadto ogólne



Rys. 2. Rozkład ciepła wydzielanego w grzejniku przetwornika termoelektrycznego przy obciążeniu prądem stałym

Δθ — różnica między ciepłem Thomsona odwracalnym a nieodwracalnym — występującym przy prądzie przemiennym

równanie przenoszenia ciepła i zbadali wpływ niesymetrii geometrycznej umiejscowienia perły szklanej na prawidłowość pracy przetwornika termoelektrycznego.

W końcu lat 50-tych w WNIIM [3] opracowano komparatory termoelektryczne oraz termowoltomierze do pomiaru napięć w zakresie częstotliwości 50 Hz ... 20 kHz [7]. Równocześnie doskonalono konstrukcję przetworników termoelektrycznych — w celu zmniejszenia błędu asymetrii i wpływu naskórkowości.

Do rozszerzenia zakresów pomiarowych zaczęto wykorzystywać rezystory dodatkowe wykonane z mikroprzewodu. W r. 1960 firma J. Fluke Manufacturing Comp. zaprojektowała wielozakresowy zestaw przetworników termoelektrycznych początkowo dla zabezpieczenia swych potrzeb produkcyjnych, a potem podjęła ich produkcję seryjną [8]. Do konstrukcji użyto termoelementów w.cz. które zamontowano współosiowo w rurach metalowych. Błąd komparacji przetworników wyznaczono metodą eksperymentalno-

[2] National Phisical Laboratory — Wielka Brytania

[3] Wszechzwiązkowy Naukowo-Badawczy Instytut im. Mendelejewa—ZSRR

analityczną. W tym celu, za pomocą kalorymetrycznego miernika mocy, dokonano pomiaru impedancji przetworników w zakresie częstotliwości 10 ... 900 MHz, a następnie przy prądzie stałym pomierzono ich rezystancję. Na podstawie uzyskanych danych i posługując się równaniami linii długiej obliczono niezbędny parametr.

Rozwój konstrukcji narzędzi pomiarowych stał się stymulatorem doskonalenia w NBS własnych zestawów przetworników wzorcowych [9]. Nowy — 14-zakresowy zestaw, składał się z dwóch przetworników wymiennych na prądy 2,5 i 5 mA oraz sześciu rezystorów dodatkowych, umożliwiających niezbędne kombinacje połączeń. Równolegle utworzono drugą grupę przetworników, z których 12 szt. wykonali różni producenci, stosując do wyrobu grzejników różne stopy metali, takie jak np. evanohm i carma[4], a dwa pozostałe miały konstrukcję wielozłączową (o czym poniżej) i otrzymano je z NPL [10]. Przy pomocy rezystorów dodatkowych uzyskano zakresy pomiarowe do 200 V. Starannie dobrane zestawy przebadano przy 20 Hz, 2 i 20 kHz. Próby trwały 3 lata i potwierdziły powtarzalność uzyskanych wyników błędu komparacji na poziomie 2.10^{-6}.

Na przełomie lat 60 i 70-tych, w ETL[5] przebadano przetworniki termoelektryczne własnej konstrukcji [11]. M.in. obiektem zainteresowań był wpływ usytuowania spoiny względem elementu grzejnego oraz oddziaływanie błędu częstotliwościowego rezystorów dodatkowych na wypadkową charakterystykę metrologiczną przetworników napięciowych. Badania przeprowadzono na 5 typach przetworników wyposażonych w rezystory warstwowe wykonane ze stopu Ni-Cr-Al i stwierdzono, że w zakresie częstotliwości 40 Hz ... ... 100 kHz, główny błąd pochodził od rezystorów dodatkowych, a występował przy 50 kHz.

Uzyskane średnie wyniki pomiarów błędu komparacji rzędu 5.10^{-5}, uznano jako zadowalający prognostyk do dalszych prac. Równocześnie kontynuowane były prace w uznanych ośrodkach: np. w NPL wykazano, że błędy przetworników jednozłączowych w zakresie częstotliwości do 20 kHz rosną znacznie wolniej niż przetworników wielozłączowych, gdyż obwody tych pierwszych są bardziej proste [12]. Wdrożono także metodę „drabinkową" porównań wzajemnych przetworników, wykorzystując w tym celu przetworniki o prądach nominalnych 5 i 10 mA wyposażone w wymienne rezystory dodatkowe.

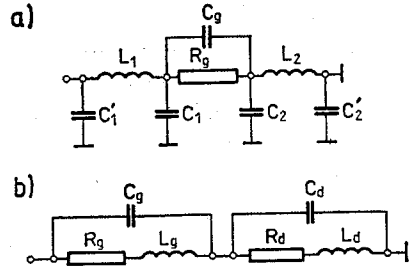
Szczegółową analizę zastosowań przetworników radzieckich typu TWB1 ... 5 w zakresie częstotliwości 10 Hz ... 200 MHz przeprowadzono w WNIIM [13]. Stwierdzono, że największy błąd częstotliwościowy występował w typach TWB1, 2, 4 i 5, których grzejniki wykonane były ze stopów żelaza i nichromu, a związany był ze zjawiskiem naskórkowości. W celu zmniejszenia wpływu zjawiska zastosowano materiały niemagnetyczne. R. F. Aknajew i T. B. Rożdestwienska opracowali pełną analizę błędu częstotliwościowego w zakresie częstotliwości 20 Hz ... 30 MHz ... [13, 14].

Podali schematy zastępcze przetworników jednozłączowych dla dwóch najczęściej spotykanych przypadków: bez rezystora dodatkowego i z zastosowaniem tego rezystora — por. rys. 3. Błąd częstotliwościowy dla przypadku podanego na rys. 3 a) przy zaniedbaniu wpływu pojemności C_1', C_2', i C_2 określa zależność:

$$\delta_f = \omega^2\left(L_1 C_g + L_2 C_g + L_1 C_1 - \frac{L_1^2}{2R_g^2} - \frac{L_2^2}{2R_g^2} - \frac{L_1 L_2}{R_g^2}\right) \tag{6}$$

[4] Modyfikowane stopy niklu i chromu

[5] Electrotechnical Laboratory — Japonia



Rys. 3. Schemat zastępczy przetwornika termoelektrycznego: a) bez rezystora dodatkowego, b) z rezystorem dodatkowym

oraz dla przypadku z rys. 3 b) przy założeniu, że $R_d \gg R_g$ określa wzór:

$$\delta_f = -\omega^2\left(\frac{1}{2}\frac{L_g^2}{R_d} - L_g C_d + \frac{1}{2} C_d^2 R_d\right). \tag{7}$$

Na początku lat 70-tych zainteresowanie przetwornikami jednozłączowymi zaczęło słabnąć, wskutek opracowania konstrukcji przetwornika wielozłączowego o doskonalszych charakterystykach metrologicznych. Po upływie około 10 lat niepowodzeń nad uzyskaniem powtarzalności i stabilności jego parametrów, podjęto przerwane prace i analizy nad przetwornikami jednozłączowymi. Między innymi w NML[6] wykazano, że błąd systematyczny występujący przy prądzie stałym może być wyeliminowany dzięki zastosowaniu nowej definicji błędu komparacji [15]:

$$\delta = -\frac{\mathrm{d}}{\mathrm{dU_{AC}}}\left[\frac{(E_{AC}-E_{DC})^2}{\frac{\mathrm{d}E_{AC}}{\mathrm{dU_{AC}}}}\right]. \tag{8}$$

Weryfikacja definicji dokonana na 11 przetwornikach jednozłączowych potwierdziła jej dokładność na poziomie rzędu 1.10^{-7} w zakresie prądów grzejnika: 20...120%. Zdaniem autorów definicja może być stosowana nawet do najdokładniej odtwarzanych wzorców.

Udowodniono też, że zaniedbywanie wpływu impedancji grzejnika przy analizie błędu komparacji może prowadzić do znaczących błędów dodatkowych [16]. Eksperyment przeprowadzono przy użyciu rezystorów dodatkowych, wyposażonych w handlowe oporniki warstwowe. Błąd komparacji o wartości 1.10^{-6} wystąpił przy częstotliwości 100 kHz, gdy impedancja grzejnika stanowiła 0,7% impedancji rezystora.

Na początku lat 80-tych prace nad utworzeniem bazy odniesienia pomiarów napięcia przemiennego podjęto w PKNMiJ [7]. Spośród zbioru handlowych przetworników termoelektrycznych typu A55 produkcji Fluke wyselekcjowano zestaw podstawowy składający się z 13 egzemplarzy o zbliżonych charakterystykach częstotliwościowych w zakresie 10 Hz...10 kHz [17]. Średni błąd komparacji nie przekraczał 5.10^{-5}, a dryft roczny nie był większy od 1.10^{-5}. Wiarygodność rezultatów potwierdziły wielokrotne badania prowadzone w ciągu 3 lat.

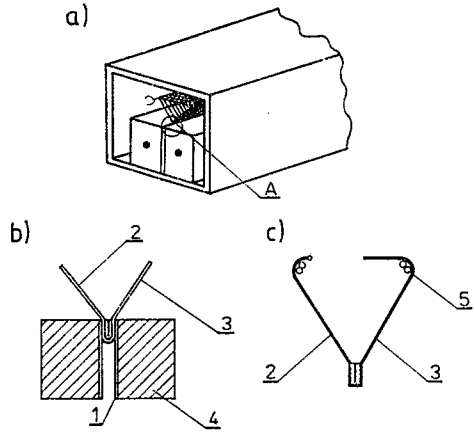
[6] National Measurements Laboratory—Australia

[7] Polski Komitet Normalizacji Miar i Jakości — Warszawa

W ostatnich latach wiele interesujących informacji o pracach nad analizą zjawisk fizycznych towarzyszących pracy termoelementu napływa z Australii. Np. przebadano tam wpływ niejednorodności materiału grzejnika na składowe błędu komparacji — zależną i niezależną od prądu obciążenia oraz wskazano na nowe źródła błędów związane z występowaniem zjawisk termoelektrycznych [18, 19]. Zasadniczo można uznać, że rozwój i doskonalenie konstrukcji przetworników termoelektrycznych jednozłączowych po okresie pewnego zastoju staje się znów przedmiotem zainteresowań badawczych.

PRZETWORNIKI WIELOZŁĄCZOWE

Na początku lat 60-tych, pod kierunkiem F. J. Wilkinsa, podjęto w NPL prace nad konstrukcją wielozłączowego próżniowego przetwornika termoelektrycznego, który charakteryzował się dużą siłą termoelektryczną wyjściową, małą symetrią, krótkim czasem odpowiedzi i niespotykanie małym błędem komparacji [20]. Zasadę konstrukcji tego urządzenia podano na rys. 4.



Rys. 4. Zasada konstrukcji przetwornika wielozłączowego (wg. Wilkinsa): a) fragment przekroju b) szczegół „A", c) wycinek pojedynczego złącza
1 — przekładki mikowe, 2 — odcinek złącza nie pokryty miedzią, 3 — odcinek złącza pokryty miedzią, 4 — bloki miedziane, 5 — przekrój grzejnika

Wielozłączowy układ termopar został wykonany techniką galwaniczną nanoszenia powłoki miedzianej na drut żelazny. Miejsca, w których powłoki te kończyły się miały wszelkie cechy złącz termoelektrycznych. Uzyskano w ten sposób 40, 80, 120, a nawet 180 złączy, które przy gradiencie temperatur 30°C wykazywały siłę elektromotoryczną do 100 mV.

Niemal równolegle do opisanych prac, pod kierunkiem W. S. Popowa w Instytucie Elektrotechnicznym w Leningradzie, skonstruowano wielozłączowy przetwornik powietrzny wykorzystywany początkowo wyłącznie do pomiaru mocy [5]. Jednostkowe wykonania przetworników angielskich tego typu stosowane były w NPL i NBS [10]. Udowodniono, że przy niskich i średnich częstotliwościach zjawisko Thomsona może mieć poważn

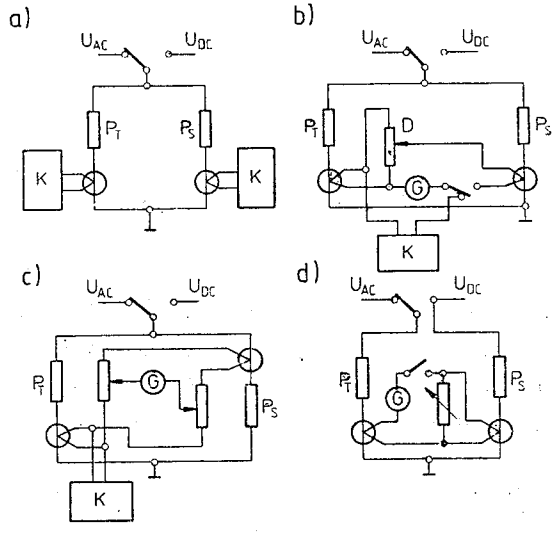
wpływ na błąd komparacji, natomiast przy wyższych częstotliwościach decydujące jest oddziaływanie indukcyjności i pojemności grzejnika [21]. Opracowano nową konstrukcję, dla której w zakresie do 100 kHz uzyskano błąd komparacji rzędu $1...2.10^{-6}$. Na bazie doświadczeń Wilkinsa, firmy Graham (W. Bryt.) i Guidline (Kanada) podjęły produkcję przetworników wielozłączowych [22, 23]. Konstrukcje radzieckie zostały wdrożone do produkcji jako typy TEM i kilka egzemplarzy używano w połowie lat 70-tych w WNIIM do pomiarów komparacyjnych napięć w zakresie częstotliwości 20 Hz...100 kHz [14].

Własne, oryginalne konstrukcje stosowano również w tym okresie w ETL przy częstotliwościach 40 Hz...100 kHz [11]. W NBS przeprowadzono porównania 8 egz. starannie wyselekcjonowanych, wielozłączowych przetworników, wyposażonych w rezystory dodatkowe [24]. Rezystory wykonano z drutu nawiniętego na płytki z mikanitu. Pomiary wykonano przy 30, 100 Hz i 10 kHz, uzyskując średni błąd komparacji rzędu $0,3...0,5.\ 10^{-6}$, z powtarzalnością lepszą od $0,3.\ 10^{-6}$, wyznaczoną w ciągu kilku miesięcy.

Zainteresowanie doskonaleniem konstrukcji i wykorzystaniem przetworników wielozłączowych trwało zasadniczo do końca lat 70-tych, kiedy okazało się, że istnieją trudności w osiągnięciu stałości i powtarzalności parametrów w wykonaniach handlowych. Ostatecznie, wady te wydatnie ograniczono, dzięki m.in. poprawie technologii usuwania gazów i skuteczniejszemu uszczelnianiu ampułek szklanych [25, 26]. Jednakże wyjaśnienia oczekuje problem wiarygodności uzyskiwania błędu komparacji na poziomie $1.\ 10^{-6}$.

KOMPARATORY PRZETWORNIKÓW TERMOELEKTRYCZNYCH

Urządzeniami służącymi do odtwarzania charakterystyk metrologicznych przetworników termoelektrycznych i porównywania ich między sobą są komparatory. Pierwsza



Rys. 5. Komparatory przetworników termoelektrycznych: a) o dwóch obwodach komparacyjnych, b) z dzielnikiem Kelvina-Varleya, c) z mostkiem oporowym, d) różnicowy

K — kompensatory sił termoelektrycznych, P_T — przetwornik badany, P_S — przetwornik odniesienia, *G* — galwanometr, *D* — dzielnik Kelvina-Varleya

informacja o stosowaniu takiego urządzenia pochodzi z 1960 r., kiedy to wymieniona już firma Fluke zastosowała je do wzorcowania produkowanych przez siebie przetworników [8] — schemat tego urządzenia podano na rys. 5a. Błąd komparacj iprzetwornika badanego wyznaczony był względem przetwornika odniesienia na podstawie algorytmu:

$$\delta_T = \left[\frac{E_{AC} - E_{DC}}{nE_{DC}}\right]_S + \delta_S \tag{9}$$

gdzie: E_{AC} — siła elektromotoryczna wyjściowa przetwornika przy zasilaniu prądem przemiennym,
E_{DC} — siła elektromotoryczna wyjściowa przetwornika przy zasilaniu prądem stałym,
n — wykładnik funkcji przetwarzania ($E = k \cdot I^n$),
δ_S — błąd komparacji przetwornika odniesienia (uwaga: indeksy „T" i „S" odnoszą się odpowiednio do przetworników badanego i odniesienia).

Komparator oparty na tej zasadzie stosowany był także w NBS, gdzie szybko dostrzeżono, że żmudne operacje manualne i obliczeniowe można zautomatyzować za pomocą komputera [27]. Wyniki pomiarów uzyskane przy pomocy tego urządzenia były zgodne z rezultatami uzyskanymi techniką manualną z dokładnością $1 \cdot 10^{-6}$. Równolegle prowadzono prace nad uproszczeniem obwodu równoważenia sił termoelektrycznych obydwu przetworników i wyeliminowaniem wpływu niestałości źródeł zasilania.

Zastosowano do tego celu układ dzielnika Kelvina-Varleya, a przykład rozwiązania podano na rys. 5b) [5]. Błąd komparacji wyznaczono z odchyleń galwanometru przy obydwu rodzajach zasilania, stosując wyrażenie:

$$\delta_T = \frac{(\alpha_{AC} - \alpha_{DC})_S}{n_S E_S} + \delta_S \tag{10}$$

gdzie: α_{AC} — odchylenie galwanometru przy zasilaniu obwodu prądem przemiennym,
α_{DC} — odchylenie galwanometru przy zasilaniu obwodu prądem stałym,
n_S — wykładnik funkcji przetwarzania przetwornika odniesienia,
E_S — siła elektromotoryczna przetwornika odniesienia.

Omawiany układ charakteryzował się błędem nie przekraczającym $\pm 2 \cdot 10^{-6}$ i umożliwiał sprawdzanie przetworników w paśmie częstotliwości do 50 kHz.

Inny układ komparatora niewrażliwy na niestabilności źródeł zasilania opracowany w NBS zawierał mostek w obwodach wyjściowych porównywanych przetworników [28]. Uproszczony schemat układu podano na rys. 5c). Błąd komparacji wyznaczono stosując zależność:

$$\delta_T = m(\alpha_{AC} - \alpha_{DC})_S + \delta_S \tag{11}$$

gdzie m — współczynnik czułości,

Za pomocą tego urządzenia dokonano porównań przetworników termoelektrycznych z dokładnością około $1 \cdot 10^{-5}$ dla napięć 0,5...500 V w paśmie częstotliwości 20 Hz...50 kHz. Nowy typ komparatora wykonano w NBS, dzięki któremu opanowano skutecznie dryft czasowy przetworników termoelektrycznych [29] — schemat ideowy urządzenia przedstawiono na rys. 5d). Wykorzystując komparator wyznaczono błąd komparacji zestawu przetworników napięciowych 0,5...500 V. Uzyskane wyniki do 20 kHz były mniejsze od $1 \cdot 10^{-5}$.

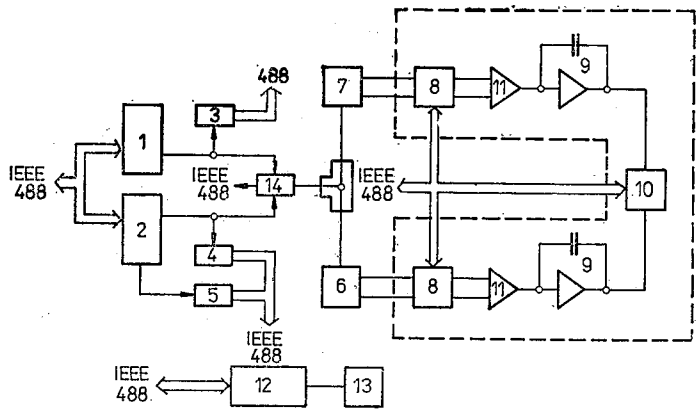
Na początku lat 70-tych skonstruowany w ETL komparator dawał możliwość porównania przetworników w układzie szeregowym i równoległym [11]. Ocenia się, że niestabilność komparatora była mniejsza od 10^{-6} przy 100 kHz. Zakres napięciowy zawierał się w przedziale 1...100 V.

Pierwsze informacje o komparatorze stosowanym w ZSRR do celów odtwarzania jednostki napięcia przemiennego pochodzą z połowy lat 70-tych [30] — schemat urządzenia był zgodny z układem podanym na rys. 5a).

W tym czasie w dziedzinę wkracza normalizacja. Norma amerykańska [31] zaleca stosowanie układu komparatora wg rys. 5a) i opisuje metodykę sprawdzania przetworników. Wydana w kilka lat później norma RWPG [32] jest merytorycznie zgodna w omawianej tematyce z normą amerykańską. Preferowanie przez obydwie normy identycznego układu pomiarowego jest uzasadnione głównie jego prostotą. Wszystkie pozostałe układy są bardziej złożone, wymagają bowiem stosowania wysokostabilnych rezystorów nastawnych i dzielników.

Komparator zgodny z zaleceniami wspomnianych norm był zastosowany w PKNMiJ do wyznaczania charakterystyk przetworników termoelektrycznych. Czułość układu nie była mniejsza od $1 \cdot 10^{-6}$. Pełną analizę pracy urządzenia podano w [33].

Jednakże prostota układowa nie rekompensuje złożoności czynności manualnych. W ostatnich latach opracowano różne systemy, umożliwiające częściowe lub całkowite zautomatyzowanie pomiarów [34—39]. Wszystkie wykorzystują sterowanie komputerowe lub mikroprocesorowe, osiągając dokładność pomiaru $1 \cdot 10^{-6}$ lub lepszą. Omówimy jeden spośród nich, którego schemat blokowy podano na rys. 6 [39] Podstawowy układ kompa-



Rys. 6. Schemat blokowy automatycznego testowania przetworników termoelektrycznych
1 — źródło napięcia stałego, 2 — źródło napięcia przeminnego, 3 — woltomierz cyfrowy napięcia stałego, 4 — woltomierz cyfrowy napięcia przemiennego, 5 — częstościomierz, 6 — przetwornik odniesienia, 7 — przetwornik badany, 8 — przetwornik A/C, 9 — układ całkujący, 10 — woltomierz cyfrowy, 11 — wzmacniacz, 12 — układ sterowania i kontroli, 13 — drukarka (ploter), 14 — przełącznik rodzajów zasilania

ratora współpracuje z blokami funkcjonalnymi połączonymi przez interfejs zgodny z normą IEEE 488. System składa się z dwóch identycznych kanałów, z których każdy zawiera: obwód równoważący przetwornika A/C, wzmacniacz operacyjny, służący do wzmacniania sygnałów nierównowagi napięciowej i obwód całkujący. Napięcie na kondensatorze in-

tegratora mierzone za pośrednictwem woltomierza cyfrowego powiązane jest z E_{SC} i E_{DC} zależnością:

$$\delta_T = \left[\frac{E_{AC}-E_{DC}}{nE_{DC}}\right]_S - \left[\frac{E_{AC}-E_{DC}}{nE_{DC}}\right]_T + \delta_S. \tag{12}$$

Wyniki pomiarów obliczane są automatycznie, gromadzone i analizowane przez komputer. Szybkość dokonania jednego pomiaru jest 2-krotnie większa w stosunku do układu z manipulacją ręczną.

UNIFIKACJA MIARY

Rozwój konstrukcji przetworników termoelektrycznych jedno- i wielozłączowych oraz komparatorów tych przetworników stworzył możliwości opracowania w wielu krajach zestawów etalonowych napięcia przemiennego oraz ich międzynarodowego porównania. Pierwszą taką komparację przeprowadzono w r. 1975 między ZSRR, USA a Wielką Brytanią [30, 40]. Uczestniczyły w niej przetworniki o zakresach: 10...100 V, a pomiarów każdego zestawu państwowego z każdym dokonano przy częstotliwościach 40 Hz...50 kHz. Niezgodności uzyskanych wyników nie przekraczały $2 \cdot 10^{-6}$. Istotnym krokiem formalnoprawnym było ustanowienie, w ZSRR normą państwową, etalonu specjalnego napięcia przemiennego [41, 42]. Ustalono skład etalonu, metodykę jego odtwarzania, zakresy napięć i częstotliwości, przy których metodyka obowiązuje, zestaw aparatury podstawowej i pomocniczej oraz sposób opracowania wyników pomiarów. Choć brak informacji o nadaniu w innych krajach etalonom napięcia przemiennego podobnego statusu, nie mniej w rzeczywistości funkcję taką one spełniają. Najczęściej stosowane są w paśmie częstotliwości 10 Hz...30 (1500) MHz z wyraźnie zarysowanym podziałem na zakresy:

— m.cz. — obejmujące 10 Hz...(50) 100 kHz,
— w.cz. — obejmujące 100 kHz...30 (1500) MHz.

Podany podział wynika z odrębności cech metrologicznych narzędzi pomiarowych pracujących w określonym zakresie częstotliwości. Np. dla pasma w.cz. cechą charakterystyczną jest praca przy niskim napięciu — praktycznie do 10 V a dla m.cz. odwrotnie — do 1000 V.

Etalon napięcia przemiennego nie ma dotychczas statusu międzynarodowego, lecz nadanie mu tej rangi niebawem nastąpi. Istotnym krokiem w tym kierunku były dwie kolejne komparacje międzynarodowe, które odbyły się w 1978 i 1980 r. między ZSRR, USA, Wielką Brytanią a Japonią [43, 44]. Porównania zestawów przetworników termoelektrycznych były prowadzone przy 10, 30 i 100 V i częstotliwościach: 40 Hz, 20 i 50 kHz. Wartość średnia błędu ostatniej komparacji określona przez każdą ze stron różniła się od średniej dla wszystkich uczestniczących o wartość $1 \cdot 10^{-5}$.

W 1984 r. inicjatywę ujednolicenia miary w ramach RWP podjęły: ZSRR, Polska, NRD, Węgry i Bułgaria. Porównania wzorcowych przetworników termoelektrycznych wspomnianych krajów były prowadzone w ramach tzw. komparacji radialnych i okrężnych [45]. Pomiary przeprowadzono w zakresie częstotliwości 20 Hz...100 kHz — dla zestawu m.cz. i w zakresie 100 kHz...30 MHz — dla zestawu w.cz. Na podstawie dokonanych

analiz ustalono, że najmniejszym błędem komparacji obarczone są zestawy przetworników radzieckich i postawiono wniosek o ustanowienie na ich bazie etalonu RWPG jednostki napięcia elektrycznego w zakresie częstotliwości 20 Hz...30 MHz [46].

PODSUMOWANIE

Wobec braku autonomicznych wzorców napięcia przemiennego, jednostka tej wielkości odtwarzana jest metodami pośrednimi przez porównanie z jednostką napięcia przy prądzie stałym.

Podstawę odtwarzania jednostki napięcia przemiennego stanowią przetworniki termoelektryczne:

— jednozłączowe, które wierniej odtwarzają jednostkę w zakresie w.cz.

— wielozłączowe, z przeznaczeniem do pracy w zakresie m.cz.

Urządzeniami służącymi do badania charakterystyk metrologicznych przetworników termoelektrycznych są komparatory. Współcześnie stosowane urządzenia tego typu, dzięki zastosowaniu komputerów i mikroprocesorów pozwalają na zautomatyzowanie procesu pomiarowego.

Rozwój i doskonalenie konstrukcji przetworników oraz wspomnianych komparatorów stworzył podstawę do ujednolicenia miary w skali międzynarodowej.

BIBLIOGRAFIA

1. A. Barański: *Wpływ konstrukcji przetworników termoelektrycznych na błąd częstotliwościowy.* Wiadomości Elektrotechniczne, 1986 r. nr 5—6.
2. A. Barański: *Termoelementy w miernictwie elektrycznym — potrzeba normalizacji.* Wiadomości Elektrotechniczne, 1986 r. nr. 1—2, str. 12—15
3. F. L. Hermach: *Thermal Corverters as AC-DC Transfer Standards for Current and Voltage Measurements at Audio Freqencies.* Journ. ef Reserch NBS, 1952 r. vol 48, nr. 2, str. 317—334
4. F. L. Hermach, E. S. Williams: *Thermal Voltage Converters for Accurate Voltage Measurements to 30 Megacyles per Second.* Comunic. and Electr. 1960 r. nr 7, str. 200—205
5. F. J. Wilkins: *Vielfach-Thermoumformer — als Weschselatrom Gleichstrom Transfer Instument.* Messtechnik, 1968 r. nr 10, str. 258—265
6. F. C. Widdis: *The Theory of Peltier-and Thomson-Effekt Errors in Thermal A. C.-D. C. Transfer Devices.* Proccedings IEE, 1962 r. vol. 1/9, nr. 16, str. 328—334
7. A. Ja. Bezikovič, D. I. Zorin, T. W. Rożdestvenskaja: *Novaja apparatura dla točnego izmerenija toka, naprjaženija i moščnosti v širokom diapazonie častot.* Acta Imeco, 1961 r. str. 156—179
8. J. D. Sleeth, W. H. Odegard: *Production Calibration and NBS Traceability of Thermal Voltage Converters and Transfer Standards.* Instrum. Society of America. 19th Annual ISA Conference and Exhibit. October 15—16. 1964, New York. repr. nr 21, 1964
9. E. S. Willams: *Termal Voltage Converters and Comparator for Very Accurate AC Voltage Measurements.* Journ. of Res. NBS 1971 r. vol. 75c, nr 3—4, str. 145—154
10. F. L. Hermach, E. S. Williams: *Thermal Converters for Audio-Frequency Voltage Measurements of High Accuracy.* IEEE Trans. Inst. Meas. 1968 r. vol. IM-15, nr 4, pp. 260—268
11. S. Iwamoto, H. Hirayama: *Convertisseur thermiques courant alternatif — courant continu.* Com. Cons. d'Electr. 13 session, 1972 r. str. 78—83

12. *Etalon et comparateur du NPL pour le passage courant alternatif courant continu.* Com. Cons. d'Electr. 13 session, 1972 r. str. 84—86
13. P. F. Aknaev, T. B. Rożdestvenskaja: *Analiz pogresnostej, vnosimych termopreobrazovateljami pri kompariowanji naprjaženij peramennogo i postojannogo tokov.* Trudy Metr. Instit. SSSR. 1987 r. Vyp. 138/198 str. 46—57
14. P. F. Aknaev, T. B. Rożdestvenskaja: *Novaja apparatura dla izmerenija diejstvujuscego značenija naprjaženija v širokom diapozonie castot.* Izm. Techn. 1970 r. nr 5, str. 55—59
15. B. D. Inglis: *Errors in AC-DC Transfer Arsing from a DC Reversal Difference.* Metrologia, 1981 r. nr 17, pp. 111—117
16. B. D. Inglis: *Evaluation of AC-DC Transfer Errors for Thermal Converter Multiplier Combinations.* Metrologia, 1980 r. nr 16, pp. 177—181
17. A. Barański: *O potrzebie ustanowienia etalonu napięcia przemiennego w pasmie częstotliwości 10Hz...10kHz.* Normalizacja 1985 r, nr 10, str. 21—26
18. B. D. Inglis: *Current-Independent AC-DC Transfer Errors in Single-Junction Thermal Couverters.* IEEE Trans. Instr. Meas. 1985 r. vol. IM-34, Nr 3, pp. 294—301
19. B. D. Inglis: *AC-DC Transfer Standards-Present Status and Future Directions.* IEEE Trans. Instr. Meas. 1985 r. vol. IM-34, nr 2, pp. 285—290
20. B. A. Wilkins, T. A. Deacon, R. S. Backer: *Multijunction thermal converter.* Proc.: IEE. 1965 r. vol. 112, nr 4. pp. 794—805
21. F. J. Wilkins: *Teoretical Analysis of the AC/DC Transfer Difference of the NPL Multijunction. Thermal Converter over the Frequency Range DC to 100 kHz.* IEEE Trans. Instr. Meas. 1972, v. IM-21, nr 4, pp. 334—340,
22. L. R. Graham and C. O. Ltd. *The „Wilkins" Multijunction Thermal Converter as Developemed in the Division of Electrical Science.* National Phisical Laboratory. Teddington. England
23. *Multi-Junction Thermal Converters (RMS to DC).* Bull. 6915. Guildine Instr. Ltd.
24. F. L. Hermach, D. R. Flach: *An Investigation of Multujunction Thermal Converters.* IEEE Trans. Instr. Meas., 1976, vol. IM-25 nr 4 pp. 524—528
25. D. Zhang, Z. Zhang: *Method for Reduction of AC-DC Transfer Error Caused by the Thomson Effect for the Multijunction Thermal Converter.* IEEE Trans. Instr. Meas. 1980 vol. IM-29, pp. 412—415
26. M. Klonz, F. J. Wilkins: *Multijunction Thermal Converter with Adiustable Output Voltage Current Characteristic.* IEEE Trans. Instr. Meas., 1980 r. vol. IM-29, str. 409—412
27. R. S. Turgel: *A Comparator for Thermal AC-DC Transfer Standards.* ISA Trans., 1967 r. vol 6, nr 4. pp. 286—292
28. F. L. Hermach, E. S. Williams: *Thermal Converters for Audio-Frequency Voltage Measurements of High Accurate.* IEEE Trans. Instr. Meas., 1956 r. vol. IM-15, nr 4, pp. 327—331
29. F. L. Hermach, J. E. Griffin, E. S. Williams: *A System for Accurate Direct and Alternating Voltage Measurements.* IEEE Trans. Instr. Meas. 1965 r. vol. IM-14, nr 4, pp. 215—224
30. R. F. Aknaev, O. P. Galachova, T. B. Rożdestvenskaja: *Metody i sredstva obespacenija edinstva izmerenij naprjazenija peremennego toka,* Izm. Tech. 1976 r. nr 4, str. 66—71
31. ANSI C 100 A-1973, *American National Standard for AC-DC Transfer Instruments and Converters*
32. ST SEV 1056—78 Metrologia. *Komparatory i preobrazovateli peremennogo i postojannogo toka. Termoelektričeskije metody poverki*
33. A. Barański: *Komparator przetworników wzorcowych.* Normalizacja, 1979 r. nr 12, str. 22—25
34. J. K. Lentner, C. B. Childers, S. G. Termaine: *A Semiautomatic System for AC/DC Difference Calibration.* IEEE Trans. Instr. Meas., 1980 r. vol. IM-29, nr 4, pp. 400—404
35. E. S. Williams: *A Thermoelement Comparator for Automatic AC-DC Difference Measurements.* IEEE Trans. Instr. Meas., 1980, v. IM-29, nr 4, pp. 405—409
36. K. J. Lentner, D. R. Flach: *Automatic System for AC/DC Calibration.* IEEE Trans. Instr. Meas. 1983 r. vol. IM-32, nr 1, pp. 51—56
37. P. Martin, R. B. Knight: *Components and Systems form AC/DC Transfer at the ppm Level.* IEEE Trans. Instr. Meas., 1983 r. v. IM-32, nr 1, pp. 63—72

38. F. Cabiati, U. Pogliano, J.C. Bosco, J. Tago;: *High Precision Thermal EMF Comparator for Automatic AC-DC Transfer Measurement.* IEEE Trans. Instr. Meas. 1983 r. vol. IM-32, nr 1, pp. 57—61
39. J.R. Kinard, E.S. Williams, T.E. Lipe: *Automated Thermal Voltage Converter Intercomparisons.* Proceedings of the IEEE, 1986 r. vol. 74, nr 1, pp. 105—107
40. T.B. Rozdhestvenskaya, O.P. Galakhova, R.F. Aknaev: *Standard Apparatus for the volt Realization in the Frequency Range from 20 Hz to 1500 MHz.* IEEE Trans. Instr. Meas., 1976, v. IM-25 nr. 4, pp. 538—541
41. *Gosudarstvennyj specjalnyj etalon i obščesojuznaja poveročnaja schema dlja sredstv izmerenij naprjaženija 0,001—1000 V v diapozone častot 20÷3. 10* . Gc. GOST 8.184-76
42. T.B. Rozdestvenskaja, R.F. Aknaev, O.P. Galachova: *Gosudarstvennyj specialnyj etalon edinicy naprjazenija peromennogo toka.* Izm, Tech. 1976 r. nr 3, str. 44—46
43. B.N. Taylor: *Rapport sur une comparaison internationale de convertisseur thermiques utilises comme ètalon de transfert courant contimme- courant alternatif.* Com. Cons. d'Electr. 15 session 1978.
44. O.P. Galakhova: *An International Comparison of Thermal Converters as AC-DC Transfer Standards.* IEEE Trans. Instr. Meas. 1980 r., vol. IM1-29, str. 396—399
46. SEV-Tem. 01.752.33-83. *Sličenija etalonov edinicy električeskogo naprjaženija v diapezonie častot ot 20 Hz do 30 MHz*
46. SEV. *Pasport etalona edinicy električeskogo naprjaźenija v diapazonie častot ot 20 Hz do 30 MHz.* II projekt.

A. BARAŃSKI

INSTRUMENTS AND METHODS OF THE STANDARDIZATION OF ALTERNATING VOLTAGE MEASUREMENTS

Summary

Development and improvement of the construction of single-junction thermo-electric converters and multi-junction thermo-electric converters to be used in high precision measurements and applied for the reproduction in the alternating voltage unit, are presented. A review of measuring systems for the determination of the above mentioned converters is carried out. The repeatability of the obtained results gave the base for works on the standardization of the measure on an international scale. In the framework of a cooperation between the national measuring offices and institutes for metrology, comparisons of special alternating voltage standards, were made. The paper deals with the Polish contribution on the background of world-wide achievements in this field.

A. BARAŃSKI

MOYENS ET MÉTHODES DE L'UNIFICATION DES MESURES DE LA TENSION ALTERNATIVE

Résumé

On a présenté les étapes du développement et du perfectionnement de la construction des transducteurs thermoélectriques à une ou à plusieurs jonctions, depuis leur application au mesurage précis de la tension alternative jusqu'à leur application à la reproduction de l'unité de mesure de la tension alternative. On a passé en revue les systèmes de mesure utilisés pour la détermination des caractéristiques des transducteurs en question. La reproductibilité des résultats obtenus a permis d'entreprendre l'initiative de l'unification de la mesure à l'échelle internationale. Dans le cadre de la coopération des bureaux de mesures et des instituts de métrologie, on a effectué des comparaisons des étalons spéciaux de la tension alternative. La contribution de la Pologne dans ce domaine a été mise en comparaison avec les résultats mondiaux.

A. BARAŃSKI

MITTEL UND METHODEN DER MAßVEREINHEITLICHUNG DER WECHSELSPANNUNG

Zusammenfassung

Es wurden Entwicklungsetappen und Vervollkommnung der Konstruktion von Thermoumformern mit einem und mehreren Übergängen dargestellt. Eine Übersicht über die Anwendungen von Thermoumformern, besonders für Präzisionsmessungen der Wechselspannung und Reproduzierbarkeit der Wechselspannungseinheit wurde getätigt sowie eine Übersicht über die Meßvorrichtungen für die Bestimmung von Charakteristiken dieser Thermoumformer. Die Wiederholbarkeit der Ergebnisse ergab den Grund für die internationale Maßvereinheitlichung. Im Rahmen der Zusammenarbeit der staatlichen Behörden und metrologischer Institute wurden internationale Vergleichmessungen der Wechselspannungs-spezialetalonen durchgeführt. Es wurde die Beteiligung der VRPolen an den Welterrungenschaften auf diesem Gebiet dargestellt.

А. БАРАНСКИ

СРЕДСТВА И СПОСОБЫ ОБЕСПЕЧЕНИЯ ЕДИНСТВА МЕРЫ НАПРЯЖЕНИЯ ТОКА

Резюме

Представлены этапы развития и совершенствования конструкции электрических одно и многоэлементных преобразователей переменного тока. Рассмотрены их применения при точных измерениях и при воспроизведении единицы напряжения переменного тока. Приведен обзор схем для определения характеристик обсуждаемых преобразователей. Повторяемость получаемых результатов явилось базой для работ по унификации единицы измерения в международном масштабе. В сотрудничестве государственных учреждений и метрологических институтов проведены сравнения специальных эталонов напряжения переменного тока. На фоне мировых достижений представлено участие П.Н.Р. в этой области.

ROZPRAWY ELEKTROTECHNICZNE, 1989, 35, z. 1, ss. 61-75

532.575.621.3.087

Metoda projektowania optymalnego kształtu cewki wzbudzającej przepływomierzy elektromagnetycznych dla kanałów otwartych

ANDRZEJ MICHALSKI, JAN SIKORA, STANISŁAW WINCENCIAK

Instytut Elektrotechniki Teoretycznej i Miernictwa Elektrycznego, Politechnika Warszawska

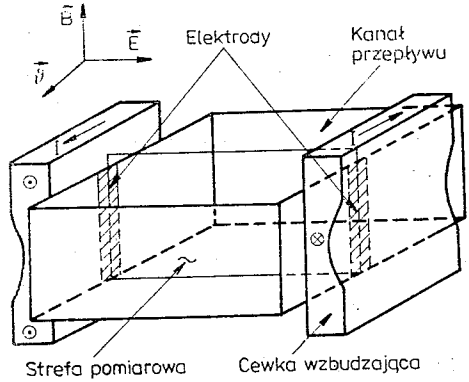
Otrzymano 1988.01.08

Autoryzowano do druku 1988.05.20

W pracy przedstawiono algorytm do optymalnego projektowania kształtu cewki wzbudzającej przetwornika pierwotnego przepływomierza elektromagnetycznego. Korzystając z metody elementów skończonych podzielono przekrój poprzeczny cewki na elementy skończone. Jako zmienne projektowe przyjęto węzły elementów skończonych leżących na jednym z brzegów przekroju poprzecznego cewki. Do minimalizacji funkcji celu zastosowano uogólniony nieliniowy algorytm Newtona. Długość cewki wyznaczono na podstawie znajomości zjawiska efektów zwarciowych.

1. WPROWADZENIE

Pomiary przepływu są jednym z podstawowych pomiarów dokonywanych w hydrologii. W ostatnim okresie czasu można zauważyć znaczny wzrost zainteresowania ze strony służb komunalnych wykorzystaniem metody elektromagnetycznej do pomiaru przepływu w małych kanałach otwartych. Idea metody elektromagnetycznej znana jakkolwiek od 1831 r. dopiero w drugiej połowie XX wieku doczekała się szerokiego rozwoju. Główne cechy tej metody, jak i podstawowe problemy występujące przy jej przemysłowym zasto-



Rys. 1. Przetwornik pierwotny przepływomierza elektromagnetycznego do kanałów otwartych

sowaniu można znaleźć w pracach [1, 2, 4, 8]. Od ponad 10 lat w Instytucie Elektrotechniki Teoretycznej i Miernictwa Elektrycznego Politechniki Warszawskiej prowadzone są prace pod kierunkiem prof. A. Chwaleby nad praktycznymi aspektami wykorzystania metody elektromagnetycznej do pomiaru przepływu w naturalnych i przemysłowych kanałach otwartych.

Tematem artykułu jest metoda obliczania optymalnego kształtu cewki wzbudzającej w przetworniku pierwotnym przepływomierza (rys. 1).

2. MODEL MATEMATYCZNY PRZEPŁYWOMIERZA

Rozkład indukowanego w kanale przepływowym potencjału elektrycznego φ opisany jest równaniem Poissone'a [1]

$$\Delta\varphi = \operatorname{div}(\vec{v}\times\vec{B}) \tag{2.1}$$

gdzie $\vec{v}$ — prędkość mierzonej cieczy

$\vec{B}$ — indukcja wzbudzonego pola magnetycznego.

Mierzoną na elektrodach różnicę potencjałów można przedstawić w postaci

$$U = \varphi_1 - \varphi_2 = -\int_V (\vec{v}\times\vec{B})\cdot\vec{W}\,\mathrm{dV} = \int_V (\vec{W}\times\vec{B})\cdot\vec{v}\,\mathrm{dV} \tag{2.2}$$

gdzie: $\vec{W}$ — wektor wagi zdefiniowany przez Bevira i Shercliffa [1, 2]

V — przestrzeń objęta przepływem.

Warunkiem przeprowadzenia poprawnego pomiaru jest uniezależnienie mierzonej różnicy potencjałów U od zmiennego rozkładu prędkości przepływającej cieczy. Zgodnie z wzorem (2.2) koniecznym staje się spełnienie warunku

$$\vec{W}\times\vec{B} = const. \tag{2.3}$$

w całej strefie pomiarowej.

Dla kanałów przemysłowych o przekroju poprzecznym w kształcie prostokąta i nieprzewodzących ściankach ograniczających przepływ, przy wykorzystaniu elektrod długich obejmujących cały przepływ rozkład wektora $\vec{W}$ jest jednorodny w całej strefie pomiarowej Spełnienie warunku (2.3) narzuca konieczność uzyskania równomiernego rozkładu indukcji $\vec{B}$.

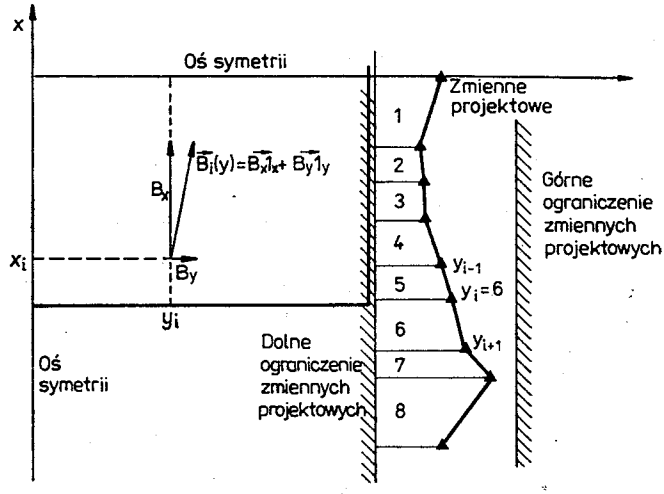
Głównym celem projektanta jest więc określenie kształtu przekroju poprzecznego cewki oraz jej długości z rys. 1. Kształt przekroju poprzecznego cewki determinowany jest głównie przez dopuszczalną nierównomierność wzbudzanego pola magnetycznego.

Długość cewki L wyznaczana jest w oparciu o eliminacje efektu zwarciowego występującego w strefie dużych zaników pola magnetycznego.

3. OPTYMALIZACJA KSZTAŁTU CEWKI PRZEPŁYWOMIERZA

3.1. METODA OPTYMALIZACJI

Rozpatrzmy przekrój poprzeczny przepływomierza, ze względu na symetrię do rozwiązań można przyjąć 1/4 przekroju (rys. 2).



Rys. 2. Przekrój poprzeczny przetwornika pierwotnego przepływomierza

Cewkę wzbudzającą pole magnetyczne należy tak ukształtować, aby wewnątrz kanału pole w każdym i-tym punkcie miało tą samą zadaną wartość. W tym celu, bazując na idei elementu skończonego, dokonano dyskretyzacji cewki na czworokątne elementy izoparametryczne. Jako zmienne projektowe przyjęto współrzędne „y" górnej krawędzi cewki.

W tym przypadku projektowanie optymalnego kształtu cewki sprowadza się do poszukiwania minimum funkcji celu o postaci

$$\min_{y \in R_0} F(y) = \frac{1}{2} \sum_{i=1}^{m} [f_i(y)]^2 \tag{3.1}$$

gdzie

$$f_i = |\vec{B}_i(y) - \vec{B}_0| = \sqrt{[B_x(x_i, y_i) - B_{0x}]^2 + [B_y(x_i, y_i) - B_{0y}]^2}$$

$\vec{B}_i(y)$ — wektor indukcji w i-tym punkcie przekroju kanału

$\vec{B}_0$ — zadany wektor indukcji.

Dodatkowo na poszukiwane zmienne y_i nałożono następujące ograniczenia wynikające z wymagań konstrukcyjnych

$$l_j \leqslant y_j \leqslant u_j \tag{3.2}$$

Ograniczenia te wynikają głównie z konieczności uzyskania zwartej konstrukcji przetwornika pierwotnego. W rezultacie można uzyskać cewkę wzbudzającą o rozmiarach nieznacznie tylko przekraczających rozmiary kanału przepływu.

Ze względu na nieliniowy charakter postawionego problemu rozwiązanie poszukiwane będzie w procesie iteracyjnym o postaci

$$\{y_{n+1}\} = \{y_n\} + \alpha\{\Delta y_n\} \tag{3.3}$$

gdzie: α — dobierane jest tak, aby w kolejnych iteracjach wartość funkcji celu (3.1) nie wzrastała.

Stosując uogólniony algorytm Newtona wektor poprawek $\{\Delta y\}$ wyznaczamy z rozwiązania algebraicznego układu równań

$$[J]\{\Delta y\} = \{f\}, \tag{3.4}$$

gdzie:

$$J_{i,j} = \frac{\frac{\partial f_i}{\partial y_i}[B_x(x_i, y_i) - B_{0x}]\frac{\partial B_x(x_i, y_i)}{\partial y_j} + [B_y(x_i, y_i) - B_{0y}]\frac{\partial B_y(x_i, y_i)}{\partial y_j}}{\sqrt{[B_x(x_i, y_i) - B_{0x}]^2 + [B_y(x_i, y_i) - B_{0y}]^2}}.$$

W przypadku ogólnym ilość punktów z zadaną wartością wektora indukcji nie musi być równa ilości poszukiwanych zmiennych. Macierz $[J]$ jest zatem macierzą prostokątną. Poszukujemy rozwiązania układu równań (3.4), które minimalizować będzie normę euklidesową [3]

$$\min \|[J]\{\Delta y\} - \{f\}\|. \tag{3.5}$$

Bardzo wygodnym do tego celu jest algorytm minimalizacji błędu średniokwadratowego (LSP) z wykorzystaniem dekompozycji macierzy $[J]$ względem jej wartości osobliwych (SVD).

3.2. WYZNACZENIE ELEMENTÓW MACIERZY $[J]$

Dla wyznaczenia elementów macierzy Jacobiego musimy określić sposób wyznaczania składowych wektora indukcji w wybranych punktach przekroju poprzecznego przepływomierza.

Jak wspomniano wcześniej w rozważaniach rozpatrujemy przekrój poprzeczny przepływomierza w połowie długości cewki wzbudzającej pole. Wyznaczenie wektora indukcji magnetycznej w przestrzeni kanału oprzemy na sumowaniu elementowych jego składników pochodzących od nieskończenie cienkich przewodów o długości L (długość cewki). Zgodnie z prawem Biotta-Savarta w punkcie położonym na osi symetrii, w odległości od niego a, długość elementarnego wektora indukcji wyraża się zależnością.

$$|\overrightarrow{\mathrm{dB}}| = \frac{\mu_0 J \mathrm{ds}}{2\pi a}\cos\left(\operatorname{arctg}\frac{2a}{L}\right), \tag{3.6}$$

gdzie: J — gęstość prądu

ds — powierzchnia przekroju elementarnego przewodu.

Oznaczając przez (x_p, y_p) współrzędne punktu w $\frac{1}{4}$ przekroju kanału, a przez (x, y) współrzędne dowolnego punktu $\frac{1}{4}$ przekroju poprzecznego cewki możemy (w obszarze I ćwiartki przyjętego układu współrzędnych) zapisać wyrażenie na składowe elementarnego wektora indukcji w postaci

$$\mathrm{dB}_x(x_p, y_p) = \frac{\mu_0 JL}{4\pi}\frac{(y - y_p)\mathrm{dx}\,\mathrm{dy}}{[(x - x_p)^2 + (y - y_p)^2]\sqrt{0.25L^2 + (x - x_p)^2 + (y - y_p)^2}} \tag{3.7}$$

$$\mathrm{dB}_y(x_p, y_p) = \frac{\mu_0 JL}{4\pi}\frac{(x - x_p)\mathrm{dx}\,\mathrm{dy}}{[(x - x_p)^2 + (y - y_p)^2]\sqrt{0.25L^2 + (x - x_p)^2 + (y - y_p)^2}}. \tag{3.8}$$

Wykorzystując symetrię przekroju poprzecznego przepływomierza możemy podać zależności na składowe elementarnego wektora $\overrightarrow{dB}$ w pierwszej ćwiartce przyjętego układu współrzędnych.

$$dB_x = -dB_{x1} + dB_{x2} + dB_{x3} - dB_{x4} \tag{3.9}$$

$$dB_y = +dB_{y1} - dB_{y2} - dB_{y3} - dB_{y4}, \tag{3.10}$$

gdzie indeksy 1, 2, 3, 4 mówią o symetrycznie położonych elementarnych przewodach z prądem odpowiednio w I, II, III i IV ćwiartce układu współrzędnych.
Zatem do wyznaczenia wartości poszczególnych składników w wyrażeniach (3.9) i (3.10) wystarczy do zależności (3.7) i (3.8) w miejsce x i y podstawić współrzędne odpowiadających sobie (symetrycznych) punktów dla przyjętego układu współrzędnych. Dlatego też w dalszych rozważaniach podamy tylko zależności dotyczące pierwszej ćwiartki uzwojenia cewki przepływomierza.

Chcąc wyznaczyć całkowitą wartość składowych wektora indukcji w wybranym punkcie (x_p, y_p) należy dokonać całkowania po obszarze cewki wyrażeń (3.9) i (3.10).

$$B_x = -\int_{S_1} dB_{x1}\,dx\,dy + \int_{S_2} dB_{x2}\,dx\,dy + \int_{S_3} dB_{x3}\,dx\,dy - \int_{S_4} dB_{x4}\,dx\,dy \tag{3.11}$$

$$B_y = \int_{S_1} dB_{y1}\,dx\,dy - \int_{S_2} dB_{y2}\,dx\,dy - \int_{S_3} dB_{y3}\,dx\,dy + \int_{S_4} dB_{y4}\,dx\,dy. \tag{3.12}$$

W celu wyznaczenia odpowiednich całek w wyrażeniach (3.11) i (3.12) dokonano dyskretyzacji przekroju poprzecznego cewki na elementy skończone. Przyjęto izoparametryczne elementy czworokątne.

Dyskretyzacja przekroju poprzecznego cewki na elementy skończone ma także na celu wyróżnienie zmiennych projektowych w postaci wybranych współrzędnych wierzchołków elementów. Takie postępowanie pozwoli nam wyrazić składowe wektora indukcji B_x i B_y jako funkcje poszukiwanych wielkości w procesie optymalizacji.

Dla izoparametrycznego elementu czworokątnego współrzędne dowolnego punktu wewnętrznego elementu można wyrazić poprzez współrzędne wierzchołków w postaci

$$x = \sum_{i=1}^{4} N_i x_i \tag{3.13}$$

$$y = \sum_{i=1}^{4} N_i y_i \tag{3.14}$$

gdzie: N_i — funkcje kształtu dla współrzędnych lokalnych
$x_i y_i$ — współrzędne wierzchołków elementów.

Dla rozpatrywanego przez nas elementu funkcje kształtu wyrażają się następująco

$$N_1 = \frac{1}{4}(1-\xi)(1-\eta)$$

$$N_2 = \frac{1}{4}(1+\xi)(1-\eta) \tag{3.15}$$

$$N_3 = \frac{1}{4}(1+\xi)(1+\eta)$$

$$N_4 = \frac{1}{4}(1-\xi)(1+\eta)$$

Wykorzystując koncepcję metody elementów skończonych możemy wyrażenia (3.11) i (3.12) zapisać w postaci

$$B_x(x_p, y_p) = -\sum_e B^e_{x1}(x_p, y_p) + \sum_e B^e_{x2}(x_p, y_p) + \sum_e B^e_{x3}(x_p, y_p) - \sum_e B^e_{x4}(x_p, y_p) \tag{3.16}$$

$$B_y(x_p, y_p) = \sum_e B^e_{y1}(x_p, y_p) - \sum_e B^e_{y2}(x_p, y_p) - \sum_e B^e_{y3}(x_p, y_p) + \sum_e B^e_{y4}(x_p, y_p) \tag{3.17}$$

gdzie: $\sum_e$ — oznacza sumowanie po wszystkich elementach przekroju cewki leżących w jednej ćwiartce przyjętego układu współrzędnych

B^e_{xi}, B^e_{yi} — są wartościami składowych wektora indukcji magnetycznej w wybranym punkcie x_p, y_p pochodzącymi od e-tego elementu cewki wzbudzającej pole.

Poniżej przedstawimy postać wyrażeń na B^e_{xi} i B^e_{yi} na przykładzie B^e_{x1} i B^e_{y1}.

$$B^e_{x1}(x_p, y_p) = \int_{s^e} \frac{\mu_0 JL}{4\pi} \frac{\sum_{i=1}^{4} N_i y_i - y_p}{M(x_p, y_p)} \, \mathrm{d}s^e \tag{3.18}$$

$$B^e_{y1}(x_p, y_p) = \int_{s^e} \frac{\mu_0 JL}{4\pi} \frac{\sum_{i=1}^{4} N_i x_i - x_p}{M(x_p, y_p)} \, \mathrm{d}s^e \tag{3.19}$$

gdzie: $M(x_p, y_p) = \left[\left(\sum_{i=1}^{4} N_i x_i - x_p\right)^2 + \left(\sum_{i=1}^{4} N_i y_i - y_p\right)^2\right] \times$

$$\sqrt{0.25 \cdot \mathrm{L}^2 + \left(\sum_{i=1}^{4} N_i x_i - x_p\right)^2 + \left(\sum_{i=1}^{4} N_i y_i - y_p\right)^2}. \tag{3.20}$$

Wykorzystując na terenie elementu całkowanie numeryczne można całki w zależnościach (3.18) i (3.19) zapisać w postaci:

$$B^e_{x1}(x_p, y_p) = \frac{\mu_0 JL}{4\pi} \sum_{j=1}^{l} \frac{\sum_{i=1}^{4} N_i y_i - y_p}{M(x_p, y_p)} w_j \det[G] \tag{3.21}$$

$$B^e_{y1} = \frac{\mu_0 JL}{4\pi} \sum_{i=1}^{l} \frac{\sum_{i=1}^{4} N_i x_i - x_p}{M(x_p, y_p)} w_j \det[G] \tag{3.22}$$

W wyrażeniach (3.21) i (3.22) wskaźnik j oznacza sumowanie po przyjętych punktach całkowania, a funkcje N_i i $M(x_p, y_p)$ wyznaczone są dla lokalnych współrzędnych punktów całkowania ξ_j i η_j. w_j oznacza wartości funkcji wagi dla przyjętej metody całkowania, natomiast $[G]$ jest macierzą tansformacji geometrii wybranego elementu do kwadratu o wierzchołkach $(-1, -1)$, $(1, -1)$, $(1, 1)$, $(-1, 1)$ w lokalnym układzie współrzędnych ξ, η. Macierz $[G]$ wyraża się następująco

$$[G] = \begin{bmatrix} \frac{\partial x}{\partial \xi} & \frac{\partial x}{\partial \eta} \\ \frac{\partial y}{\partial \xi} & \frac{\partial y}{\partial \eta} \end{bmatrix} \tag{3.23}$$

Przedstawione powyżej rozważania pozwalają numerycznie wyznaczyć $B_x(x_i, y_i)$ oraz $B_y(x_i, y_i)$ w wyrażeniu (3.4), a także na tej podstawie można określić odpowiednie pochodne [3]

$\frac{\partial B_x(x_i, y_i)\partial y}{\partial y_j}$ i $\frac{\partial B_y(x_i, y_i)}{\partial y_j}$. Są dwa sposoby wyznaczania występujących w zależności pochodnych. Pierwszy z nich korzysta z aproksymacji pochodnej przez iloraz różnicowy dając wyrażenie o postaci:

$$\frac{\partial B_x(x_i, y_i)}{\partial y_j} = \frac{B_x[(x_i, y_i)(y_j+\Delta y)] - B_x[(x_i, y_i), y_j]}{\Delta y}, \tag{3.24}$$

$$\frac{\partial B_y(x_i, y_i)}{\partial y_j} = \frac{B_y[(x_i, y_i), y_j+\Delta y)] - B_y[(x_i, y_i), y_j]}{\Delta y}. \tag{3.25}$$

Drugi sposób polega na wyprowadzeniu zależności na odpowiednie pochodne przez różniczkowanie wyrażeń na B_x i B_y względem zmiennych y_j. Poniżej przedstawiono wyrażenie na jeden ze składników zależności (3.24) uzyskane na drodze różniczkowania numerycznej postaci wzoru (3.21)

$$\frac{\partial B_x^e(x_i, y_i)}{\partial y_j} = \frac{\mu_0 JL}{4\pi} \sum_{k=1}^{I} W_n \left[\frac{N_j M - \left(\sum_{k=1}^{4} N_k y_k - y_i\right) M'}{M^2} \det[G] + \right.$$

$$\left. + \frac{\sum_{k=1}^{4} M_k y_k - y_i}{M} \frac{\partial[\det[G]]}{\partial y_j} \right], \tag{3.26}$$

gdzie:

$$M' = \frac{\partial M}{\partial y_j} = 2\left(\sum_{k=1}^{4} N_k y_k - y_i\right) N_j \cdot \sqrt{0{,}25 \cdot L^2 + \left(\sum_{k=1}^{4} N_k y_k - y_i\right)^2 + \left(\sum_{k=1}^{4} N_k y_k - y_i\right)^2} +$$

$$+ \left[\left(\sum_{k=1}^{4} N_k x_k - x_i\right)^2 + \left(\sum_{k=1}^{4} N_k y_k - y_i\right)^2\right] \frac{\left(\sum_{k=1}^{4} N_k y_k - y_i\right) N_j}{\sqrt{0.25 \cdot L^2 + \left(\sum_{k=1}^{4} N_k x_k - x_i\right)^2 + \left(\sum_{k=1}^{4} N_k y_k - y_i\right)^2}} \tag{3.27}$$

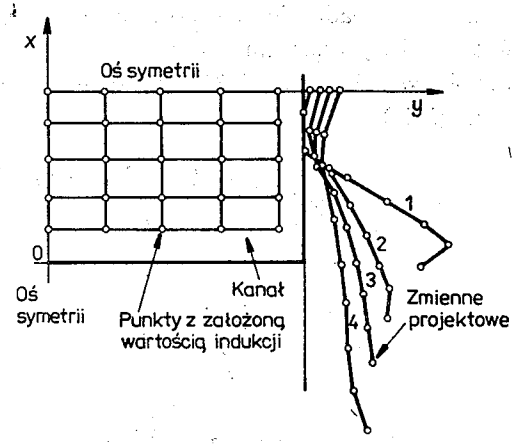
oraz

$$\frac{[\partial \det[G]]}{\partial y_j} = \left(\sum_{k=1}^{4} \frac{\partial N_k}{\partial \xi} x_k\right) \cdot \frac{\partial N_j}{\partial \eta} - \left(\sum_{k=1}^{4} \frac{\partial N_k}{\partial \eta} x_k\right) \cdot \frac{\partial N_j}{\partial \xi} \tag{3.28}$$

W wyrażeniach (3.26)—(3.28) N_j oznacza funkcję kształtu odpowiadającą numeracji lokalnej j-tego węzła w danym elemencie. Analizując przedstawione rozważania można stwierdzić, że wybór metody różniczkowania pomiędzy zastosowaniem perturbacji numerycznej a wykorzystaniem wyprowadzonych zależności numerycznych wiąże się z wyborem pomiędzy większą ilością obliczeń wektora indukcji według prostszych zależności lub wykorzystaniem do obliczeń jednorazowych bardzo złożonej postaci wzorów.

4. WYNIKI OPTYMALIZACJI

$\frac{1}{4}$ przekroju cewki podzielono na osiem elementów skończonych o dziewięciu węzłach, dla których współrzędna y była wielkością projektową. Przyjęto 25 punktów wewnątrz $\frac{1}{4}$ przekroju kanału, w których wartość indukcji zadanej wynosiła $B_{0x} = -0{,}02$ T i $B_{0y} = 0$T. Wymiary przekroju kanału wynosiły $0{,}1 \times 0{,}15$ m. Rezultaty obliczeń dla cewek o różnych wymiarach na oś 0x przedstawiono na rys. 3. Zadanie to rozwiązano przyjmując tylko ograniczenie z dołu na zmienną y.



Rys. 3. Rezultaty obliczeń dla cewki o różnych wymiarach na osi 0x

Współczynnik równomierności rozkładu wektora indukcji w kanale zdefiniowany jak następuje:

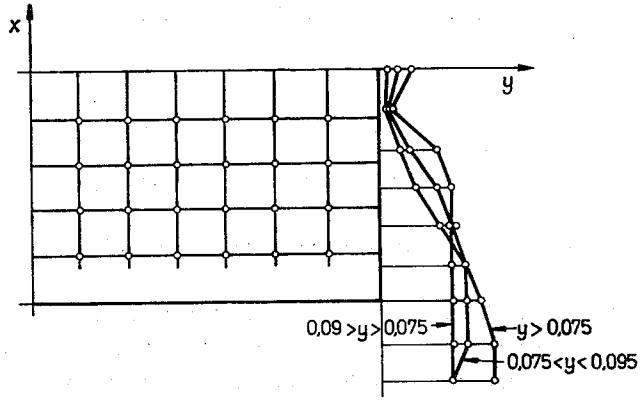
$$\eta = \sqrt{\frac{\sum_{i=1}^{m} [(B_{xi} - B_{0xi})^2 + (B_{y_j} - B_{0yi})^2]}{\sum_{i=1}^{m} (B_{0xi}^2 + B_{0y_i}^2)}} \, 100\% \tag{4.1}$$

dla odpowiednich przypadków wynosi

1 – 11,7%
2 – 7,7%
3 – 5,2%
4 – 2,4%.

Ze względów konstrukcyjno-ekonomicznych do budowy przepływomierza wybrano kształt przekroju cewki nr 2.

Kolejne obliczenia zostały wykonane dla ustalonego wymiaru na oś 0x cewki wzbudzającej pole magnetyczne, przy przyjęciu za zmienne projektowe ograniczenia z dołu oraz różnych wartości ograniczenia z góry. Ograniczenia te wynikły z przyjętych maksymalnych rozmiarów przetwornika pierwotnego. Na rys. 4 przedstawiono rezultaty tych obliczeń.



Rys. 4. Rezultaty obliczeń kształtu cewki dla różnych wartości ograniczeń

Przedstawione wyniki obliczeń dla tego przypadku realizują założone parametry pola magnetycznego z różną dokładnością. Współczynnik równomierności dla odpowiednich przypadków tego zadania wynosi:

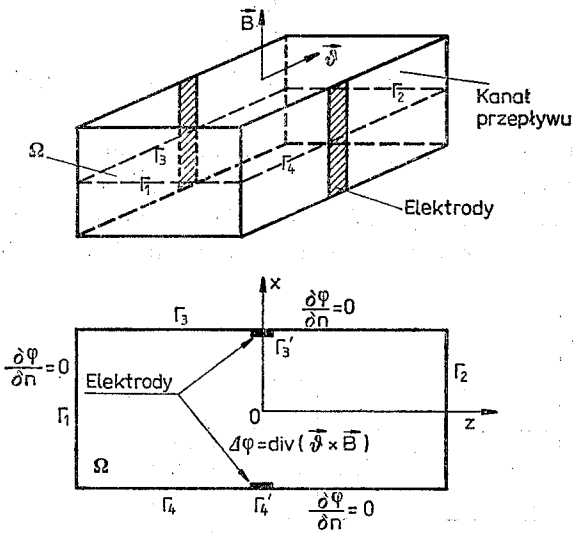
$$1 - 7{,}7\%;\ 2 - 9{,}0\%;\ 3 - 10{,}3\%.$$

Z powyższych obliczeń wynika, że przyjęcie krótszej cewki lub ostrzejszych ograniczeń na zmienne projektowe powodują, że realizujemy zadany rozkład pola magnetycznego w kanale z mniejszą dokładnością.

5. METODA WYZNACZANIA DŁUGOŚCI „L" CEWKI WZBUDZAJĄCEJ

Długość cewki wzbudzającej określona jest przy wykorzystaniu analizy efektów brzegowych występujących w strefie szybkich zaników pola magnetycznego. Efekty te noszą nazwę zwarciowych [4]. W strefach granicznych indukowane jest zakłócające pole elektryczne, które powoduje zmniejszenie sygnału użytecznego indukowanego na elektrodach pomiarowych. Eliminacja efektu zwarciowego polega na odsunięciu stref dużych zaników

pola magnetycznego, przez zwiększenie długości cewki „L", od strefypomiarowej. Strefa pomiarowa wyznaczona jest przez dwie elektrody. Schemat ideowy modelu, który był wykorzystywany do oceny ilościowej efektów zwarciowych przedstawiony jest na rys. 5.



Rys. 5. Schemat ideowy modelu wykorzystywanego przy ocenie ilościowej efektów zwarciowych

Rozwiązując równanie Poissone'a (2.1) w płaszczyźnie A umieszczonej w połowie głębokości kanału pomiarowego, uzyskano selektywną zależność indukowanej różnicy potencjałów od długości cewki przy założeniu stałej wartości prądu w niej płynącego. Pole magnetyczne wzbudzane było przez układ cewki, której przekrój poprzeczny podlegał optymalizacji. Dla przedstawionego problemu rozwiązanie równania Poissona (2.1) przy wykorzystaniu metody elementów skończonych, sprowadza się do rozwiązania algebraicznego układu równań

$$[\boldsymbol{H}]\{\varphi\} = \{\boldsymbol{F}\}. \tag{5.1}$$

Po przyjęciu dyskretyzacji obszaru na elementy trójkątne o liniowej aproksymacji rozkładu potencjału na terenie elementu zależności pozwalające wyznaczyć macierz $[H]$ są powszechnie znane w literaturze [5], natomiast osobnych rozważań wymaga ustalenie postaci wyrażeń określających elementy wektora $\{F\}$. Załóżmy, że znamy wartości wektora indukcji magnetycznej oraz prędkości cieczy w obszarze kanału i przyjmujemy, że są one stałe na terenie elementu.

Wyznaczenie składowych wektora indukcji w środkach ciężkości trójkątów obszaru przeprowadzamy podobnie jak to miało miejsce przy optymalizacji kształtu przekroju cewki.

Przyjęty rodzaj elementów skończonych, na które podzielono przekrój kanału narzuca sposób wyznaczenia wartości funkcji źródła na terenie elementu. Jak wynika z równania (2.1) funkcja źródła ma postać

$$Q = \operatorname{div}(\vec{v} \times \vec{B}). \tag{5.2}$$

Dla skrócenia zapisu do dalszych rozważań przyjmiemy

$$\vec{A} = \vec{v} \times \vec{B}. \tag{5.3}$$

W celu wyznaczenia na terenie każdego z elementów funkcji źródła Q korzystamy z definicji operatora dywergencji

$$Q = \operatorname{div} \vec{A} = \lim_{\Delta V \to 0} \frac{\oint_s \vec{A} \cdot \overrightarrow{\mathrm{ds}}}{\Delta V}. \tag{5.4}$$

Dla obszaru płaskiego zależność (5.4) przyjmie postać

$$Q = \lim_{\Delta s \to 0} \frac{\oint_L A_n \mathrm{dl}}{\Delta s} \tag{5.5}$$

Korzystając z metody elementów skończonych możemy i-ty składnik wektora $\{F\}$ pochodzący od e-tego elementu wyrazić jako

$$F_i^e = \int_{S^e} N_i Q \mathrm{ds}. \tag{5.6}$$

Wykonując całkowanie w wyrażeniu (5.5) wzdłuż boków elementu oraz całkowanie po powierzchni elementu w wyrażeniu (5.6) otrzymamy

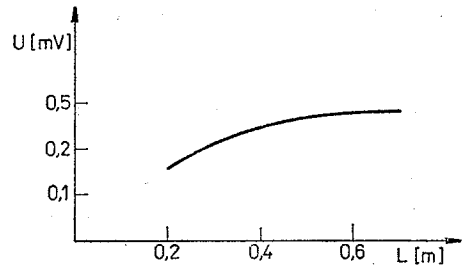
$$F_i^e = \frac{1}{2}[A_x(y_k - y_j) + A_y(x_k - x_j)], \tag{5.7}$$

gdzie: (x_j, y_j) i (x_k, y_k) współrzędne wierzchołków trójkąta (elementu) zawierającego i-ty węzeł.

Rozciągając sumowanie na wszystkie elementy przyległe do i-tego węzła otrzymamy i-ty element wektora $\{F\}$ o postaci

$$F_i = \sum_{e=1}^{\mathrm{NE}} F_i^e, \tag{5.8}$$

gdzie NE jest liczbą elementów obszaru. Zależność indukowanej różnicy potencjałów od długości cewki, przy założeniu stałej gęstości prądu w niej płynącego, pokazano na rys. 6.



Rys. 6. Zależność indukowanej różnicy potencjałów U od długości cewki L, przy stałej gęstości prądu w niej płynącego

Z przedstawionych wyników można wnioskować, że optymalnym rozwiązaniem byłoby przyjęcie odpowiednio długiej cewki, bo wtedy możemy uzyskać możliwie duży sygnał pomiarowy. Rzeczywisty układ pomiarowy nie jest zasilany ze źródła prądu, jak to oznaczono powyżej, a ze źródła napięcia E = 24 V. Zatem przedstawione na rys. 6 wyniki należy skorygować uwzględniając rezystancję uzwojeń cewki. Sposób przeliczenia omawianych wyników dla zasilania układu pomiarowego ze źródła napięcia przedstawimy poniżej.

Dla określonej długości l i danego kształtu przekroju poprzecznego cewki wzbudzającej pole magnetyczne napięcie na elektrodach pomiarowych proporcjonalne jest do prądu płynącego w cewce. A zatem dla wymuszenia w postaci źródła prądu otrzymamy

$$U_1 = k \cdot I \cdot z \tag{5.9}$$

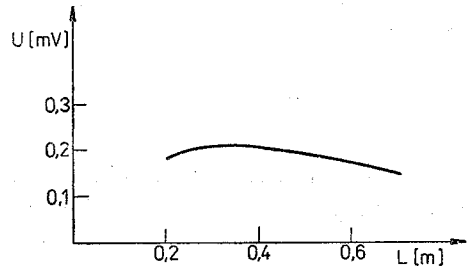
i dla wymuszenia w postaci źródła napięcia także otrzymamy

$$U_2 = k \cdot \frac{E}{R(L)} z. \tag{5.10}$$

Podstawiając $k \cdot z$ z wyrażenia (5.9) do (5.10) uzyskamy zależność określającą napięcie indukowane na elektrodach pomiarowych dla określonej długości cewki, przy wymuszeniu napięciowym, o postaci

$$U_2 = \frac{E}{I} \frac{u_1}{R(L)} \tag{5.11}$$

Wykorzystując zależność (5.11) możemy przeskalować charakterystykę przedstawioną na rys. 6 i otrzymamy nowy kształt zależności napięcia indukowanego na elektrodach pomiarowych od długości cewki. Zależność ta przedstawiona została na rys. 7. W tym

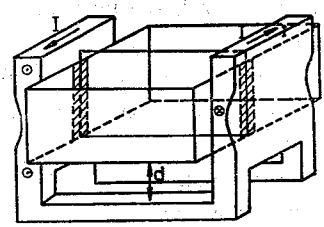


Rys. 7. Zależność indukowanej różnicy potencjałów U od długości cewki L, przy uwzględnieniu rzeczywistej rezystancji cewki

przypadku widzimy wyraźne ekstremum przedstawionej funkcji dla długości cewki około 0.2—0.3 m.

Jak wynika z przedstawionych rozważań przyjęcie rzeczywistych warunków wymuszenia napięciowego, w porównaniu dla przyjętego do obliczeń metodą elementów skończonych wymuszenia prądowego, zmienia zdecydowanie kształt krzywej uzależniającej wartości sygnału pomiarowego od długości cewki i pozwala ustalić kryterium doboru długości cewki dającej maksymalny sygnał pomiarowy. Kolejnym etapem urealnienia

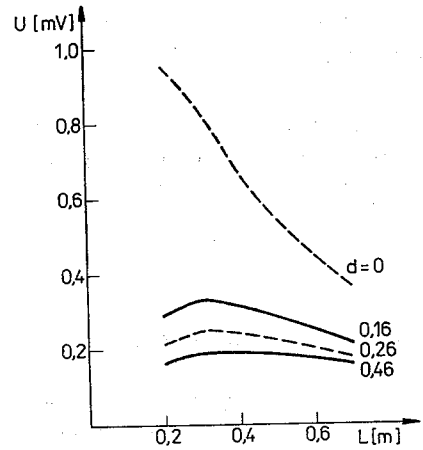
wyników obliczeń (projektowania ostatecznego kształtu cewki) jest uwzględnienie połączenia obu części cewki przedstawionej na rys. 1. Pełny kształt projektowanej cewki pokazano na rys. 8.



Rys. 8. Schemat ideowy projektowanego układu cewki wzbudzającej

Uwzględniając ten fakt, w obliczeniach wartości napięcia indukowanego na elektrodach (przy niezmienionym kształcie przekroju poprzecznego części głównych cewki), koniecznym staje się dodanie do wyznaczonej funkcji źródła pola Q składowej wektora indukcji pochodzącej od odcinków uzwojeń ułożonych poprzecznie do kierunku przepływu cieczy w kanale.

Wyniki obliczeń przeprowadzono dla kilku przypadków usytuowania części poprzecznych cewki względem dna kanału chcąc przekonać się o wpływie nowych elementów analizowanego modelu na wartość indukowanego sygnału pomiarowego. Wyniki tych obliczeń dla wymuszenia napięciowego przedstawiono na rys. 9.



Rys. 9. Wpływ usytuowania poprzecznego połączenia pomiędzy bokami cewki na amplitudę indukowanej różnicy potencjałów U (d — odległość połączeń od dna kanału)

Przedstawione powyżej rezultaty obliczeń potwierdzają występowanie ekstremum badanego sygnału dla określonej długości cewki pomiarowej, a również wskazują na duży wpływ uwzględnionej obecnie części cewki na wartość napięcia indukowanego na elektrodach w przypadku, gdy jest ona bezpośrednio usytuowana pod dnem kanału.

Potwierdzają się tu jednocześnie fakty oczywiste, że opuszczenie części łączącej cewki poniżej dna kanału, a także zwiększanie długości cewki zmniejsza wpływ oddziaływania

poprzecznych odcinków cewki. Ze względu na ograniczenia konstrukcyjne konieczne jest usytuowanie odcinków poprzecznych łączących oba boki cewki bezpośrednio pod dnem kanału. Zbytnie oddalenie połączeń od kanału przepływu uczyniłoby konstrukcję cewki zbyt rozbudowaną przestrzennie, co w praktyce uniemożliwiłoby jej instalację na rzeczywistym kanale. Przyjmując omawiane wcześniej kryterium maksymalnej wartości sygnału pomiarowego wydawałoby się, że oddziaływanie połączeń poprzecznych sprzyja założeniom projektowym. W rzeczywistości jednak może okazać się, że uzyskany rozkład indukcji pola magnetycznego nie spełnia warunku $B = const$. Niespełnienie tego warunku z zadaną dokładnością może przyczynić się do powstania nadmiernych błędów pomiaru przepływu przy występowaniu zmiennego rozkładu prędkości. Dopiero przeprowadzenie badań laboratoryjnych przepływomierza z cewką wykonaną wg projektu opracowanego zgodnie z zaproponowaną metodą pozwoli na pełną ocenę błędu pomiaru. Tym niemniej wydaje się, że zaproponowaną metodę projektowania cewek do przepływomierzy należałoby rozszerzyć o analizę i syntezę zjawisk przestrzennych występujących w rzeczywistych układach cewek. Zagadnienia te będą tematem dalszych prac.

6. PODSUMOWANIE

W pracy przedstawiono zastosowanie uogólnionej metody Newtona do syntezy kształtu przekroju cewki przepływomierza elektromagnetycznego. Wykorzystanie metody elementów skończonych do aproksymacji numerycznej przekroju poprzecznego cewki pozwoliło przyjąć zmienne projektowe w postaci współrzędnych (na osi 0y) wybranych węzłów dyskretyzacji, a także umożliwiło wyprowadzenie zależności na elementy macierzy Jacobiego.

Otrzymane wyniki obliczeń, dla cewki o różnej długości na osi 0x i przy różnych ograniczeniach wymiarów na osi 0y, pozwalają podjąć decyzję konstrukcyjną. Wybór projektanta związany jest z rezygnacją z równomierności rozkładu pola magnetycznego w obszarze kanału na rzecz zmniejszenia wymiarów cewki.

BIBLIOGRAFIA

1. I. A. Shercliff: *The theory of electromagnetic flow measurement.* Cambridge University Press 1962
2. M. K. Bevir: *The theory of Electromagnetic Flowmeter.* J. Fluid Mech. nr 43 1970
3. S. Wincenciak: *Wyznaczanie macierzy Jacobiego dla syntezy kształtu obszaru.* X Seminarium z Podstaw Elektrotechniki i Teorii Obwodów Gliwice—Wisła 1987 ss. 479—486
4. A. Michalski: *Analiza głównych źródeł zakłóceń występujących przy pomiarach przepływu w ciekach naturalnych metodą elektromagnetyczną.* PAK, 1987 nr 3, ss. 52—54
5. O. C. Zienkiewicz: *Metoda elementów skończonych.* Arkady, Warszawa 1972
6. A. Michalski, J. Sikora, St. Wincenciak: *Optimal shape design of an electromagnetic flow gauge.* IEE Trans. on Magnetics, 1988 January, vol 24, no 1, pp. 565—568
7. A. Michalski., J. Sikora, St. Wincenciak: *Optimal shape design of an electric coil.* Cavtat/Dubrownik—Jugosławia IX Int. Symp. „Computer at the University" 18—22 May 1987
8. A. Michalski: *Wektor wagi w zagadnieniu pomiaru metodą elektromagnetyczną średniej prędkości wody w kanałach otwartych.* Rozprawy Elektrotechniczne 1987, 33 z. 1, ss. 219—228

A. MICHALSKI, J. SIKORA, S. WINCENCIAK

ALGORITHM FOR DESIGNING THE OPTIMUM SHAPE OF THE EXCITING COIL FOR AN ELECTROMAGNETIC FLOWMETER

Summary

The algorithm for designing the optimum shape of the exciting coil of the primary transducer of an electromagnetic flowmeter, is presented. Basing on the finite element idea, the crossection of the coil is discretized. Some of the nodes of the, thus, obtained finite element, are admitted as design variables. The generalized nonlinear Newton algorithm is applied for minimization of the objective function. The length of the coil is calculated by means of the End-Shorting idea.

A. MICHALSKI, J. SIKORA, S. WINCENCIAK

ALGORITHME POUR L'ÉTABLISSEMENT DU PROJET D'UNE BOBINE INDUCTRICE DU DÉBITMÈTRE ÉLECTROMAGNÉTIQUE

Résumé

L'article présente l'algorithme de la réalisation de l'optimal projet de la forme d'une bobine inductrice du transducteur d'un débitmètre électromagnétique. En se servant de la méthode des éléments finis on a divisé la section droite de la bobine en éléments finis. Pour variables de projet on a pris les noeuds des éléments finis, situés sur un des bords de la section droite de la bobine. Pour minimiser la fonction de but on a appliqué le non-linéaire algorithme généralisé de Newton. La longueur de la bobine a été déterminée à partir de la connaissance du phénomène des effets du court-circuit.

A. MICHALSKI, J. SIKORA, S. WINCENCIAK

METHODE FÜR DIE OPTIMALE PROJEKTIERUNG DER FORM EINER WANDLERERREGELSPULE FÜR ELEKTROMAGNETISCHEN DURCHLAßVOLTMETERS

Zusammenfassung

In diesem Artikel wurde der Algoritmus für die optimale Projektierung der Formeiner Wandlererregelspule des primären elektromagnetischen Durchlaßvoltmeters geschildert. Unter Nutzung der Methode beendeter Elemente wurde der Spulenquerschnitt in beendete Elemente eingeteilt. Als Projektierungsveränderliche wurden Knoten der beendeten Elemente angenommen, die an einem der Querschnittränder der Spule liegen. Für die Minimalisierung der Zielfunktion wurde der verallgemeinerte nichtlineare Newton-Algorithmus angewandt. Die Spulenlänge wurde auf Grund des Phänomens des Kurzschlußeffektes bestimmt.

А. МИХАЛЬСКИ, Я. СИКОРА, С. ВИНЦЕНЦЯК

МЕТОД ПРОЕКТИРОВАНИЯ ОПТИМАЛЬНОЙ ВОЗБУЖДАЮЩЕЙ КАТУЩКИ ЭЛЕКТРОМАГНИТНЫХ РАСХОДОМЕРОВ ДЛЯ ОТКРЫТЫХ КАНАЛОВ

Резюме

Представлен алгоритм оптимального проектирования возбуждающей катушки первичного преобразователя электромагнитного расходомера. Используя метод конечных элементов разделено поперечное сечение катушки на конечные элементы. В качестве переменных проектирования приняты узлы конечных элементов расположенных на одном краю поперечного сечения катушки. Для минимизации функции цели применен обобщенный нелинейный алгоритм Ньютона. Длина катушки определена при использовании явления эффектов короткого замыкания.

ROZPRAWY ELEKTROTECHNICZNE, 1989, 35, z. 1, ss. 77-88

537.226.1:534.414

Zastosowanie cylindrycznych rezonatorów quasi TE_{0nm} i quasi TM_{010} do pomiaru zespolonej przenikalności elektrycznej cylindrycznych próbek dielektryków o dowolnych wymiarach

ANNA HARASIEWICZ, JERZY KRUPKA

Instytut Mikro- i Optoelektroniki, Politechnika Warszawska

Otrzymano 1988.01.05

Autoryzowano do druku 1988.05.06

Praca dotyczy metody pomiaru w paśmie mikrofalowym przenikalności elektrycznej i stratności próbek dielektryka przy zastosowaniu cylindrycznych rezonatorów quasi TE_{0nm} oraz quasi TM_{010}. Badaną próbkę w kształcie cylindra umieszcza się centrycznie na denku rezonatora. Wymiary (średnica i wysokość) próbki mogą być dowolne, tzn. nie muszą być dopasowane do wymiarów rezonatora. Rozwiązania zagadnienia brzegowego równań Helmholtza dokonano przybliżoną metodą sklejania podobszarów. Problem sprowadza się do rozwiązania układu N jednorodnych równań liniowych. Z obliczeń numerycznych wynika, że metoda jest szybkozbieżna, tzn. dla $N = 9$ błędy spowodowane niedokładnością metody są mniejsze od błędów pomiarowych. Dobroć rezonatora obliczono metodą rachunku zaburzeń. Metoda ta może być zastosowana w odniesieniu do rezonatora quasi TE_{0nm}.

1. WSTĘP

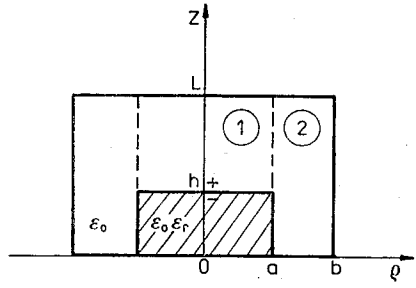
Rezonatory cylindryczne typu TE_{0nm} i TM_{010} są jednymi z częściej stosowanych rezonatorów do pomiaru zespolonej przenikalności elektrycznej dielektryków [1÷4]. Mierzone próbki mają zazwyczaj kształt cylindrów, których jeden z wymiarów całkowicie wypełnia wysokość lub przekrój poprzeczny rezonatora. Ograniczenie to związane jest z możliwością uzyskania w takich wypadkach ścisłych rozwiązań (metodą rozdzielenia zmiennych) zagadnienia brzegowego dla równań Helmholtza w obszarze rezonatora z próbką. Jeśli obydwa wymiary próbki tzn. średnica i wysokość są mniejsze od odpowiednich wymiarów rezonatora, zagadnienie brzegowe można rozwiązać jedynie metodami przybliżonymi.

Istnieje kilka metod przybliżonych umożliwiających uzyskanie dużej dokładności rozwiązania tego problemu, między innymi: zmodyfikowana metoda Rayleigha-Ritza [5], metoda elementów skończonych [6], metoda różnic skończonych [7], czy metoda sklejenia podobszarów (the mode metching method) [8÷9]. Ostatnia z wymienionych metod charakteryzuje się szybką zbieżnością, w związku z tym istnieje możliwość przeprowadzenia obliczeń numerycznych na mikrokomputerze o małej szybkości obliczeniowej. Z tego względu zostanie ona zastosowana w tej pracy.

2. TEORIA

2.1. OBLICZENIA PULSACJI REZONANSOWEJ REZONATORÓW

Ze względów praktycznych wygodnie jest mierzyć cylindryczne próbki dielektryków umieszczając je na jednym z denek rezonatora, jak pokazano na rys. 1.



Rys. 1. Rezonator cylindryczny z próbką dielektryka umieszczoną na jego denku

W metodzie sklejania podobszarów dzielimy obszar rezonatora na podobszary, w których można znaleźć rozwiązania elementarne (rozwiązania równań Maxwella spełniające wszędzie, z wyjątkiem powierzchni podobszarów, odpowiednie warunki brzegowe). Z rozwiązań elementarnych tworzymy w podobszarach rozwiązanie w postaci szeregów. Następnie żądamy spełnienia odpowiednich warunków brzegowych przez składowe pola elektromagnetycznego styczne na powierzchni rozdziału podobszarów, co prowadzi do równania charakterystycznego.

Dla rezonatorów analizowanych w tym artykule podziału na podobszary dokonano powierzchnią $\varrho = a$, jak zaznaczono na rys. 1.

W celu wyznaczenia równań elementarnych w podobszarach: pierwszym ($0 \leqslant \varrho \leqslant a$) i drugim ($a \leqslant \varrho \leqslant b$) zastosujemy metodę potencjałów wektorowych [10]. Zgodnie z tą metodą wszystkie składowe pól elektromagnetycznych (stanowiące rozwiązania elementarne) mogą być wyznaczone poprzez operacje różniczkowe na funkcjach ψ i $\overline{\psi}$.

Funkcje skalarne ψ i $\overline{\psi}$ spełniają w każdym z podobszarów równania Helmholtza:

$$\begin{aligned} &\nabla^2\psi_\alpha + k^2\psi_\alpha = 0, \\ &\nabla^2\overline{\psi}_\alpha + k^2\overline{\psi}_\alpha = 0, \end{aligned} \qquad \alpha = 1, 2 - \text{numer podobszaru} \tag{1}$$

gdzie:

$$k^2 = \begin{cases} \omega^2/c^2 \cdot \varepsilon_r & \text{w obszarze } 0 \leqslant z \leqslant h \cap 0 \leqslant \varrho \leqslant a \\ \omega^2/c^2 & \text{w pozostałej części rezonatora} \end{cases} \tag{2}$$

Wektory pól elektromagnetycznych można wyrazić poprzez funkcje ψ i $\overline{\psi}$ w następujący sposób [10]:

$$E_z = \frac{1!}{j\omega\varepsilon}\left(k^2 + \frac{\partial^2}{\partial z^2}\right)\psi, \tag{3}$$

$$E_\varrho = \frac{1}{j\omega\varepsilon}\frac{\partial^2\psi}{\partial\varrho\,\partial z} - \frac{1}{\varrho}\frac{\partial\overline{\psi}}{\partial\varphi}, \tag{4}$$

$$E_\varphi = \frac{1}{j\omega\varepsilon\varrho}\frac{\partial^2\psi}{\partial\varphi\,\partial z} + \frac{\partial\bar{\psi}}{\partial\varrho}, \tag{5}$$

$$H_z = \frac{1}{j\omega\mu}\left(k^2 + \frac{\partial^2}{\partial z^2}\right)\bar{\psi}, \tag{6}$$

$$H_\varrho = \frac{1}{j\omega\mu}\frac{\partial^2\bar{\psi}}{\partial\varrho\,\partial z} + \frac{1}{\varrho}\frac{\partial\psi}{\partial\varphi}, \tag{7}$$

$$H_\varphi = \frac{1}{j\omega\mu\varrho}\frac{\partial^2\bar{\psi}}{\partial\varphi\,\partial z} - \frac{\partial\psi}{\partial\varrho}. \tag{8}$$

Dla struktury rezonatora przedstawionej na rys. 1, w wypadku pól o symetrii osiowej, mogą istnieć rodzaje quasi TE_{0nm} zależne tylko od funkcji $\bar{\psi}$ i rodzaje quasi TM_{0nm} zależne tylko od funkcji ψ. Na powierzchniach granicznych rezonatora pola elektromagnetyczne spełniają odpowiednie warunki brzegowe (ciągłość składowych stycznych pól elektromagnetycznych na granicy rozdziału dielektryków oraz znikanie składowej stycznej pola elektrycznego na granicy z idealnym przewodnikiem). W konsekwencji funkcje $\bar{\psi}$ i ψ muszą także spełniać odpowiednie warunki. Są one następujące:

Dla rodzajów quasi TE_{0nm}:

$$\bar{\psi}_\alpha(z = 0) = \bar{\psi}_\alpha(z = L) = 0, \quad \alpha = 1,2 \tag{9}$$

$$\frac{\partial\bar{\psi}_2}{\partial\varrho}(\varrho = b) = 0, \tag{10}$$

$$\bar{\psi}_1(z = h_+) = \bar{\psi}_1(z = h_-), \tag{11}$$

$$\frac{\partial\bar{\psi}_1}{\partial z}(z = h_+) = \frac{\partial\bar{\psi}_1}{\partial z}(z = h_-), \tag{12}$$

$$\left(k^2 + \frac{\partial^2}{\partial z^2}\right)\bar{\psi}_1(\varrho = a) = \left(k^2 + \frac{\partial^2}{\partial z^2}\right)\bar{\psi}_2(\varrho = a), \tag{13}$$

$$\frac{\partial\bar{\psi}_1}{\partial\varrho}(\varrho = a) = \frac{\partial\bar{\psi}_2}{\partial\varrho}(\varrho = a). \tag{14}$$

Dla rodzajów quasi TM_{0nm}:

$$\frac{\partial\psi_\alpha}{\partial z}(z = 0) = \frac{\partial\psi_\alpha}{\partial z}(z = L) = 0, \quad \alpha = 1, 2 \tag{15}$$

$$\psi_2(\varrho = b) = 0, \tag{16}$$

$$\frac{1}{\varepsilon_r}\frac{\partial\psi_1}{\partial z}(z = h_-) = \frac{\partial\psi_1}{\partial z}(z = h_+), \tag{17}$$

$$\psi_1(z = h_-) = \psi_1(z = h_+), \tag{18}$$

$$\frac{1}{\varepsilon_r(z)}\left(k^2 + \frac{\partial^2}{\partial z^2}\right)\psi_1(\varrho = a) = \left(k^2 + \frac{\partial^2}{\partial z^2}\right)\psi_2(\varrho = a), \tag{19}$$

$$\frac{\partial\psi_1}{\partial\varrho}(\varrho = a) = \frac{\partial\psi_2}{\partial\varrho}(\varrho = a). \tag{20}$$

W dalszym ciągu wyprowadzimy jedynie równanie charakterystyczne dla rodzajów quasi TE_{0nm}. (Zależności dla rodzajów quasi TM_{0nm} wyprowadza się analogicznie). Rozwiązania elementarne spełniające równanie (1) i warunki brzegowe (9÷12) są następujące:

$$\bar{\psi}_{1n} = R_{1n}(\varrho) \cdot Z_{1n}(z), \tag{21}$$

$$\bar{\psi}_{2m} = R_{2m}(\varrho) \cdot Z_{2m}(z), \tag{22}$$

gdzie:

$$R_{1n}(\varrho) = \begin{cases} J_0(k_{\varrho n}^{(1)}\varrho), & \text{dla} \quad (k_{\varrho n}^{(1)})^2 > 0 \\ I_0(|k_{\varrho n}^{(1)}|\varrho), & \text{dla} \quad (k_{\varrho n}^{(1)})^2 < 0 \end{cases}, \tag{23}$$

$$R_{2m}(\varrho) = \begin{cases} J_0(k_{\varrho m}^{(2)}\varrho) \cdot Y_1(k_{\varrho m}^{(2)}b) - Y_0(k_{\varrho m}^{(2)}\varrho) \cdot J_1(k_{\varrho m}^{(2)}b) & (k_{\varrho m}^{(2)})^2 > 0 \\ I_0(|k_{\varrho m}^{(2)}|\varrho) \cdot K_1(|k_{\varrho m}^{(2)}|b) + K_0(|k_{\varrho m}^{(2)}|\varrho) \cdot I_1(|k_{\varrho m}^{(2)}|b) & (k_{\varrho m}^{(2)})^2 < 0 \end{cases} \tag{24}$$

$$(k_{\varrho m}^{(2)})^2 = \left(\frac{\omega}{c}\right)^2 - \left(\frac{m\pi}{L}\right)^2, \tag{25}$$

$$Z_{1n}(z) = \begin{cases} \sin(k_{zn}^{(\varepsilon)} z) & \text{dla} \quad 0 \leqslant z \leqslant h \\ \sin[k_{zn}^{(0)}(L-z)] \dfrac{\sin(k_{zn}^{(\varepsilon)} h)}{\sin[k_{zn}^{(0)}(L-h)]}, & \text{dla} \quad h \leqslant z \leqslant L \end{cases} \tag{26}$$

$$Z_{2m}(z) = \sin\left(\frac{m\pi}{L} z\right). \tag{27}$$

Wielkości $k_n^{(1)}$, $k_{zn}^{(\varepsilon)}$, $k_{zn}^{(0)}$ spełniają następujący układ równań:

$$\left.\begin{aligned} &\frac{\tan(k_{zn}^{(\varepsilon)} h)}{k_{zn}^{(\varepsilon)}} + \frac{\tan[k_{zn}^{(0)}(L-h)]}{k_{zn}^{(0)}} = 0, \\ &(k_{\varrho n}^{(1)})^2 + (k_{zn}^{(\varepsilon)})^2 = \left(\frac{\omega}{c}\right)^2 \varepsilon_r, \\ &(k_{\varrho n}^{(1)})^2 + (k_{zn}^{(0)})^2 = \left(\frac{\omega}{c}\right)^2. \end{aligned}\right\} \tag{28}$$

W każdym z podobszarów przedstawiamy rozwiązanie w postaci szeregów:

$$\bar{\psi}_1 = \sum_{n=1}^{\infty} a_n \bar{\psi}_{1n}, \tag{29}$$

$$\bar{\psi}_2 = \sum_{m=1}^{\infty} b_m \bar{\psi}_{2m}. \tag{30}$$

W praktyce szeregi (29) i (30) musimy ograniczyć do sum złożonych z N pierwszych wyrazów.

$$\bar{\psi}_1 \approx \sum_{n=1}^{N} a_n \bar{\psi}_{1n}, \tag{31}$$

$$\bar{\psi}_2 \approx \sum_{m=1}^{N} b_m \bar{\psi}_{2m}. \tag{32}$$

Warunki brzegowe (13) i (14) mogą być spełnione przez skończone sumy (31) i (32) jedynie w przybliżeniu. Współczynniki rozwinięcia a_n i b_m $(n, m = 1, 2, \ldots, N)$ wyznacza się minimalizując błąd zszycia pól elektromagnetycznych na granicy $\varrho = a$. W rozważanym przypadku wygodnie jest to zrobić metodą momentów [11] przyjmując jako funkcje wagowe — funkcje $Z_{2j}(z)$. Prowadzi to do następujących zależności:

$$\sum_{n=1}^{N} a_n \cdot (k_{\varrho n}^{(1)})^2 \cdot R_{1n}(a) \cdot (Z_{1n}, Z_{2j})_L = b_j (k_{\varrho j}^{(2)})^2 \cdot R_{2j}(a) \cdot ||Z_{2j}||^2, \tag{33}$$

$$\sum_{n=1}^{N} a_n \cdot R'_{1n}(a) \cdot (Z_{1n}, Z_{2j})_L = b_j \cdot R'_{2j}(a) \cdot ||Z_{2j}||^2, \tag{34}$$

$$R'_{1n}(a) = \frac{\partial R_{1n}}{\partial \varrho}(\varrho = a), \qquad R'_{2j}(a) = \frac{\partial R_{2j}}{\partial \varrho}(\varrho = a)$$

gdzie:

$$(Z_{1n}, Z_{2j})_L \overset{\text{df}}{=} \int_0^L Z_{1n} \cdot Z_{2j}^* \, \mathrm{d}z,$$

$$||Z_{2j}|| = [(Z_{2j}^0, Z_{2j})]^{1/2}.$$

Z równań (33) i (34) otrzymuje się po prostych przekształceniach układ N jednorodnych równań liniowych względem współczynników a_n. Jest on następujący:

$$\sum_{n=1}^{N} a_n \left[\frac{(k_{\varrho n}^{(1)})^2 \cdot R_{1n}(a)}{(k_{\varrho j}^{(2)})^2 \cdot R_{2j}(a)} - \frac{R'_{1n}(a)}{R'_{2j}(a)} \right] \cdot (Z_{1n}, Z_{2j})_L = 0, \quad j = 1, 2, \ldots, N. \tag{35}$$

Układ równań (35) ma niezerowe rozwiązanie, jeśli jego wyznacznik znika, co ma miejsce dla wartości będących przybliżeniami pulsacji rodzajów quasi TE_{0nm} dla struktury rezonatora z rys. 1:

$$\det[W] = 0 \tag{36}$$

gdzie $[W]$ jest macierzą $N \times N$ o następujących elementach:

$$W_{nj} = \left[\frac{(k_{\varrho n}^{(1)})^2 \cdot R_{1n}(a)}{(k_{\varrho j}^{(2)})^2 \cdot R_{2j}(a)} - \frac{R'_{1n}(a)}{R'_{2j}(a)} \right] \cdot (Z_{1n}, Z_{2j})_L. \tag{37}$$

Zagadnienie numerycznego wyznaczania wartości własnych opisaną wyżej metodą jest dość skomplikowane. Dla każdej ustalonej wartości ω wyznacza się najpierw wartości $k_{\varrho n}^{(1)}$, $k_{zn}^{(\varepsilon)}$, $k_{zn}^{(0)}$ znajdując N pierwszych rozwiązań układu równań (28). Następnie oblicza się elementy macierzy $[W]$ i jej wyznacznik. Lokalizacja zer wyznacznika nie przedstawia zazwyczaj większych trudności numerycznych, chociaż trzeba zwracać uwagę na istnienie biegunów wyznacznika, w których występuje także zmiana jego znaku.

2.2. OBLICZENIA DOBROCI REZONATORÓW

Najbardziej ogólną metodą wyznaczania dobroci rezonatora jest metoda zespolonej pulsacji, która występuje bezpośrednio w równaniach Helmholtza [14]. Metodą tą mogą być analizowane wszystkie rodzaje strat równocześnie [15], ale wymaga ona znacz-

nych nakładów obliczeniowych. W praktyce stosuje się najczęściej przybliżone metody obliczeń dobroci, które w wypadku rezonatorów zawierających dielektryki i przewodniki dają równie dobre, jak metoda zespolonej pulsacji, rezultaty. Metody te polegają na analizie rezonatora bezstratnego prowadzącej do wyznaczenia pulsacji rezonansowej i rozkładu pól elektromagnetycznych.

Następnie dobroć oblicza się z definicji zakładając, że rozkład pól elektromagnetycznych w rezonatorze rzeczywistym jest taki sam jak w rezonatorze bezstratnym. Korzysta się przy tym z następujących zależności:

$$Q^{-1} = Q_d^{-1} + Q_c^{-1}, \tag{38}$$

$$Q_d = \frac{W_d + W_a}{W_d} \frac{1}{\tan\delta} = \frac{\int\limits_{Vd} \varepsilon_0 \varepsilon_r \cdot |\vec{E}|^2 \mathrm{dv} + \int\limits_{Va} \varepsilon_0 |\vec{E}|^2 \mathrm{dv}}{\int\limits_{Vd} \varepsilon_0 \varepsilon_r \cdot |\vec{E}|^2 \mathrm{dv}} \cdot \frac{1}{\tan\delta}, \tag{39}$$

$$Q_c = \frac{\omega \int\limits_{V} \mu_0 \cdot |\vec{H}|^2 \mathrm{dv}}{R_s \cdot \int\limits_{S} |\vec{H}_t|^2 \mathrm{ds}}, \tag{40}$$

gdzie:

V_d — objętość próbki dielektryka

V_a — objętość części rezonatora wypełnionej powietrzem

V — całkowita objętość rezonatora ($V = V_d + V_a$)

S — powierzchnia ścianek rezonatora

R_s — rezystancja ścianek rezonatora

$\vec{H}_t$ — składowa pola magnetycznego styczna do ścianek rezonatora.

Jak widać, jeśli rozkłady pól elektromagnetycznych i parametry materiałowe (R_s i $\tan\delta$) są znane, dobroć może być wyznaczona na drodze obliczenia całek występujących w zależnościach (38)÷(40). Dla obliczenia wyrażeń (39) i (40) można też zastosować metodę perturbacyjną [12], [13].

W obliczeniach dobroci Q_d wykorzystuje się metodę zaburzenia wartości parametrów dielektryka wypełniającego rezonator. Zgodnie z metodą rachunku zaburzeń zmianie przenikalności elektrycznej dielektryka o $\Delta\varepsilon_r$ odpowiada następująca zmiana pulsacji rezonansowej:

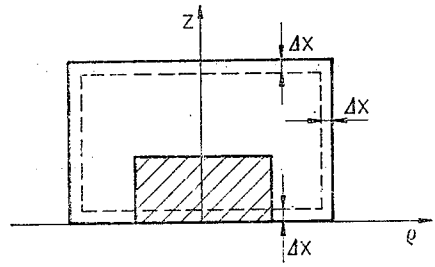
$$\frac{\Delta\omega}{\omega} = -\frac{\Delta\varepsilon_r}{2\varepsilon_r} \frac{W_d}{W_d + W_a}. \tag{41}$$

Z porównania wzorów (39) i (41) otrzymamy:

$$Q_d = \frac{1}{2\left|\frac{\Delta\omega}{\omega} \Big/ \frac{\Delta\varepsilon_r}{\varepsilon_r}\right|} \cdot \frac{1}{\tan\delta}. \tag{42}$$

Dobroć Q_d możemy więc obliczyć numerycznie korzystając jedynie z procedury określenia wartości ω. Dla wyznaczenia dobroci Q_c możemy zastosować metodę rachunku zaburzeń

jedynie dla takich rodzajów elektromagnetycznych, dla których na wszystkich metalowych powierzchniach rezonatora znika składowa normalna pola elektrycznego. Załóżmy, że zmieniono jednocześnie wszystkie wymiary geometryczne rezonatora o wartość Δx, jak na rys. 2.



Rys. 2. Ilustracja zmiany wymiarów geometrycznych rezonatora

Zgodnie z metodą rachunku zaburzeń zmianie objętości rezonatora o ΔV odpowiada następująca zmiana pulsacji rezonansowej rezonatora:

$$\frac{\Delta\omega}{\omega} = \frac{\int\limits_{\Delta V} \mu_0 \cdot |\vec{H}|^2 \mathrm{dv} - \int\limits_{\Delta V} \varepsilon_0 \varepsilon_r |\vec{E}|^2 \mathrm{dv}}{2 \cdot \int\limits_{V} \mu_0 |\vec{H}|^2 \mathrm{dv}}. \tag{43}$$

W związku z tym, że na ściankach rezonatora znika składowa styczna pola elektrycznego, a analizujemy tylko takie rodzaje pól, dla których składowa normalna pola $\vec{E}$ także znika, zatem druga całka w liczniku wyrażenia (43) równa jest zero. Ponieważ na ściankach rezonatora znika także składowa normalna pola magnetycznego, a składowa styczna osiąga ekstremum, więc:

$$\int\limits_{\Delta V} \mu_0 |\vec{H}|^2 \mathrm{dv} \approx \Delta x \cdot \mu_0 \cdot \int\limits_{s} |\vec{H}_t|^2 \mathrm{ds}. \tag{44}$$

Podstawienie wyrażeń (43) i (44) do wyrażenia (40) prowadzi do następującego wzoru na dobroć Q_c:

$$Q_c = \frac{\omega \mu_0 \Delta x}{2 \cdot \left| \frac{\Delta\omega}{\omega} \right| \cdot R_s}, \tag{45}$$

gdzie $\Delta\omega$ jest zmianą pulsacji, odpowiadającą zmianie wymiarów rezonatora o Δx. Tak więc dla rodzajów, dla których na metalowych ściankach rezonatora znika składowa normalna pola elektrycznego (takimi rodzajami są rodzaje quasi TE_{0nm}) dobroć rezonatora może być wyznaczona w oparciu o procedurę obliczeń pulsacji rezonansowych.

W wypadku, gdy wartość pochodnych $\Delta\omega/\omega$ może być wyznaczona analitycznie (np. dla pustego rezonatora TE_{0nm}) wartość dobroci Q_d i Q_c obliczona ze wzorów (42) i (45) zgadza się dokładnie z wartościami obliczonymi z zależności (39) i (40) przez całkowanie pól elektromagnetycznych.

3. WYNIKI OBLICZEŃ NUMERYCZNYCH I EKSPERYMENTÓW

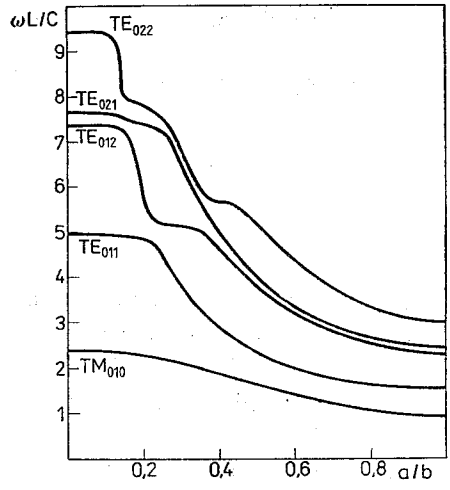
W tabeli 1 przedstawiono wyniki obliczeń wartości zredukowanych pulsacji $(\omega L/c)$ sześciu pierwszych rodzajów quasi TE_{0nm} oraz rodzaju quasi TM_{010} w funkcji stopnia wyznacznika N dla rezonatora o $h/L = 0.25$, $\varepsilon_r = 10$, $a/h = b/L = 1.00$.

Tabela 1

Wyniki obliczeń wartości zredukowanych pulsacji $\omega L/c$

N	$\omega L/c$						
	TE_{011}	TE_{012}	TE_{021}	TE_{022}	TE_{013}	TE_{031}	TM_{010}
1	4.07981	—	—	—	—	—	2.02046
2	4.32412	5.15487	—	—	—	—	2.16842
3	4.40270	5.15801	—	7.49754	7.90860	8.19561	2.20876
4	4.42289	5.16582	7.15565	7.45886	7.91231	8.21483	2.21949
5	4.42405	5.16837	7.23677	7.46290	7.90435	8.21437	2.22236
6	4.42393	5.16866	7.27404	7.46977	7.89954	8.21781	2.22068
7	4.42488	5.16866	7.28842	7.47364	7.89744	8.22207	2.21915
8	4.42560	5.16869	7.29109	7.47431	7.89711	8.22418	2.22043
9	4.42575	5.16872	7.29072	7.47413	7.89722	8.22449	2.22014

Wyniki pozwalają na oszacowanie zbieżności i dokładności metody obliczeń. Jak widać, dla stopnia wyznacznika $N = 9$ ustalają się praktycznie 4 miejsca znaczące, co pozwala na oszacowanie dokładności obliczeń na 10^{-4}. W praktyce dokładność pomiarów geometrycznych dielektryka, a także samego rezonatora jest o rząd wielkości gorsza, co pozwala na stwierdzenie, iż błędy spowodowane niedokładnością metody obliczeniowej są mniejsze od błędów wynikających z niedokładności pomiarów.



Rys. 3. Charakterystyki ilustrujące zmiany zredukowanej pulsacji rezonansowej rodzaju quasi TM_{010} oraz rodzajów quasi TE_{0nm} wskutek wzrostu stopnia wypełniania rezonatora badanym dielektrykiem

Na rysunku 3 przedstawiono zależności zredukowanej pulsacji rodzaju quasi TM_{010} oraz czterech pierwszych rodzajów quasi TE_{0nm} od stopnia wypełnienia rezonatora badanym dielektrykiem. Obliczenia przeprowadzono dla rezonatora o parametrach $b/L = a/h = 1.00$, $\varepsilon_r = 10$ przy stopniu wyznacznika $N = 9$.
Powyższe wykresy mają charakter uniwersalny, są aktualne dla dowolnego zakresu częstotliwości uwarunkowanego wymiarami rezonatora. Z przebiegu charakterystyk wynika, iż bardziej czułymi na zmiany parametrów próbki dielektryka są rodzaje typu TE_{0nm}.

Tabela 2

Wyniki obliczeń zredukowanej pulsacji rodzaju quasi TE_{011}, zredukowanej dobroci zależnej od strat w dielektryku ($Q_d \cdot \tan\delta$) i zredukowanej dobroci zależnej od strat przewodnictwa ($Q_c \cdot R_s$) w funkcji wymiarów próbki dielektrycznej

a/b	$\omega L/c$	$Q_d \cdot \tan\delta$	$Q_c \cdot R_s$
0.15	4.95100	364.37	660.6
0.20	4.90228	14.12	623.3
0.25	4.42575	1.443	519.1
0.30	3.75732	1.171	514.7
0.35	3.24484	1.118	516.5
0.40	2.86326	1.092	515.3
0.45	2.56482	1.074	510.8
0.50	2.32909	1.060	501.9
0.60	1.98858	1.038	464.1
0.70	1.76926	1.021	395.8
0.80	1.63782	1.008	311.4
0.90	1.57744	1.001	240.0
1.00	1.56549	1.000	210.1

W tabeli 2 przedstawiono wyniki obliczeń zredukowanej pulsacji rodzaju quasi TE_{011}, zredukowanej dobroci zależnej od strat w dielektryku ($Q_d \cdot \tan\delta$) i zredukowanej dobroci zależnej od strat przewodnictwa ($Q_c \cdot R_s$) w funkcji wymiarów próbki dielektrycznej. Obliczenia przeprowadzono dla rezonatora o parametrach $b/L = a/h = 1.00$, $\varepsilon_r = 10$ przy stopniu wyznacznika $N = 9$.

Tabela 3

Wyniki pomiarów kilku próbek dielektrycznych w rezonatorze quasi TE_{012}

Materiał	a (mm)	h (mm)	L (mm)	Q	ε_r	$\tan\delta$
kwarc	8.60	14.34	45.80	23000	4.436	$3.8 \cdot 10^{-5}$
teflon	18.78	12.52	49.14	13450	2.047	$1.7 \cdot 10^{-4}$
ceramika Al_2O_3	7.00	7.00	38.37	10400	9.55	$2.0 \cdot 10^{-4}$

$L_0 = 65.97$ mm $b = 25.01$ mm $f = f_0 = 8607.00$ MHz $Q_0 = 28400$
L_0, f_0, Q_0 — oznaczają parametry rezonatora pustego
L, f, Q — oznaczają parametry rezonatora z badaną próbką

Jak widać z przedstawionych w Tabeli 2 wyników, dobroć zależna od strat w dielektryku już dla stosunkowo niewielkich wartości $h/L \approx 0.25$ osiąga wartości bliskie granicznym (graniczna wartość dla $h/L = 1$ wynosi $Q_d = 1/\tan\delta$). Jednocześnie dla małych wartości h/L dobroć zależna od strat w ściankach rezonatora $Q_c \cdot R_s$ pozostaje stosunkowo duża, bliska wartościom charakterystycznym dla pustego rezonatora. Stwarza to dogodne warunki dla pomiarów $\tan\delta$ materiałów małostratnych, dla których dokładność pomiaru jest tym większa, im większa jest wartość $Q_c \cdot R_s$ i im mniejsza jest wartość Q_d.

W tabeli 3 przedstawiono wyniki pomiarów kilku próbek dielektrycznych w rezonatorze quasi TE_{012}. Pomiarów dokonano metodą stałej częstotliwości rezonansowej (po włożeniu próbki zmieniono długość rezonatora tak, aby jego częstotliwość rezonansowa była taka sama jak w przypadku rezonatora pustego).

Wyniki pomiarów parametrów materiałowych zgadzają się z danymi literaturowymi [16÷17] i wynikami badań własnych na próbkach, których średnica była równa średnicy rezonatora.

BIBLIOGRAFIA

1. A. Brandt: *Isledowanije dielektrikow na svierchvysokich czastotach.* Fiz. matgiz., Moskwa, 1963
2. M. Jaworski: *Teoretyczne podstawy pomiaru zespolonej przenikalności elektrycznej w paśmie mikrofalowym.* Praca doktorska, Inst. Fizyki PAN, Warszawa, 1976
3. J. Krupka: *Zastosowanie cylindrycznych rezonatorów H_{01n} do pomiaru zespolonej przenikalności elektrycznej w paśmie mikrofalowym.* Rozprawa doktorska, Politechnika Warszawska, 1977
4. J. Krupka, A. Kędzior: *Optimization of the permittivity and loss tangent measurement of low loss dielectrics in cylindrical TE_{01n} mode cavities.* Electron Technology, 1981, vol. 14, no 1
5. J. Krupka: *Optimization of an electrodynamic basis for determination of the resonant frequencies of microwave cavities partially filled with a dielectric.* IEEE Trans. Microwave Theory Techn. 1983, vol. MTT-31, pp. 302—305
6. F. H. Gil and J. Gisermo: *Finite element analysis of dielectric resonators on microstrip structures.* XXI st General Assembly of URSI, Florence, Italy, 1984 Aus. 28-Sept. 5
7. J. Delaballe, P. Guillon and Y. Garault: *Local complex permittivity measurement on MIC substrates.* Arch. Elektron. Ubertragung (AEU) 1981, vol. 35, pp. 80—83, no. 2
8. U. Crombach and R. Michelfeit: *Resonans Frequenzen und Feldstarken in geschirmten dielektrischen Sheiben und Ringresonatoren.* Frequenz, 1981, no 12, vol. 35, pp. 324—328
9. Y. Kobayashi, N. Fukuoka and S. Yoshida: *Resonant modes for a shielded dielectric rod resonator.* Electronics and Communication (in Japan), 1981 vol. 64-B, no 11, pp. 46—51
10. R. F. Harrington: *Time-harmonic Electromagnetic Fields.* New York: Mc Graw-Hill, 1961, ch. 5
11. R. F. Harrington: *Field Computation by Moment Methods.* New York: Macmillan, 1968 ch. 9
12. D. Kajfez: *Incremental frequency rule for computing the Q-factor of a shielded TE_{0mp} dielectric resonator.* IEEE Trans. Microwave Theory Techn., 1984, vol. MTT-32
13. Y. Kobayashi, T. Aoki and Y. Kabe: *Influence of conductor shields on the Q-factors of a TE_0 dielectric resonator.* IEEE MTT-S Int. Microwave Symposium Digest., 1985, St. Louis, pp. 281—284
14. Hsi Tech. Hsieh: *A resonant cavity study of semiconductor.* J. Appl. Phys., 1954. no 3

15. J. Krupka, Sz. Maj: *Application of* $TE_{01\delta}$ *mode dielectric resonator for the complex permittivity measurements of semiconductors.* CPEM'86 Conference Digest, Gaithersburg, June 1986, pp. 154—155
16. R. G. Jones: *The measurement of dielectric anisotropy using a microwave open resonator.* J. Phys. D: Appl. Phys. 1976, vol. 9
17. I. Wolff, N. Schwab: *Messung der Dielektrizitätszahl anisotroper dielektrischer Materialien im Mikrowellenbereich.* Westdeutcsher Verlag, 1980

A. HARASIEWICZ, J. KRUPKA

APPLICATION OF QUASI-TE_{0nm} AND QUASI-TM_{010} CYLINDIRCAL RESONATORS TO MEASURING THE COMPLEX PERMITTIVITY OF CYLINDRICAL DIELECTRIC SAMPLES OF ARBITRARY DIMENSIONS

Summary

The paper deals with a measuring method of the permittivity and loss-factor of dielectric samples in the microwave band by means of cylindrical quasi-TE_{0nm} and quasi-TM_{010} resonators. The tested cylindrical sample is coaxially placed on the bottom of the resonator. The dimensions (diameter and height) of the sample are arbitrary, i.e. they are not necessarily adapted to the dimensions of the resonator. The solution of the boundary problem of Helmholtz equations is carried out by the mode matching method. The problem is reduced to solving N uniform linear equations. Digital computations prove that the method is rapidly convergent, i.e. for $N = 9$ the errors due to the inaccuracy of the method are lower than the measuring errors. The quality factor of the resonator is calculated by means of the perturbation method. The latter method is applicable to the quasi-TE_{0nm} resonator.

A. HARASIEWICZ, J. KRUPKA

APPLICATION DES RÉSONATEURS CYLINDRIQUES QUASI TE_{0nm} ET QUASI TM_{010} À LA MESURE DE LA PERMITTIVITÉ COMPLEXE DES ÉCHANTILLONS CYLINDRIQUES DES DIÉLECTRIQUES À DIMENSIONS ARBITRAIRES

Résumé

L'étude concerne la méthode de mesure, dans la bande micro-ondes, de la permittivité et des pertes en watts des échantillons d'un diélectrique lors de l'application des résonateurs cylindriques quasi TE_{0nm} et quasi TM_{010}. L'échantillon examiné en forme d'un cylindre est placé au centre du fond du résonateur. Les dimensions (le diamètre et la hauteur) de l'échantillon peuvent être arbitraires, c'est à dire qu'elles ne doivent pas être ajustées aux dimensions du résonateur. La résolution du problème au limite des équations de Helmholtz a été faite par méthode approximative de collage des sousespaces. Le problème est reduit à la résolution du système N des équations homogènes linéaires. Les calculs numériques montrent, que la méthode se distingue d'une convergence rapide, c'est à dire que pour $N = 9$ les erreurs dues au manque de précision de la méthode sont inférieures à celles de mesures. Cette méthode peut être appliqueé au résonateur quasi TE_{0nm}.

A. HARASIEWICZ, J. KRUPKA

ANWENDUNG ZYLINDRISCHER QUASI TE_{0nm}- UND QUASI TM_{010}-RESONATOREN ZUR MESSUNG DER ELEKTRISCHEN GESAMTDURCHLÄSSIGKEIT VON DIELEKTRISCHEN ZYLINDRISCHEN PROBEN MIT BELIEBIGEN AUSMAßEN

Zusammenfassung

Die Bearbeitung betrifft Meßmethoden im Mikrowellenband der elektrischen Durchlässigkeit und der Probenverlustzahl unter Anwendung zylindrischer quasi TE_{0nm}- und quasi TM_{010}-Resonatoren. Die Probengrößen (Durchmesser und Höhe) können beliebig sein, d.h. daß sie den Abmessungen des Resonators nicht angepaßt werden müssen; sie werden in Zylinderform auf dem Resonatordeckel angebracht. Die Lösung des Grenzproblems der Helmholtz-Gleichungen wurde mittels der Annäherungsmethode unter Verkleben von Untergebieten gewonnen. Das Problem wird auf die Lösung des N-Systems homogener Lineargleichungen zurückgeführt. Aus numerischen Berechnungen ergeht, daß die Methode schnellkonvergent ist, d.h. bei $N = 9$ die wegen der Ungenauigkeit der Methode verursachten Fehler geringer, als die Meßfehler sind.

Die Resonatorgüte wurde mit Hilfe der Störungrechnung berechnet. Diese Methode kann in Bezug auf den quasi TE_{0nm}-Resonator angewandt werden.

А. ХАРАСЕВИЧ, Е. КРУПКА

ПРИМЕНЕНИЕ ЦИЛИНДРИЧЕСКИХ РЕЗОНАТОРОВ КВАЗИ $TE_{0нм}$ И КВАЗИ TM_{010} К ИЗМЕРЕНИЮ СОПРЯЖЕННОЙ ЭЛЕКТРИЧЕСКОЙ ПРОНИКАЕМОСТИ ЦИЛИНДРИЧЕСКИХ ОБРАЗЦОВ ДИЭЛЕКТРИКОВ ПРОИЗВОЛЬНЫХ РАЗМЕРОВ

Резюме

Обсуждены методы измерения в свч диапазоне электрической проницаемости и потерь диэлектрика при применении цилиндрических резонаторов квази $TE_{0нм}$ и квази TM_{010}. Исследуемый образец установливается в центре крышки резонатора. Размеры (диаметр и высота) образца могут быть произвольными, т.е. не требуется их подбора к размерам резонатора. Решение краевых условий уравнения Гельмгольца проведено приближенным методом склеивания подпространств. Проблема сводится к решению системы N однородных линейных уравнений. Из числовых расчетов вытекает, что метод быстро сходящийся т.е. для $N = 9$ ошибки от неточности метода меньше ошибок измерения. Добротность резонатора определена методом расчета возмущений. Метод применяет к резонатору квази $TE_{0нм}$.

ROZPRAWY ELEKTROTECHNICZNE, 1989, 35, z. 1, ss. 89-99

621.3.087.92

Zastosowanie DPF w dynamicznym testowaniu torów przetwarzania analogowo-cyfrowego

ADAM KRZYWAŹNIA, JANUSZ OCIEPKA

Instytut Metrologii Elektrycznej, Politechnika Wrocławska

Otrzymano 1987.06.12

Autoryzowano do druku 1988.03.04

W pracy przedstawiono zagadnienie testowania torów przetwarzania analogowo-cyfrowego w warunkach dynamicznych. Metoda pomiarowa polega na analizie cyfrowego sygnału wyjściowego z przetwornika a/c za pomocą dyskretnego przekształcenia Fouriera przy wzorcowym sinusoidalnym sygnale wejściowym. W wyniku analizy wyznacza się efektywną liczbę bitów badanego toru oraz stosunek mocy sygnału do mocy zastępczego błędu skutecznego toru. Przedstawiono wyniki pomiarów 10-bitowego woltomierza próbkującego.

1. WPROWADZENIE

Rozwój technologii układów scalonych bardzo wielkiej skali integracji wpłynął w znacznym stopniu na ukierunkowanie ewolucji systemów pomiarowych w stronę zastosowań techniki cyfrowego przetwarzania sygnałów (CPS). Tendencję tę obserwuje się np. w rozpoznawaniu obrazów, analizie widmowej, analizie korelacyjnej, itp. [1]. Podstawowym blokiem struktury systemu z zakresu tych zastosowań jest procesor sygnałów cyfrowych współpracujący z układami wejściowymi i wyjściowymi oraz kontrolerem systemu. Nieodzownym blokiem układów wejściowych jest przetwornik analogowo-cyfrowy, umożliwiający sprzężenie procesora sygnałów cyfrowych z otoczeniem. Obecnie obserwuje się wyraźny rozwój konstrukcji próbkujących przetworników a/c w kierunku zwiększenia szybkości przetwarzania i polepszenia dokładności [2].

Zagadnienie pomiarów parametrów metrologicznych próbkujących przetworników a/c, jak i zbudowanych na ich bazie przyrządów pomiarowych, nie jest jeszcze w pełni rozwiązane. W niektórych przypadkach, jak np. w zagadnieniach analizy widma, problemy metrologiczne w zakresie oceny dokładności nie mają ostatecznego rozwiązania, słusznego np. dla wszystkich rodzajów mierzonych sygnałów. W niniejszej pracy podjęto próbę przedstawienia jednej z metod doświadczalnej oceny właściwości woltomierza próbkującego przystosowanego do pracy w systemie CPS. Jest to zagadnienie o dużym znaczeniu praktycznym, podjęte przez kilku badaczy zagranicznych [3, 4].

2. WŁAŚCIWOŚCI METROLOGICZNE WOLTOMIERZY PRÓBKUJĄCYCH

Woltomierz próbkujący można zdefiniować jako cyfrowy przyrząd do pomiaru wartości chwilowych napięcia, którego wielkością wejściową jest napięcie $u(t)$, a wielkością wyjściową skwantowana wartość chwilowa napięcia zmierzona w chwili t_k i przedstawiona za pomocą przyjętego kodu cyfrowego:

$$u_k = Q[u(t_k)], \tag{1}$$

gdzie: $Q[\cdot]$ jest symbolem operacji kwantowania.
W zdecydowanej większości przypadków wartości u_k są reprezentowane za pomocą naturalnych kodów binarnych. Proces idealnego kwantowania polega wtedy na przypisaniu do wartości chwilowej próbki $u(t_k)$ liczby binarnej z zakresu $[-2^{Q-1}, 2^{Q-1})$ w taki sposób, aby błąd kwantowania e_k, rozumiany jako różnica pomiędzy wartością rzeczywistą $u(t_k)$ i jej liczbową reprezentacją u_k był nie większy od połowy kroku kwantowania:

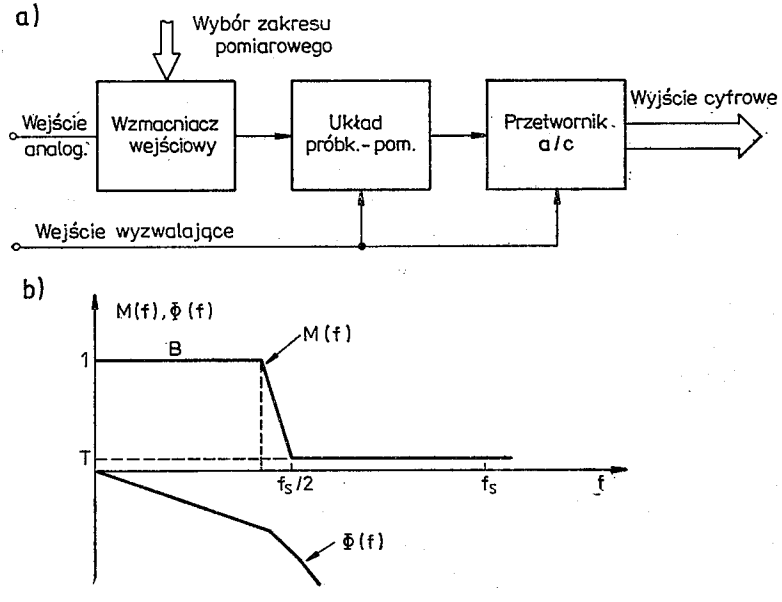
$$e_k = u_k - u(t_k), \tag{2}$$

przy czym

$$|e_k| \leqslant Z \cdot 2^{-(Q-1)}, \tag{3}$$

gdzie: Z jest zakresem jednopolarnym przetwornika a/c;
Q — liczbą bitów reprezentacji próbki $u(t_k)$.

Woltomierze próbkujące budowane są na bazie próbkujących przetworników a/c, uzupełnionych o dodatkowe bloki funkcjonalne, umożliwiające pomiary napięcia w pewnym zakresie jego wartości, jak wzmacniacze, dzielniki napięcia, filtry dolnopasmowe, itp. Struktury części analogowej woltomierzy próbkujących są podobne do struktur wolto-



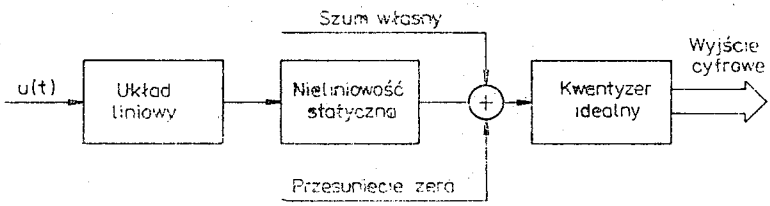
Rys. 1 a) Schemat blokowy woltomierza próbkującego do pomiaru małych wartości napięcia, b) Uproszczone pożądane charakterystyki częstotliwościowe toru analogowego woltomierza próbkującego

mierzy analogowych napięcia przemiennego [5]. Na rys. 1 pokazano przykładowy schemat blokowy woltomierza próbkującego do pomiaru małych wartości napięć (struktura miliwoltomierza), a także pożądane charakterystyki częstotliwościowe układów wejściowych, wymaganych w typowych zastosowaniach tych woltomierzy.

W określaniu właściwości metrologicznych woltomierzy próbkujących konieczne jest uwzględnienie właściwości ich części analogowej oraz części cyfrowej. Właściwości te można określać w sposób analityczny, rozpatrując zjawiska zachodzące w poszczególnych blokach funkcjonalnych. Takie podejście jest charakterystyczne na etapie konstruowania przyrządu, natomiast do oceny gotowego urządzenia bardziej przydatne jest charakteryzowanie jego właściwości w sposób globalny, bez wnikania w istotę zjawisk w układzie. Na przykład w typowym woltomierzu próbkującym z układem próbkująco-pamiętającym i kompensacyjnym przetwornikiem a/c można wymienić następujące zjawiska w poszczególnych blokach funkcjonalnych, mające wpływ na jego globalne właściwości:

— analogowe układy wejściowe: nierównomierność amplitudowych i fazowych charakterystyk częstotliwościowych, błąd współczynnika wzmocnienia, błąd przesunięcia zera i jego zmiany, szumy, podatność na zakłócenia,
— układ próbkująco-pamiętający: proces ładowania i rozładowania kondensatora pamiętającego, parametry resztkowe kluczy oraz ich nieliniowości,
— kompensacyjny przetwornik analogowo-cyfrowy: błędy odtworzenia wartości stosunku poszczególnych wag przy kompensacji wagowej, stany przejściowe komparatora, błędy kodowania.

Mnogość wymienionych zjawisk mogących podowować błędy pomiaru w woltomierzach próbkujących uzasadnia stosowanie uproszczonych modeli metrologicznych, które odwzorowują globalne właściwości takiego woltomierza. Taki uproszczony model pokazany jest na rys. 2. Składa się on z kwantyzera idealnego o rozdzielczości określonej nominalną liczbą



Rys. 2. Blokowy schemat zastępczy woltomierza próbkującego do potrzeb analizy jego właściwości metrologicznych

bitów, poprzedzonego układem nieliniowości statycznej (reprezentującej błędy nieliniowości charakterystyki statycznej całego toru) i układem liniowym odpowiedzialnym za zmiany wzmocnienia w torze analogowym w funkcji częstotliwości. Wyróżniono także napięcie przesunięcia zera oraz napięcie szumów własnych jako istotne parametry metrologiczne woltomierza próbkującego. Znaczenie szumu w woltomierzach próbkujących związane jest z tym, że odpowiednio dobrany poziom szumu w przetwornikach a/c zmniejsza wynikowy błąd pomiaru wielkości uśrednionych, wyliczanych ze zbioru próbek [6].

W dalszej części pracy rozważane będzie zagadnienie doświadczalnego wyznaczania efektywnej liczby bitów zastępczego kwantyzera w modelu rzeczywistego woltomierza próbkującego.

3. DYNAMICZNE TESTOWANIE PRÓBKUJĄCYCH PRZETWORNIKÓW A/C

W literaturze przedmiotu wyróżnia się statyczne i dynamiczne metody testowania przetworników a/c [7]. Najczęściej stosowaną w praktyce metodą statyczną jest metoda wzorcowego przetwornika c/a, której istota polega na zamianie cyfrowego kodu wyjściowego badanego przetwornika a/c na napięcie i porównanie jego wartości z wzorcową wartością napięcia wejściowego. Różnica tych dwóch napięć jest miarą błędu przetwornika a/c, jeśli rozdzielczość przetwornika c/a jest znacznie lepsza niż rozdzielczość badanego przetwornika a/c. Pomiar tego błędu w funkcji wejściowego napięcia wzorcowego pozwala określić błędy statycznej charakterystyki przetwarzania przetwornika a/c [8].

W metodach dynamicznych analizowana jest cyfrowa odpowiedź przetwornika a/c na wzorcowy analogowy sygnał testujący. Do analizy odpowiedzi przetwornika stosuje się dyskretne przekształcenie Fouriera (DPF) lub estymację nieliniową sygnału wyjściowego metodą minimalizacji błędu średniokwadratowego [9].

Metody stosowane w badaniach przetworników a/c mogą być zastosowane w badaniach woltomierzy próbkujących. Najczęściej stosuje się te metody dynamiczne, w których wykorzystuje się sinusoidalne sygnały testujące. Wynika to z właściwości tych sygnałów, wyróżniających się następującymi cechami na tle innych sygnałów testujących:

— deterministyczny charakter sygnału,
— kontrolowany kształt przebiegu dzięki bardzo małym osiągalnym zniekształceniom (np. współczynnik zawartości harmonicznych może osiągnąć wartości mniejsze od 0,005%),
— możliwość dokładnego pomiaru parametrów sygnału (wartość skuteczna, średnia) za pomocą przyrządów analogowych.

Sposób pomiaru parametrów metrologicznych woltomierza próbkującego opiera się na numerycznej analizie zbioru próbek $\vec{U}$ w postaci cyfrowej, otrzymanego na wyjściu badanego woltomierza, przy soinusoidalnym wzorcowym sygnale wejściowym:

$$\vec{U} = \left\{u\left(k\frac{T_w}{N}\right)\right\}, \quad k = 0, 1, \ldots, N-1, \tag{4}$$

gdzie: N — liczba próbek, T_w — szerokość okna czasowego, w którym dokonywane są pomiary.

Elementy zbioru $\vec{U}$ są obarczone błędami występującymi w torze przetwarzania sygnału, wyszczególnionymi w części 2 niniejszej pracy. Na podstawie tego zbioru można wyznaczyć efektywną liczbę bitów Q_e charakteryzującą wynikową rozdzielczość toru przetwarzania a/c.

W celu wyznaczenia wartości Q_e w badanym systemie zbiór próbek $\vec{U}$ poddaje się dyskretnemu przekształceniu Fouriera, w wyniku czego uzyskuje się zbiór próbek widma sygnału wyjściowego $\vec{Y}$. Elementami zbioru $\vec{Y}$ są zespolone próbki widma Y_n liczone ze wzoru:

$$Y_n = P_n + jQ_n = \frac{1}{N}\sum_{k=0}^{N-1} u\left(k\frac{T_w}{N}\right)\cdot w(k)\cdot e^{-j\frac{2\pi}{N}kn}, \tag{5}$$

$$n = 0, 1, \ldots, \frac{N}{2} - 1,$$

gdzie: P_n — część rzeczywista Y_n, Q_n — część urojona Y_n,

$w(k)$ — wartości funkcji wagowej minimalizującej przeciek widma (okno czasowe).

Zbiór $\vec{Y}$ jest podstawą dalszej analizy metrologicznej. Wartości modułów próbek widma służą do wyznaczania wartości zastępczego błędu rozdzielczości oraz efektywnej liczby bitów charakteryzującej błąd rozdzielczości toru przetwarzania a/c.

4. WYZNACZANIE EFEKTYWNEJ LICZBY BITÓW WOLTOMIERZA PRÓBKUJĄCEGO

Efektywna liczba bitów woltomierza próbkującego, charakteryzującą jego rzeczywistą rozdzielczość, jest związana z wartością skuteczną zastępczego błędu rozdzielczości, który ujmuje wszystkie błędy cząstkowe powstające w torze przetwarzania a/c. Efektywną liczbę bitów Q_e definiuje się wzorem [4].:

$$Q_e = Q - \log_2 \frac{e_{sr}}{e_{si}}, \tag{6}$$

gdzie: Q — nominalna liczba bitów, e_{sr} — zmierzona wartość skuteczna błędu rozdzielczości, e_{si} — wartość skuteczna błędu kwantowania, wynikająca z nominalnej liczby bitów.

Idealny kwantyzer o rozdzielczości wyrażonej liczbą bitów Q wprowadza błąd skuteczny e_{si} o wartości:

$$e_{si} = \frac{Z}{\sqrt{12}\, 2^{Q-1}}, \tag{7}$$

gdzie: Z jest zakresem przetwornika a/c.

W praktyce do wyznaczania Q_e wykorzystuje się zbiór próbek widma $\vec{Y}$, wyliczonych ze wzoru (5). Z otrzymanego zbioru wyznacza się najpierw zbiór modułów widma $\vec{A}$, z którego wyznacza się z kolei wartość stosunku sygnał/szum (SNR) w systemie według wzoru:

$$\mathrm{SNR}_z = 10 \lg \frac{\sum_l A_l^2}{\sum_p A_p^2}, \tag{8}$$

gdzie: SNR_z — oznacza zmierzoną wartość współczynnika SNR w badanym systemie, l — numery próbek modułu widma związane ze składową podstawową sygnału, p — obejmuje wszystkie pozostałe próbki modułu widma, ale bez wartości średniej (składowej stałej).

Wybór wartości l i p zależy od zastosowanej funkcji wagowej (okno czasowe). Wyznaczona w powyższy sposób wartość SNR_z zależy od wypadkowego szumu w systemie i od stopnia

wykorzystania zakresu przetwornika a/c. Efektywną liczbę bitów Q_e wyznacza się, przy założeniu pełnego wykorzystania zakresu przetwornika, ze wzoru:

$$Q_e = \frac{\mathrm{SNR}_z - 10 \log 6}{20 \log 2} + 1 = 3{,}322 \frac{\mathrm{SNR}_z}{20} - 0{,}293, \tag{9}$$

gdzie: $\log(\cdot)$ — oznacza logarytm o podstawie 10.

Stosowanie funkcji wagowej $w(k)$ występującej we wzorze (5) ma na celu minimalizację efektu przecieku widma [10]. Zazwyczaj wybiera się cosinusoidalną funkcję wagową, np. tzw. okno Hanna [9].

W tabeli 1 zawarto wartości SNR w funkcji Q. Warto zauważyć, że SNR ma istotne znaczenie praktyczne, gdyż jest on miarą zakresu dynamicznego systemu cyfrowego przetwarzania sygnałów w dziedzinie widma. Określa on maksymalny odstęp między po-

Tabela 1

Wartości e_{st}, SNR oraz hg dla najczęściej spotykanych nominalnych wartości rozdzielczości przetworników a/c

Liczba bitów Q	e_{st} $\times 10^{-6}$	SNR dB	hg %
8	2260	49,93	0,1
9	1130	55,95	0,05
10	564	61,97	0,025
11	282	67,99	$1{,}26 \cdot 10^{-2}$
12	141	74,01	$6{,}3 \cdot 10^{-3}$
13	70,5	80,03	$3{,}16 \cdot 10^{-3}$
14	35,2	86,05	$1{,}58 \cdot 10^{-3}$
15	17,6	92,07	$7{,}94 \cdot 10^{-4}$
16	8,81	98,09	$3{,}94 \cdot 10^{-4}$

ziomem mocy szumu i mocy sygnału sinusoidalnego w danym systemie. W tabeli tej podano również dopuszczalne wartości współczynnika zniekształceń nieliniowych sygnału testującego, przy założeniu, że poziom zniekształceń powinien być o 10 dB mniejszy od SNR osiągalnego w badanym torze.

W praktyce szybkie testowanie woltomierza próbkującego polega na określaniu wartości Q_e w funkcji częstotliwości sygnału testującego. Kryterium jakości może mieć postać:

$$\min_{f_i} Q_e(f_i) \geqslant Q_{\min}, \tag{10}$$

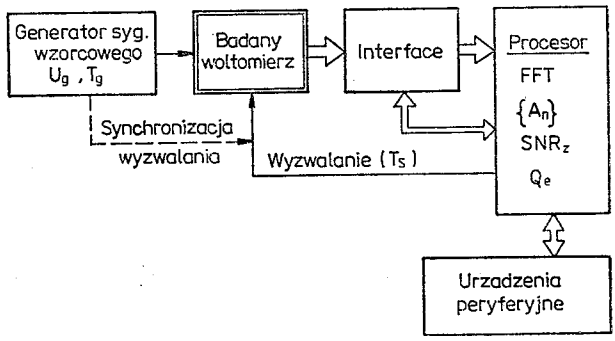
gdzie: $Q_{\min}$ jest dopuszczalną minimalną wartością Q_e,
$\{f_i\}$ jest zbiorem częstotliwości pomiarowych.

Koncepcja efektywnej liczby bitów cechuje się prostotą, a wartość Q_e pozwala szybko ocenić jakość woltomierza próbkującego. Jest wygodniej posługiwać się jedną liczbą Q_e niż zbiorem błędów cząstkowych wymienionych w 2 części pracy. Słabością koncepcji jest trudność w ujęciu wpływu zmian wzmocnienia i przesunięcia fazowego w funkcji częstotliwości w torze analogowym woltomierza, co powoduje, że w praktyce wartość Q_e można

wyznaczyć tylko przy wymuszeniach sinusoidalnych. Podkreśla to Linnenbrink [11], który udowodnił silny wpływ charakterystyk częstotliwościowych na błąd powstający w torze przetwarzania a/c przy wymuszeniach poliharmonicznych. Oznacza to, że w zależności od rodzaju sygnału testującego można otrzymać różne wartości Q_e.

5. PRAKTYCZNA REALIZACJA SYSTEMU DO DYNAMICZNEGO TESTOWANIA WOLTOMIERZY PRÓBKUJĄCYCH I WYNIKI BADAŃ WŁASNYCH

Schemat blokowy systemu do dynamicznego testowania woltomierzy próbkujących (rys. 3) zawiera generator wzorcowego sygnału sinusoidalnego o regulowanej amplitudzie



Rys. 3. Schemat blokowy systemu do testowania woltomierzy próbkujących za pomocą sygnału sinusoidalnego z zastosowaniem dyskretnego przekształcenia Fouriera

i częstotliwości oraz komputerowy system cyfrowego przetwarzania sygnałów, który służy do analizy zbiorów próbek na wyjściu woltomierza. Oprogramowanie systemu musi umożliwić wykonanie następujących operacji obliczeniowych na zbiorze próbek:

— przemnożenie każdej próbki $y\left(k\frac{T_w}{N}\right)$ przez wartość funkcji wagowej $w(k)$,

— obliczanie dyskretnej transformaty Fouriera,

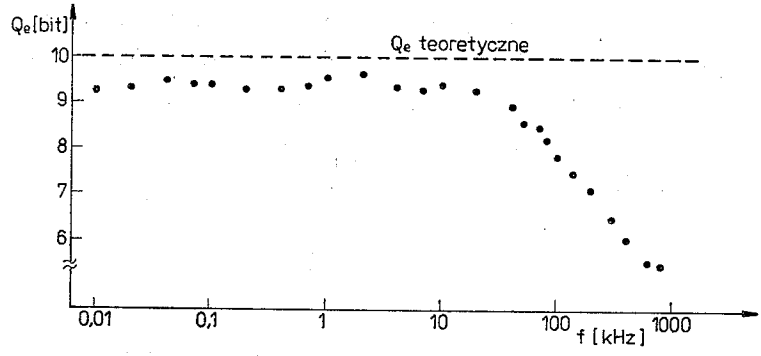
— obliczanie modułu widma oraz SNR i Q_e.

Jedną z zalet opisanej metody jest jej szybkość. Czas wyznaczenia mierzonych parametrów zależy od czasu pomiarów bezpośrednich, który jest przeznaczony na zebranie próbek, oraz czasu przetwarzania tego zbioru. Pomiary bezpośrednie trwają z reguły kilkanaście okresów sygnału testującego, stąd czas ich trwania jest pomijalny w stosunku do czasu przetwarzania. O czasie przetwarzania w dużym stopniu decyduje czas realizacji dyskretnego przekształcenia Fouriera, gdyż zbiory próbek do analizy w opisanej metodzie muszą zawierać co najmniej 256 elementów. Teoretycznie można obecnie realizować algorytm FFT za pomocą niedostępnych w kraju specjalizowanych procesorów scalonych, uzyskując czasy przetwarzania rzędu kilkunastu ms, jednak w praktyce bardziej realne jest stosowanie mikrokomputerów domowych lub osobistych, które pozwalają na uzyskanie czasów przetwarzania rzędu kilku do kilkudziesięciu sekund.

W pracach doświadczalnych badano woltomierz próbkujący opracowany w Instytucie

Metrologii Elektrycznej. Był to woltomierz 10-bitowy zbudowany w oparciu o przetwornik c/a typu DPCA-10 (OBREUS Toruń), o maksymalnej częstotliwości próbkowania równej 125 kHz. Woltomierz badano w systemie pomiarowym jak na rys. 3. W pomiarach stosowano sygnał sinusoidalny o współczynniku zniekształceń nieliniowych mniejszych od 0,02%.

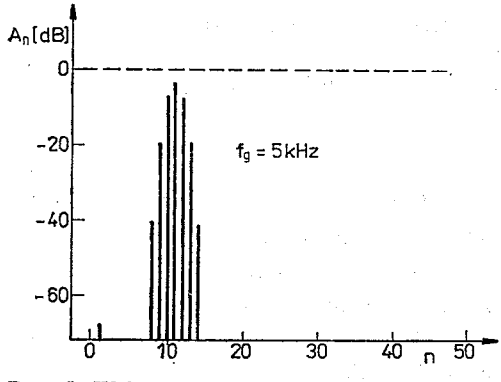
Przy częstotliwościach sygnału testującego większych od 0,5 kHz stosowano próbkowanie stroboskopowe. W celu minimalizacji przecieku widma stosowano okno Blackmana o tłumieniu listków bocznych 92 dB [10]. Na rys. 4 pokazano uzyskane wyniki pomiarów



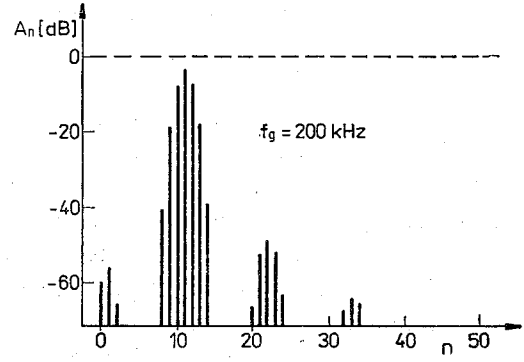
Rys. 4. Wyniki pomiarów efektywnej rozdzielczości 10-bitowego woltomierza próbkującego zbudowanego na bazie przetwornika c/a typu DPCA-10

efektywnej rozdzielczości w funkcji częstotliwości. W całym badanym paśmie częstotliwości efektywna rozdzielczość jest niższa od nominalnej (10 bitów) z tym, że do 30 kHz ten spadek jest mniejszy od 1 bitu.

Na rys. 5 i 6 pokazano wykresy modułu widma dla dwóch różnych częstotliwości.



Rys. 5. Wykres modułu widma A_n odpowiedzi badanego woltomierza przy $f_g = 5$ kHz z oknem Blackmana



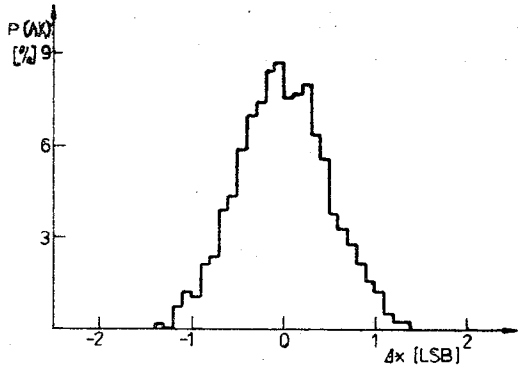
Rys. 6. Wykres modułu widma odpowiedzi woltomierza przy $f_g = 200$ kHz

Na rys. 5 widać charakterystyczny dla zastosowanego okna tylko jeden zbiór prążków odpowiadający częstotliwości sygnału testującego, natomiast na rys. 6 wyraźnie widać występowanie zniekształceń odtworzonego z próbek sygnału testującego. Świadczą o tym

dwie dodatkowe grupy prążków odpowiadające drugiej i trzeciej harmonicznej sygnału testującego. Prawdopodobną przyczyną tych zniekształceń jest różnica czasów narostów zboczy sygnału w układzie wejściowym.

Istnieje możliwość analizy modułu widma otrzymanego w wyniku przetwarzania pod względem jego kształtu w celu określenia przyczyn powodujących zniekształcenia w torze. Jest to jednak analiza jakościowa, a do jej praktycznego zastosowania należałoby stworzyć atlas typowych zniekształceń modułu widma (jak np. na rys. 6) oraz powiązać te zniekształcenia z odpowiednimi zjawiskami w torze przetwarzania a/c.

Na rys. 7 pokazano histogram błędu woltomierza próbkującego w dziedzinie czasu



Rys. 7. Histogram błędu woltomierza próbkującego w dziedzinie czasu w serii pomiarów przy $f_g = 1$ kHz

przy częstotliwości 1 kHz. Wyniki te uzyskano w zmodyfikowanym systemie pomiarowym z rys. 3, w którym zastosowano synchronizację wyzwalania badanego woltomierza z sygnałem testującym (linia przerywana na rys. 3), co pozwoliło na wyliczenie wartości chwilowych błędu w dziedzinie czasu.

W przypadku, gdyby kwantyzer był jedynym źródłem błędów (rys. 2), to rozkład ich powinien zawierać się w przedziale $-0{,}5$ do 0,5 LSB. Przedstawiony histogram błędów (rys. 7) ma rozkład zbliżony do rozkładu normalnego i zajmuje większy przedział. Oznacza to, że znaczący udział mają również inne źródła błędów np. szum i nieliniowość. Istotnie, obserwacje oscyloskopowe wykazały występowanie w torze analogowym szumu o wartości międzyszczytowej odpowiadającej w przybliżeniu 1 LSB.

Z otrzymanych wyników można wnioskować o użyteczności badanego woltomierza w zastosowaniach pomiarowych. Do tego celu może służyć zwłaszcza zależność $Q_e(f)$. Z przykładowych wyników pomiarów $Q_e(f)$ zamieszczonych na rys. 4 można bezpośrednio określić obszar częstotliwości, w którym można stosować badany woltomierz. Do bardziej szczegółowej analizy zjawisk występujących w torze przetwarzania a/c można wykorzystywać wykresy modułu widma oraz histogramy błędów.

6. WNIOSKI

Testowanie torów przetwarzania a/c metodą dynamiczną z zastosowaniem DPF jest przydatne w badaniu gotowych urządzeń i systemów CPS, jak i indywidualnych przetwor-

ników a/c. Metoda nadaje się zwłaszcza do testowania torów o rozdzielczości do 12 bitów, gdyż dla takich torów wamagania co do zniekształceń sygnału testującego i liczby próbek do analizy są stosunkowo łatwe do spełnienia.

Niedogodnością opisanej metody jest brak informacji w jawnej postaci o błędzie charakterystyki przetwarzania, wyrażonej jako zależność wypadkowego błędu badanego toru od wartości napięcia wejściowego. Uzyskana informacja pozwala natomiast na bezpośrednią ocenę jakości toru przetwarzania w dziedzinie widma (SNR) oraz jej transformację do dziedziny amplitud sygnału (Q_e). Wartości obu tych parametrów toru wystarczają do oceny jego przydatności w pomiarach wielkości uśrednionych, jak np. widmo sygnału czy wartość skuteczna, a zależność $Q_e(f)$ dobrze charakteryzuje właściwości częstotliwościowe toru pod względem utraty jego zdolności rozdzielczej ze wzrostem częstotliwości.

Opisana metoda testowania powinna znaleźć zastosowanie jako standardowy sposób kontroli jakości przetworników a/c i torów przetwarzania a/c.

BIBLIOGRAFIA

1. A. V. Oppenheim: *Applications of Digital Signal Processing*. Prentice-Hall, New Jersey, 1978
2. F. Goodenough: *Focus on Hybrid A-D Converters*. Electronic Design, 1985 vol. 33, nr 24, s. 191—199
3. C. Clayton, J. A. McClean, G. McCarra: *FFT Performance Testing of Data Acquisition Systems*. IEEE Trans. Instrum. Meas. 1986, vol. 35, nr 2, s. 212—215
4. T. M. Souders, D. R. Flack, T. C. Wong: *An Automated test set for the dynamic characterisation of A/D converters*. IEEE Trans. Instrum. Meas., 1983, vol. 32, nr 1, s. 180—186
5. A. Jellonek, Z. Karkowski: *Miernictwo radiotechniczne*, Warszawa: WNT, 1972
6. J. Vanderkooy, S. P. Lipshitz: *Resolution Below the Least Significant Bit in Digital Systems with Dither*. J. Audio Eng. Soc., 1984, vol. 32, nr 3, s. 106—113
7. D. Seitzer, G. Pretzl, N. A. Hamdy: *Electronic Analog-to-Digital Converters*, J. Wiley, 1983
8. W. Burney: *High Resolution Converter Cuts Linearity Test to 12 Seconds*. Electronics, 1981, vol 54, nr 19, s. 142—146
9. B. Pratt: *Test A/D Converters Digitally*. Electronic Design, 1975, vol. 23, nr 25, s. 86—88
10. J. Domagała: *Własności okien danych do analizy skończonych szeregów czasowych metodą dyskretnej transformacji Fouriera*. Rozprawy Elektrotechniczne, 1983, vol. 29, z. 2, s. 563—610
11. T. E. Linnenbrink: *Effective Bits: Is That All There Is*? IEEE Trans. Instrum. Meas., 1984, vol. 33, nr 3, s. 184—187

A. KRZYWAŹNIA, J. OCIEPKA

APPLICATION OF THE DFT TO DYNAMIC TESTING OF ANALOG-TO-DIGITAL CONVERTING SYSTEMS

Summary

The problem of dynamic testing of analog-to-digital converting systems is presented. A measuring method based on the analysis of the digital output of an a/d converter by means of a discrete Fourier transform in response to high purity sinusoidal input is described. In result, the number of effective bits and ratio of the signal to the equivalent noise are determined. Measuring results for a 10-bit sampling voltmeter are quoted.

A. KRZYWAŹNIA, J. OCIEPKA

APPLICATION DE LA DISCRÈTE TRANSFORMÉE DE FOURIER (DFT) AU TEST DYNAMIQUE DES CHAÎNES DE CONVERSION ANALOGIQUE-DIGITALE

Résumé

Dans l'étude on a présenté le problème des tests des chaînes de conversion analogique-digitale dans les conditions dynamiques. La méthode de mesure consiste en analyse du signal digital de sortie du convertisseur analogique-digital à l'aide de la discrète transformée de Fourier, pour le sinusoïdal signal standard de sortie. A partir de l'analyse, on détermine le nombre effectif des bits de la chaîne examinée et le ratio entre la puissance du signal et la puissance de lneffective erreur résultante de la chaîne. On a présenté les résultats des mesures d'un voltmètre d'échantillonnage à 10 bits.

A. KRZYŻAŹNIA, J. OCIEPKA

ANWENDUNG DER DFT BEI DYNAMISCHER TESTMETHODE DES ANALOG-DIGITALEN MEßVERARBEITUNGSPFADES

Zusammenfassung

In der Bearbeitung ist das Problem des analog-digitalen Meßpfadtestes unter dynamischen Bedingungen erörtert worden. Die Meßmethode und Analyse beruhen auf der diskreten Fourier-Transformation. Das Eingangssignal wird mittels eines A/D-Wandlers umgestaltet, wobei das sinusförmige Signal als Mustereingangssignal dient. Die Analyse bestimmt die effektive Bitzahl des Meßpfades sowie das Verhältnis der Signalleistung und der Leistung des effektiven Ersatzfehlers des Meßpfades. Der Autor hat auch die Meßergebnisse unter Anwendung eines 10-bit-Abtastvoltmeters dargestellt.

А. КЖИВАЗНЯ, Я. ОЦЕПКА

ИСПОЛЬЗОВАНИЕ ДПФ В ДИНАМИЧЕСИКХ ТЕСТАХ СИСТЕМ АНАЛОГО-ЦИФРОВОГО ПРЕОБРАЗОВАНИЯ

Резюме

Приведена проблема тестования систем аналого-цифрового преобразования в динамических условиях. Метод измерений заключается в цифровой обработке ответа системы на входный синусоидальный сигнал при помощи дискретного преобразования Фурье. Вычисляется эффективное число битов системы и отношение мощности сигнала к мощности эквивалентного синала погрешности. Приведены результаты тестования 10-битового выборочного вольтметра.

ROZPRAWY ELEKTROTECHNICZNE, 1989.35, z. 1, ss. 101-124

53.087.354:535.332.5

Problemy automatyzacji projektowania optycznych układów obrazujących

JERZY M. WOŹNICKI

Instytut Mikroelektroniki i Optoelektroniki, Politechnika Warszawska

Otrzymano 1987.02.12

Autoryzowano do druku 1988.05.07

Artykuł jest wprowadzeniem w zagadnienia automatyzacji analizy i projektowania optycznych układów obrazujących. Omówiono metodę promieni jako podstawę modelowania zjawisk w układach. Przedstawiono propozycję systematyki parametrów charakteryzujących optyczne przetwarzanie obrazu. Sformułowano tezę o możliwości jednolitej metodyki analizy i projektowania układów optyki światła i optyki elektronowej. Wskazano na atrakcyjną możliwość opracowania wspólnego, uniwersalnego oprogramowania. Przedstawiono ogólną koncepcję bazowego systemu automatyzacji projektowania elektronowych układów przetwarzania obrazu.

1. WPROWADZENIE

W przeszłości optyka była traktowana jako nauka o świetle i jego oddziaływaniu z materią. Wnosiło to ograniczenie zakresu badawczego optyki do widzialnej części widma i to właśnie mogło stanowić kryterium wyróżnienia jej jako dyscypliny. Obecnie optyka falowa weszła w nadfiolet i podczerwień a także osiągnęła swą dojrzałość optyka korpuskularna. Różnica między optyką i innymi zbliżonymi dyscyplinami naukowymi tkwi więc już nie w zakresie częstotliwości, ale raczej w całokształcie specyficznych metod badawczych tradycyjnie związanych głównie z obserwacją i wykorzystaniem światła [1]. Podejście takie, które jak się wydaje, intuicyjnie jest dość powszechnie akceptowane, może być podstawą kwalifikacji problemów, metod lub parametrów jako optycznych lub mających inny charakter.

Postępująca ekspansja elektroniki w kierunku wysoko-częstotliwościowego zakresu widma znalazła swój wyraz w wielu rozwiązaniach technicznych z pogranicza elektroniki i optyki. Rozwój ten doprowadził do wykształcenia się optoelektroniki obrazowej jako dyscypliny zajmującej się analizą, syntezą i przetwarzaniem obrazów. Obraz jest przestrzennym rozkładem informacji niezależnie od fizycznej natury jej nośnika, którym mogą być: promieniowanie elektromagnetyczne, wiązka elektronowa, ale także np. potencjał elektrostatyczny, chemiczny, itd. Termin *informacja przestrzenna* jest więc szerszy od pojęcia *informacja optyczna*, które zakłada optyczny charakter nośnika. Nośnik informacji ma charakter optyczny w myśl przyjętej tu zasady kwalifikacji, jeśli metody właściwe

do jego opisu i analizy są wzorowane na metodach tradycyjnie związanych z obserwacją i wykorzystaniem światła. Optyczny charakter nośnika przesądza więc istnienie sprecyzowanej analogii jego opisu i właściwości z opisem i właściwościami światła.

Podstawowymi optycznymi nośnikami informacji są: wiązka fal elektromagnetycznych, a w tym zwłaszcza światło, oraz wiązka elektronowa jako nośnik elektrono-optyczny[1]. Zjawiska i zagadnienia związane z wykorzystaniem tych nośników informacji przestrzennej składają się na tzw. optyczne przetwarzanie informacji. Termin ten oznacza dziedzinę zajmującą się informacją przenoszoną przez przestrzenną a nie czasową modulację fali [2, 3].

Układ obrazujący lub przetwarzający obraz jest układem optycznym jeśli jego działanie opiera się na wykorzystaniu optycznego nośnika informacji lub mówiąc inaczej, jeśli przetwarza on informację optyczną. Z obrazowaniem mamy do czynienia wówczas, gdy informacja o stanie fizycznym, chemicznym itd. dowolnego przedmiotu punktowego (ściśle jego niewielkiego otoczenia będącego obszarem lokalego uśrednienia funcji rozkładu opisującej ten stan) z pewnej umownej przestrzeni przedmiotowej zostaje przeniesiona i zlokalizowana w innym niewielkim obszarze przestrzeni odwzorowania, wokół wyróżnionego w tym obszarze umownego obrazu punktowego. Niezależnie od konieczonści istnienia odpowiedniego nośnika informacji efekt obrazowania uwarunkowany jest właściwościami układu obrazującego. Postęp technologiczny umożliwiający wykonywanie elementów układów o pożądanych parametrach spowodował, że o jakości układu decyduje stopień jego optymalizacji konstrukcyjnej. Wysoka jakość współcześnie produkowanych optycznych układów obrazujących jest wynikiem postępu w zakresie automatyzacji analizy i projektowania tych układów oraz perfekcji technologicznej. Prace nad systemami komputerowego wspomagania projektowania optycznego CAOD — (ang. computer aided optical design) są już od lat prowadzone z narastającym rozmachem (patrz np. [4—8]).

Opracowanie niniejsze jest wprowadzeniem w zagadnienia CAOD w odniesieniu do układów obrazujących. Zawiera ono m.in.: charakterystykę ogólną optycznego układu obrazującego oraz optycznych nośników informacji, opis metody promieni jako narzędzia do analizy układów, systematykę parametrów optycznych, uwagi metodologiczne o modelowaniu i optymalizacji układów. W końcowej części przedstawiono koncepcję aktualnie opracowywanego przez autora we współpracy z zespołem, bazowego systemu automatyzacji projektowania dla elektronowych układów przetwarzania obrazu EPOS1.0.

2. OPTYCZNE UKŁADY OBRAZUJĄCE

Wśród optycznych układów obrazujących wyróżnić można układy z wiązką światła, a w tym układy konwencjonalne, układy holograficzne i układy gradientowe oraz układy z wiązką elektronów, a w tym układy z przetwarzaniem równoległym i sekwencyjnym.

Cechą charakterystyczną układów konwencjonalnych jest to, że składają się one z

[1] także wiązka jonowa, neutronowa itp., co jednak ma mniejsze znaczenie z punktu widzenia materii tego artykułu.

ciągu różnych ośrodków optycznych, z których każdy ma w zasadzie stały współczynnik załamania. Granice między ośrodkami, na których następuje skokowa zmiana współczynnika załamania są odpowiednio ukształtowane, stosownie do wymagań wynikających z obrazującej roli układu (patrz np. [9, 10]). Przykładami układów konwencjonalnych są zwierciadła i soczewki klasyczne. Od strony technologicznej układy te są bardzo dobrze rozwinięte.

Układy holograficzne są to układy optyczne z elementami holograficznymi, które mogą realizować różne funkcje, w tym także analogiczne do funkcji układów konwencjonalnych. Zwłaszcza jednak mogą one być elementami obrazującymi (patrz np. [11]). Istotne jest to, że funkcje te układy holograficzne realizują inaczej — za pośrednictwem efektów dyfrakcyjnych na ich siatkowej strukturze, a nie zjawisk odbicia i załamania, jak to ma miejsce np. w soczewkach klasycznych. Przykładami rozważanych tu układów holograficznych mogą być po prostu różnego rodzaju i przeznaczenia hologramy realizujące funkcje obrazujące. Stan wiedzy i techniki w tej dziedzinie, po okresie wielkiego rozwoju, jest już bardzo zaawansowany.

Układy gradientowe (układy optyki gradientowej[2]) są to optyczne ośrodki niejednorodne charakteryzujące się zmiennym przestrzennie współczynnikiem załamania światła (patrz np. [12]). Pwoduje to, że efekt ogniskowania wiązki światła emitowanej z punktowego źródła uzyskiwany jest w wyniku zmiany współczynnika załamania na całej drodze wiązki w obszarze niejednorodności przestrzennej. Przykładami układów gradientowych są: soczewka oka ludzkiego, atmosfera ziemska, soczewka Maxwella typu „*rybie oko*" lub soczewka Luneburga. Optyka GRIN weszła obecnie w fazę intensywnego rozwoju.

Układy obrazujące z wiązką elektronową można podzielić na układy z wiązką szerokokątną, która w sposób równoległy przenosi informację optyczną oraz układy z wiązką wąskokątną, która jest elementem rysującym lub wybierającym w przetworniku pracującym sekwencyjnie (patrz np. [3]). Przykładami przetworników z równoległym przenoszeniem informacji są wzmacniacze obrazu i konwertery obrazu, zaś z sekwencyjnym — widikony i kineskopy. Od strony technologicznej elektronooptyczne układy obrazujące są dobrze rozwinięte.

3. OPTYCZNE NOŚNIKI INFORMACJI, KONCEPCJA PROMIENI

Historycznie ukształtowały się dwa podstawowe modele światła: — model korpuskularny (kwantowy) i model falowy. Oba te podejścia różnią się zasadniczo zarówno pod względem formy jak i idei przewodniej modelu. Dla nas istotne jest to, że model falowy oferuje w odróżnieniu od kwantowego, determinizm pojęć, co umożliwia wygodny opis światła jako sterowanego nośnika informacji. Pełny opis wszystkich zjawisk związanych z tworzeniem i przenoszeniem obrazu wymaga konsekwentnego posługiwania się modelem falowym, jakkolwiek w większości przypadków wystarczające jest przybliżenie skalarne, oparte na równaniu Helmholtza:

$$\Delta U + k^2 U = 0 \tag{1}$$

[2] ang. gradient index optics (GRIN)

dla amplitudy zespolonej U fali świetlnej. Przybliżenie to umożliwia określenie rozkładu intensywności światła z uwzględnieniem efektów interferencyjnych i dyfrakcyjnych (patrz np. [13]).

Charakterystyczną cechą światła jest jego mała długość fali w porównaniu z liniowymi rozmiarami układów optycznych. Pozwala to na dalsze uproszczenie modelu światła jako nośnika informacji w tych układach. Uproszczenie to zapewnia optyka geometryczna światła będąca granicznym przypadkiem optyki falowej przy $\lambda \to 0$ (patrz np. [14]). Podstawowym równaniem optyki geometrycznej jest tzw. równanie eikonału:

$$\left(\frac{\partial S}{\partial x}\right)^2 + \left(\frac{\partial S}{\partial y}\right)^2 + \left(\frac{\partial S}{\partial z}\right)^2 = n^2(x, y, z), \tag{2}$$

gdzie $n(x, y, z)$ jest współczynnikiem załamania, zaś rodzina funkcji $S(x, y, z) = conts$ określa ciąg powierzchni stałej fazy. Przybliżenie optyki geometrycznej zachowuje słuszność, gdy amplituda A fali świetlnej spełnia warunek $\nabla A/A \ll 1/\lambda^2$ (gdzie ∇ jest symbolem laplasjanu), co oznacza wymóg dostatecznie powolnej zmiany intensywności światła. W przybliżeniu optyki geometrycznej pole świetlne jest określone w przestrzeni w sensie geometrycznym.

Duże znaczenie praktyczne ma pojęcie trajektorii ortogonalnych do geometrycznych frontów falowych $S(x, y, z) = const$. Trajektorie te nazywają się promieniami świetlnymi. Pojęcie to ściśle nie obejmuje niczego prócz kierunku, który jest zgodny z lokalnym kierunkiem propagacji fali (przenoszenia energii, przenoszenia informacji). Możliwe jest jednak przypisanie promieniom świetlnym pewnych parametrów falowych, takich jak długość fali, prędkość propagacji, intensywność. Pojęcie pojedynczego promienia z fizycznego punktu widzenia jest niekonsekwentne. Przybliżenie optyki geometrycznej abstrahuje od zjawisk interferencji i dyfrakcji.

Podstawowe równania optyki geometrycznej mogą być wyprowadzone z równania Helmholtza lub z zasady Fermata. To drugie podejście wskazuje na związek między optyką promieni i mechaniką klasyczną w ujęciu Hamiltona. Porównanie zasady najmniejszej drogi optycznej Fermata z zasadą najmniejszego działania pozwala wprowadzić analogię między wiązką promieni świetlnych i wiązką elektronów (patrz np. [15, 16]). Analitycznym wyrazem tej analogii jest elektrono-optyczny współczynnik załamania μ zależny w ogólnym przypadku od rozkładu pól elektrycznego i magnetycznego w układzie zgodnie ze wzorem:

$$\mu = \sqrt{\frac{V(1+\varepsilon V)}{V_0}} - \frac{e}{p_0}\vec{A}\cdot\vec{s}_0 \tag{3}$$

gdzie V jest potencjałem pola elektrycznego, V_0 wartością potencjału odpowiadającą prędkości początkowej elektronu startującego z punktu o potencjale zerowym, $p_0 = \sqrt{2em_0V_0}$ jest pędem początkowym elektronu, $\vec{A}$ jest potencjałem wektorowym pola magnetycznego, $\vec{s}_0$ jest wektorem jednostkowym stycznym do toru elektronu, $\varepsilon = e/2m_0c^2$ jest współczynnikiem poprawki relatywistycznej, zaś m_0, e są masą spoczynkową i ładunkiem elektronu.

Podobnie jak to ma miejsce w przypadku światła istnieją także dwa modele opisu ruchu cząstek materialnych: model korpuskularny i model falowy. Granice stosowalności

balistyki elektronowej w modelu korpuskularnym wynikają z relacji Heisenberga, która wnosi ograniczenia dla jednoczesnej oznaczoności pędu i położenia, co tkwi u podstaw modelu korpuskularnego. Szczegółowe rozważania wskazują na pełną stosowalność tego modelu w przypadku rozważanych tutaj układów elektrono-optycznych przetworników obrazu (nie mówimy tu o mikroskopii elektronowej). W ten sposób dochodzimy do modelu geometrycznej optyki elektronowej, w ramach której przyjmuje się, że wiązka elektronowa składa się ze zbioru torów elektronów, który jest zbiorem krzywych geometrycznych, wzdłuż których poruszają się elektrony [17]. Te krzywe są odpowiednikiem trajektorii określających promienie świetlne. Dlatego często operuje się pojęciem promieni elektronowych. I tak oto na gruncie ogólnej optyki geometrycznej uzyskuje się możliwość jednolitości opisu zjawisk związanych z optycznym przetwarzaniem informacji. Należy jednak podkreślić, że rozciągnięcie zasady jednolitości podejścia na metody symulacji zjawisk fizycznych w układach obrazujących wymaga ograniczenia rozważań do przetworników elektrono-optycznych bez pól magnetycznych wprowadzających składnik anizotropowy do elektrono-optycznego współczynnika załamania, co nie ma analogii w rozpowszechnionych układach obrazujących optyki światła. Nie jest to jednak ograniczenie krytyczne ze względu na to, że najbardziej interesujące tutaj układy elektrono-optyczne z przetwarzaniem równoległym (np. wzmacniacze obrazu) są współcześnie zdominowane przez konstrukcje z polem elektrostatycznym (EWO).
W dalszej części artykułu mówić będziemy zatem o układach bez pola magnetycznego.

Analogia między propagacją światła w ośrodkach bezstratnych w przybliżeniu optyki geometrycznej i ruchem cząstek naładowanych w polach elektrycznym i magnetycznym jest klasycznym i dobrze opisanym zagadnieniem w optyce (np. [15—17]).

4. METODA PROMIENI JAKO NARZĘDZIE DO ANALIZY UKŁADÓW OBRAZUJĄCYCH

Metoda promieni, zwana także, zwłaszcza w odniesieniu do układów elektrono-optycznych, metodą torów, polega na analizie zbiorów promieni w rozważanym układzie obrazującym. Każdy promień scharakteryzowany jest przez: swoją trajektorię w układzie lub przez wybrane jej punkty charakterystyczne i wersor styczny, przez związane z nim parametry fizyczne, a także przez wartość funkcji wagowej wynikającej z zasady reprezentacji wiązki. Trajektorie poszczególnych promieni określane mogą być metodami analitycznymi, doświadczalnymi, numerycznymi lub metodami kombinowanymi. Nie tylko materia tego opracowania, ale przede wszystkim zupełnie jednoznaczne tendencje światowe pozwalają ograniczyć, rozważania do metod numerycznych, które właściwie wyparły wszystkie inne. Metody te stosuje się do rozwiązywania równań trajektorii. Trajektorie te mogą być określone wychodząc z równania promieni (Eulera) wyprowadzonego najpierw dla promieni świetlnych [13]:

$$\frac{\mathrm{d}}{\mathrm{d}s}\left(n\frac{\mathrm{d}\vec{r}}{\mathrm{d}s}\right) = \nabla n, \tag{4}$$

gdzie $n = n(\vec{r})$ jest współczynnikiem załamania, $\vec{r}$ — jest wektorem położenia, zaś s jest zmienną położenia wzdłuż trajektorii, określającą jej długość od zadanego punktu. Równanie to w postaci (4) jest bezpośrednio przydatne w przypadku ośrodków o ciągłej zmianie współczynnika załamania. W odniesieniu do klasycznych układów optyki światła w praktyce stosowana jest wektorowa postać prawa załamania (Snella):

$$\vec{e}_n x \left[n_1 \left(\frac{d\vec{r}}{\mathrm{ds}} \right)_1 \right] = \vec{e}_n x \left[n_2 \left(\frac{\mathrm{d}\vec{r}}{\mathrm{ds}} \right)_2 \right], \tag{5}$$

gdzie $\vec{e}_n$ jest wektorem jednostkowym lokalnie prostopadłym do granicy (zadanej lub hipotetycznej) dwóch ośrodków o współczynnikach załamania n_1, n_2 (patrz np. [18]). Równanie (5) przy założeniu infinitezymalnej różnicy współczynników załamania w ośrodku 1 i 2 może być przekształcone do postaci (4) [15].

Dla trajektorii elektronowych równaniem wyjściowym jest oczywiście równanie ruchu (Newtona):

$$\frac{\mathrm{d}}{\mathrm{dt}} \left(m \frac{\mathrm{d}\vec{r}}{\mathrm{dt}} \right) = e \nabla V \tag{6}$$

gdzie $m = m_0 / \sqrt{1 - (v/c)^2}$. Z równania ruchu (6) oraz z równania

$$m_0 c^2 \sqrt{1 - \left(\frac{v}{c} \right)^2} - m_0 c^2 = eV \tag{7}$$

wynikającego z zasady zachowania energii, po podstawieniu $\vec{v} = v \cdot \mathrm{d}\vec{r}/\mathrm{ds}$ otrzymuje się równanie promieni elektronowych w postaci:

$$\frac{\mathrm{d}}{\mathrm{ds}} \left(\sqrt{V(1 + \varepsilon V)} \frac{\mathrm{d}\vec{r}}{\mathrm{ds}} \right) = \frac{1 + 2\varepsilon V}{2\sqrt{V(1 + \varepsilon V)}} \nabla V \tag{8}$$

Nietrudno sprawdzić, że równanie (8) przechodzi w równanie (4) po podstawieniu wielkości $\mu = \sqrt{V(1 + \varepsilon V)}/V_0$ (przypadek $\vec{A} = 0$ — brak pola magnetycznego) w miejsce współczynnika n.

Oznacza to, że równanie (4) jak i równania (6) i (8) są słuszne w przybliżeniu optyki geometrycznej zarówno dla układów optyki światła jak i optyki elektronowej. Współczynnik załamania dla promieni elektronowych jest proporcjonalny do prędkości elektronów. Współczynnik załamania dla promieni świetlnych jest proporcjonalny do odwrotności prędkości propagacji frontu falowego. Przy czym podkreślenia wymaga fakt, że prędkość elektronu jest tu prędkością grupową paczki falowej fal materii de Broglie'a reprezentowanej przez elektron. Prędkość światła zaś jest prędkością fazową frontu falowego. W przypadku określania drogi promieni poprzez całkowanie równania ruchu (6) konieczne jest więc odpowiednie przeliczenie czasowego kroku procedury numerycznej całkowania (bazując na równości przyrostu drogi) w zależności od tego, czy mamy do czynienia z trajektorią elektronową, czy świetlną.

Zagadnienie numeryczne dla metody promieni w zastosowaniu do układów niejednorodnych (elektrono-optycznych, gradientowych i holograficznych z falami niejednorodnymi [19, 20]) od strony matematycznej polega na rozwiązaniu zagadnienia początkowego dla równania różniczkowego, zwyczajnego rzędu drugiego, do którego daje się

sprowadzić każde z równań (4), (6), (8). Zakłada się, że rozkład przestrzenny uogólnionego współczynnika załamania w układzie jest znany. Dla układów elektronooptycznych wymaga to wyprzedzającego obliczenia rozkładu potencjału pola elektrycznego w węzłach zadanej sieci obliczeniowej, a następnie aproksymacji składowych gradientu tego potencjału. Istnieje metoda wymagająca określenia także drugich pochodnych rozkładu potencjału dzięki czemu uzyskuje się lepsze narzędzie do wyznaczania aberracji [21].

Problem numeryczny dla metody promieni jest zagadnieniem klasycznym i dobrze opisanym w literaturze (patrz np. [22] oraz pozycje książkowe w języku polskim np. [23—25]). Mimo to jednak nadal pojawiają się doniesienia o nowym podejściu do tego problemu, przynoszącym konkretne korzyści (np. [21, 26]). Typowe, stosowane metody numeryczne to metoda Rungego-Kutty, metody ekstrapolacyjno-interpolacyjne, a w tym metody zmodyfikowane, metoda ekstrapolacji wymiernej itp. Algorytmy muszą przewidywać możliwość automatycznej zmiany długości przestrzennego kroku całkowania w zakresie do 6 rzędów wielkości. Stwarza to problem optymalizacji procedur i programów [27].

W zastosowaniu do klasycznych układów obrazujących procedura metody promieni jest odmienna. Nie jest ona oparta na metodach numerycznych rozwiązywania równań różniczkowych, lecz wykorzystuje wektorową postać prawa załamania (patrz r-nie(5)). Droga promienia wyznaczana jest jedynie poprzez wyliczanie jego położenia i kierunku na kolejnych powierzchniach załamujących (odbijających) w układzie aż do osiągnięcia przez promień płaszczyzny (powierzchni) obrazu. Punkty na trajektorii leżące między poszczególnymi powierzchniami nie są obliczane, ponieważ powierzchnie te są oddzielone przez jednorodny ośrodek optyczny (n = const), w którym tor promienia jest prostoliniowy. Zagadnienie wytyczania biegu promienia w układach klasycznych nie jest trywialne, albowiem dotyczy promieni dowolnych i powierzchni asferycznych.

Istnieją dwie podstawowe przyczyny, dla których prace nad poszukiwaniem coraz doskonalszych narzędzi numerycznych dla metody promieni są ciągle kontynuowane. Są to:

— krytyczne wymagania na dokładność obliczeń,

— istotne ograniczenia na czas obliczeń.

Dopuszczalny błąd określenia położenia trajektorii w płaszczyźnie obrazowej układu musi być co najmniej o rząd wielkości mniejszy od średnich rozmiarów krążka rozproszenia. Oznacza to, że w układach o rozmiarach rzędu 10^{-1} m błąd ten powinien być rzędu $10^{-7}-10^{-6}$ m.

Ograniczenia na czas obliczeń pojedynczej trajektorii promienia wynikają z konieczności obliczania 10^4—10^5 i więcej trajektorii w procesie optymalizacyjnym. Szybkość obliczeń syntetycznie wyraża się nieco inaczej w odniesieniu do układów klasycznych i układów niejednorodnych (w tym elektrono-optycznych). W przypadku układów klasycznych miarą szybkości jest liczba przeliczonych powierzchni na drodze promienia w jednostce czasu. W przypadku układów niejednorodnych miarą szybkości jest liczba przeliczonych punktów trajektorii w jednostce czasu. Liczba tych punktów zawierać się może, ze względu na wymagane dokładności, w przedziale między 100 a 300, jak to ma miejsce w układach elektrono-optycznych o dużych gradientach pól (np. EWO). Ilość tych punktów jest więc znacząco większa od liczby powierzchni w układach klasycznych. Jeśli uwzględnić ponadto,

że średnia czasochłonność obliczania punktu trajektorii w układzie elektrono-optycznym przewyższa co najmniej o rząd wielkości czasochłonność obliczania parametrów trajektorii promienia świetlnego na jednej powierzchni, to otrzymuje się miarę stopnia trudności dla twórcy systemów CAOD dla układów optyki elektronowej.

Osobnym zagadnieniem związanym z praktycznym stosowaniem metody torów w celu wyznaczania optycznych parametrów układów, jest symulacja generacji wiązki promieni określana także jako symulacja emisji w odniesieniu do układów elektrono-optycznych. Generacja wiązki obejmuje następujące zagadnienia szczegółowe:

— narzucanie warunków początkowych dla pojedynczego promienia,

— zapewnienie reprezentacji wiązki fizycznej, selekcja promieni,

— określenie funkcji wagowych dla wszystkich parametrów początkowych dla poszczególnych promieni,

— organizacja procedur generacji wiązki.

Wszystkie wymienione problemy są wspólne dla obrazujących układów optycznych ze świetlnym i elektronowym nośnikiem informacji. Wybór metody promieni jako podstawy modelowania umożliwia więc przyjęcie jednolitego podejścia do projektowania optycznych układów obrazujących bez względu na rodzaj użytej wiązki. Oczywiście dany rodzaj nośnika informacji wnosi także pewne zagadnienia specyficzne, takie jak np. dyfrakcja dla układów świetlnych, która nie ma odpowiednika w układach wzmacniaczy lub konwerterów obrazu (nie mówimy tu o mikroskopii elektronowej) lub ładunek przestrzenny w układach elektrono-optycznych. Powoduje to jednak jedynie konieczność uzupełnienia ogólnie obowiązujących algorytmów o pewne procedury specyficzne.

5. SYSTEMATYKA PARAMETRÓW CHARAKTERYZUJĄCYCH OPTYCZNE PRZETWARZANIE OBRAZU

Systematyka przedstawiona niżej ma charakter ogólny i nie zależy od rodzaju układu przetwarzania informacji przestrzennej oraz od fizycznej natury samego obrazu. Obejmuje ona także układy z czasowym przetwarzaniem obrazów.

Zaproponowana systematyka nawiązując do pewnych propozycji literaturowych [28] stanowi ich znaczne rozwinięcie. Parametry charakteryzujące proces przetwarzania obrazu rozdzielone zostały na trzy grupy odnoszące się do:

— nośnika informacji przestrzennej (wiązki) lub jego źródła fizycznego lub umownego, jakim może być obiekt fizyczny, źrenica wejściowa układu lub fotokatoda; są to więc parametry odnoszące się do wiązki i jej generacji,

— układu przetwarzania obrazu, np. zbioru soczewek, układu elektrod itp., oraz

— samego obrazu.

Niezależnie od fizycznej natury obrazu i jego nośnika, a także od procesów, w wyniku których jest on przetwarzany, nośnik obrazu, układ przetwarzania jak również sam obraz mogą być charakteryzowane za pomocą parametrów należących do trzech grup:

— grupy parametrów przestrzennych związanych z rozkładem i upakowaniem informacji strukturalnej,

— grupy parametrów intensywnościowych określających ilość informacji metrycznej,

Systematyka parametrów charakteryzujących optyczne przetwarzanie obrazu

Triada	Grupa	Parametry nośnika obrazu-wiązki, lub jej źródła fizycznego/umownego			Parametry układu przetwarzania obrazu				Parametry obrazu	
przedziałów	przestrzenna	kształt i rozmiary wiązki przy pełnym pobudzeniu	rozkłady energetyczne lub fotometryczne (światłość, emitancja, luminancja) przy pełnym pobudzeniu	funkcja źrenicy we	rozmiary pola widzenia	powiększenie	statystyczna powierzchnia obrazowa (SIS)	dystorsja	rozmiary geometryczne	
		geometria apertury we			zakres dynamiczny	OTF		przestrzenno-czasowa funkcja przenoszenia	kontrast	
	intensywnościowa	graniczne wartości energetyczne, czułość								
		wydajność energetyczna	intensywność w szerokokątnej wiązce							
	czasowa	czas przenoszenia obrazu			czas pracy	zmiana skali czasu			czas istnienia	
		czas pracy, trwałość								
wartości początkowych	przestrzenna	współrzędne początkowe źródła	przestrzenno-intensywnościowe rozkłady początkowe		współrzędne początkowe pola widzenia				współrzędne początkowe	
	intensywnościowa	próg czułości źródła			próg czułości określany na wyjściu				intensywność tła	
	czasowa	parametr bezwładności wiązki			moment zaistnienia rejestracji (przetworzenia)				moment zaistnienia	
szczegółów	przestrzenna	stopień koherencji przestrzennej	rozkłady przestrzenno-intensywnościowe w wiązce przy pobudzeniu punktowym		aberracje geometryczne, rozmiary krążka rozproszenia, przestrzenna zdolność rozdzielcza				rozmiar elementu rozkładu (analizy)	ostrość, liczba elementów obrazu
	intensywnościowa	rozkład chromatyczny			fizyczna zdolność rozdzielcza, PSF, LSF, ESF				poziom szumu przestrzennego	gradacja szarości
					progowy kontrast					
	czasowa	stopień koherencji czasowej			czas rejestracji kadru	czas magazynowania informacji	czas wymazywania informacji		czas trwania kadru	liczba kadrów
		częstotliwość (długość fali)								

Uwaga: Termin „parametry" użyty został w szerokim sensie i odnosi się także do funkcji charakterystycznych

— grupy parametrów czasowych.

Wśród parametrów należących do każdej z tych grup wyróżnić można parametry:

— odnoszące się do przedziałów pewnych wielkości,

— określające wartości progowe,

— charakteryzujące szczegóły.

Pozwala to na wprowadzenie zgodnie z tym podziałem triad składających się z wcześniej wprowadzonych grup parametrów (Tablica 1).

Zaproponowana systematyka wprowadza pewien porządek, co może ułatwić specjalistom z różnych dziedzin współpracę nad zaprojektowaniem i wykonaniem nowej konstrukcji układu obrazującego. Wszystkie wymienione parametry mogą być zdefiniowane w sposób ścisły, jakkolwiek w niektórych przypadkach umowny. Przedstawienie definicji parametrów wykracza poza ramy tego artykułu.

W tablicy I nie wymieniono wszystkich możliwych parametrów ani nie wprowadzono wszystkich możliwych podziałów parametrów. Jest oczywiste, że każda klasyfikacja zależy od przyjętego kryterium i, co za tym idzie, ma charakter umowny.

W procesie projektowania wykorzystywane są odpowiednie kryteria optymalizacyjne. Przyjmowane funkcje celu niekiedy charakteryzuje złożoność formalna i brak interpretacji fizycznej. Często jednak jest tak, że za podstawę optymalizacji przyjmuje się funkcję lub wyrażenie interpretacyjne jasne z punktu widzenia jakości układu obrazującego. Tego rodzaju wielkości mają charakter parametrów syntetycznych układu. Przykładem parametru opartego na MFT może być całka [29]:

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \mathrm{MTF}(N_x, N_y)\mathrm{dN}_x\mathrm{dN}_y \tag{9}$$

Przykładami parametrów syntetycznych opartych na pojęciu aberracji falowej W, mogą być jej wariancja i wartość średnia [30]. Jeszcze innym często używanym parametrem jest radialny rozkład poprzecznych błędów (lub aberracji) promienia[3)] (np. [31]). W większości przypadków przyjmowane kryteria optymalizacyjne wykorzystują parametry optyczne charakteryzujące jakość układu z punktu widzenia podstawowej jego roli jaką jest funkcja obrazująca. Zgodnie z przyjętą wcześniej podstawą kwalifikacji, przez parametry optyczne rozumiemy także parametry, które zostały wprowadzone i są tradycyjnie używane w optyce światła lub są na nich wzorowane. Przykładami parametrów optycznych, które są argumentami dla złożonych funkcji celu kryteriów optymalizacyjnych są funkcje rozmycia, funkcje przenoszenia, aberracje, rozdzielczość itp.

W tablicy I pominięto parametry syntetyczne podobnie jak i inne parametry niestandardowe jakie bywają stosowane do oceny jakości układu np. parametry proponowane przez metody statystyczne analizy rozkładów fizycznych (momenty funkcji PSF i inne).

6. PARAMETRY WYKORZYSTYWANE W METODZIE PROMIENI

Metoda promieni posługuje się wyselekcjonowaną grupą parametrów. Klasyfikację tych parametrów zestawiono w tablicy 2. Podział zaproponowany w tej tablicy wynika

[3)] ang. transverse ray errors, transverse ray aberrations

z ogólnej procedury metody promieni w zastosowaniu do analizy i projektowania optycznych układów obrazujących. Stąd klasyfikacja wyróżnia parametry związane z generacją wiązki, oraz parametry wyznaczone numerycznie — wyjściowe w procesie modelowania. Obok parametrów konwencjonalnych tablica 2 zawiera pewne parametry nowe lub mniej znane.

Tablica 2

Parametry stosowane w metodzie promieni

Rodzaj parametrów	Parametry przestrzenne	Parametry intensywnościowe lub przestrzenno-intensywnościowe	Parametry syntetyczne
Parametry nośnika obrazu — wiązki, lub jej źródła fizycznego/ /umownego	Kształt i wymiary apertury źrenicy wejściowej: Parametry geometryczne źródła	Funkcja źrenicy. Rozkłady: powierzchniowy, kierunkowy i chromatyczny promieni w wiązce na powierzchni źródła wiązki	Liczba promieni w wiązce
Parametry optycznych układów obrazujących — wejściowe w procesie analizy	Dane konstrukcyjne, wymiary	Rozkład (uogólnionego) współczynnika załamania	Liczba elementów układu np. soczewek, elektrod itp.
Parametry optycznych układów obrazujących — wyjściowe w procesie analizy	Miary poszczególnych aberracji geometrycznych. Rozmiary krążka rozproszenia. Przestrzenna zdolność rozdzielcza. Powiększenie, dystorsja.	Funkcje rozmycia PSF, LSF, ESF Funkcje przenoszenia OTF (MTF, PTF) Statystyczna powierzchnia obrazowa SIS. Fizyczna rozdzielczość. Miara aberracji chromatycznej. Parametry fazowe.	Całkowy parametr jakości przenoszenia modulacji. Aberacja falowa. Parametry statystyczne funkcji rozkładów fizycznych.

Należą do nich:

— statystyczna powierzchnia obrazowa (SIS) zdefiniowana formalnie w [32, 33]

— statystyczne parametry wybranych funkcji rozkładów wielkości fizycznych charakteryzujących jakość obrazowania (np. drugi moment PSF [34])

— całkowy parametr jakości przenoszenia modulacji przestrzennej (parametr syntetyczny określony wzorem (9))

Zagadnienie wyboru kryterium optymalizacji w automatycznym projektowaniu układów optycznych zostało szczegółowo przedyskutowane w [35].

7. MODELOWANIE I OPTYMALIZACJA W SYSTEMACH AUTOMATYZACJI PROJEKTOWANIA OPTYCZNYCH UKŁADÓW OBRAZUJĄCYCH

Metoda promieni jest podstawą modelowania matematycznego w systemach automatyzacji projektowania optycznych układów obrazujących. Modelowanie to dotyczy zjawisk fizycznych w tych układach, zaś jego celem jest obliczenie ich parametrów optycznych. Funkcjonalnie, modelowanie obejmuje następujące kroki:

— generację wiązki promieni

— wyznaczenie kształtu i położenia trajektorii wygenerowanych promieni oraz odpowiednio wybranych punktów charakterystycznych na tych trajektoriach,

— analizę wyznaczonych punktów charakterystycznych (obróbkę statystyczną wyników),

— określenie pożądanego, optycznego parametru układu.

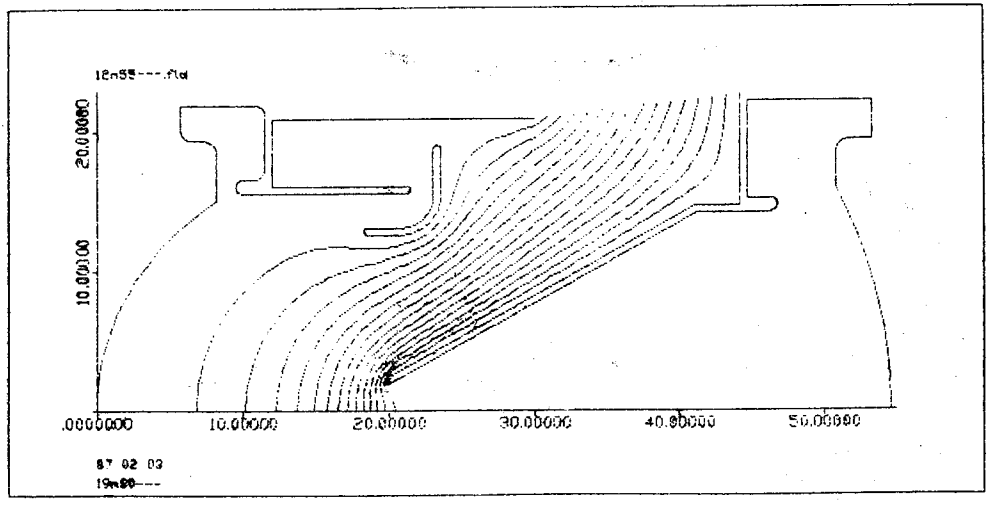
Generacja wiązki promieni odgrywa bardzo istotną rolę w procesie modelowania. Wybór promieni analizowanych następnie w kolejnych krokach postępowania decyduje o wydajności algorytmu. Zbiór promieni reprezentatywnych dla wybranej metody modelowania danego parametru optycznego jest scharakteryzowany przez parametry wymienione w tablicy 2. Wyznaczanie kształtu i położenia trajektorii następuje poprzez rozwiązanie równania różniczkowego — jak to ma miejsce w ośrodkach niejednorodnych, np. elektrono-optycznych lub gradientowych, lub metodami wytyczania biegu promienia, metodami algebraicznymi (macierzowymi) itd. w innych przypadkach.

Analiza punktów charakterystycznych zbiorów lub podzbiorów promieni dotyczy najczęściej punktów końcowych trajektorii na zadanej płaszczyźnie obrazu. Procedura ta może obejmować wstępne, aproksymacyjne rozszerzanie zbyt małych zbiorów tych punktów uzyskanych metodą czasochłonnych obliczeń numerycznych [36]. W ten sposób możliwa staje się analiza dziesiątków tysięcy punktów na podstawie obliczonych setek punktów. Przykładem wydajnej procedury tego rodzaju jest algorytm obliczania MTF wzmacniacza obrazu dla przypadku oświetlenia dowolnego [37]. Uzyskano tu 20—30-krotne skrócenie czasu obliczeń.

Analizowane punkty końcowe mogą być bazą danych m.in. dla procedur przekształcania jednych dwuwymiarowych rozkładów w inne, mogą być podstawą do obliczenia współczynników aberracyjnych układów, lub też są bezpośrednio zliczane w elementarnych obszarach określonych lokalnie w płaszczyźnie obrazowej, co umożliwia wyznaczenie dowolnej funkcji rozmycia (np. PSF). Optyczną funkcję przenoszenia (OTF) układu określa transformata Fouriera funkcji PSF. Parametry obliczane dla przypadku oświetlenia polichromatycznego określane są jako średnia ważona wartości parametrów obliczanych kolejno dla różnych długości fali. Efekty dyfrakcyjne w układach z wiązką światła powinny być uwzględnione dodatkowo.

Opisana procedura ma charakter ogólny i może być podstawą modelowania zarówno w odniesieniu do układów z wiązką światła jak i z wiązką elektronów. Przykładem możliwości stosowania jednolitej metodyki modelowania własności optycznych układów obrazujących jest numeryczne obliczanie funkcji MTF, która jest podstawowym kryterium jakości układu obrazującego (np. [38]). Dla wybranego rodzaju układu (np. układu soczewek klasycznych, układu wzmacniacza obrazu (EWO)) o znanym (obliczonym wcześ-

niej) współczynniku załamania, realizowana jest generacja wiązki promieni z wejściowej powierzchni generacji (źródła). Dla układu soczewek powierzchnię tę stanowi źrenica wejściowa z zadanym rozkładem powierzchniowym promieni świetlnych. Dla układu wzmacniacza obrazu powierzchnią tą jest fotokatoda z zadanym punktem generacji promieni elektronowych o określonych rozkładach początkowych. Następnie obliczane są trajektorie poszczególnych promieni oraz ich punkty końcowe w płaszczyźnie obrazowe układu. Kolejno punkty te (być może po rozszerzeniu ich zbioru) są zliczane w elementarnych obszarach analizy, co pozwala określić PSF i na tej podstawie MTF. Tak wyznaczona charakterystyka MTF winna być odpowiednio skorygowana. W układzie soczewek klasycznych wymaga to przemnożenia jej przez funkcję MTF będącą skutkiem jedynie zjawisk dyfrakcyjnych w danym układzie. Funkcję tę określają: średnica źrenicy wejściowej i odległość płaszczyzny obrazowej od źrenicy wejściowej a także długość fali [29]. W układzie wzmacniacza korekcja wiąże się z uwzględnieniem innych elementów w torze przetwarzania niż sam układ elektrono-optyczny (np. płytki światłowodowa i mikrokanalikowa, ekran). Obliczoną funkcję MTF należy więc przemnożyć przez funkcję MTF syntetycznie ujmującą wpływ tych innych elementów. Przykładowe wyniki praktycznego zastosowania metody promieni do modelowania [39] zawierają Rys. 1-4, na których przedstawiono kon-

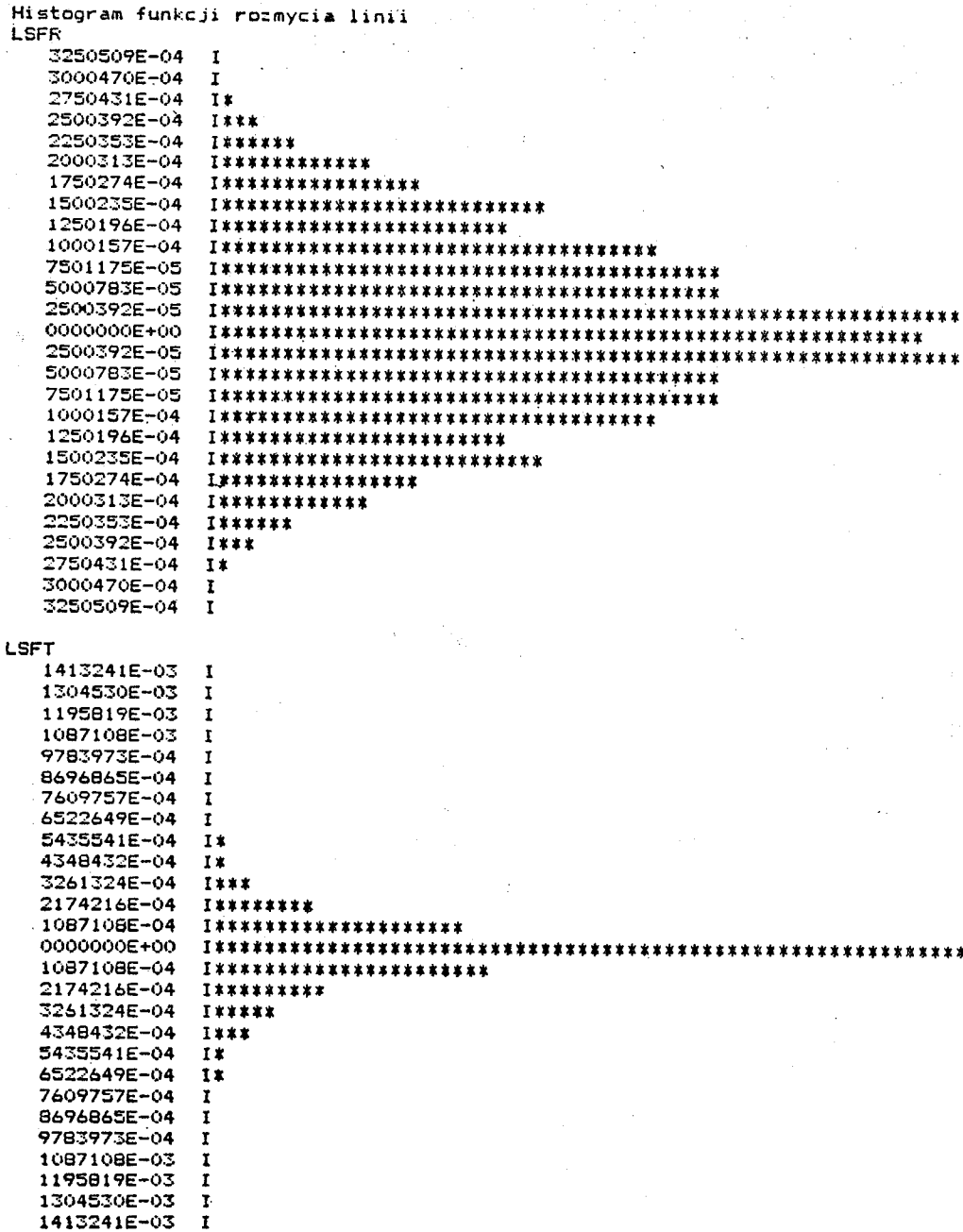


Rys. 1. Konfiguracja modelowanego elektrono-optycznego wzmacniacza obrazu (EWO) wraz z rozkładem linii ekwipotencjalnych ogniskującego pola elektrostatycznego

figurację EWO, dla którego przeprowadzono obliczenia (Rys. 1) oraz otrzymane wyniki: charakterystyki LSF i MTF dla kierunku radialnego i tangencjalnego (Rys. 2, 3) graficzną reprezentację aberracyjnego krążka rozmycia (Rys. 4).

Określenie parametrów optycznych dla ustalonej konfiguracji konstrukcyjnej układu jest zadaniem podsystemów analizy w systemach CAOD. Jednak właściwym przeznaczeniem całego systemu CAOD jest automatyzacja projektowania [40—42]. Funkcjonalnie obejmuje to:

— optymalizację konstrukcji układu tj. dobór jego parametrów konstrukcyjnych,



Rys. 2. Obliczone histogramy funkcji rozmycia linii (LSF) dla kierunku radialnego i tangencjalnego odpowiednio

takich jak kształt elementu, wymiary podstawowe np. odległość, promienie krzywizny, kąty, typ szkła itp.,

— określenie tolerancji (na podstawie analizy wymiarowej) elementów konstrukcyjnych,

— oszacowanie wpływu asymetrii w układzie, decentrycznego montażu elementów itp.

Rozwiązywanie tych zagadnień wymaga stosowania odpowiednich procedur optymalizacyjnych, które mogą być wprost oparte na znanych metodach matematycznych takich jak na przykład metody gradientowe. Rozwiązaniem koncepcyjnie najprostszym jest organizacja pętli, w której układ jest przeliczany wielokrotnie ze sterowaną zmianą poszczególnych parametrów. Jednak wielka czasochłonność takiego podejścia zwłaszcza w przypadku optymalizacji układów elektrono-optycznych skłania do poszukiwania szybszych metod. Osiągnąć to można poprzez wspieranie postępowania optymalizacyjnego elementami analizy układów od strony fizycznej. Wykorzystuje się m.in. metody rachunku zaburzeń, metody szeregów dla parametrów syntetycznych itp.

```
Zakres czestotliwosci przestrzennych NMAX =  62
Podaj wartosc przyrostu czestotliwosciowej zmiennej niezaleznej
dla obliczania MTF:              2.0000000

        W Y N I K I   O B L I C Z A N I A   M T F

w kierunku radialnym

  FREQ    MTF     ....................................................................
   .00  1.000     .####################################################################
  2.00   .991     .####################################################################
  4.00   .965     .###################################################################
  6.00   .922     .#################################################################
  8.00   .865     .##############################################################
 10.00   .796     .###########################################################
 12.00   .719     .#####################################################
 14.00   .635     .##############################################
 16.00   .550     .#######################################
 18.00   .465     ############################
 20.00   .384     .#####################
 22.00   .309     .################
 24.00   .243     .############
 26.00   .185     .#########
 28.00   .138     .#######
 30.00   .100     .#####
 32.00   .073     .###
 34.00   .054     .##
 36.00   .042     .##
 38.00   .037     .#
 40.00   .036     .#
 42.00   .039     .#
 44.00   .047     .##
 46.00   .048     .##
 48.00   .053     .##
 50.00   .057     .##
 52.00   .060     .###
 54.00   .061     .###
 56.00   .061     .###
 58.00   .059     .###
 60.00   .057     .##
 62.00   .053     .##
```

```
w kierunku tangencjalnym

  FREQ    MTF     .....................................................................
   .00  1.000     .#####################################################################
  2.00   .967     .###################################################################
  4.00   .883     .##############################################################
  6.00   .781     .#########################################################
  8.00   .686     .#################################################
 10.00   .607     .#############################################
 12.00   .542     .#########################################
 14.00   .487     .#####################################
 16.00   .441     .#################################
 18.00   .400     .#############################
 20.00   .362     .############################
 22.00   .330     .##########################
 24.00   .304     .########################
 26.00   .286     .#######################
 28.00   .273     .######################
 30.00   .262     .######################
 32.00   .251     .#####################
 34.00   .239     .####################
 36.00   .227     .####################
 38.00   .217     .###################
 40.00   .211     .###################
 42.00   .209     .###################
 44.00   .210     .###################
 46.00   .210     .###################
 48.00   .210     .###################
 50.00   .209     .###################
 52.00   .211     .###################
 54.00   .217     .###################
 56.00   .227     .####################
 58.00   .239     .####################
 60.00   .251     .#####################
 62.00   .262     .######################

Czas obliczen:  49 min  25 sek

Czy kontynuacja obliczen i wykresy dla nowej wartosci
przyrostu czestotliwosciowej zmiennej niezaleznej (T/F)?:   F
```

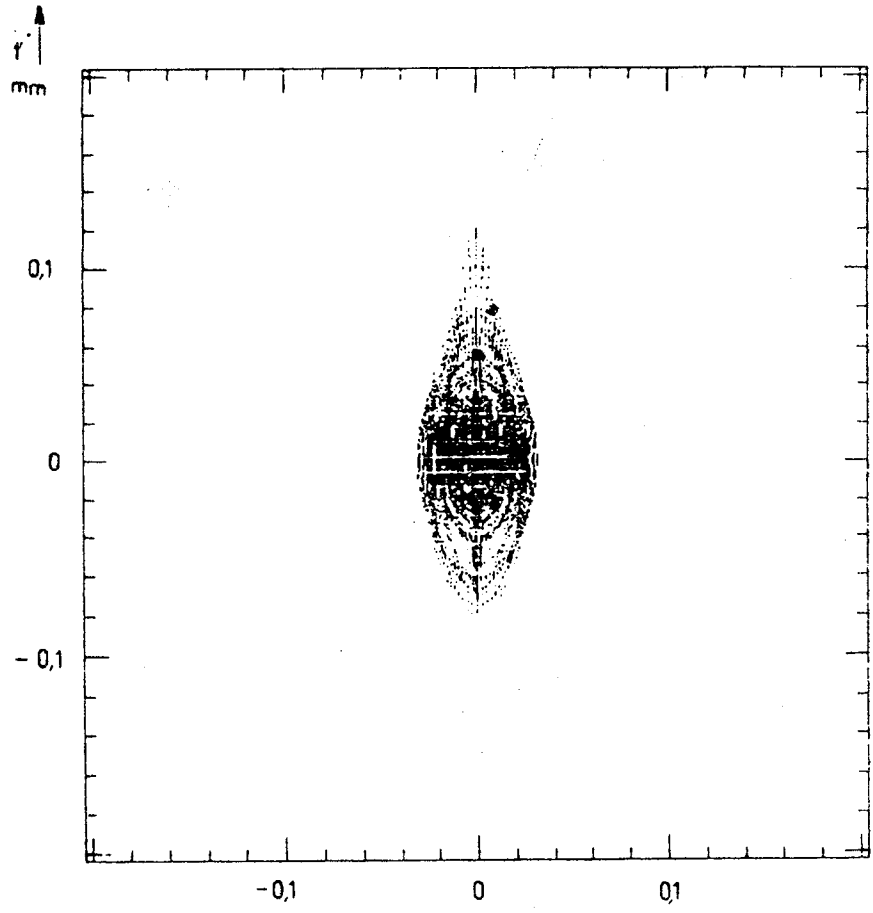
Rys. 3. Obliczone histogramy funkcji przenoszenia modulacji MTF a) dla kierunku radialnego, b) dla kierunku tangencjalnego

(patrz np. [43]). W przypadku układów o mniejszym stopniu złożoności, metody numeryczne bywają wspomagane przez metody analityczne.

W ogólności, projektowanie układów obrazujących z wiązką światła charakteryzuje się w porównaniu z projektowaniem układów z wiązką elektronów, znacznie większą liczbą parametrów, w tym zwłaszcza abberracyjnych (nawet np. 40 zmiennych × 70 aberracji) przy relatywnie szybkim obliczaniu układu danego. Z kolei trudności związane z projektowaniem układów elektrono-optycznych wynikają przede wszystkim z większej złożoności i znacznie większej czasochłonności obliczania zadanego układu.

8. OGÓLNA KONCEPCJA SYSTEMU AUTOMATYZACJI PROJEKTOWANIA ELEKTRONOWYCH UKŁADÓW PRZETWARZANIA OBRAZU

Przykładem systemu CAOD może być bazowy system automatyzacji projektowania dla elektronowych układów przetwarzania obrazu EPOS1.0, opracowywany aktualnie w IMiO PW, z przeznaczeniem dla konstruktorów EWO [44]. System ma uniwersalne



Os	L [mm]	min [j]	max [j]	Pdz [j]	Dod [j]	Mod [j]	Wskali
<y>	150.0	-2.04E-04	2.04E-04	.00E+00	1.00E-04	2.00E-05	7.68E+05
<x>	150.0	-2.04E-04	2.04E-04	.00E+00	1.00E-04	2.00E-05	7.68E+05

Os <x> skierowana jest od lewej do prawej strony.
Os <y> skierowana jest w kierunku wysuwu papieru.

Rys. 4. Graficzna reprezentacja zdeformowanego aberracyjnie krążka rozmycia, uzyskana metodami modelowania

przeznaczenie dla EWO z ogniskowaniem w polu elektrycznym bez względu na rodzaj konstrukcji.

CHARAKTERYSTYKA FUNKCJONALNA SYSTEMU

System ma umożliwić wyznaczenie dowolnych spośród niżej wymienionych parametrów użytkowych kompletnego EWO, przeprowadzenie symulacji syntetycznego testu dla

wszystkich optycznych i elektrono-optycznych elementów w torze przetwarzania obrazu w EWO, oraz prowadzenie dwuparametrowej analizy tolerancji (analizy wymiarowej) i dwuparametrowej optymalizacji konstrukcji układów optyki elektronowej EWO. Funkcjonalnie system obejmuje:

1) Analizę wpływu na optyczne parametry użytkowe kompletnego EWO:

— parametrów wejściowego układu optycznego takich jak MTF, aberracje, powiększenie, otrzymanych na podstawie wyników pomiarów fizycznych lub danych katalogowych i literaturowych,

— parametrów układów optyki włóknistej takich jak charakterystyka transmisji, liczba aperturowa, MTF, otrzymanych na podstawie wyników pomiarów fizycznych lub danych katalogowych i literaturowych,

— parametrów powielaczy mikrokanalikowych takich jak wzmocnienie, MTF, rozdzielczość, rozproszenie wyjściowe, otrzymanych na podstawie wyników pomiarów fizycznych, danych katalogowych i literaturowych, lub na drodze modelowania matematycznego,

— optycznych i elektronowych parametrów fotokatod takich jak charakterystyki widmowe i energetyczno-przestrzenne fotoemisji, MTF, otrzymanych na drodze wyników pomiarów fizycznych lub danych katalogowych i literaturowych,

— elektronowych i optycznych parametrów ekranów luminescencyjnych takich jak charakterystyki widmowe i energetyczne, MTF i rozdzielczość otrzymywanych na podstawie pomiarów fizycznych lub danych katalogowych i literaturowych.

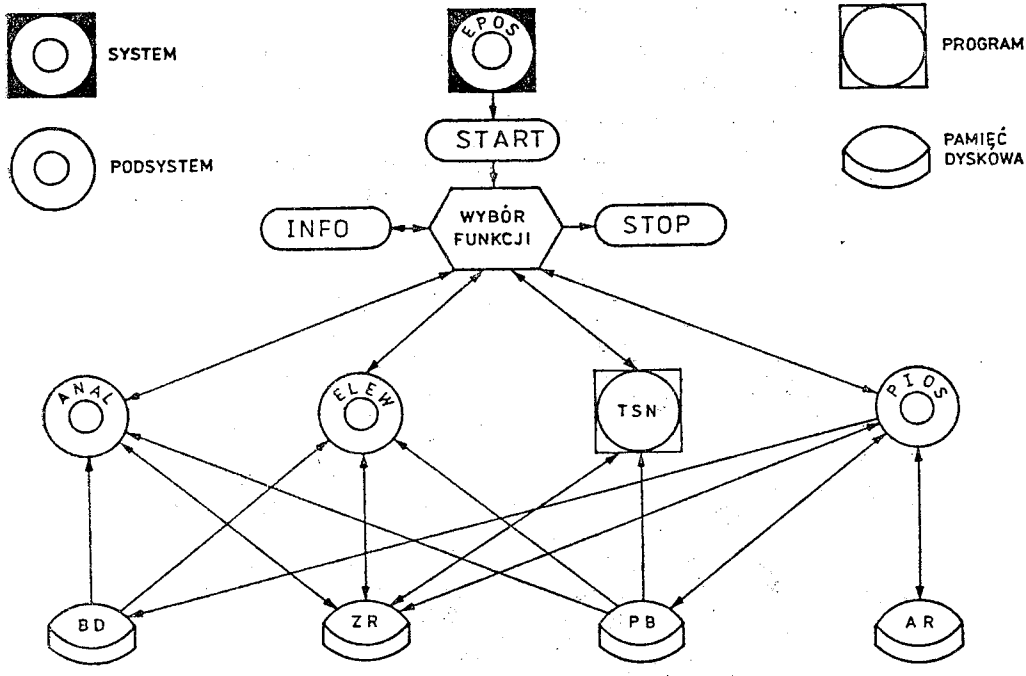
2) Wyznaczanie i analizę statystycznej powierzchni obrazowej (SIS) oraz parametrów układów optyki, elektronowej EWO takich jak powiększenie, dystorsja, rozdzielczość, aberracje, MTF, dla przypadku oświetlenia mono i polichromatycznego, na podstawie modelowania matematycznego tych układów.
3) Jedno i dwuparametrową analizę tolerancji układów optyki elektronowej EWO prowadzoną w odniesieniu do ich optycznych parametrów użytkowych, realizowaną na drodze modelowania matematycznego.
4) Jedno i dwuparametrową optymalizację cząstkową wymiarów konstrukcyjnych i napięć zasilających układów optyki elektronowej EWO, prowadzoną ze względu na kryterium optymalizacyjne będące funkcją optycznych parametrów użytkowych tych układów, realizowaną na drodze modelowania matematycznego.
5) Symulację testu syntetycznego, którego wyniki będą miarą jakości całego EWO z uwzględnieniem wszystkich optycznych i elektrono-optycznych elementów w torze przetwarzania EWO, realizowaną metodą matematycznej syntezy parametrów poszczególnych elementów.
6) Informatyczną obsługę systemu.

System ma umożliwiać równoczesną pracę 2-4 użytkowników przy racjonalnym wykorzystaniu zasobów sprzętowych w lokalnej sieci mikrokomputerowej składającej się m.in. z trzech jednostek typu IBM PC, zarządzanej i nadzorowanej przez mikrokomputer IBM-PC/AT [45].

OGÓLNA STRUKTURA SYSTEMU

System EPOS1.0 będzie zorganizowany jako zbiór podsystemów elastycznie powiązanych ze sobą przez wspólne sterowanie i wspólne bazy danych.

Przewidywana sieć powiązań funkcjonalnych systemu na poziomie 1 przedstawiona jest na Rys. 5. Na tym poziomie należy wyróżnić opisane niżej moduły i zbiory systemowe (przepływ informacji zaznaczono strzałkami):



Rys. 5. Schemat funkcjonalny Bazowego Systemu Automatyzacji Projektowania dla Elektronowych Układów Przetwarzania Obrazu — EPOS1.0 (poziom 1)

— główny moduł systemu EPOS, sterujący wyborem funkcji systemu, udostępnieniem zbiorów i umożliwiający organizację cyklu obliczeniowego,

— podsystem ANAL przeznaczony do analizy wpływu parametrów optycznych elementów składowych EWO na jego parametry użytkowe,

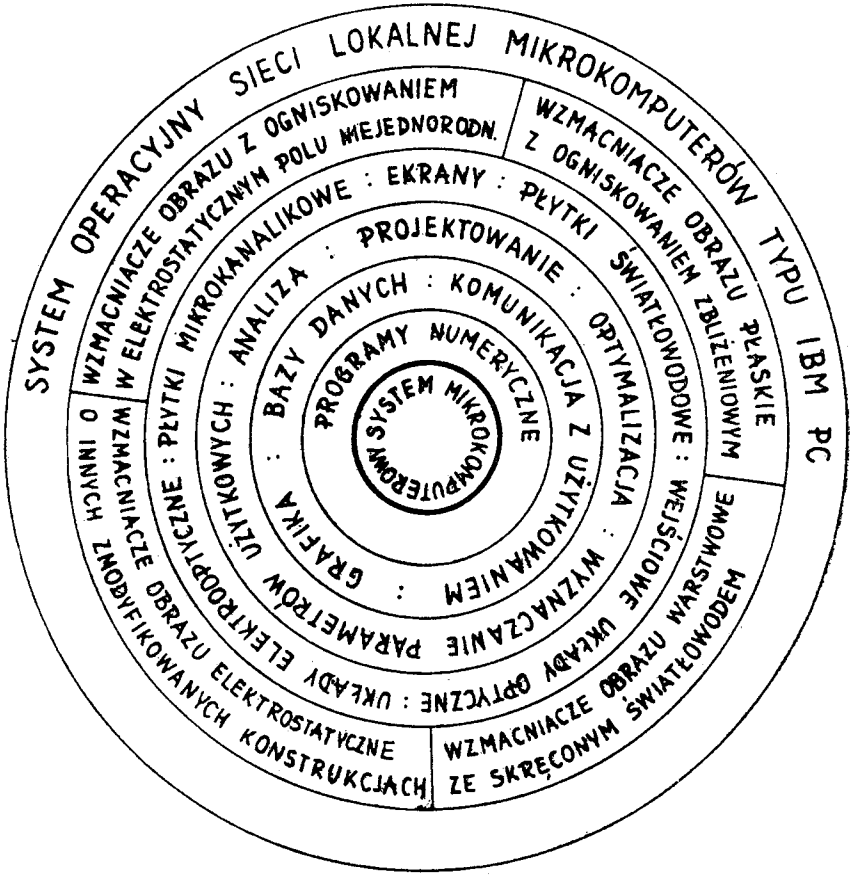
— podsystem ELEW modelowania układów optyki elektronowej EWO, przeznaczony do komputerowego wspomagania analizy i projektowania układów optyki elektronowej EWO,

— program TSN symulacji testu syntetycznego kompletnego EWO pozwalający na matematyczne modelowanie przenoszenia określonego testu obrazowego przez kolejne, składowe elementy konstrukcji EWO, w oparciu o wyznaczone wcześniej parametry op-

tyczne tych elementów. (Ilość i kolejność występowania elementów konstrukcji EWO określać będzie użytkownik programu. Efektem działania programu będzie określenie optycznych parametrów użytkowych kompletnego EWO modelowanych matematycznie w podsystemie ELEW).

Zbiory systemu organizowane w pamięci dyskowej sieci lokalnej obejmować będą (Rys. 5):

1) BD — Zbiory bazy danych systemu, które zawierać będą opisy i parametry układów wzorcowych i katalogowych wykorzystywanych przez system.
2) ZR — Zbiory robocze systemu, które będą zakładane i wykorzystywane w danym cyklu przetwarzania i kasowane na końcu tego cyklu.
3) PB — Bibliotekę programów binarnych systemu, która zawierać będzie oprogramowanie użytkowe (własne) i narzędziowe (pochodzenia obcego) niezbędne do prawidłowego działania systemu.
4) AR — Zbiory archiwalne, które zawierać będą dokumentację systemu oraz dzienniki przebiegów przetwarzania niezbędne dla programu informatora systemu.



Rys. 6. Schemat ideowy systemu EPOS1.0

Schemat ideowy bazowego systemu automatyzacji projektowania dla elektronowych układów przetwarzania obrazu EPOS1.0 zamieszczono na Rys. 6. Na schemacie wymieniono m.in. rodzaje, oraz elementy konstrukcji EWO, które zostaną objęte przez system.

9. PODSUMOWANIE

Przedstawione metody analizy i projektowania optycznych układów obrazujących są stosowane w systemach automatyzacji projektowania. Obecnie, znamienną tendencją jest opracowywanie systemów tego rodzaju implementowanych na sprzęcie mikrokomputerowym 16- i 32-bitowym [46]. Równolegle z rozwojem sprzętu następuje postęp w zakresie algorytmów i narzędzi programowych umożliwiający rozszerzenie zakresu automatyzacji projektowania [47]. Niezależnie od ułatwiania eksploatacji systemów (grafika, wprowadzanie danych, emisja wyników) można się spodziewać, że mikrokompuetrowe systemy CAOD w niedalekiej przyszłości będą mogły obejmować swym zasięgiem układy obrazujące wszystkich rodzajów. Dotyczy to: układów o zmiennej ogniskowej, układów wielo-konfiguracyjnych, układów nie spełniających wymagań w zakresie symetrii itp. Systemy CAOD umożliwiać będą pełne korygowanie charakterystyk przenoszenia układów oraz optymalizację ze względu na coraz większą liczbę parametrów.

BIBLIOGRAFIA

1. *Encyklopedia Fizyki.* Warszawa: PWN 1972, hasło: Optyka tom 2 s. 579
2. W. T. Cathey: *Optyczne przetwarzanie informacji i holografia.* Warszawa: WPN, 1978
3. I. I. Cukkerman: *Prieobrazowania elektronnych izobrażenij.* Energia, 1972
4. *Computer-Aided Optical Design*: SPIE Proc. 1978, vol. 147
5. *Assesment of Imaging Systems*: Visible and Infrared. SPIE 1981, Proc/Vol. 274
6. *Optical Systems Design, Analysis, and Production*: SPIE Proc. 1983 Vol. 399
7. P. W. Hawkes: *Computer-aided design in electron optics.* Comp.-Aid. Design, Oct. 1973 vol. 5, No 4, pp. 200—214
8. P. W. Hawkes: *Image Processing and Computer-aided Design in Electron Optics.* Acad. Press. 1973
9. J. R. Meyer-Arendt: *Wstęp do optyki.* Warszawa: WPN, 1977
10. R. Jóźwicki: *Optyka instrumentalna.* Warszawa: WNT, 1970
11. M. Pluta: *Holografia optyczna,* (praca zbiorowa). Warszawa: PWN 1980
12. E. W. Marchand: *Gradient Index Optics.* Acad. Press., 1978
13. M. Born, E. Wolf: *Principles of Optics.* Perg. Press, 1964
14. M. Kline, I. W. Kay: *Electromagnetic theory and geometrical optics.* J. Wiley and Sons, 1965
15. W. Glaser: *Grundlagen der Elektronenoptik.* Wien: Springer-Verlag, 1952
16. R. K. Luneburg: *Mathematical Theory of Optics.* Univ. of Calif. Press., 1964
17. H. Szymański, A. Mulak, A. Duda, A. Romanowski: *Optyka elektronowa,* Wyd. II zmienione. Warszawa: WNT, 1984
18. K. J. van Oostrum: *CAD in light optics and electron optics.* Philips Technical Review, Oct. 1985, Vol. 42, No 3, pp. 69—84
19. J. M. Woźnicki: *Geometry of Recording and Color Sensitivity for Evanescent Wave Holography Using Gaussian Beam.* Appl. Opt. 1980, Vol. 19, No 13
20. J. M. Woźnicki: *Reference Gaussian Beam in Evenescent Wave Holography.* Optik 1980, Vol. 55, No 4

21. E. Kasper: *On the numerical determination of electron-optical focusing properties and aberrations.* Optik, 1985, Vol. 69, pp. 117—125
22. R. A. Buckingham: *Numerical methods.* I. Pitman etc. (first published 1958)
23. A. Ralston: *Wstęp do analizy numerycznej.* Warszawa: PWN, 1971
24. J. M. Jankowscy: *Przegląd metod i algorytmów numerycznych* Cz. 1, Warszawa: WNT, 1981
25. Z. Fortuna, B. Macukow, J. Wąsowski: *Metody numeryczne.* Warszawa: WNT, 1982
26. H. A. Van Hoof: *A new method for numerical calculation of potentials and trajectories in systems of cylindrical symmetry.* J. Phys. E. Sci. Instr. 1980, Vol. 13
27. J. M. Woźnicki: *Metoda ekstrapolacyjno-interpolacyjna w zastosowaniu do komputerowej analizy odwzorowań elektrono-optycznych.* Elektronika (XXVI) 1985, Nr 8
28. M. W. Fok: *Obszczije woprosy wizualizacji izobrażenij.* Met. wiz. izobr. vol. 129 Moskwa: Izd. Nauka, 1981
29. I. Juvells, S. Callmitjana, D. Ros, J. R. de F. Moneo: *Numerical evaluation of the two-dimensional MTF by means of spot diagram.* Comparison with experimental mesurements. J. Optics (Paris) 1983, vol. 14, No 6, pp. 293—297
30. H. H. Hopkins: Optica Acta, 1966, Vol. 4, p. 343
31. I. Powell: *Optical design and analysis program.* Appl. Opt. 1979, Vol. 17, No 121, pp. 3361—3367
32. J. M. Woźnicki, S. Jeszka, Z. Kupisz-Andrzejewski: *The method for numerical determining the focusing properties of electron optical system in image converter.* Proc. of Intern. Conf. „Image Deterction and Quality", Paryż, 1986
33. J. M. Woźnicki: *Statistical-Image-Surface Conception in Analysis of Electrostatic Electron-Optical Systems.* Proc. of the 14-th Congress of the International Commission for Optics, Quebec 1987
34. N. Bareket: *Second moenmt of the diffraction point spread functions as an image quality criterion.* J. Opt. Soc. Am. 1979, Vol. 69, No 9
35. K. J. Rosenbruch: *Use of OTF-Based criteria in automatic optical design.* Optica Acta 1979, vol. 22, No 4, pp. 291—300
36. J. M. Woźnicki: *Evaluation of Polychromatic MTF for Image Converter with Real Electrostatic Lenses and Photocathodes of Real Characteristics.* Proc. of Intern. Conf. „Image Detection and Quality", Paryż, 1986
37. J. M. Woźnicki: *New procedures considerably reducing computing time of image converter polychromatic MTF calculation.* Proc. of SPIE s 31st Annual Intern. Techn. Symp. on Opt. and Optoelectr. Appl. Sci. and Eng. San Diego, 1987
38. Z. Felczak, S. Jeszka, J. Kaczmarczyk: *Modulacyjna funkcja przekazu jako kryterium oceny optoelektronowych przetworników obrazu.* Rozprawy Elektrotechniczne. 1980, vt. 26, z. 4.
39. J. M. Woźnicki, S. Jeszka, Z. Kupisz-Andrzejewski: *Opracowanie wybranych procedur systemu kompuetrowego wspomagania analizy i projektowania elektrono-optycznych układów obrazujących z fotokatodą.* Sprawozdanie z II etapu pracy umownej. ITE PW 1986
40. A. J. Kirkham: *The use of computers in optical system design.* Phys. Techn., 1982, vol. 13, pp. 210—216
41. G. Schwierz: *Improvement of the Electron Optics of X-Ray Image-Intensifiers.* Siemens Forsch. u Entwickl. 1973, Ber. Bd. 2, No 5, pp. 305—315
42. M. O. Lidwell: *Optical design on a multi-terminal computer.* SPIE Proc., vol. 399 „Optical Systems Design, Analysis and Production", 1983
43. D. S. Grey: *The inclusion of tolerance sinsitivities in the merit function for lens optimization.* SPIE Proc. Vol. 147 „Computer-Aded Optical Design". 1978, pp. 63—65
44. J. M. Woźnicki, W. Jeute: *Bazowy system automatyzacji projektowania dla elektronowych układów przetwarzania obrazu EPOS1.0 — Założenia Techniczne.* CPBR 8.8, ITE PW, 1987
45. J. M. Woźnicki: *System mikrokomputerowy EPOS1.0 — dokumentacja wstępnej wersji sieci lokalnej ... CPBR 8.8.* ITE PW, 1987

46. M. J. Kidger: *Optical design with small computers.* SPIE Proc. vol. 399,,Optical Systems Design, Analysis and Production", 1983
47. A. G. du Woit: *Microcomputer programs for the design of image intensifier tubes.* IEEE Trans. on Magnetics 1985, Vol. MAG-21, No 6, pp. 2527—2530

M. WOŹNICKI

PROBLEMS OF AUTOMATION OF DESIGNING OPTICAL IMAGING SYSTEMS

Summary

The paper contains an introduction to problems of automation of the analysis and the designing of optical imaging systems. The ray method as the basis for modelling of the system phenomena is described. Systematics for the optical image processing parameters are proposed. The possibility of a general procedure for the analysis and the designing of light and electron systems is pointed out. An interesting possibility of employing a common software is indicated. A general concept of CAD integrated programs for electron-image conversion and intensification systems is proposed.

M. WOŹNICKI

PROBLÈMES DE L'AUTOMATISATION DE L'ÉTABLISSEMENT DES PROJETS DES SYSTÈMES OPTIQUES D'IMAGES

Résumé

L'article est une initiation au problème de l'automatisation et de l'établissement du projet des systèmes optiques d'images. On a discuté la méthode des rayons constituant la base du modelage des phénomènes dans les systèmes. On a suggéré le mode de systématiser les paramètres caractérisant la transformation optique de l'image. On a formulé la thèse de la possibilité de l'application de la méthodologie uniforme à l'analyse et à l'établissement du projet des systèmes de l'optique de lumière et de l'optique électronique. On a montré l'intéressante possibilité de l'élaboration d'un projet commun, universelle programmation. On a présenté la conception générale du système de base de l'automatisation de l'établissement du projet des systèmes électroniques de la transformation de l'image.

M. WOŹNICKI

PROBLEME DER AUTOMATISIERUNG DER PROJEKTIERUNG VON OPTISCHEN ABBILDUNGSSYSTEMEN

Zusammenfassung

Der Artikel ist eine Einführung zum Problem der Automatisierung der Analyse und Projektierung von optischen Abbildungssystemen. Erörtert wurde die Methoden von Strahlen als Grundlage für die Modellierung von Phänomenen in Systemen. Es wurde der Vorschlag einer Systematisierung der die Bildverarbeitung charakterisierenden Parameter unterbreitet. Auch wurde die These formuliert, es sei möglich, eine einheitliche Analysen- und Projektierungsmethode für Systeme der Licht- und elektronischen Optik zu schaffen. Es wurde auf die straktive Möglichkeit der Bearbeitung einer gemeinsamen Universalprogrammierung hingewiesen. Eine allgemeine Konzeption eines Basissystemes für die Projektierungsautomatisierung von elektronischen Systemen der Bildverarbeitung wurde dargestellt.

Е. М. ВОЗНИЦКИ

ПРОВЛЕМЫ АВТОМАТИЗАЦИИ ПРОЕКТИРОВАНИЯ ОПТИЧЕСКИХ ИЗОБРАЖАЮЩИХ СИСТЕМ

Р е з ю м е

Приведено введение в проблемы автоматизации анализа и проектирования оптических изображающих систем. Обсужден метод луча как базу моделирования явлений в системах. Указано предложение систематики параметров характеризующих оптическое переобразовывание изображения. Сформулированы тезисы о возможности однородной методики анализа и проектирования систем оптики света и электронной оптики. Обращено внимание на заманчивую возможность совместной разработки универсального программного обеспечения. Указана общая концепция базовой системы автоматизации проектирования электронных систем переобразовывания изображения.

ROZPRAWY ELEKTROTECHNICZNE, 1989, 35, z. 1, ss. 125–131

53.087.354:621.397.331.2

Image Processing for Shape Defects Detection

BOGUSŁAW WIĘCEK

Technical University of Łódź

Received 1988.02.27

Authorized 1988.04.15

The algorithm of binary pattern recognition and comparison with the use of the so-called Fourier descriptors method is presented. The possibilities of its application for measuring the shape complexity, the quality of its representation, the distance and the figure likeness are described. Some practical results are presented.

1. INTRODUCTION

During the last several years, the multidimensional signal processing has an increasing interest. It is a direct consequence of the significant progress in the technology of IC's production and the implementation possibilities of new, sophisticated methods.

In many applications the binary pattern recognition may be employed. In this paper, the binary pattern analysis for the estimation of shape representation quality is presented. There are several well-known algorithms that are based on the converting 2-D signal into 1-D sequence domain. The majority of them originates from the methods being applied in handwritten character recognition [1].

Unfortunately, many of them require the determined position and orientation of the considered image. More useful methods are based on so-called Freeman chaincode [2]. Although this method is very effective, especially from the data compression point of view, it is too much sensitive to the raster representation of the image for the shape quality testing.

In both 1-D and 2-D signal processing, there are known methods that consist in forming the figures or solids of easy-described elements: lines, arcs, planes, cylindric, spherical surfaces etc. [3, 4]. It is a kind of approximation that is very similar to the algorithm using spline functions. Even omitting the problem of approximation smoothness, the numerical procedure is complex.

The another approach to the binary pattern recognition employs, well-known in 1-D signal theory, Wiener-type filtering. The first attempts in this field have been already made [5]. Some modifications of 2-D adaptive filtering to increase the convergence rate, have been presented lately [6, 7].

The modified, so-called Fourier descriptor method is proposed for the practical realization. It assumes that the lengths of vectors fastened to the mass center of the considered object are taken to the spectral analysis. Chosen results of data conversion with the use of the described method are presented.

2. PRINCIPLES OF THE METHOD

The most essential feature of the method being mentioned is the algorithm independence from both the position and the orientation of the considered object. A binary pattern is transformed into the 1-D sequence of vectors lengths that have the origins in the mass center and their ends on the image contour.

Let $x(i)$, $y(i)$, for $i = 1, 2, \ldots, n$, denote the integer values of the contour coordinates in the cartesian system. The mass center of the image is expressed typically

$$xc = \frac{\sum_{i=1}^{n} x(i)}{n}; \quad yc = \frac{\sum_{i=1}^{n} y(i)}{n}. \tag{1}$$

The desired vector lengths sequence $s(i)$ takes the form

$$s(i) = \sqrt{(x(i)-xc)^2 + (y(i)-yc)^2} \tag{2}$$

where:

$$\arctan\frac{y(i)-yc}{x(i)-xc} - \arctan\frac{y(i+1)-yc}{x(i+1)-xc} = const, \tag{3}$$

$$y(n+i) = y(i), \quad i = 1, 2, \ldots, n.$$

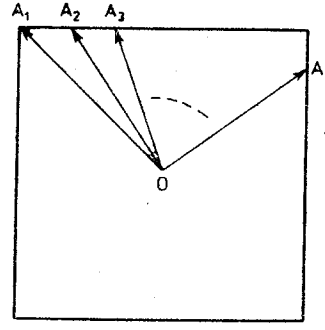


Fig. 1. General method concept

From the last equation it is noticeable, that the samples of $s(i)$ are taken for a multiple of certain angle. It means that the contour is transformed into the planar coordinates domain and in consequence, the different positions of the images involve the changes of signal $s(i)$ phase. The phase differences of $s(i)$ are eliminated by the Fast Fourier Transform. The FFT gives another advantage that concerns the noise influence decreasing. It is

a well-known fact that the Fourier analysis operates as a filter for an uncorrelated signal [8]. For more effective noise cancelling it is necessary to prolong the computational procedure and take into account more than one signal period.

In the presented method each contour can be expressed by the geometrical sum of vectors that describes the elementary figures. It means that the mean value of $s(i)$ describes the size of the recognized image and its details depend upon the other harmonics. The mean value is represented by a circle and the further spectral elements by the leaflets as it is shown in Fig. 2.

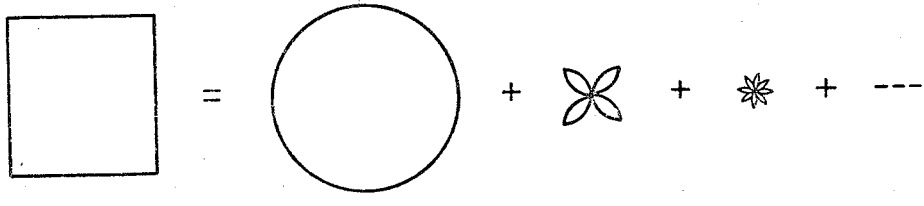


Fig. 2. Shape decomposition

For the spectral analysis, one of the known method can be used. To achieve the implementation simplicity and the effectiveness of its operation the periodogram formula is recommended. The unknown coefficients of the trigonometric series are obtained by using the least-square approximation. For more sophisticated applications the faster version of the spectral analysis has to be implemented, e.g. Cooley-Tukey algorithm.

For the recognition process, one applies the threshold criterion in the spectral coefficient domain. To compare two patterns the likeness scales are created by considering the Fourier data.

Let $f(j)$ denote the amplitude of j-th harmonic of $s(i)$

$$s(i) = \sum_{j=0}^{N} f(j)*\sin\left(j\frac{2\pi}{N}i+\varphi(j)\right). \tag{4}$$

To determine a relationship between the sizes of two images one can consider the mean values. If the further spectrum corresponding elements are equal, the similitude transform is identified. The main parameter of such a transform-similitude factor takes the form

$$k = \frac{f(0)}{f'(0)}; \quad f(j) = f'(j) \quad \text{for} \quad j = 1, 2, \ldots, N \tag{5}$$

where: $f(j)$ and $f'(j)$, $(j = 0, 1, \ldots, N)$ denote the spectrum of the main and recognized patterns respectively.

In many practical cases a distance between two functions as a norm in the discerete function space can be used. According to this criterion, both the size and image details are compared. The mean-square error as the distance of functions is expressed by

$$d = \frac{\sum_{j=0}^{N} (f(j)^2 - f'(j)^2)*w(j)}{N}, \tag{6}$$

where: $w(j)$-weight factors that gain the algorithm sensitive for a chosen part of spectrum, i.e. for special type of shape.

The next comparison formula is applied to the polygon separation. It assumes that the existence of the certain harmonics depends on the number of angles, e.g.:

regular polygon (N angles) — $j*N, j = 1, 2, \ldots$
rectangle, rhomb — $2, 4, 6, \ldots$

In general, it is easy to estimate the number of angles by determining all of such points of $s(i)$ in which the derivatives are not continuous.

Sometimes, there is a real need of an area measurement. In such case the following relation is very useful

$$S = A + \frac{B}{2} - 1 \tag{7}$$

where: A and B denote the number of raster pixels inside the figure and number of pixels on its edge respectively. In this way the area is expressed in the raster units.

In addition, the presented method allows to determine the image orientation. By analysing the phases of harmonics, it is possible to obtain the angle of figure rotation on condition of the same start point in the traced contours. The rotation angle is expressed as difference of phases for corresponding harmonics.

$$\vartheta(j) = const; \quad \vartheta(j) = (\varphi(j) - \varphi'(j))*j \qquad j = 1, 2, \ldots, N. \tag{8}$$

In a variety of applications, the distance measurement is required. For that case, the computer (robot) is a learning automat that knows both the size and the distance of a certain object in its field of operation. Knowing the size, one can directly determine the distance from the similitude factor k (Eqn 4).

One contour image conversion is an essential limitation of the method mentioned above. To compare the whole image (2-D or 3-D), the new approach of the multidimensional pattern recognition has been introduced lately [6, 7]. This concept is based on expanding the Wiener-Hopf filtering theory into two-dimensional signal processing. The Wiener-type filtering consists in the adjustments of the filter parameters to minimize the mean-square error between two images. In these terms, two signals (patterns) are equal if and only if the transfer function of the filter described by impulse response is a simple Kronecker delta [6]. One can create easy-to-implementation algorithm of filtering using the least-square approximation. To achieve very effective adaptive process, the well-known weight gradient method has been proposed.

In general, one of the significant advantages of Wiener adaptive process is the possibility of its application in N-dimensional signal domain, e.g.: in the tomography, seismic data processing, image satellite communication etc.

3. SOME PRACTICAL RESULTS

Some tests of the method mentioned above have been carried out on general-purpose IBM-compatible machine. The recognized images are sent via a camera and the digitizing

interface to the system memory in DMA channel. To optimize the conversion time, software has been written with the use of *C*-compiler tools. For illustrating the operation of the algorithm and monitoring all of its parts some final and indirect results are presented in a graphic form.

At first, some examples of signals $s(i)$ and their spectrums are shown. The next

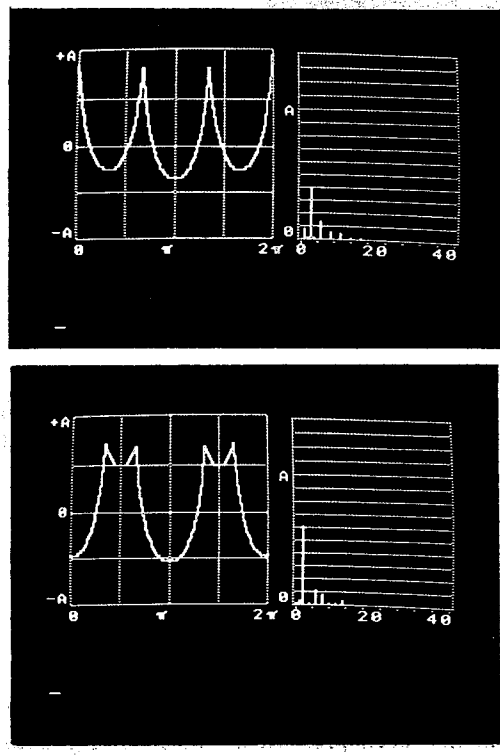


Fig. 3. Signal s_t and its spectrum (triangle, rectangle)

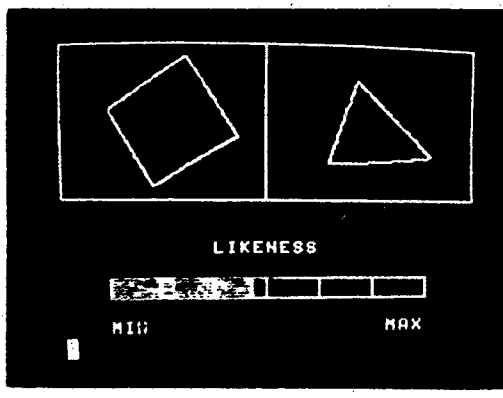


Fig. 4. Polygon separation

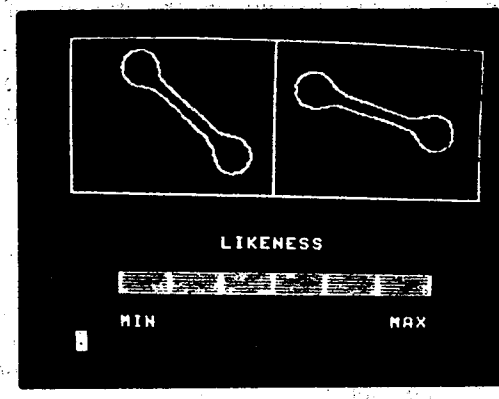


Fig. 5. Algorithm independence from the figure orientation

diagrams illustrate the final results of binary pattern comparison for a likeness scale defined by the eqn. (6). MAX denotes the images equivalence. As it is shown, for such an application algorithm ignores the figure orientation.

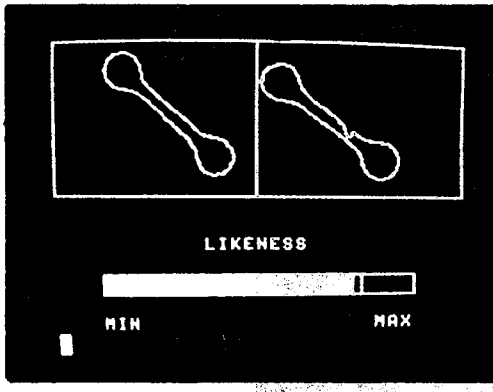


Fig. 6. Shape defect detection

For patterns separation and the shape accuracy testing the great attention has to be paid while choosing the weight factors $w(j)$ in the eqn. (6). The weight factor variation depends upon the concrete application and their defining should be preceded by practical investigations. For example, to achieve more information on the images sizes one must gain the prime spectrum descriptors $f(i)$. The image details analysis involve the further coefficients increasing.

CONCLUSION

The method of binary pattern recognition for one-contour figures, suitable for microcomputer implementation has been presented. The most significant practical aspects of discussed algorithm are both the possibility of creation different likeness scales for images comparison and the independence from the object orientation. The possibilities of the size and the details recognition, the attempts of the algorithm application to measure image orientation and the distance as well, reflect the flexibility and usability of described methods. For more sophisticated tasks the Wiener-type adaptive filtering is recommended. Both methods give the high level noise immunity. It means that they can be employed in real, noisy environment.

REFERENCES

1. S. Mori, K. Yamamoto, M. Yasuda: *Research on Machine Recognition of Handprinted Characters.* IEEE Trans. Pattern Anal. Machine Intell., 1984 vol. PAMI-6
2. H. Li, C. Tsang, Y. Cheung: *Pattern Recognition for Automated Wire Bonding.* Proc. IEE, Jan. 1984 vol. 131
3. L. Dorst, A. Smeulders: *Discrete Representation of Straight Lines.* IEEE Trans. Pattern Anal. Machine Intell., July 1984 vol. PAMI-6

4. R. Bolle, D. Cooper: *Bayesan Recognition of Local 3-D Shape by Approximating Image Intensity Functions with Quadric Polynomials.* IEEE Trans. Pattern Anal. Machine Intell. July 1984 vol, PAMI-6
5. J. Justice: *A Levinson-Type Algorithm for Two-Dimensional Wiener Filtering Using Bivariete Szego Polynomials.* Proc. IEEE, July 1977 vol. 65
6. B. Więcek: *3-D Pattern Recognition via Adaptive Filtering*, Proc. ECCTD'85, Prague, Sept. 1985
7. B. Więcek: *Binary Pattern Recognition and Comparison.* Proc. Int. Conf. Robotics, Artif. Intell., Identif. and Pattern Recog,. Toulouse, June 1986
8. J. Cadzow: *High Performance Spectral Estimation — A New ARMA Method.* IEEE Trans. Acoust., Speech, Signal Processing, No 2, May 1980

B. WIĘCEK

ROZPOZNAWANIE KONTURÓW OBRAZÓW PRZEDMIOTÓW PŁASKICH

Streszczenie

W artykule przedstawiono algorytm rozpoznawania i porównywania konturów obrazów przedmiotów płaskich przy wykorzystaniu metody tzw. podpisów Fouriera. Pokazano możliwość zastosowania algorytmu do pomiaru złożoności kształtu, jakości jego odwzorowania, odległości i podobieństwa figur. Przedstawiono wybrane wyniki praktyczne.

B. WIĘCEK

RECONNAISSANCE DES CONTOURS DES IMAGES DES OBJETS PLATS

Résumé

Dans l'article on a présenté l'algorithme de la reconnaissance et de la comparaison des contours des images plates à l'aide de la méthode des descriptions de Fourier. On a montré la possibilité de l'application de l'algorithme pour la mesure de la complexité du contour, de la qualité de sa représentation, de la distance et de la similitude des figures. On a présenté les résultats pratiques choisis.

B. WIĘCEK

UMRIßABTASTUNG VON BILDERN FLACHER GEGENSTÄNDE

Zusamm enfassung

Im Artikel wurde der Algorithmus der Umrißabtastung und -vergleichung von Bildern flacher Gegenstände unter Anwendung der Fourier-Methode beschrieben. Es wurden Anwendungsmöglichkeiten des Algorithmus für Messung der Formvielfalt, Abbildungsgüte, Entfernung und Ähnlichkeit der Bilder gezeigt. Auch einige praktische Ergebnisse wurden gebracht.

Б. ВЕНЦЕК

РАСПОЗНАВАНИЕ КОНТУРОВ ПЛОСКИХ ИЗОБРАЖЕНИЙ

Резюме

Представлен алгоритм распознавания и сравнивания контуров плоских изображений при использовании метода т.н. подписей Фурье. Показана возможность применения алгоритма для измерения сложности формы обьекта, качества его отображения, растояния, подобия фигур. Представлены некоторые практические результаты.

ROZPRAWY ELEKTROTECHNICZNE, 1989, t. 35, z. 1, ss. 133–148

537.228.1.53.082.731

Piezorezystywne właściwości rezystorów grubowarstwowych

STANISŁAW PASZCZYŃSKI

Instytut Automatyki i Metrologii, Politechnika Rzeszowska

Otrzymano 1988.02.24

Autoryzowano do druku 1988.03.22

Przedstawiono opis analityczny zjawiska piezorezystywnego oraz metodykę pomiarów i wyniki badań piezorezystywnych właściwości rezystorów grubowarstwowych. Przeanalizowano korelacje pomiędzy właściwościami piezorezystywnymi materiału warstwy a możliwymi do zdefiniowania i pomiaru czułościami odkształceniowymi rezystancji rezystora. Omówiono podstawowe metody pomiaru czułości odkształceniowych rezystancji i wskazano źródła możliwych błędów pomiaru. Do wykonania rezystorów testowych zastosowano prototypową serię rutenianowych past rezystywnych typu PS i dostępne handlowo pasty rezystywne nr 1841, 8049 i 8039 oraz przewodzącą nr 6120 produkcji firmy Du Pont. Zbadano wpływ zawartości rutenianu bizmutu w warstwie oraz długości rezystora na czułości odkształceniowe restystancji oraz na rezystancję na kwadrat. Stwierdzono, że istnieje silna zależność zarówno czułości odkształceniowych jak i rezystancji na kwadrat od tych czynników. Sformułowano wnioski dotyczące najbardziej prawdopodobnych uwarunkowań przebiegu badanych charakterystyk rezystorów grubowarstwowych.

1. WPROWADZENIE

Zjawisko piezorezystywne w rezystorach grubowarstwowych stało się przedmiotem zainteresowań na początku lat siedemdziesiątych [1]. Stwierdzono, że czułość odkształceniowa rezystancji rezystorów grubowarstwowych zdefiniowana jako stosunek względnego przyrostu rezystancji do względnego wydłużenia rezystora, jest znacznie większa niż w przypadku cienkich, ciągłych warstw metalicznych, a stabilność czasowa i temperaturowa tego parametru znacznie lepsza niż dla nieciągłych warstw cienkich [2—5]. Zależność względnych zmian rezystancji od wydłużenia względnego jest liniowa i pozbawiona histerezy w takim zakresie wydłużeń, przy których nie występują jeszcze nieodwracalne zmiany struktury rezystora grubowarstwowego [6]. Te pozytywne, obiecujące właściwości piezorezystywne spowodowały wzrost zainteresowania zarówno stroną fizyczną tego zjawiska [7—9] jak również możliwościami wykorzystania rezystorów grubowarstwowych do konstrukcji czujników ciśnienia lub siły [10—14]. Jak dotychczas brak jest modelu, który wyjaśniałby mechanizm tego zjawiska w elementach grubowarstwowych, chociaż przypuszcza się, że odpowiedzialnym za jego występowania może być zjawisko tunelowania elektronów w mikrobarierach typu MIM (Metal-Izolator-Metal) istniejących w postaci

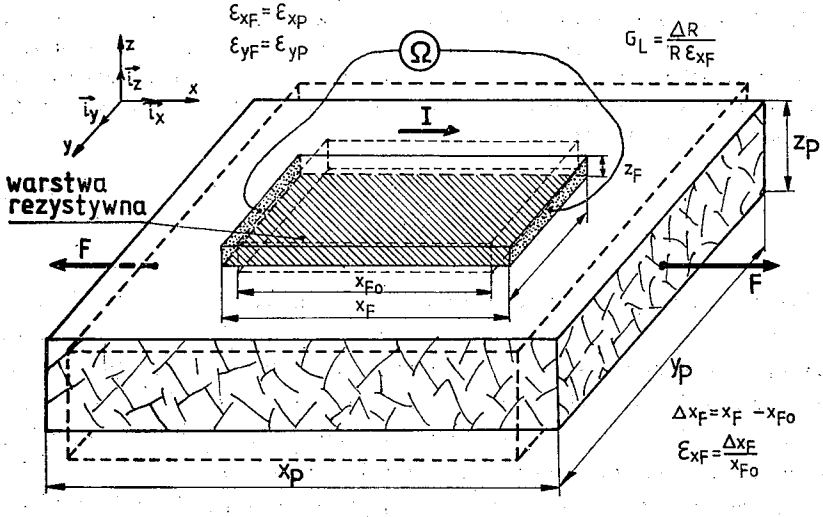
przewodzących ziaren odizolowanych cienką warstwą szkliwa [8]. Interesującym jest fakt występowania zjawiska piezorezystywnego również w innych elementach grubowarstwowych jak np. termistorach [9].

W pracy przedstawiono analityczny opis zjawiska piezorezystywnego w rezystorach grubowarstwowych z uwzględnieniem specyfiki struktury tych elementów, zaprezentowano metody jego badania, a także podano wyniki badań piezorezystywnych właściwości rezystorów grubowarstwowych. Do badań wybrano prototypową serię past rutenianowych PS opracowaną w Politechnice Rzeszowskiej oraz, w celach porównawczych, handlowo dostępne pasty nr 8039, 8049 oraz 1841 produkcji firmy Du Pont.

2. OPIS ANALITYCZNY ZJAWISKA PIEZOREZYSTYWNEGO

Zjawisko piezorezystywne w rezystorach grubowarstwowych może być opisane za pomocą zależności względnych zmian rezystancji elementu $\left(\frac{dR}{R}\right)$ od jego wydłużenia względnego ε. Kształt oraz stabilności czasowa i temperaturowa tej charakterystyki będą natomiast określać ewentualny obszar zastosowań rezystorów.

Rezystancję jednorodnego rezystora warstwowego (rys. 1) o kształcie prostokątnym



Rys. 1. Rezystor warstwowy połączony trwale, w płaszczyźnie xy z podłożem ceramicznym poddawany wydłużeniu w kierunku x zgodnym z kierunkiem przepływu prądu I; F — siła, ε_{xF} — wydłużenie względne warstwy

i kontaktach w płaszczyźnie yz oraz połączonego w sposób trwały z podłożem ceramicznym w płaszczyźnie xy można wyznaczyć ze znanej zależności:

$$R = \varrho \frac{x_F}{y_F z_F}, \tag{1}$$

gdzie ϱ jest rezystywnością warstwy, natomiast x_F, y_F i z_F są rozmiarami rezystora. W równaniu (1) pominięto wpływ kontaktów. Jeżeli rezystor poddamy wydłużeniu ε w jakimś kierunku, to spowodujemy zmianę zarówno jego rezystywności jak i rozmiarów. Zapisując to zagadnienie ogólnie otrzymamy następującą postać równania (1):

$$R(\varepsilon) = \varrho[\varepsilon_x(\varepsilon), \varepsilon_y(\varepsilon), \varepsilon_z(\varepsilon)]x_F[\varepsilon_x(\varepsilon)]\{y_F[\varepsilon_y(\varepsilon)]z_F[\varepsilon_z(\varepsilon)]\}^{-1} \tag{2}$$

gdzie ε_x, ε_y, ε_z oznaczają składowe wektora wydłużenia w kierunkach x, y, z odpowiednio. Wyznaczmy dla takiego rezystora współczynnik czułości odkształceniowej rezystancji G_ε zdefiniowany jako [7]:

$$G_\varepsilon = \frac{1}{R}\frac{\mathrm{d}R(\varepsilon)}{\mathrm{d}\varepsilon} \cong \frac{\Delta R(\varepsilon)}{R\varepsilon}. \tag{3}$$

Po wstawieniu równania (2) do (3) i zróżniczkowaniu otrzymuje się:

$$G_\varepsilon = \frac{1}{\varrho}\frac{\partial\varrho}{\partial\varepsilon_x}\frac{\mathrm{d}\varepsilon_x}{\mathrm{d}\varepsilon} + \frac{1}{\varrho}\frac{\partial\varrho}{\partial\varepsilon_y}\frac{\mathrm{d}\varepsilon_y}{\mathrm{d}\varepsilon} + \frac{1}{\varrho}\frac{\partial\varrho}{\partial\varepsilon_z}\frac{\mathrm{d}\varepsilon_z}{\mathrm{d}\varepsilon} + \\ + \frac{1}{x_F}\frac{\partial x_F}{\partial\varepsilon_x}\frac{\mathrm{d}\varepsilon_x}{\mathrm{d}\varepsilon} - \frac{1}{y_F}\frac{\partial y_F}{\partial\varepsilon_y}\frac{\mathrm{d}\varepsilon_y}{\mathrm{d}\varepsilon} - \frac{1}{z_F}\frac{\partial z_F}{\partial\varepsilon_z}\frac{\mathrm{d}\varepsilon_z}{\mathrm{d}\varepsilon}. \tag{4}$$

Ponieważ

$$\frac{1}{x_F}\frac{\partial x_F}{\partial\varepsilon_x} = \frac{1}{y_F}\frac{\partial y_F}{\partial\varepsilon_y} = \frac{1}{z_F}\frac{\partial z_F}{\partial\varepsilon_z} = 1,$$

więc oznaczając

$$\frac{1}{\varrho}\frac{\partial\varrho}{\partial\varepsilon_x} = G_x, \quad \frac{1}{\varrho}\frac{\partial\varrho}{\partial\varepsilon_y} = G_y, \quad \frac{1}{\varrho}\frac{\partial\varrho}{\partial\varepsilon_z} = G_z, \tag{5}$$

gdzie współczynniki G_x, G_y, G_z reprezentują tzw. wkład fizyczny do wartości G_ε zależny od mechanizmu przewodnictwa rezystora warunkującego wartość ϱ, zależność (4) przyjmie postać:

$$G_\varepsilon = G_x\frac{\mathrm{d}\varepsilon_x}{\mathrm{d}\varepsilon} + G_y\frac{\mathrm{d}\varepsilon_y}{\mathrm{d}\varepsilon} + G_z\frac{\mathrm{d}\varepsilon_z}{\mathrm{d}\varepsilon} + \frac{\mathrm{d}\varepsilon_x}{\mathrm{d}\varepsilon} - \frac{\mathrm{d}\varepsilon_y}{\mathrm{d}\varepsilon} - \frac{\mathrm{d}\varepsilon_z}{\mathrm{d}\varepsilon}. \tag{6}$$

Dla ciał izotropowych zachodzi równość:

$$G_x = G_y = G_z = Q. \tag{7}$$

Wykorzystując równość (7) w równaniu (6), otrzymamy ostatecznie:

$$G_\varepsilon = Q\left(\frac{\mathrm{d}\varepsilon_x}{\mathrm{d}\varepsilon} + \frac{\mathrm{d}\varepsilon_y}{\mathrm{d}\varepsilon} + \frac{\mathrm{d}\varepsilon_z}{\mathrm{d}\varepsilon}\right) + \frac{\mathrm{d}\varepsilon_x}{\mathrm{d}\varepsilon} - \frac{\mathrm{d}\varepsilon_y}{\mathrm{d}\varepsilon} - \frac{\mathrm{d}\varepsilon_z}{\mathrm{d}\varepsilon}. \tag{8}$$

W zależności od relacji pomiędzy kierunkami prądu $\vec{I}$ płynącego w rezystorze oraz jego wydłużeniem $\vec{\varepsilon}$, a także uwzględniając uwarunkowania konstrukcyjne rezystora grubowarstwowego można zdefiniować następujące czułości odkształceniowe [7, 11]:

$$\left.\begin{aligned} &\text{— podłużną} && -G_L \text{ gdy } \varepsilon = \varepsilon_x \text{ oraz } \vec{I} \parallel \vec{i}_x, \\ &\text{— poprzeczną} && -G_T \text{ gdy } \varepsilon = \varepsilon_y \text{ oraz } \vec{I} \parallel \vec{i}_x, \\ &\text{— prostopadłą} && -G_P \text{ gdy } \varepsilon = \varepsilon_z \text{ oraz } \vec{I} \parallel \vec{i}_x \text{ lub } \vec{I} \parallel \vec{i}_y, \\ &\text{— prostopadłą} && -G_Q \text{ gdy } \varepsilon = \varepsilon_z \text{ oraz } \vec{I} \parallel \vec{i}_z. \end{aligned}\right\} \tag{9}$$

Aby można było wyznaczyć tak zdefiniowane czułości odkształceniowe na podstawie równania (8) należy uwzględnić fakt, że warstwa rezystywna jest trwale połączona z podłożem w płaszczyźnie xy. Ponieważ podłoże z ceramiki alundowej (96÷99% Al_2O_3) posiada znacznie większy moduł sprężystości podłużnej niż warstwa rezystywna [15], a jego grubość jest około dwa rzędy wartości większa niż grubość warstwy, więc można z bardzo dobrym przybliżeniem przyjąć, że wydłużenie warstwy rezystywnej (ε_F) oraz podłoża (ε_P) w kierunkach x oraz y są sobie równe:

$$\varepsilon_{xF} \cong \varepsilon_{xP}, \qquad \varepsilon_{yF} \cong \varepsilon_{yP}.$$

Jednocześnie zachodzą równości:

$$\frac{\mathrm{d}\varepsilon_{yP}}{\mathrm{d}\varepsilon_{xP}} = \frac{\mathrm{d}\varepsilon_{xP}}{\mathrm{d}\varepsilon_{xP}} = -\nu_P, \qquad \frac{\mathrm{d}\varepsilon_{zF}}{\mathrm{d}\varepsilon_{xF}} = -\nu_F \tag{10}$$

gdzie ν_P i ν_F są współczynnikami Poissona podłoża i warstwy odpowiednio.
Uwzględniając również fakt, że w przypadku czułości odkształceniowych prostopadłych G_P i G_Q spełniony jest warunek:

$$\frac{\mathrm{d}\varepsilon_{xP}}{\mathrm{d}\varepsilon_{zF}} = \frac{\mathrm{d}\varepsilon_{yP}}{\mathrm{d}\varepsilon_{zF}} \cong 0, \tag{11}$$

można łatwo otrzymać, korzystając z równania (8) oraz podanych wcześniej definicji czułości odkształceniowych następujące wyrażenia:

$$\left.\begin{aligned} G_L &= Q(1-\nu_P-\nu_F)+1+\nu_P+\nu_F, \\ G_T &= Q(1-\nu_P-\nu_F)-1-\nu_P+\nu_F, \\ G_P &= Q-1, \\ G_Q &= Q+1. \end{aligned}\right\} \tag{12}$$

Ponieważ wartość współczynnika Poissona dla większości materiałów mieści się w zakresie 0,22—0,40, więc dla $Q \geqslant 4$ czułości odkształceniowe opisane zależnościami (12) można uporządkować następująco:

$$G_Q > G_P > G_L > G_T. \tag{13}$$

Relacja ta jest użyteczna przy projektowaniu czujników siły lub ciśnienia bazujących na zjawisku piezorezystywnym w rezystorach grubowarstwowych [11]. Biorąc pod uwagę to, że typowe, handlowo dostępne materiały rezystywne pozwalają otrzymywać rezystory, których współczynnik G_L osiąga wartości kilkanaście przy rezystancji na kwadrat równej około 1 MΩ/□ [8, 9], a w niektórych przypadkach nawet 25÷30 [16], więc wartość współczynnika Q może wynosić nawet około 40.

Jeżeli założymy, że $\nu_P = \nu_F = \nu$ wówczas na podstawie (12) mamy:

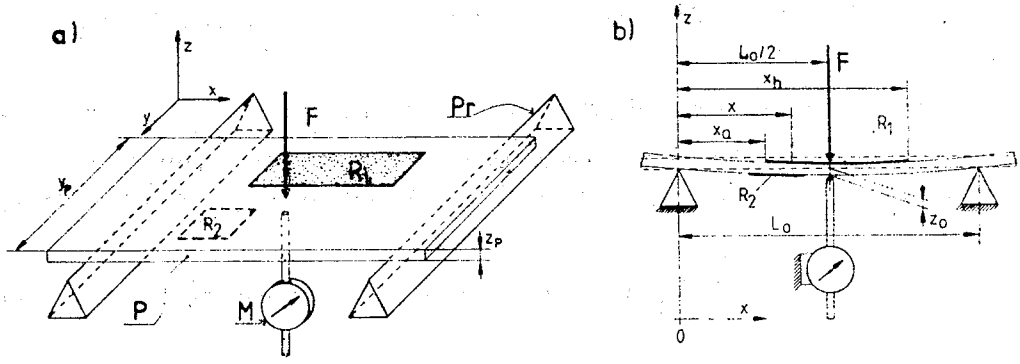
$$G_L - G_T = 2(1+\nu). \tag{14}$$

Wartość tej różnicy powinna się mieścić w przedziale 2,4÷2,8 i pozwala ona, podobnie jak wyrażenia (12) zweryfikować słuszność przyjętych założeń o izotropowości warstwy i braku wpływu warstw kontaktowych na współczynnik Q.

3. METODYKA POMIARU CZUŁOŚCI ODKSZTAŁCENIOWEJ REZYSTANCJI

W literaturze są przedstawione metody pomiaru czułości odkształceniowych poprzecznej G_T oraz podłużnej G_L [6, 9]. Pomiaru tych współczynników można dokonać stosunkowo łatwo, gdyż kierunki wydłużeń warstwy leżą w płaszczyźnie xy (rys. 1) i są praktycznie równe wydłużeniu podłoża na powierzchni.

Zasadę najczęściej stosowanej metody pomiaru czułości G_T i G_L (tzw. metodę ugięcia trójpunktowego) zilustrowano schematycznie na rys. 2. Strzałka ugięcia (z_0) płytki podło-



Rys. 2. Metoda kontrolowanego wydłużania rezystora za pomocą tzw. trójpunktowego ugięcia płytki podłożowej podczas pomiaru czułości odkształceniowych rezystancji; a) schemat układu, b) przekrój układu w płaszczyźnie zx uwidaczniający strzałkę zgięcia (z_0) podłoża (P);
Pr — pryzma stalowa, M — czujnik mikrometryczny, F — siła, R_1,R_2 — rezystory warstwowe

żowej z odpowiednio naniesionymi rezystorami warstwowymi, podpartej na dwóch równoległych ostrzach, jest związana z wydłużeniem względnym podłoża na powierzchni w punkcie x zależnością postaci [17, 18]:

$$\varepsilon_x(x) \cong \frac{12xz_P z_0}{L_0^3} = cxz_0 \quad \text{dla} \quad z_0 \ll z_P \quad \text{i} \quad x \leqslant L_0/2 \tag{15}$$

gdzie

$$z_0 = \frac{L_0^3 F}{4E_P y_P z_P^3}, \tag{16}$$

natomiast: c jest stałą przy ustalonych warunkach pomiarowych, E_P, y_P i z_P są modułem sprężystości podłużnej, szerokością i grubością płytki podłożowej odpowiednio, L_0 jest odległością ostrzy, a F—wartością siły przyłożonej w połowie odległości pomiędzy ostrzami, prostopadle do podłoża powodującej jego ugięcie. Wartość ugięcia wyznacza się na ogół za pomocą czujnika mikrometrycznego (M), natomiast zależność (16) jest użyteczna przy projektowaniu czujników siły. Ponieważ warstwa rezystywna ma długość x_F, więc średnia

wartość wydłużenia względnego będzie opisana, w zależności od położenia rezystora na podłożu, następującymi wyrażeniami:

dla $0 \leqslant x_a < L_0/2$ i $L_0/2 \leqslant x_b \leqslant L_0$

$$\bar{\varepsilon}_x = \frac{\Delta x_F}{x_F} = \frac{1}{x_F}\int_{x_a}^{x_b} \varepsilon(x)\mathrm{d}x = \frac{1}{2} cz_0 L_0 \left\{1 - \left(\frac{a}{L_0}\right)\left[1 + \left(\frac{b}{a}\right)^2\right] \Big/ \left(1 + \frac{b}{a}\right)\right\}, \tag{17}$$

dla $0 \leqslant x_a < x_b \leqslant L_0/2$

$$\bar{\varepsilon}_x = \frac{1}{2} cz_0 (x_a + x_b), \tag{18}$$

dla $L_0/2 \leqslant x_a < x_b \leqslant L_0$

$$\bar{\varepsilon}_x = \frac{1}{2} cz_0 [2L_0 - (x_b + x_a)] \tag{19}$$

gdzie $b = x_b - L_0/2$, $a = L_0/2 - x_a$ natomiast x_a i x_b są współrzędnymi kontaktów warstwy rezystywnej.

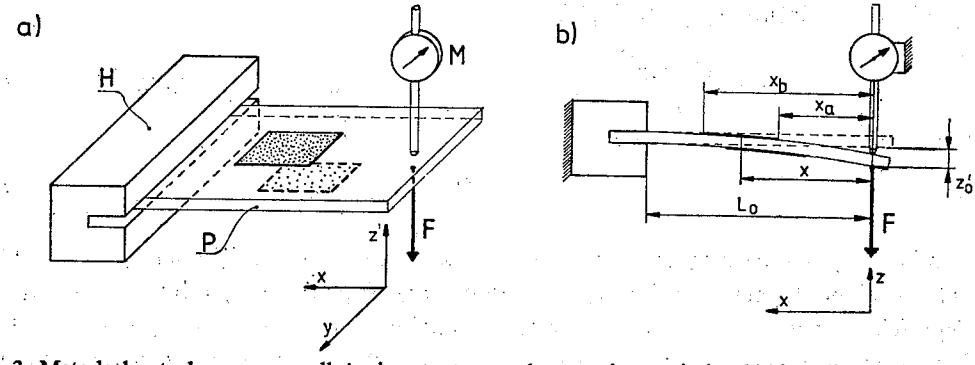
Po dokonaniu pomiaru względnych zmian rezystancji rezystora $\left(\frac{\Delta R}{R}\right)$ spowodowanych wydłużeniem określonym za pomocą jednej z zależności (17)÷(19) dla znanego ugięcia, można wyznaczyć, korzystając z definicji (3) i (9) czułości odkształceniowe G_L i G_T jako:

$$G_L = \frac{\Delta R_x[\bar{\varepsilon}_x(z_0)]}{R_x \bar{\varepsilon}_x(z_0)} \quad \text{(kontakty leżą w płaszczyźnie } yz\text{)}, \tag{20}$$

$$G_T = \frac{\Delta R_y[\bar{\varepsilon}_x(z_0)]}{R_y \bar{\varepsilon}_x(z_0)} \quad \text{(kontakty leżą w płaszczyźnie } xz\text{)}. \tag{21}$$

W wyrażeniach tych dodatni znak wydłużenia odpowiada naprężeniom rozciągającym w warstwie, natomiast ujemny — ściskającym.

Alternatywną metodę pomiaru czułości odkształceniowych poprzecznej i podłużnej pokazano schematycznie na rys. 3. W metodzie tej płytka podłożowa z naniesionymi re-



Rys. 3. Metoda kontrolowanego wydłużania rezystora wykorzystująca ugięcie płytki podłożowej umocowanej trwale jednym końcem w uchwycie H; a) schemat układu, b) przekrój układu w płaszczyźnie xz ukazujący maksymalne ugięcie (z_0') mierzone czujnikiem mikrometrycznym

zystorami warstwowymi jest mocowana trwale jednym końcem w uchwycie (H) i uginana poprzez przyłożenie siły (F) w pobliżu przeciwległej, swobodnej krawędzi. Strzałka ugięcia (z_0') i położenie rezystora na podłożu pozwalają na wyznaczenie średniego wydłużenia. Odpowiednie zależności umożliwiające wyznaczenie tej wielkości są bardzo zbliżone do podanych wcześniej, a mianowicie:

$$\varepsilon_x(x) = \frac{3}{2}\frac{z_P x z_0'}{L_0^3} = c' x z_0', \tag{22}$$

gdzie

$$z_0' = \frac{4FL_0^3}{E_P y_P z_P^3}, \tag{23}$$

$$\bar{\varepsilon}_x = \frac{1}{2} c' z_0' (x_a + x_b). \tag{24}$$

Pomiaru współczynników G_L i G_T można również dokonać powodując wydłużenie płytki podłożowej w kierunkach x i y odpowiednio poprzez przyłożenie naprężeń rozciągających lub ściskających w tych kierunkach. Wymaga to jednak stosowania dość znacznych sił, a pomiar wydłużenia warstwy jest bardziej złożony. Taki sposób wywoływania naprężeń w warstwach rezystywnych stosowano przykładowo w pracy Szymańskiego [19].

Współczynniki G_P i G_Q mogą być oszacowane na podstawie zależności (12), jeżeli znane są wartości współczynników G_L i G_T oraz jeden ze współczynników Poissona ν_P lub ν_F. Mogą one też być zmierzone w układzie, w którym będzie można wywoływać i kontrolować wydłużenie warstwy rezystywnej w kierunku z, przy czym w przypadku G_Q rezystor musi posiadać strukturę kondensatorową z kontaktami w płaszczyźnie xy.

Dokładność wyznaczenia czułości odkształceniowych rezystancji zależy głównie od dokładności określenia wydłużenia $\bar{\varepsilon}_x$, gdyż pomiar rezystancji z dokładnością nawet $\pm 0{,}005\%$ nie stwarza już obecnie większych trudności. Dokładność wyznaczenia $\bar{\varepsilon}_x$ wynika bezpośrednio z zależności (15)÷(19), przy czym decydującym jest błąd określenia odległości ostrzy (L_0), położenia rezystora na płytce (x_a, x_b), grubości podłoża (z_P) oraz strzałki ugięcia (z_0). Przy właściwie skonstruowanej przystawce umożliwiającej kontrolowane uginanie płaskorównoległych podłoży nie mających wad kształtu, maksymalny błąd względny w określeniu średniego wydłużenia rezystora dla $L_0 \geqslant 20$ mm, może być mniejszy od 1% [6]. Należy podkreślić, że dostępne podłoża z ceramiki alundowej posiadają szereg wad kształtu takich jak zwichrowanie, ugięcie, brak płaskorównoległości oraz niejednorodność struktury, co powoduje dodatkowe błędy trudne do oszacowania, mogące zwiększyć sumaryczny błąd w określeniu średniego wydłużenia rezystora nawet do kilku procent. Będzie to powodować również wzrost odchyleń standardowych wartości G_L i G_T.

4. PROCES PRZYGOTOWANIA REZYSTORÓW TESTOWYCH I METODYKA BADAŃ

Do badań piezorezystywnych właściwości rezystorów grubowarstwowych wykonano trzy serie rezystorów testowych. Do pierwszej wykorzystano prototypową serię past rutenianowych typu PS opracowanych w Politechnice Rzeszowskiej. Charakteryzowała się

ona tym, że zawartość (m_r) proszku rutenianu bizmutu w paście zmieniano kosztem szkliwa w zakresie od 20 do 60% wagowych z przyrostem równym 10%, przy czym w celu poprawy stabilności parametrów rezystorów w czasie i przy zmianach temperatury, wprowadzono dodatkowo proszek tlenku kadmowego w ilości 1% w stosunku do masy szkliwa [20]. Pasty oznaczono symbolami odpowiednio od PS20 do PS60, tak, że cyfry po symbolu literowym oznaczają wartość m_r w procentach wagowych. Rutenian bizmutu zastosowany w pastach uzyskano w wyniku reakcji w stanie stałym tlenków Bi_2O_3 i RuO_2, a średni rozmiar agregatów w tym proszku określony za pomocą mikroskopu i analizatora obrazu był równy około 1,12 μm, natomiast wyznaczony na podstawie pomiarów powierzchni właściwej proszku[1] rozmiar cząstki był równy 150 nm. Jako szkliwo zastosowano w pastach proszek szkła ołowiowo-boro-krzemowego o składzie w procentach wagowych: 65% PbO, 25%SiO_2 i 10%B_2O_3. Średni rozmiar cząstek proszku szkliwa określony zarówno metodą mikroskopową jak i sedymentacyjną był równy około 1,7 μm. W celu otrzymania past o odpowiednich właściwościach reologicznych wysuszone proszki materiałów wyjściowych, odważone w odpowiednich proporcjach, mieszane były w agatowym, planetarnym młynku kulowym w obecności acetonu, a następnie po odparowaniu acetonu do mieszaniny wprowadzono nośnik organiczny (terpineol + 3,5% etylocelulozy) w takiej ilości, aby pasta uzyskała lepkość o wartości w przedziale $20 \div 25$ Pa · s w temperaturze 20°C.

Druga seria rezystorów testowych wykonana była z dostępnych handlowo past rezystywnych nr 8049 i 8039 produkcji firmy Du Pont, w których ośrodkiem przewodzącym jest również rutenian bizmutu. Uzyskiwane z tych past rezystory grubowarstwowe mają rezystancje na kwadrat odpowiednio równe 200 kΩ/□ i 20 kΩ/□ [21].

Trzecią serię rezystorów testowych wykonano stosując dostępną handlowo pastę nr 1841 również produkcji firmy Du Pont, która była opracowywana między innymi w aspekcie maksymalizacji wartości czułości odkształceniowej rezystancji [16].

Opracowano wzór testowy, który zawierał 4 rezystory do pomiaru czułości odkształceniowej podłużnej o długościach 0,5, 1, 3 i 6 mm i parametrze x_a odpowiednio równym 21, 20,5, 18,5 i 15,5 mm oraz odległości L_0 równej 46 mm, a także identyczne 4 rezystory umieszczone poprzecznie do odpowiednich do pomiaru czułości odkształceniowej poprzecznej. Do wykonania kontaktów użyto przewodzącą pastę palladowo-srebrową nr 6120 produkcji firmy Du Pont.

Pasty nanoszono metodą sitodruku[2] na podłoża z ceramiki alundowej o rozmiarach $50{,}8 \times 15 \times 0{,}65$ mm^3 produkcji firmy Kyocera i wypalano w piecu tunelowym w ciągu 60 min w optymalnych dla danego typu past profilu temperaturowym. Warstwy kontaktowe były nanoszone i wypalane przed nanoszeniem warstw rezystywnych. W sumie wykonano 32 serie rezystorów testowych o liczności od 12 do 16.

Po dolutowaniu do pól kontaktowych elastycznych doprowadzeń rezystory testowe poddano badaniom umożliwiającym wyznaczenie czułości odkształceniowych rezystancji. Zastosowano w tym celu, opisaną w rozdziale 3, metodę trójpunktowego ugięcia płytki podłożowej. Opracowana przystawka umożliwiała ugięcie płytki podłożowej zgodnie

[1] jest to powierzchnia 1 grama proszku

[2] Proces przygotowania rezystorów testowych oraz część badań zrealizowano w Laboratoriach ESAT, Wydziału Elektrotechniki Uniwersystetu Leuven w Belgii podczas stażu naukowego autora

z wymaganiami metody i pomiar tego ugięcia z dokładnością ± 1 μm. Pomiar ugięcia realizowano za pompocą czujnika indukcyjnego z odczytem cyfrowym typu MDNF-fa o dokładności $\pm 0{,}5\%$ i rozdzielczości 0,1 μm. Pomiar rezystancji wykonano za pomocą omomierza cyfrowego typu YEW 2803 z dokładnością $\pm 0{,}02\%$.

Korzystając z zależności (20) i (21) obliczono dla każdego rezystora czułości odkształceniowe podłużną i poprzeczną, a następnie dla każdej serii wyznaczono ich wartości średnie ($\bar{G}_{L,T}$) oraz ich estymatory odchyleń standardowych (σ). Jednocześnie, ponieważ zaobserwowano silną zależność rezystancji na kwadrat warstw, wyznaczonej jako:

$$R_{\square} = R\frac{y_F}{x_F}, \tag{25}$$

gdzie R — rezystancja rezystora, x_F i y_F — długość i szerokość rezystora, od długości rezystora, to dla każdej płytki testowej określono współczynnik $r_{\square}$ zdefiniowany jako

$$r_{\square}(x_F) = \frac{R_{\square}(x_F)}{R_{\square}(x_{F0})}, \tag{26}$$

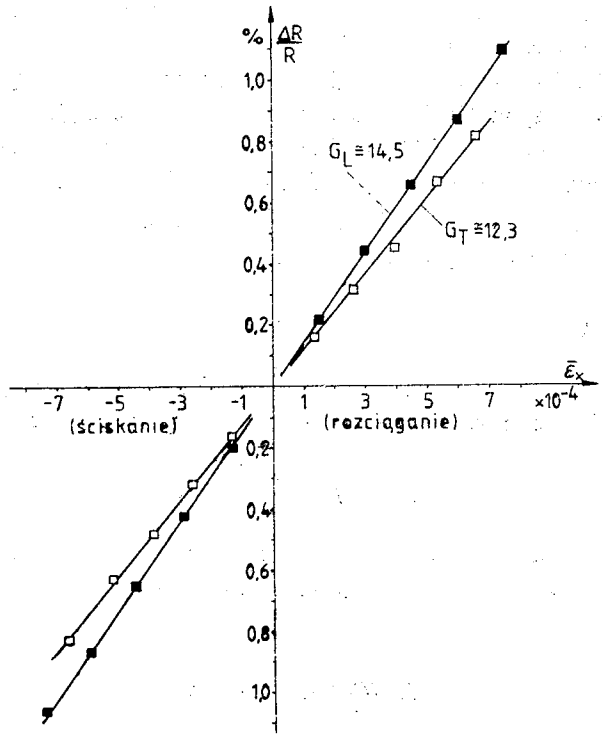
gdzie $x_{F0} = 0{,}5$ mm, a następnie wyznaczono wartości średnie i estymatory odchyleń standardowych tego współczynnika w poszczególnych seriach.

5. WYNIKI BADAŃ

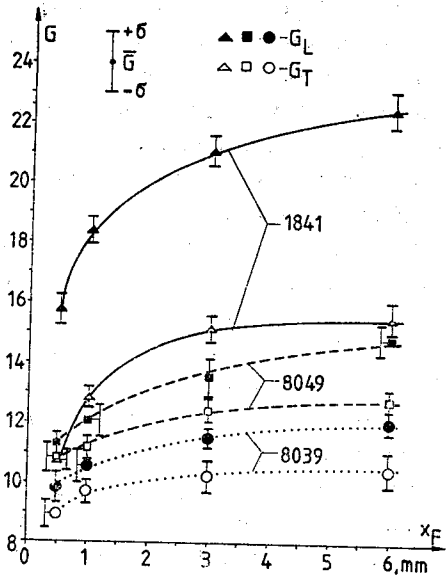
Na rys. 4 przedstawiono, jako przykładowe, charakterystyczne dla badanych rezystorów zależności względnych zmian rezystancji od wydłużenia uzyskane przy pomiarze czułości odkształceniowych rezystancji rezystorów o długości $x_F = 6$ mm wykonanych z pasty nr 8049. Charakterystyki te są praktycznie liniowe i pozbawione histerezy w badanym zakresie zmian wydłużenia i wskazują na niezależność współczynników G_L i G_T od wydłużenia. Właściwość ta nie była obserwowana w przypadku rezystorów wykonanych z pasty PS60 oraz niektórych wykonanych z pasty PS50, gdzie występowały pewne niewielkie zmiany nieodwracalne rezystancji podczas dokonywania pomiarów.

Zależności G_L i G_T od długości x_F rezystorów wykonanych z past nr 1841, 8049 i 8039 przedstawiono na rys. 5. Współczynniki te rosną wraz z x_F i obserwuje się stabilizację ich wartości dla $x_F \gtrsim 6$ mm. Rezystory z pasty nr 1841 charakteryzuje prawie dwukrotnie większa wartość czułości odkształceniowej podłużnej niż rezystory z past nr 8049 i 8039. Różnica wartości czułości odkształceniowych podłużnej i poprzecznej ($G_L - G_T$) rezystorów wykonanych z past nr 8049 i 8039 jest niewielka i rośnie ze wzrostem x_F osiągając wartości odpowiednio około 2,2 i 1,7. Różnica ta dla rezystorów wykonanych z pasty nr 1841 jest względnie duża i rośnie od około 5 dla $x_F = 0{,}5$ mm do 7 dla $x_F = 6$ mm.

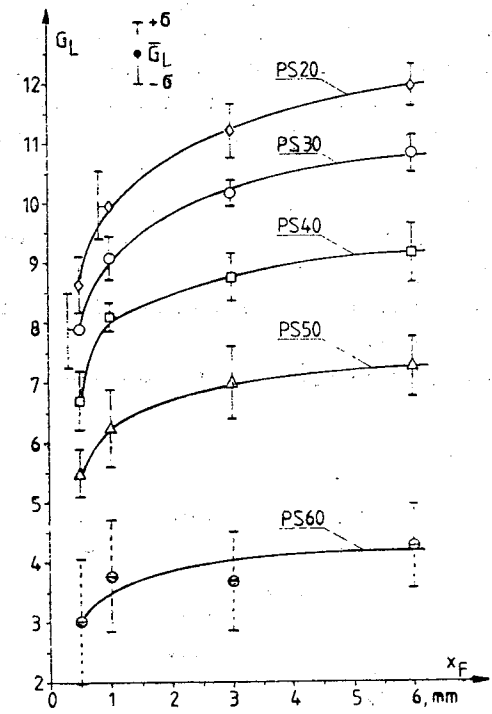
Rysunek 6 przedstawia zależności $G_L(x_F)$ dla rezystorów wykonanych z past serii PS. Kształt tych zależności jest bardzo podobny do tych, które pokazano na rys. 5. Współczynnik G_L maleje ze wzrostem zawartości rutenianu bizmutu w paście od wartości około 12 dla $m_r = 20\%$ od około 4 dla $m_r = 60\%$ (dla $x_F = 6$ mm). Względne zmiany współczynnika G_L wynoszą, w badanym zakresie zmian x_F, około 30% i są praktycznie niezależne od zawartości rutenianu bizmutu w paście. Wartości współczynników G_T dla rezystorów



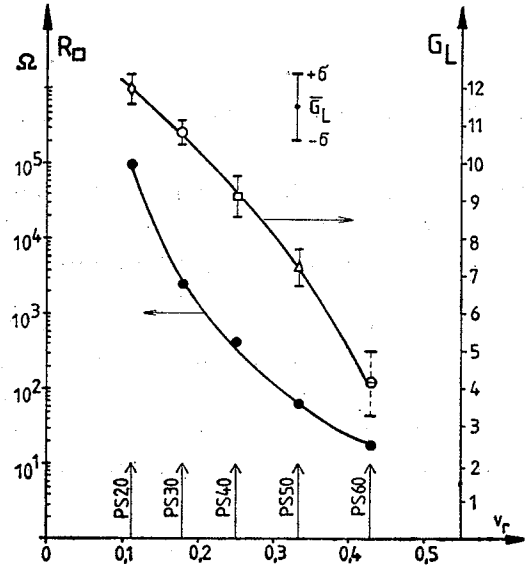
Rys. 4. Zależności względnych zmian rezystancji od względnego wydłużenia otrzymane podczas pomiaru czułości odkształceniowych podłużnej (G_T) i poprzecznej (G_L), reprezentatywne dla rezystorów o długości równej 6 mm wykonanych z pasty nr 8049



Rys. 5. Zależności podłużnej (G_L) i poprzecznej (G_T) czułości odkształceniowych rezystancji od długości rezystora dla rezystorów wykonanych z past nr 1841, 8049 i 8039



Rys. 6. Zależności podłużnej czułości odkształceniowej rezystancji G_L od długości rezystora x_F dla rezystorów otrzymanych z past serii PS

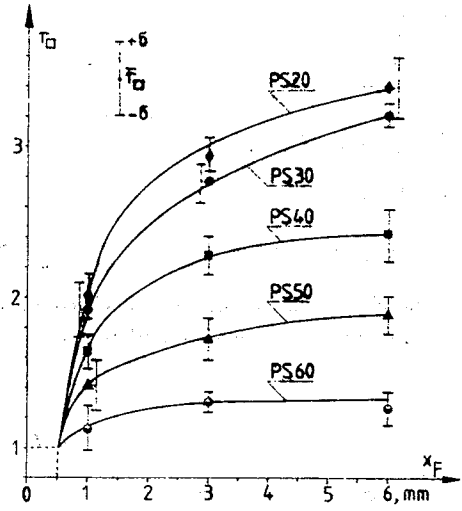


Rys. 7. Zależności rezystancji na kwadrat $R_□$ i podłużnej czułości odkształceniowej rezystancji G_L od ułamka objętościowego v_r rutenianu bizmutu dla rezystorów o długości 6 mm wykonanych z past serii PS

wykonanych z past PS były mniejsze od wartości G_L, a ich różnice dla poszczególnych wartości m_r zawierają się w zakresie 1,5—2,5 dla $x_F = 6$ mm, przy czym nie zaobserwowano tutaj korelacji pomiędzy wartością tej różnicy a wartością m_r. Charakterystyki $G_T(x_F)$ nie zostały przedstawione na rys. 6, ze względu na zachowanie jego czytelności.

Na rys. 7 przedstawiono zależności rezystancji na kwadrat ($R_\square$) oraz G_L od ułamka objętościowego[3] (v_r) rutenianu bizmutu dla rezystorów wykonanych z past serii PS. Ułamek v_r wyznaczono na podstawie znajomości składu pasty i gęstości materiałów wyjściowych przy założeniu, że ułamek objętościowy porów jest równy zero [15]. Estymatory odchyleń standardowych rezystancji w poszczególnych seriach były mniejsze od 8%. Z rys. 7 wynika, że w badanym zakresie zmian v_r wartość $R_\square$ maleje około 4 rzędy wartości, a współczynnik G_L — trzykrotnie.

Zależność $r_\square(x_F)$ dla rezystorów wykonanych z past serii PS przedstawiono na rys. 8.



Rys. 8. Zależności współczynnika $r_\square$ od długości rezystora dla rezystorów wykonanych z past serii PS

Wynika z niego, że rezystancja na kwadrat tych rezystorów zależy silnie od długości zwłaszcza w przypadku warstw o małych wartościach m_r. Przykładowo, rezystory wykonane z pasty PS20 charakteryzował ponad 3-krotny wzrost wartości $R_\square$ w badanym zakresie zmian x_F. Kształt pokazanych charakterystyk $r_\square(x_F)$ wskazuje, że wartości $R_\square$ stabilizują się dla dużych wartości x_F, przy czym następuje to szybciej w przypadku warstw o większych wartościach m_r, a więc małych wartościach $R_\square$. Jednocześnie, bezwzględne przyrosty wartości współczynnika $r_\square$ wyraźnie maleją i dla rezystorów z past serii PS60 przyrost ten wynosi już tylko około 0,3. Podobne jak dla past serii PS20 przebiegi zależności $r_\square(x_F)$ obserwowano również dla serii rezystorów wykonanych z past nr 8039, 8048 i 1841.

[3] Ułamek objętościowy danego ośrodka w warstwie jest zdefiniowany jako stosunek objętości tego ośrodka do objętości warstwy

6. PODSUMOWANIE

Zaprezentowane wyniki badań wskazują, że istnieje szereg uwarunkowań właściwości piezorezystywnych rezystorów grubowarstwowych zarówno technologicznych jak i konstrukcyjnych. Czułości odkształceniowe rezystancji zależą od składu pasty, a w szczególności od zawartości ośrodka przewodzącego (rys. 7), przy czym maleją one ze wzrostem v_r i jest to obserwowane nie tylko w rezystorach rutenianowych ale także w takich, które są wykonane z użyciem innych materiałów przewodzących, jak np. RuO_2, IrO_2 i in. [8]. Wykazano jednocześnie, że G_L i G_T zależą silnie od długości rezystora. Charakter zależności $G_{L,T}(x_F)$ pokazanych na rys. 5 i 6 oraz zależności $r_\square(x_F)$ zaprezentowanych na rys. 8 świadczy, że są one wynikiem wpływu warstw kontaktowych, gdyż następuje wyraźna stabilizacja wartości badanych współczynników G_L, G_T i $r_\square$ dla dużych wartości x_F. Materiał warstw kontaktowych dyfundując w czasie procesu wypalania w głąb warstwy rezystywnej modyfikuje jej właściwości w obszarach przykontaktowych. Należy przy tym podkreślić, że ze względu na specyficzne właściwości warstw grubych oraz konstrukcję rezystora, powierzchnie kontaktowe warstw rezystywnej i przewodzących mają złożone kształty i nie leżą w płaszczyźnie yz jak to założono na rys. 1. Jednakże takie przybliżenie jest w przypadku rozważanych badań dopuszczalne ze względu na to, że grubość warstwy rezystywnej jest znacznie mniejsza od jej długości oraz szerokości obszaru „zachodzenia" warstwy kontaktowej. Powoduje to, że w rozpatrywanym wpływie warstw kontaktowych polegającym na migracji metalu do warstwy rezystywnej, można rzeczywiste pola kontaktowe sprowadzić do płaskich kontaktów, jak to pokazano na rys. 1. Proces migracji jonów srebra z warstw kontaktowych przeanalizowano w pracy Cattaneo i innych [22] i wykazano, że zasięg ich oddziaływania na właściwości warstwy zależy od parametrów procesu wypalania rezystora i może być rzędu nawet $0{,}6 \div 1$ mm. Wyniki badań zależności $R_\square(x_F)$ rezystorów grubowarstwowych z kontaktami typu PdAg podane w pracy Naguiba [23] są zbieżne z podanymi na rys. 8 charakterystykami $r_\square(x_F)$ dla rezystorów wykonanych z past serii PS.

Dobra zgodność uzyskanych w wyniku badań wartości różnic $G_L - G_T$ (gdy $x_F \gtrsim 6$ mm) ze wzorem (14) dla rezystorów z past serii nr 8039, 8049 oraz PS, wskazuje, że rezystory te z punktu widzenia ich właściwości piezorezystywnych mogą być traktowane jako materiały izotropowe, jeżeli wpływ kontaktów jest pomijalny. Nie można tego natomiast powiedzieć o rezystorach wykonanych z pasty nr 1841, gdyż wyniki badań ich czułości odkształceniowych rezystancji nie potwierdzają słuszności wzoru (14). Ich struktura musi charakteryzować się dodatkowymi niejednorodnościami związanymi najprawdopodobniej z oddziaływaniem podłoża [24], a obserwowane znacznie większe wartości współczynników G_L i G_T w stosunku do tych, które charakteryzują rezystory z past serii PS20 i nr 8049 (o porównywalnych wartościach $R_\square$) mogą świadczyć o tym, że do wykonania pasty nr 1841 zastosowano materiały wyjściowe o innych właściwościach.

Przedstawione wyniki badań wskazują również, iż konstruktor czujników ciśnienia lub siły, działających na podstawie zjawiska piezorezystywnego w rezystorach grubowarstwowych, musi uwzględnić nie tylko bezwzględne wartości czułości odkształceniowych rezystancji, ale również ich zależności od rozmiarów rezystora i typu warstw kontaktowych oraz wyraźne ich skorelowanie z wartością rezystancji na kwadrat warstw. Za-

gadnienie to jest istotne z punktu widzenia jednoczesnej maksymalizacji czułości i zdolności rozdzielczej czujników [25].

Pomiary stabilności czułości odkształceniowych rezystancji rezystorów rutenianowych w czasie i przy zmianach temperatury będą stanowiły kontynuację zaprezentowanych w tej pracy wyników badań, natomiast pomiary rozkładu pierwiastków w objętości warstwy rezystywnej powinny ostatecznie określić obszar oddziaływania warstw kontaktowych.

BIBLIOGRAFIA

1. P. J. Holmes: *Changes in thick-film resistor value due to substrate flexure.* Microelectronics and Reliability, 1973, V. 12, ss. 395—396
2. G. R. Witt: *The electromechanical properties of thin films and the thin film strain gauge.* Thin Solid Films, 1974, V. 22, ss. 133—156
3. C. R. Tellier: *Thin metal film sensors.* Active and Passive. Elec. Comp, 1985, V. 12, ss. 9—12
4. R. Dell'Acqua, G. Dell'Orto, A. Simonetta, C. Canali: *Long term stability of thick film resistors under strain,.* Int. J. for Hyb. Microelectron., 1982, V. 5, ss. 82—85
5. C. Canali, D. Malavasi, B. Morten, M. Prudenziati, A. Taroni: *Strain sensitivity in thick film resistors.* IEEE Trans. on Comp., Hyb. and Manuf. Technol., 1980, CHMT-3, ss. 421—423
6. J. S. Shah: *Strain sensitivity of thick film resistors.* IEEE Trans. on Comp., Hyb. and Manuf. Technol., 1980, CHMT-3, ss. 554—564
7. B. Morten, L. Pirozzi, M. Prudenziati, A. Taroni: *Strain sensitivity in film and cermet resistors: measured and physical quantities.* J. Phys. D: Appl. Phys., 1979, V. 12, ss. L51-L54
8. C. Canali, D. Malavasi, B. Morten, M. Prudenziati, A. Taroni: *Piezoresistive effects in thick-film resistors.* J. Appl. Phys., 1980, V. 51, ss. 3282—3288
9. S. Paszczyński, J. Potencki, B. Rząsa: *Piezoresistive effects in thick-film structures.* Prace Naukowe ITE Politechniki Wrocławskiej, 1985, V. 30, (Proc. 5th and 6th Conf. of the ISHM — Poland Chapter, Bystre 1984), ss. 103—114
10. Z. H. Meiksin, W. P. Humbel: *Thick-film strain gauges and pressure transducers.* Electronic Letters, 1980, V. 16, ss. 242—243
11. M. Prudenziati, B. Morten, A. Taroni: *Characterization of thick-film resistor strain gauges on enamel steel.* Sensors and Actuators, 1981/82, V. 2, ss. 17-27
12. B. Rząsa, J. Potencki, S. Paszczyński: *Realisierungsmöglichkeiten von Temperatur- und Druckfühlern in Dickschichttechnik.* Wiss. Beiträge der IH Zwickau, 1985, V. 11, ss. 66—73
13. S. Paszczyński, B. Peurs, W. Sansen: *Some new constructions of force (pressure) sensors based on piezoresistive effects in thick film resistors.* Proc. of 4th Czechoslovac Conf. on Microelectronics and Microsystem, Bratysława 1986, ss. 143—150
14. R. Dell'Acqua, G. Dell'Orto, P. Vicini: *Thick-film pressure sensors: performances and practical applications.* Proc. of 3rd European Hyb. Microelectron. Conf., Avignon 1981, ss. 121—134
15. S. Paszczyński: *Wpływ wybranych czynników technologicznych na właściwości warstw grubych z rutenianem bizmutu.* Rozprawa doktorska, Instytut Technologii Elektronowej, Politechnika Warszawska, Warszawa 1979
16. *Informacja prywatna firmy Du Pont*
17. S. P. Timoshenko, J. N. Goodier: *Theory of elasticity.* Mc Graw-Hill Book Comp., New York 1973
18. S. Timoshenko: *Strength of materials: Part I- elementary theory and problems.* D. van Nostrand Comp., New York 1958, ss. 143—146
19. D. Szymański, S. Achmatowicz, J. Bekisz, B. Szczytko: *Modelowanie temperaturowych zmian rezystancji warstw metodą wprowadzania naprężeń mechanicznych.* Elektronika, 1984, V. XXV, nr 5, ss. 19—22

20. S. Paszczyński: *Niektóre aspekty wpływu tlenku kadmowego na właściwości rezystorów grubowarstwowych z rutenianem bizmutu.* Rozprawy Elektrotechniczne, 1983, t. 29, ss. 217—255
21. J. M. Wheeler: *Thic film conductors and resistors on dielectrics for high reliability applications.* Proc. of 5th European Hyb. Microelectron. Conf., Stresa 1985, ss. 51—60
22. A. Cattaneo, M. Cocito, F. Forlani, M. Prudenziati: *Influence of the metal migration from screen and-fired terminations on the electrical characteristics of thick-film resistors.* Electrocomp. Sci. and Technol., 1977, V. 4, ss. 205—211
23. H. M. Naguib: *The characterization of thick film resistors terminated with Pd/Ag conductors of various compositions.* Proc. of the 1977 International Microelectron. Symp., Baltimore 1977, ss. 48—59
24. R. W. Vest: *The effects of substrate composition on thick film circuit reliability.* Final Technical Reports, Purdue University, U.S.A., 1977—1979
25. S. Paszczyński, J. Potencki: *Thick film force (pressure) sensor: method of optimization of its construction.*, Vortragserihe „Konstruktive und technologische Probleme electronischer Funktionsblöcke" der 32. Intern. Wiss. Koll. TH Ilmenau, 1987, ss. 71—74

S. PASZCZYŃSKI

PIEZORESISTIVE PROPERTIES OF THICK FILM RESISTORS

Summary

An analytical description, the procedure and the results of research concern piezoresistive properties of a thick film resistor, are presented. The correlations between piezoresistive properties of the film material and possible for determining and measuring tensometric factors of the resistance of this resistor are analysed. The basic measuring methods of tensometric factors and sources of possible errors are displayed. For preparing the test resistor prototype, pastes of the range PS and Du Pont products 1841, 8049 and 8039 have been used. The influence of bismuth ruthenate concentration in the film and of the resistor length on the longitudinal and transverse tensometric factors of resistance and on the sheet resistivity has been measured. It was found that, in the case of using a PdAg contact film, a strong dependence of both the tensometric factors and the sheet resistivity on the resistor length occurred and that for lengths exceeding 5—6 mm their values become stable. It was, moreover, recorded that the tensometric factors decreased with the increased bismuth ruthenate concentration in film and that the difference between the longitudinal and transverse tensometric factors could be a measure of the isotropy of a resistive film. Conclusions concerning the most probable reasons of the measured changes of the resistor characteristics are drawn.

S. PASZCZYŃSKI

PIÉOZRÉSISTIVES PROPRIÉTÉS DES RÉSISTORS À COUCHE ÉPAISSE

Résumé

On a présenté la description analytique du phénomène piézorésistif, la méthode de mesure et les résultats des recherches piézorésistives des propriétés des résistors à couche épaisse. On a analysé la corrélation existant entre les proprétés piézorésistives du matériau de la couche et les oossibles de définir et de mesurer sensibilites de déformations de la résistance du résistor. On a discuté les principales méthodes de mesure de la sensibilité de deformation des resistances et on a indiqué les sources des erreurs de mesure possibles. Pour la production des résistors tests on s'est servi de la série prototype de pâtes résistives du type PS à base de ruthénate et d'accessibles dans le commerce pates résistives nr. 1841, 8039 et 8049, ainsi que de pate conductible nr. 6120 produite par la firme Du Pont. On a examiné l'influence de la teneur en ruthénale de bismuth dans la couche ainsi que celle de la longueur du résistor sur les sensibilités de déformation de la résistance et sur la résistance carrée. On a constaté, qu'il existe une forte dépendence entre des facteurs et les sensibilités de déformation et de la résistance carrée. On a formulé les conclusions sur les plus probables conditions influant sur le parcours des caractéristiques examinées des résistors à couche épaisse.

S. PASZCZYŃSKI

PIEZORESISTIVE EGIENSCHAFTEN VON DICKSCHICHTWIDERSTÄNDEN

Zusammenfassung

Im Beitrag wurden analytische Beschreibung, Vorgehensweise und Untersuchungsergebnisse von piezoresistiven Dickschichtwiderständen dargestellt. Es wurden Zusammenhänge zwischen den piezoresistiven Eigenschaften von Schichtmaterial und den noch definierbaren und messbaren Verformungsempfindlichkeiten des Widerstandes analysiert. Es wurden die Grundmethoden der Messung von Verformungsempfindlichkeiten eines Widerstandes besprochen sowie die Quellen möglicher Messfehler angezeigt. Zur Herstellung der Prüflinge wurde eine Serie von neuentwickelten Ruthenat-Pasten Typ PS und die handelsüblichen Widerstanspasten Typ 1841, 8049 und 8039 der Firma Du Pont angewendet. Es wurde der Einfluß des Gehalts von Wismut-Ruthenat in der Schicht und der Widerstandslänge auf die Längs- und Querverformungsempfindlichkeit des Widerstandes sowie auf den Widerstand pro Quadrat untersucht. Es konnte festgestellt werden, daß im Falle der Anwendung von Kontaktschichten Typ PdAg eine große Abhangigkeit der Verformungsempfindlichkeiten und des Widerstandes pro Quadrat von der Widerstandslänge existiert, wobei beide Parameter bei der Vergrößerung der Widerstandslänge zunehmen und ihre Werte werden erst ab Länge 5—6 mm stabil. Es konnte ermittelt werden, daß die Verformungsempfindlichkeiten mit der Zunahme des Wismut-Ruthenat-Gehalts in der Schicht immer kleiner werden und die Differenz zwischen den Werten der Längs- und Querverformungsempfindlichkeit ein Maß für die Isotropie der Widerstandsschicht sein kann. Es wurden schließlich Schußfolgerungen zu den höchstwahrscheinlichen Bedingungen des Verlaufs von Kennlinien der untersuchten Dickschichtwiderstände formuliert.

С. ПАЩИНЬСКИ

ПЬЕЗОРЕЗИСТИВНЫЕ СВОЙСТВА ТОЛСТОПЛЕНОЧНЫХ РЕЗИСТОРОВ

Резюме

Представлены аналитическое описание, методика и результаты исследования пьезорезистивных свойств толстопленочных резисторов. Проведен анализ корреляции между пьезорезистивными свойствами материала плёнки и возможными к определению и измерению чувствительностей деформационных сопротивлений резистора. Представлены основные методы измерения чувствительностей деформационных сопротивлений и источники возможных ошибок. Для приготовления испытуемых резисторов использовано прототипную серю паст типа PS и пасты 1841, 8049, 8039 производства фирмы Du Pont. Исследовано влияние содержания рутената висмута в плёнке и длины резистора на продольную и поперечную чувствительности деформационных сопротивлений и плёночное удельное сопротивление. Констатировано, что для использованных контактных плёнок типа PdAg, существует сильная зависимость между длиной резистора и деформационными чувствительностями сопротивления и плёночным удельным сопротивлением и эти параметры растут когда возрастает длина резистора а их значение стабилизируются для длин больше 5—6 мм. Замечено что чувствительности деформационные сопротивления уменшаются когда возрастает содержание рутената висмута в плёнке а разница между продольной и поперечной деформационными чувствительностями может быть мерой изотропии резистивной плёнки. Сформулированы выводы связаны с провдоподобными причинами формы иследованых характеристик толстоплёночных резисторов.

ROZPRAWY ELEKTROTECHNICZNE, 1989, 35, z. 1, ss. 149–158

537.311.33.2

Degradacja lutu stosowanego do mocowania kryształów laserów złączowych i diod elektroluminescencyjnych

WŁODZIMIERZ NAKWASKI

Instytut Fizyki, Politechnika Łódzka

Otrzymano 1987.11.16

Autoryzowano do druku 1988.04.08

W pracy przedstawiono najczęstsze przyczyny degradacji lutu łączącego z obudową kryształy laserów złączowych i diod elektroluminescencyjnych, tj. tworzenie się w nim warstwy międzymetalicznej oraz wzrost na jego powierzchni kryształów włoskowatych.

1. WPROWADZENIE

Rozwój technologii otrzymywania wielowarstwowych struktur półprzewodnikowych (metal-organic chemical vapour deposition (MO-CVD), molecular-beam epitaxy (MBE), atomic layer epitaxy (ALE), electroepitaxy etc.) doprowadził do sytuacji, w której o prawidłowym działaniu przyrządów półprzewodnikowych w dużej mierze zaczęła decydować jakość ich kontaktów oraz lutu łączącego kryształy tych przyrządów z ich obudowami. Dotyczy to szczególnie wymagań długotrwałej i stabilnej pracy przyrządów. W ten sposób zagadnienie to, spychane dotąd często na plan dalszy, a w każdym razie traktowane marginesowo, niespodziewanie zaczęło odgrywać wiodącą rolę. Jego prawidłowe rozwiązanie stało się kluczowym działaniem prowadzącym do dalszej optymalizacji konstrukcji przyrządów półprzewodnikowych. Obserwuje się to na przykład w optoelektronice półprzewodnikowej, gdzie problem odpowiedniego doboru i wykonania kontaktów oraz połączenia przyrząd-obudowa stał się zagadnieniem pierwszoplanowym.

W niniejszej pracy przedstawiono wyjaśnienie podstawowych przyczyn degradacji lutu łączącego z obudową kryształy laserów złączowych i diod elektroluminescencyjnych, tj. tworzenie się w nim warstwy międzymetalicznej oraz wzrost na jego powierzchni kryształów włoskowatych.

2. WPŁYW LUTU INDOWEGO NA DEGRADACJĘ

Badania [1, 2] obszarów w pobliżu kontaktów przyrządów wykonanych z półprzewodników $A^{III}B^{V}$ wykazały, że praktycznie nie ma ostrej granicy oddzielającej metal od półprzewodnika. Oba ośrodki w wyniku dyfuzji i reakcji chemicznych wzajemnie się

mieszają i przenikają. W szczególności nałożenie warstwy metalu na półprzewodnik powoduje częściową dysocjację tego drugiego ośrodka i w rezultacie tworzenie stopów metal-anion i metal-kation. Grubość warstwy przejściowej w temperaturze pokojowej zawiera się w przedziale od kilku do kilkudziesięciu warstw atomowych. Jak stąd więc widać, proces osadzania kontaktów metalicznych oraz mocowania przyrządu do jego obudowy może mieć poważny wpływ na elektryczne charakterystyki wytwarzanych przyrządów.

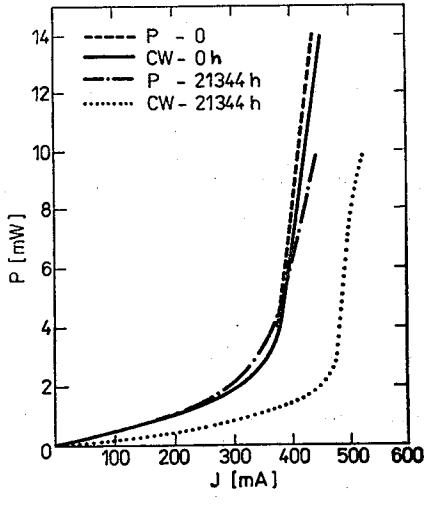
Niekiedy oporność cieplna R_T laserów złączowych bądź diod elektrolumienescencyjnych rośnie w miarę upływu czasu [3—6]. Świadczą o tym zachodzące w wyniku degradacji zmiany charakterystyk emisyjnych $P = f(J)$ (gdzie P — moc emitowanego promieniowania, a J — prąd zasilania) tych przyrządów w przypadku długotrwałej pracy z falą ciągłą (cw operation) (rys. 1). W trakcie pracy impulsowej takich zmian się nie obserwuje, co dowodzi, że wzrostowi oporności cieplnej nie towarzyszy analogiczny wzrost oporności elektrycznej.

Proces degradacji połączenia kryształu półprzewodnikowego z obwodem zewnętrznym zależy od stosowanego lutu i metalizacji kontaktowej. Wzrost R_T w laserach GaAs/(AlGa)As jest zwykle spowodowany pogarszaniem się przewodnictwa cieplnego lutu indowego między przyrządem a obudową, przy czym wzrost ten zależy tylko od temperatury i występuje również w diodach niespolaryzowanych [7].

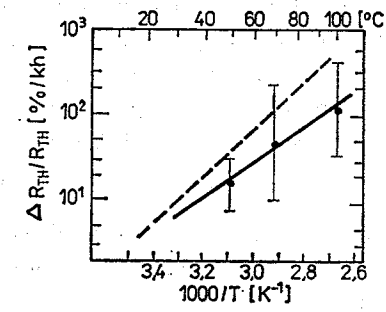
W pracy [7] przedstawiono wyniki obserwacji lutu indowego w sąsiedztwie zwierciadeł rezonatora za pomocą elektronowego mikroskopu skaningowego (SEM) po przyśpieszonym starzeniu w podwyższonej temperaturze (1000 godzin, 70°C lub 100°C). Stwierdzono, że lut ten stawał się granulkowaty i zawierał wiele bąbli.

Degradacja lutu indowego zachodzi w wyniku reakcji między atomami indu i złota [7]. Wynikiem dyfuzji atomów złota, której sprzyja wzrost temperatury, do warstwy indowej, są m.in. związki międzymetaliczne (Au_3In, AuIn, $AuIn_2$) powstające zgodnie z wykresem fazowym. Przewodność cieplna warstw międzymetalicznych jest stosunkowo niska, np. dla $AuIn_2$ wynosi ona w temperaturze pokojowej około 11 W/mK [8, 9].

Ponieważ, jak stwierdzono powyżej, jakość połączenia kryształu lasera z obudową



Rys. 1. Charakterystyki emisyjne $P = f(J)$ lasera złączowego przed (0 h) i po degradacji (21 344 h) dla pracy z falą ciągłą (cw) i pracy impulsowej (p) [3]



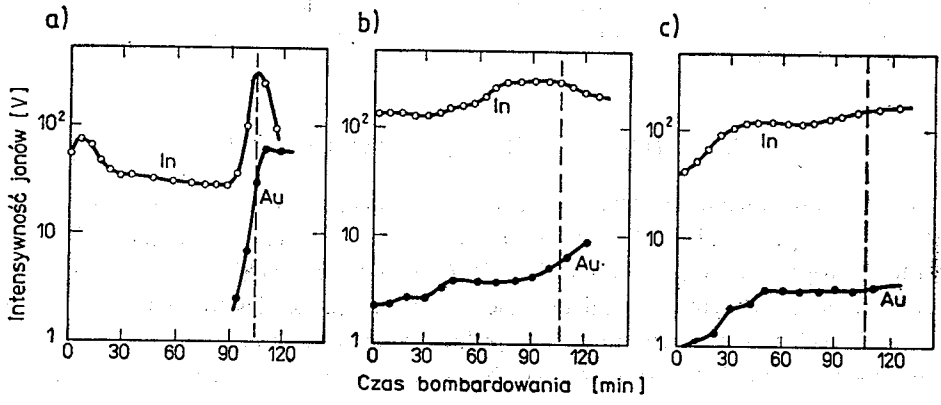
Rys. 2. Wzrost oporności cieplnej $\Delta R_{TH}/R_{TH}$ niespolaryzowanych [7] (linia ciągła) i spolaryzowanych [10] (linia przerywana) laserów złączowych podczas przyśpieszonego starzenia

zależy w dużym stopniu od jakości lutu indowego, jakość tego lutu odbija się na wartości oporności cieplnej. W pracy [7] badano oporność cieplną niespolaryzowanych laserów GaAs/(AlGa)As podczas przyśpieszonego starzenia w temperaturach 50, 70 i 100°C. Wyniki pomiarów przedstawiono na rys. 2 razem z wynikami pomiarów dla laserów spolaryzowanych napięciem w kierunku przewodzenia. Energia aktywacji dla obserwowanego wzrostu oporności cieplnej wynosiła około 0,6 eV, co zgadza się w granicach błędu z wcześniej opublikowaną wartością 0,5 eV [11].

W przypadku laserów (InGa)(AsP)/InP analogiczny proces degradacji jest wynikiem reakcji między materiałem diody laserowej (InP) i lutu (SnPb) [12].

W celu przeanalizowania mechanizmu degradacji lutu indowego zbadano w pracy [7] zmiany rozkładu atomów złota w warstwie indowej. Pomiar przeprowadzono w sposób następujący: kryształ podłoża został rozłupany na trzy części, z których dwie zostały poddane przyśpieszonemu starzeniu (1: 1000 h i 70°C, 2: 1000 h i 100°C). Następnie analizowano rozkłady atomów złota w warstwie In za pomocą metody IMA. Wyniki pomiarów, przedstawione na rys. 3, potwierdzają wcześniejsze wnioski o dominującym wpływie dyfuzji atomów złota na degradację warstwy lutu indowego.

Niezwykle istotny wpływ na działanie przyrządu ma jednorodność warstw jego kontaktu i lutu. Niestety zwykle często występują w nich pęcherzyki powietrza. Kobayashi



Rys. 3. Profile rozkładów atomów Au w warstwie indowej dla [7]: a) momentu połączenia kryształu półprzewodnikowego z obudową, b) po starzeniu 1000 h w temperaturze 70°C, c) po starzeniu 1000 h w temperaturze 100°C

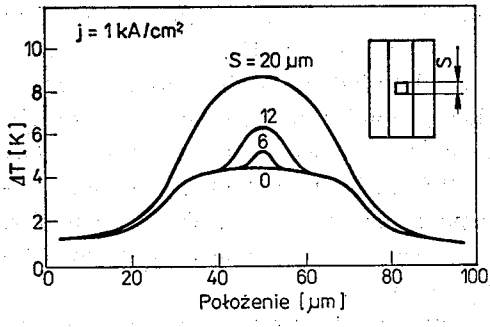
Linią przerywaną zaznaczono granicę warstwy indu tuż po wykonaniu połączenia kryształ — obudowa

i Iwane [13] badali teoretycznie wpływ obecności takiego pęcherzyka w warstwie lutu na wzrost temperatury warstwy czynnej lasera GaAs/(AlGa)As. Wyniki ich obliczeń przedstawiono na rys. 4. Z kolei na rys. 5 pokazano zależność maksymalnego przyrostu temperatury T_{max}, spowodowanego przez obecność pęcherzyka, od jego średnicy S. Zależność tę można dla gęstości prądu $j = 1$ kA/cm^2 przedstawić prostym wzorem [13]:

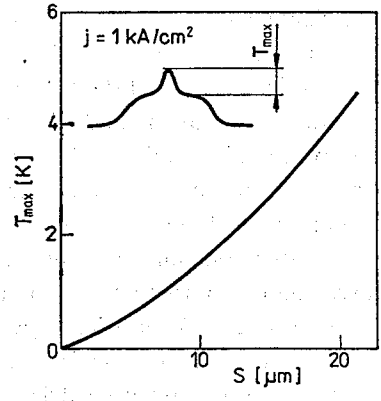
$$T_{max}[\mathrm{K}] = 0{,}0475\ (\mathrm{S}[\mu\mathrm{m}])^{1{,}5}. \tag{1}$$

Z obliczeń wynika, że aby zmniejszyć T_{max} do wartości poniżej 10% przyrostu temperatury dla przypadku jednorodnego lutu, średnica pęcherzyka nie powinna być większa niż 4 μm.

Podobne wyniki obliczeń otrzymał również Lynch [14], który stwierdził, że nawet w przypadku, gdy tylko nieznaczny ułamek kryształu lasera ma kiepski kontakt cieplny z obudową, laser taki wykazuje ogromną cieplną wrażliwość i może nastąpić zerwanie jego pracy z falą ciągłą nawet wówczas, gdy poza tym struktura przyrządu jest bez zarzutu.



Rys. 4. Obliczone rozkłady przyrostu temperatury ΔT w obszarze czynnym lasera GaAs/(AlGa)As zasilanego prądem gęstości $j = 1$ kA/cm^2 dla różnych rozmiarów S pęcherzyka w lucie indowym [13]



Rys. 5. Obliczony maksymalny przyrost temperatury T_{max} obszaru czynnego w funkcji rozmiaru pęcherzyka [13]. Rysunek w lewym górnym rogu definiuje pojęcie T_{max}

3. METODY POPRAWY JAKOŚCI LUTU

Trwałość lutowanego połączenia przyrząd półprzewodnikowy — obudowa można pozornie zwiększyć stosując jako lut zamiast indu luty twarde, czyli materiały o wyższej temperaturze topnienia, np. AuGe [10, 13, 15], AuSn [10, 15], AuSi [15], czy też SnPb [16]. Niekorzystnym zjawiskiem pojawiającym się w tym przypadku jest niestety wzrost naprężeń μ powstałych w heterostrukturze w wyniku różnej rozszerzalności liniowej materiału obudowy i kryształu półprzewodnikowego. Naprężenia te można przedstawić wzorem [17]:

$$\mu = [\mathrm{E}/(1-\mathrm{P})](k_1 - k_2)(T_c - T_A), \tag{2}$$

gdzie E jest modułem Younga, P — współczynnikiem Poissona, k_1 i k_2 — współczynnikami rozszerzalności liniowej materiałów odpowiednio obudowy i kryształu lasera, T_A — tempe-

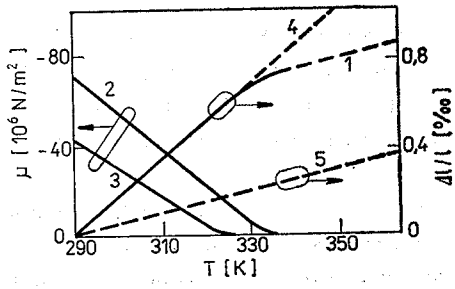
ratura otoczenia, a T_c — temperatura, w której, w trakcie ochładzania lutu łączącego kryształ lasera z podstawą, następuje ich trwałe związanie się. W przypadku lasera GaAs/(AlGa)As umocowanego do miedzianej obudowy za poomocą lutu indowego, wartości powyższych parametrów są następujące:

$$k_1 = 16{,}7 \cdot 10^{-6}\mathrm{K}^{-1} \quad [18]$$
$$k_2 = 5{,}82 \cdot 10^{-6}\mathrm{K}^{-1} \quad [18]$$
$$E/(1-\mathrm{P}) = 1{,}5 \cdot 10^{11}\ \mathrm{N/m^2} \quad [19]$$
$$T_c = 319\ \mathrm{K} \quad [17]$$

i wzór (2) redukuje się do postaci:

$$\mu = 1.6 \cdot 10^6(319 - T_A),\ \mathrm{N/m^2}, \tag{3}$$

co ilustruje rys. 6.



Rys. 6. Temperaturowa zależność względnego wydłużenia $\Delta l/l$ (krzywa (1)) kryształu lasera GaAs/(AlGa)As przymocowanego do obudowy miedzianej lutem SnPb grubości 5 μm [16]. Krzywa (2) przedstawia mechaniczne naprężenia μ powstałe w wyniku połączenia lasera z obudową lutem SnPb [16], a krzywa (3) — analogiczne naprężenia w wyniku użycia lutu indowego [17]; proste (4) i (5) ilustrują temperaturowe zależności względnego wydłużenia odpowiednio dla kryształu GaAs i dla kryształu miedzianej podstawy [16]

Jak widać, w temperaturze pokojowej wartość powyższych naprężeń jest tego samego rzędu, co naprężeń wynikających z różnych stałych sieci składowych heterostruktury [20]. Naprężenia te znacznie rosną w przypadku stosowania innych lutów ze względu na większe wartości parametru T_c, np. dla lutu SnPb: $T_c = 335$ K [16], co w przypadku powyższej struktury daje naprężenia $\mu(T_A = 290\ \mathrm{K}) = 7{,}5 \cdot 10^7\ \mathrm{N/m^2}$. Z drugiej zaś strony wiadomo, że szybkość degradacji stopniowej jest proporcjonalna do μ [16]. Stąd też, aby uniknąć nadmiernych naprężeń mechanicznych w krysztale, należy stosować miękki lut, na przykład ind. Wróciliśmy więc do punktu wyjścia.

Kompromisowe rozwiązanie zostało zaproponowane przez Hayakawa'ę i in. [11]. Użyli oni mianowicie lutu indowego, stosując jednocześnie warstwę molibdenu w charakterze bariery zabezpieczającej lut przed oddziaływaniem z GaAs i miedzianą obudową. Jednocześnie uniknęli formowania się warstwy międzymetalicznej (In, Au) podczas starzenia poprzez zastosowanie warstwy Au o grubości co najmniej 10 razy mniejszej od grubości warstwy indu. Wówczas w trakcie procesu łączenia, gdy cały ind jest w stanie stopionym, następuje całkowite rozpuszczenie Au w lucie indowym i stąd nie jest możliwe

późniejsze formowanie się warstwy międzymetalicznej. Innym rozwiązaniem może być zamiana w kontaktach warstwy Au przez warstwę Pt [21], bądź wykonanie innych kontaktów, nie zawierających warstwy Au, np. PtNi [22].

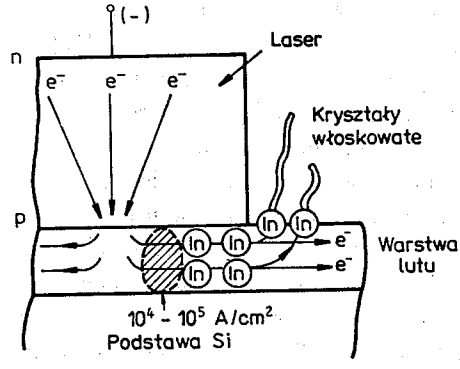
4. WZROST KRYSZTAŁÓW WŁOSKOWATYCH

Niekiedy, zarówno w przypadku laserów GaAs/(AlGa)As [23], jak i laserów (InGa)(AsP)/InP [24], obserwuje się pewien rodzaj degradacji raptownej, zachodzącej niespodziewanie w trakcie długotrwałego, stabilnego działania lasera. Degradacja ta jest zwykle spowodowana przez metalurgiczne defekty struktury krystalicznej indukowane w warstwie lutu w trakcie starzenia. Spośród tych defektów najczęściej obserwuje się:

— migracje atomów w głąb kryształu lasera,
— wzrost włoskowatych kryształów In (wąsów — *whiskers*) w wyniku elektromigracji w lucie indowym oraz
— wzrost włoskowatych kryształów Sn, gdy stosuje się luty Sn, AuSn lub SnPb.

Migracja atomów lutu ma miejsce w obszarze lutu w pobliżu zwierciadeł rezonatora z powodu lokalnego wzrostu temperatury. Proces ten jest przyśpieszany zarówno przez wzrost temperatury, jak i wzrost płynącego prądu. Tworzy on ścieżki niskooporowe w poprzek złącza p-n diody.

W laserach mocowanych do podstawy wykonanej z półizolacyjnego krzemu (Si) za pomocą lutu indowego, obserwuje się wzrost włoskowatych kryształów indu na powierzchni lutu w pobliżu paskowego obszaru czynnego. Z drugiej strony nie zanotowano wzrostu kryształów włoskowatych w laserach mocowanych do podstawy miedzianej (Cu). Jest to spowodowane tym, iż gęstość prądu w warstwie lutu na podłożu miedzianym jest znacznie niższa niż w analogicznej warstwie na podłożu półizolacyjnego krzemu. Dowodzi to elektromigracyjnego mechanizmu wzrostu kryształów włoskowatych indu, co schematycznie pokazano na rys. 7. Wzrost kryształów o średnicy 1—2 μm i długości 100—200 μm powoduje w końcu elektryczne zwarcie lasera.

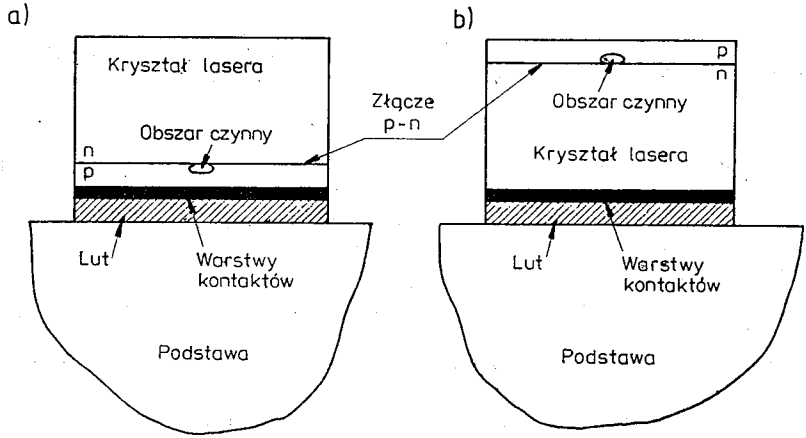


Rys. 7. Ilustracja elektromigracyjnego mechanizmu wzrostu kryształów włoskowatych w laserach mocowanych za pomocą lutu indowego do podstawy wykonanej z półizolacyjnego krzemu [23]

5. KONFIGURACJA ODWROTNA

Proces nakładania kontaktów, szczególnie po stronie *p*, oraz mocowania kryształu lasera do obudowy jest z uwagi na bliskość obszaru czynnego procesem mającym istotny wpływ na przebieg degradacji. Defekty generowane na powierzchni metal-półprzewodnik mogą bowiem migrować do pobliskiego obszaru czynnego, tworząc centra niepromieniste, bądź absorpcyjne. Z tego względu powierzchnia kryształu powinna być przed nałożeniem kontaktu odpowiednio przygotowana.

Najprostszą metodą radykalnego zmniejszenia kłopotów związanych z degradacją kontaktów i warstwy lutu jest zastosowanie tzw. konfiguracji odwrotnej [12, 25] (*p-side up*), tj. mocowanie laserów do podstawy stroną *n* (rys. 8). Oddzielenie obszaru czynnego o około



Rys. 8. Konfiguracja zwykła (a) i konfiguracja odwrotna (b) lasera złączowego

100 μm od warstwy lutu jest zazwyczaj wystarczającym zabezpieczeniem przed migracją atomów lutu, czy też wzrostem kryształów włoskowatych. Konfiguracja taka ułatwia przy tym sprzężenie lasera ze światłowodem włóknistym.

Rola powyższej konfiguracji wzrosła szczególnie w ostatnim czasie, kiedy to wspomniany na wstępie rozwój technik epitaksjalnych (MO-CVD, MBE, ALE, electroepitaxy etc.) umożliwił masową produkcję laserów złączowych o stosunkowo niskich prądach progowych. Wówczas wady konfiguracji odwrotnej (*p-side up*) stają się mniej dotkliwe. Jest więc ona obecnie stosowana coraz powszechniej przez wiele firm produkujących lasery.

Główną wadą konfiguracji odwrotnej jest znaczny wzrost oporności cieplnej spowodowany odprowadzaniem strumienia ciepła przez grubą warstwę podłoża [25]. Prowadzi to do pogorszenia jego charakterystyk eksploatacyjnych, przede wszystkim zmniejszając jego trwałość, a przy silnym pobudzeniu lasera może nawet doprowadzić do przerwania akcji laserowej.

PODSUMOWANIE

Wzrost wymagań odnośnie stabilności i trwałości pracy przyrządów półprzewodnikowych, w tym półprzewodnikowych przyrządów optoelektronicznych, spowodował koniecz-

ność radykalnej poprawy jakości połączenia kryształ półprzewodnikowy — obudowa. Temat ten stał się nagle zagadnieniem pierwszoplanowym. Celem niniejszej pracy było zwrócenie uwagi polskiego czytelnika na niebezpieczeństwo traktowania kontaktów i lutu łączącego kryształ przyrządu z obudową jako najbardziej stabilnych elementów przyrządów półprzewodnikowych.

Przedstawione mechanizmy degradacji warstwy lutu nie są jedynymi. Wybrano je jednakże, gdyż spotyka się je najczęściej i stanowią największe zagrożenie dla trwałości optoelektronicznych przyrządów półprzewodnikowych.

BIBLIOGRAFIA

1. J. Nogami, T. Kendelewicz, I. Lindau, W. E. Spicer: *Binding-energy shift from alloying at metal-compound-semiconductor interfaces.* Physical Review B, 1986, vol. 34, pp. 669—674
2. J. Wuerftl, H. C. Hartnagel: *Field and temperature dependent life-time limiting effects of metal-GaAs interfaces of device structures studied by XPS and electrical measurements.* Proc. 1986 Intern. Reliability Physics Symposium, Anaheim, CA, USA, April 1—3, 1986
3. H. Kressel, M. Ettenberg, I. Ladany: *Accelerated step temperature aging of $Al_x Ga_{1-x}$ As heterojunction laser diodes.* Appl. Phys. Lett., 1978, vol. 32, pp. 305—308
4. S. Ritchie, R. F. Godfrey, B. Wakefield, D. H. Newman: *The temperature dependence of degradation mechanisms in long lived (AlGa)As DH lasers.* J. Appl. Phys., 1978, vol. 49, pp. 3127—3132
5. M. Ettenberg, H. Kressel: *The reliability of (AlGa)As cw laser diodes,* IEEE J. Quantum Electron., 1980, vol. QE-16, pp. 186—196
6. M. Ettenberg, I. Ladany: *Metallization for diode lasers.* J. Vac. Sci. Technol., 1981, vol. 19, pp. 799—802
7. K. Fujiwara, H. Imai, T. Fujiwara, K. Hori, M. Takusagawa: *Analysis of deterioration in In solder for GaAlAs DH lasers.* Appl. Phys. Lett., 1979, vol. 35, pp. 861—863
8. W. Both, R. Schliesser: *Metallurgische Reaktion von Gold in Weichloten — Ursache der Kontaktdegradation optoelektronischer Bauelemente.* Feingeraetetechnik, 1986, vol. 35, s. 410—413
9. W. Both, R. Schliesser: *Thermische Leitfaehigkeit von Weichloten und ihren intermetallischen Phasen mit Gold,* Schweisstechnik, 1987, vol. 37 pp. 111—112
10. K. Fujiwara, T. Fujiwara, K. Hori, M. Takusagawa: *Aging characteristics of $Ga_{1-x} Al_x$ As double-heterostructure lasers bounded with gold eutectic alloy solder.* Appl. Phys. Lett., 1979, vol. 34, pp. 668—670
11. T. Hayakawa, S. Yamamoto, S. Matsui, T. Sakurai, T. Hijikata: *Long-lived (GaAl)As DH lasers bonded with In produced by eliminating deterioration of In solder.* Jpn. J. Appl. Phys., 1982, vol. 21, pp. 725—727
12. S. Tsuji, K. Mizuishi, Y. Nakayama, M. Shimaoka, M. Hirao: *InGaAsP/InP laser diodes mounted on semi-insulating SiC ceramics.* Jpn. J. Appl. Phys., 1983 vol. 22, Supplement, pp. 239—242
13. T. Kobayashi, G. Iwane: *Three dimensional thermal problems of double-heterostructure semiconductor lasers.* Jpn. J. Appl. Phys., 1977, vol. 16, pp. 1403—1408
14. R. T. Lynch, Jr.: *Effect of inhomogeneous bonding on output of injection lasers.* Appl. Phys. Lett. 1980, vol. 36, pp. 505—506
15. T. Hayakawa, N. Miyauchi, S. Yamamoto, H. Hayashi, S. Yano, T. Hijikata: *Improved lifetimes of (GaAl)As visible (740 nm) lasers by reducing bonding stress.* Appl. Phys. Lett., 1983, vol. 42, pp. 23—24
16. T. B. Zhizdyuk, N. D. Zhukov, V. M. Kalinichenko, A. N. Kuźmin,

G. I. Ryabtsev, S. N. Sokolov: *About degradation of GaAs-(Al, Ga)As heterolasers attached to a heat sink with the aid of a hard solder.* Zhurnal Tekhnicheskoj Fiziki, 1985, vol. 55, pp. 1463—1465
17. L. E. Bataj, Yu. L. Bessonov, V. F. Voronin, A. N. Kuźmin, G. T. Pak, G. I. Ryabtsev, S. M. Sapozhnikov, L. V. Tanin: *An investigation of mechanical stress in a system: injection heterolaser — heat sink.* Pisma v Zh. T. F., 1983, vol. 9, pp. 6—10
18. S. I. Novikova: *Thermal Expansion of Solid States*, Moskva, 1974
19. W. A. Brantley: J. Appl. Phys., 1973, vol. 44, pp. 534
20. G. H. Olsen, M. Ettenberg: *Calculated stresses in multilayered heteroepitaxial structures.* J. Appl. Phys., 1977, vol. 48, pp. 2543—2547
21. C. Lindstroem, P. Tihanyi: *Kink-free narrow stripe proton isolated GaAlAs/GaAs injection lasers.* Electron. Lett., 1980, vol. 16, pp. 121—123
22. C. Lindstroem, P. Tihanyi: *Temperature stable heat sink metallisation for improved bonding of c.w. stripe lassers*, Electron. Lett., 1980, vol. 16, pp. 97—98
23. K. Mizuishi: *Some aspects of bonding-solder deterioration observed in long-lived semiconductor lasers: Solder migration and whisker growth.* J. Appl. Phys., 1984, vol. 55, pp. 289—295
24. M. Fukuda, O. Fujita, G. Iwane, *Failure modes of InGaAsP/InP lasers due to adhesives.* IEEE Trans. Components, Hybrids and Manufact. Technol., 1984, vol. CHMT-7, pp. 202—206
25. P. G. Eliseev, M. Sh. Kobyakova, G. T. Pak, V. V. Popovicher, S. N. Sokolov: *An injection heterolaser of the mesastripe design with heat removal through the substrate.* Kvantovaya Elektronika, 1980, vol. 7, pp. 2504—2506

W. NAKWASKI

DEGRADATION OF A SOLDER USED FOR BONDING LASER DIODES AND LIGHT-EMITTING DIODES

Summary

The most often met causes of solder deterioration in laser diodes and light-emitting diodes, i.e. the formation of the intermetallic layer and the so-called whiskers growth, are presented in the paper.

W. NAKWASKI

DÉGRADATION DE LA SOUDURE UTILISÉE À LA FIXATION DES CRISTAUX DES LASERS DE JONCTION ET DES DIODES ÉLECTROLUMINESCENTS AU BOÎTIER

Résumé

Dans l'étude on cité les causes les plus fréquentes de la dégradation de la soudure unissant au boîtier les cristaux des lasers de jonction et des diodes électroluminescentes, c'est à dire la création au dedans de la soudure de la couche intermétallique et la croissance sur sa surface des cristaux capillaires.

W. NAKWASKI

WEICHKLOTENDEGRADIERUNG VON INJEKTIONSLASERN UND ELEKTROLUMINESZENZDIODEN

Zusammenfassung

Der Aufsatz nennt die öftesten Ursachen der Weichlotendegradierung von Injektionslasern und Elektrolumineszenzdioden, nämlich die Bildung der intermettalischen Phasen, das heisst das Schnurrbartphänomen.

В. НАКВАСКИ

ДЕГРАДАЦИЯ ПРИПОЯ СВЯЗЫВАЮЩЕГО КРИСТАЛЛЫ ЛАЗЕРНЫХ ДИОДОВ И СВЕТОИЗЛУЧАЮЩИХ ДИОДОВ С КОРПУСОМ

Р е з ю м е

Представлены самые частые причины деградации припоя связывающего кристаллы лазерных диодов и светоизлучающих диодов с корпусом, а именно образование интерметаллического слоя а также так называемое явление усов.

ROZPRAWY ELEKTROTECHNICZNE, 1989, 35, z. 1, ss. 159–174

621.317.791:621.382.2

Aproksymacja charakterystyki statycznej diody tunelowej funkcją wykładniczą

LECH J. WEISS, TADEUSZ WYSOCKI, TADEUSZ WYSOCKI JR
Akademia Techniczno-Rolnicza w Bydgoszczy

FELICJA WYSOCKA
Wyższa Szkoła Pedagogiczna w Bydgoszczy

Otrzymano 1987.03.15
Autoryzowano do druku 1988.04.27

W artykule opisano funkcję wykładniczą nadającą się do aproksymowania statycznej charakterystyki diody tunelowej oraz sposób wyznaczania parametrów tej funkcji, gdy znane są wartości parametrów diody. Podano również dwa przykłady liczbowe ilustrujące zastosowanie zaproponowanej metody.

1. WPROWADZENIE

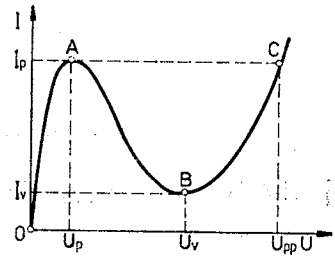
Dla dokonania analizy układu zawierającego diodę tunelową konieczne jest przedstawienie charakterystyki statycznej tej diody w postaci analitycznej. Producenci przyrządów półprzewodnikowych nie zamieszczają w katalogach wykresów charakterystyk statycznych konkretnych diod, lecz jedynie jakościowy przebieg $I = f(U)$ z zaznaczeniem na nim charakterystycznych punktów (rys. 1). Podstawowe parametry diody tunelowej, podawane w katalogach, to:

I_p — prąd szczytu,
U_p — napięcie szczytu,
I_v — prąd doliny,
U_v — napięcie doliny,
U_{pp} — napięcie „przerzutowanego" punktu szczytowego,

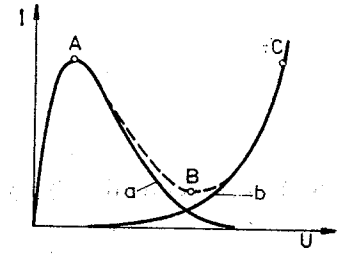
a ponadto:

R_s — rezystancja szeregowa,
R_N — rezystancja ujemna, mierzona w punkcie przegięcia opadającego odcinka charakterystyki statycznej.

Na charakterystykę statyczną diody tunelowej składają się dwie składowe: składowa tunelowa (krzywa a, rys. 2) oraz składowa nośników wstrzykniętych prądu złącza, zwana składową dyfuzyjną (krzywa b, rys. 2).



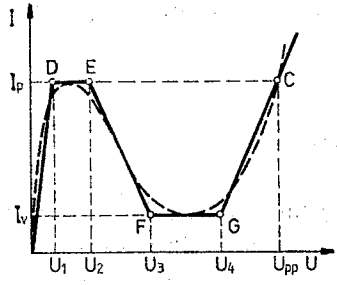
Rys. 1. Wykres przykładowej charakterystyki statycznej diody tunelowej



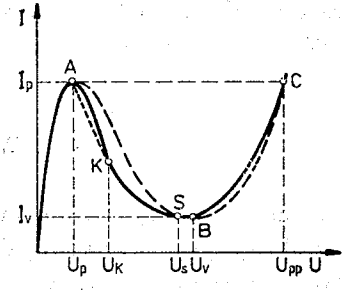
Rys. 2. Wykres zależności prądów składowych diody tunelowej od napięcia: a — składowa tunelowa, b — składowa dyfuzyjna

2. APROKSYMACJE DOTYCHCZAS STOSOWANE

Stosowane są różne sposoby aproksymacji charakterystyk statycznych diody tunelowej, jak np.: aproksymacja odcinkami linii prostych, aproksymacja odcinkami łuków paraboli kwadratowych, aproksymacja odcinkiem linii prostej i odcinkiem łuku paraboli kwadratowej, aproksymacja funkcjami potęgowymi, aproksymacja wielomianami oraz aproksymacja funkcjami wykładniczymi [2, 3].



Rys. 3. Odcinkowo-liniowa aproksymacja charakterystyki statycznej diody tunelowej



Rys. 4. Aproksymacja charakterystyki statycznej diody tunelowej dwoma łukami paraboli kwadratowej

W przypadku odcinkowo-liniowej aproksymacji, rzeczywistą charakterystykę diody przybliża się linią łamaną, jak to pokazano na rysunku 3. Analityczne wyrażenie aproksymujące charakterystykę diody ma postać:

$$I = \begin{cases} a_1 U & \text{dla} \quad 0 \leqslant U \leqslant U_1 \\ I_p & \text{dla} \quad U_1 < U \leqslant U_2 \\ I_p - a_2(U-U_2) & \text{dla} \quad U_2 < U \leqslant U_3 \\ I_v & \text{dla} \quad U_3 < U \leqslant U_4 \\ I_v + a_3(U-U_4) & \text{dla} \quad U_4 < U \end{cases} \tag{1}$$

gdzie:

$$a_1 = I_p/U_1; \tag{2}$$

$$a_2 = \frac{I_p - I_v}{U_3 - U_2}; \tag{3}$$

$$a_3 = \frac{I_p - I_v}{U_{pp} - U_4} \tag{4}$$

Zaletą aproksymacji liniowo-odcinkowej jest zastosowanie funkcji liniowych do opisu zależności $I = f(U)$. Jednakże taka aproksymacja, nawet przy użyciu linii łamanej złożonej z pięciu odcinków prostych, jest mało dokładna. Ponadto położenie punktów D, E, F, G, stanowiących naroża linii łamanej, nie jest jednoznacznie określone; współrzędne U tych punktów nie są powiązane ze współrzędnymi U ekstremalnych punktów charakterystyki diody (punktów A i B, rys. 1). Dodatkową wadę stanowi fakt, że przy tego rodzaju aproksymacji ujemna rezystancja diody na odcinku EF (rys. 3) ma stałą wartość, podczas gdy w rzeczywistości na odcinku AB (rys. 1) maleje od nieskończoności (pkt. A) do R_N i następnie znów wzrasta do nieskończoności (pkt. B).

Sposób aproksymacji charakterystyki diody dwoma łukami paraboli kwadratowej zilustrowany jest rysunkiem 4. Funkcja aproksymująca ma w tym przypadku postać

$$I = \begin{cases} I_p - b_1(U - U_p)^2 & \text{dla} \quad 0 \leqslant U \leqslant U_k \\ I_v + b_2(U - U_s)^2 & \text{dla} \quad U_k < U, \end{cases} \tag{5}$$

gdzie:

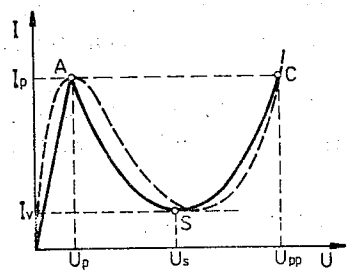
$$U_s = \frac{U_p + U_{pp}}{2}; \tag{6}$$

$$b_1 = \frac{I_p}{U_p^2} \tag{7}$$

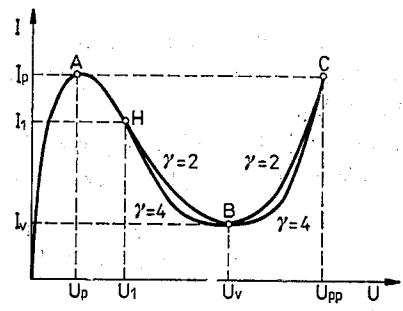
$$b_2 = \frac{I_p - I_v}{(U_p - U_s)^2}; \tag{8}$$

$$U_k = \frac{2(b_2 U_s + b_1 U_p)}{b_1 + b_2} - U_p. \tag{9}$$

Zaletą takiej aproksymacji, w porównaniu z aproksymacją odcinkowo liniową, jest to, że funkcja przybliżająca składa się tylko z dwu, a nie pięciu odcinków. Uzyskiwana dokładność przybliżenia jest podobna jak przy aproksymacji odcinkowo-liniowej. Wadą natomiast jest nieliniowość funkcji (5).



Rys. 5. Aproksymacja charakterystyki statycznej diody tunelowej odcinkiem prostej i łukiem paraboli kwadratowej



Rys. 6. Aproksymacja charakterystyki statycznej diody tunelowej czterema łukami paraboli drugiego wzgl. czwartego stopnia

W przypadku aproksymacji charakterystyki diody odcinkiem prostej i łukiem paraboli kwadratowej (rys. 5) funkcja aproksymująca opisana jest w następujący sposób

$$I = \begin{cases} aU & \text{dla} \quad 0 \leqslant U \leqslant U_p \\ I_v + b(U - U_s)^2 & \text{dla} \quad U_p < U, \end{cases} \tag{10}$$

gdzie:

$$a = \frac{I_p}{U_p}; \quad (11) \qquad b = \frac{I_p - I_v}{(U_p - U_s)^2}; \quad (12) \qquad U_s = \frac{U_p + U_{pp}}{2}. \quad (13)$$

Dobre przybliżenie statycznej charakterystyki diody tunelowej można otrzymać stosując czteroodcinkową aproksymację funkcjami potęgowymi, jak to pokazano na rysunku 6. Łuki OA oraz AH są aproksymowane łukami paraboli kwadratowej, a łuki HB oraz BC łukami paraboli kwadratowej lub paraboli wyższego stopnia. Na rysunku 6 pokazano wykresy łuków HB oraz BC dla $\gamma = 2$ oraz $\gamma = 4$, gdzie γ jest wykładnikiem potęgi. Funkcja aproksymująca ma postać

$$I = \begin{cases} I_p - \dfrac{I_p}{U_p^2}(U - U_p)^2, & \text{dla} \quad 0 \leqslant U \leqslant U_p \\ I_p - \dfrac{I_p - I_1}{(U_1 - U_p)^2} \cdot (U - U_p)^2, & \text{dla} \quad U_p < U \leqslant U_1 \\ I_v + (I_1 - I_v) \cdot \left(\dfrac{U_v - U}{U_v - U_1}\right)^{\gamma}, & \text{dla} \quad U_1 < U \leqslant U_v \\ I_v + (I_p - I_v) \cdot \left(\dfrac{U - U_v}{U_{pp} - U_v}\right)^{\gamma}, & \text{dla} \quad U_v < U \leqslant U_{pp} \end{cases} \quad (14)$$

gdzie: γ — wykładnik potęgowy ($\gamma = 2 \div 4$), H — punkt przegięcia łuku AB charakterystyki.

W ogólnym przypadku pochodna funkcji aproksymującej nie jest ciągła.

Charakterystykę statyczną diody tunelowej można również przybliżać wielomianami. Funkcja aproksymująca ma wówczas postać

$$I = \sum_{j=1}^{n} a_j U^j, \quad (15)$$

przy czym stopień wielomianu n jest zależny od wymaganej dokładności przybliżenia. Gdy zachodzi konieczność przeprowadzenia dokładnych obliczeń, trzeba przyjmować $n = 9 \div 11$ [2].

Stosowane jest również przybliżanie statycznej charakterystyki diody tunelowej z wykorzystaniem funkcji wykładniczych. W [2] podana jest postać funkcji aproksymującej

$$I = I_p \left(\frac{U}{U_p}\right)^{\gamma} \cdot \exp\left[\gamma\left(1 - \frac{U}{U_p}\right)\right] + I_0[\exp(\beta U) - 1], \quad (16)$$

gdzie pierwszy składnik sumy reprezentuje składową prądu diody wywołaną efektem tunelowym, a drugi — składową spowodowaną wstrzykiwaniem nośników. Wartość współczynnika γ, dla którego błąd aproksymacji charakterystyki w obszarze szczytu prądu jest możliwie mały, wynosi [2]:

— dla diod germanowych $1{,}4 \div 1{,}6$

— dla diod z arsenku galu $1{,}6 \div 1{,}8$.

Wartości współczynników I_0 oraz β zależą od typu diody. W pracy [2] nie podano ani sposobu ich wyznaczenia, ani typowych ich wartości. Stwierdza się mianowicie, że funkcja

(16) dobrze aproksymuje statyczną charakterystykę diody tunelowej „w obszarze maksimum prądu". Stwierdzenie to jest zresztą zilustrowane wykresem, z którego wynika, że najmniejszy błąd aproksymacji otrzymuje się dla $U \approx (1 \div 1{,}5)U_p$, a ze wzrostem napięcia powyżej 1,5 U_p błąd ten, stale dodatni, wyraźnie rośnie. Np. dla $U \approx 2\ U_p$ błąd wynosi już około +20%. Natomiast w [2] nie podano, że pierwszy człon funkcji (16), aproksymujący składową tunelową, wykazuje w początku układu współrzędnych nachylenie równe zeru, podczas gdy dla rzeczywistej charakterystyki diody występuje w tym punkcie lokalne maksimum nachylenia.

3. NOWA APROKSYMACJA WYKŁADNICZA

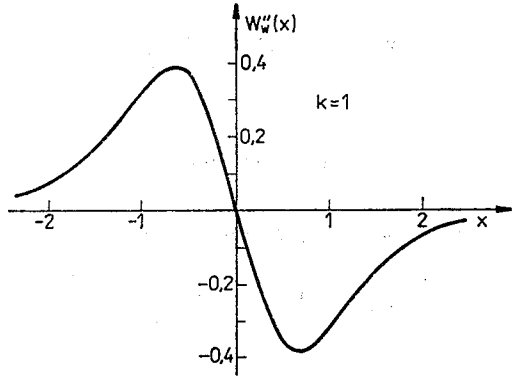
Zdaniem autorów niniejszej pracy lepsze przybliżenie składowej tunelowej prądu diody, decydującej o kształcie łuku OAB charakterystyki statycznej, można uzyskać przy użyciu drugiej pochodnej funkcji

$$W_w(x) = \frac{1}{1+\exp(-2kx)} \tag{17}$$

stosowanej do aproksymacji jednostkowej funkcji skokowej [5]. Druga pochodna tej funkcji jest równa

$$W''_{(w)}(x) = \frac{4k^2\exp(-2kx)[\exp(-2kx)-1]}{[1+\exp(-2kx)]^3}, \tag{18}$$

a jej wykres jest przedstawiony na rysunku 7. Funkcja (18), tak jak rzeczywista charakterystyka diody tunelowej, wykazuje największe nachylenie w początku układu współ-



Rys. 7. Wykres funkcji $W''_w(x)$ dla $k = 1$

rzędnych. Wartości funkcji aproksymującej składową tunelową prądu diody z wykorzystaniem zależności (18), jak to zostanie zilustrowane przykładem 1, są dla $U > 1{,}1\ U_p$ nieco mniejsze od wartości funkcji aproksymującej tę składową według zależności (16). Np. dla $U = 2\ U_p$ wynoszą odpowiednio: wg (16) $I(2U_p) = 1{,}010$ mA, a wg (18) $I(2U_p) = 0{,}899$ mA, czyli około 10% mniej. Daje to podstawę do stwierdzenia, że wyko-

rzystując zależność (18) można uzyskać lepsze przybliżenie składowej tunelowej niż według zależności (16).

W celu wykorzystania funkcji (18) do aproksymacji składowej tunelowej prądu diody trzeba wyznaczyć

— współrzędne ekstremów funkcji,

— współrzędne punktów przegięcia wykresu funkcji,

— wartość pierwszej pochodnej funkcji (18) w punkcie przegięcia.

Pierwsza pochodna funkcji (18) jest równa

$$W_w'''(x) = \frac{(2k)^3\exp(-2kx)}{[1+\exp(-2kx)]^4}\,[\exp(-4kx)-4\exp(-2kx)+1]. \tag{19}$$

Dla wyznaczenia współrzędnych ekstremów funkcji (18) przyrównujemy $W_w'''(x)$ do zera i otrzymujemy rozwiązanie równania

$$W_w'''(x) = 0 \tag{20}$$

jako

$$\exp(-2kx) = 2 \pm \sqrt{3} \tag{21}$$

i stąd

$$x_1 = -\frac{\ln(2+\sqrt{3})}{2k} \tag{22}$$

oraz

$$x_2 = -\frac{\ln(2-\sqrt{3})}{2k}. \tag{23}$$

Po wyliczeniu wartości liczbowych mamy

$$x_1 = -0{,}65894789 \cdot \frac{1}{k} \tag{24}$$

oraz

$$x_2 = 0{,}65894789 \cdot \frac{1}{k}. \tag{25}$$

Dla tych wartości dostajemy odpowiadające im wartości rzędnych

$$W_w''(x_1) = \frac{4k^2(2+\sqrt{3})(2+\sqrt{3}-1)}{(1+2+\sqrt{3})^3} \tag{26}$$

oraz

$$W_w''(x_2) = \frac{4k^2(2-\sqrt{3})(2-\sqrt{3}-1)}{(1+2-\sqrt{3})^3} \tag{27}$$

i po wyliczeniu wartości liczbowych

$$W_w''(x_1) = 0{,}3849002 \cdot k^2 \tag{28}$$

$$W_w''(x_2) = -0{,}3849002 \cdot k^2. \tag{29}$$

Dla wyznaczenia współrzędnych punktów przegięcia krzywej opisanej funkcją (18) oraz wyznaczenia nachylenia krzywej w tym punkcie, obliczamy drugą pochodną tej funkcji. Jest ona równa

$$W_w^{IV}(x) = \frac{16k^4\exp(-2kx)}{[1+\exp(-2kx)]^5} \cdot [\exp(-6kx) - 11\exp(-4kx) + 11\exp(-2kx) - 1]. \tag{30}$$

Rozwiązanie równania

$$W_w^{IV}(x) = 0 \tag{31}$$

otrzymujemy w postaci

$$\exp(-2kx) = 1 \tag{32}$$

oraz

$$\exp(-2kx) = 5 \pm \sqrt{24} \tag{33}$$

i stąd

$$x_3 = 0 \tag{34}$$

$$x_4 = -\frac{\ln(5+\sqrt{24})}{2k} \tag{35}$$

$$x_5 = -\frac{\ln(5-\sqrt{24})}{2k}. \tag{36}$$

Po wyznaczeniu wartości liczbowych mamy

$$x_4 = -\frac{1{,}1462158}{k} \tag{37}$$

$$x_5 = \frac{1{,}1462158}{k} \tag{38}$$

Dla odciętych x_4 oraz x_5 otrzymujemy rzędne punktów przegięcia krzywej równe

$$W_w''(x_4) = 0{,}2721655 \cdot k^2 \tag{39}$$

$$W_w''(x_5) = -0{,}2721655 \cdot k^2. \tag{40}$$

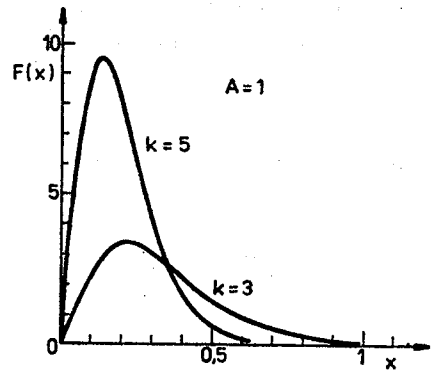
Są one równe

$$\frac{0{,}2721655}{0{,}3849002} = \frac{1}{\sqrt{2}}$$

wartości ekstremów. Nachylenie krzywej opisanej funkcją (18) jest w punktach przegięcia równe

$$W_w'''(x_4) = 8k^3(5+\sqrt{24})\frac{(5+\sqrt{24})^2 - 4(5+\sqrt{24}) + 1}{(1+5+\sqrt{24})^4} = \frac{k^3}{3}, \tag{41}$$

$$W_w'''(x_5) = 8k^3(5-\sqrt{24})\frac{(5-\sqrt{24})^2 - 4(5-\sqrt{24}) + 1}{(1+5-\sqrt{24})^4} = \frac{k^3}{3}. \tag{42}$$



Rys. 8. Wykres funkcji $F(x) = -A \cdot W_w''(x)$ dla $k = 3$ oraz $k = 5$

Na rysunku 8 pokazany jest wykres funkcji

$$F(x) = -A \cdot W_w''(x) \tag{43}$$

dla $A = 1$, $k = 3$ oraz 5. Funkcję tę, po stosownym dobraniu parametrów A oraz k, można wykorzystać do aproksymacji składowej tunelowej prądu diody. Uwzględniając składową dyfuzyjną prądu diody (równ. 16), otrzymujemy funkcję aproksymującą charakterystykę diody tunelowej w postaci

$$I = I_n \frac{\exp(-2kU)[1-\exp(-2kU)]}{[1+\exp(-2kU)]^3} + I_0[\exp(\beta U)-1] \tag{44}$$

gdzie:

$$I_n = 4Ak^2. \tag{45}$$

Dla napięć nie przekraczających $2U_p$, drugi składnik sumy występującej w wyrażeniu (44) po prawej stronie znaku równości jest tak mały w porównaniu z pierwszym składnikiem, że jego wpływ na wartość prądu I można pominąć, nie popełniając przy tym znaczącego błędu. Można więc przyjąć, że charakterystykę diody tunelowej na odcinku od początku układu współrzędnych do punktu przegięcia krzywej, leżącego na opadającym zboczu, aproksymuje wyrażenie

$$I \approx I_n \frac{\exp(-2kU)[1-\exp(-2kU)]}{[1+\exp(-2kU)]^3} \tag{46}$$

wartości parametrów I_n oraz k występujących w (46) wyznaczymy dla dwu przypadków:

1. gdy znana jest wartość prądu I_p oraz napięcia U_p,
2. gdy znana jest wartość prądu I_p oraz największa wartość ujemnej rezystancji R_N przypadająca dla punktu przegięcia krzywej. W pierwszym przypadku z (25) dostajemy

$$U_p = 0{,}65894789 \cdot \frac{1}{k} \approx 0{,}6589 \cdot \frac{1}{k} \tag{47}$$

oraz

$$k = \frac{0{,}6589}{U_p} \tag{48}$$

a z (29) i (43)

$$I_p = A \cdot 0{,}3849 \cdot k^2 \tag{49}$$

oraz

$$Ak^2 = \frac{I_p}{0{,}3849}. \tag{50}$$

Z (45) i (50) otrzymujemy

$$I_n = 4 \cdot Ak^2 = \frac{4I_p}{0{,}3849}. \tag{51}$$

W drugim przypadku wg (46) i (42) mamy dla punktu przegięcia krzywej

$$\frac{\mathrm{d}I}{\mathrm{d}U} = \frac{1}{R_N} = \frac{Ak^3}{3} \tag{52}$$

Z (52) oraz (49) otrzymujemy

$$\frac{1}{R_N I_p} = \frac{Ak^3}{3 \cdot 0{,}3849 \cdot Ak^2} = \frac{k}{3 \cdot 0{,}3849} \tag{53}$$

oraz

$$k = \frac{1{,}155}{R_N \cdot I_p}. \tag{54}$$

Wartość I_n wyznaczamy z (51).

Wartości prądu I_D oraz współczynnik β, występujące w wyrażeniu opisującym składową dyfuzyjną prądu diody, wyznaczamy w następujący sposób. Dla napięcia U_{pp}, jak to wynika z teorii [6, 7], składowa tunelowa prądu diody ma wartość równą zeru. Mamy więc

$$I_D(U_{pp}) = I_p = I_0[\exp(\beta U_{pp}) - 1] \tag{55}$$

gdzie I_D oznacza składową dyfuzyjną. Dla napięcia U_v zachodzi zależność

$$I(U_v) = I_v = I_t(U_v) + I_D(U_v). \tag{56}$$

Mamy więc

$$I_D(U_v) = I_v - I_t(U_v). \tag{57}$$

Występujący w równaniu składowej dyfuzyjnej

$$I_D = I_0[\exp(\beta U) - 1] \tag{58}$$

współczynnik β jest równy

$$\beta = \frac{1}{\gamma U_T}, \tag{59}$$

gdzie: $U_T = \frac{kT}{e} \approx 26$ mV dla $T = 300$ K, a $\gamma \approx 2$ dla małych wartości prądu.

Dla napięcia $U = U_v \gg U_T$, a tym bardziej dla $U = U_{pp}$, wyrażenie $\exp(\beta U)$ jest o kilka rzędów wielkości większe od 1, więc można przyjąć, że

$$I_D(U_v) = I_0 \exp(\beta U_v) \tag{60}$$

$$I_D(U_{pp}) = I_0 \exp(\beta U_{pp}) = I_p. \tag{61}$$

Stąd otrzymujemy

$$\frac{I_p}{I_D(U_v)} = \frac{\exp(\beta U_{pp})}{\exp(\beta U_v)} \tag{62}$$

oraz po obustronnym zlogarytmowaniu

$$\ln\left[\frac{I_p}{I_D(U_v)}\right] = \beta(U_{pp} - U_v) \tag{63}$$

i wreszcie

$$\beta = \frac{\ln\left[\frac{I_p}{I_D(U_v)}\right]}{U_{pp} - U_v}. \tag{64}$$

Z (61) dostajemy

$$I_0 = \frac{I_p}{\exp(\beta U_{pp})} \tag{65}$$

Opisana metoda aproksymacji zostanie zilustrowana poniższymi przykładami.

4. PRZYKŁADY

Przykład 1

Dla germanowej diody tunelowej typu AEY 30 A (Siemens) o następujących danych katalogowych: $U_p = 75$ mV, $U_v = 350$ mV, $I_p = 1{,}6$ mA (1,4÷1,8 mA), $I_p/I_v = 9 > 6$, $R_s = 5{,}5\ \Omega$, $R_N = 60 \div 110\ \Omega$, należy ustalić przybliżone równanie charakterystyki statycznej i wykonać wykres funkcji aproksymującej.

Wśród danych katalogowych brak jest wartości napięcia U_{pp}. Według [2], charakterystyczne wartości parametrów germanowych diod tunelowych najczęściej zawarte są w przedziałach $U_p = 40 \div 70$ mV, $U_v = 250 \div 300$ mV, $U_{pp} = 400 \div 600$ mV. Ponieważ dla rozpatrywanej diody wartości U_p oraz U_v nieznacznie przekraczają górne wartości podanych przedziałów, przyjmujemy $U_{pp} = 600$ mV.

Parametry diody wyznaczamy dla dwu przypadków danych wyjściowych.

Przypadek 1: Dane $U_p, I_p, U_{pp}, I_p/I_v$.

Według (48) mamy

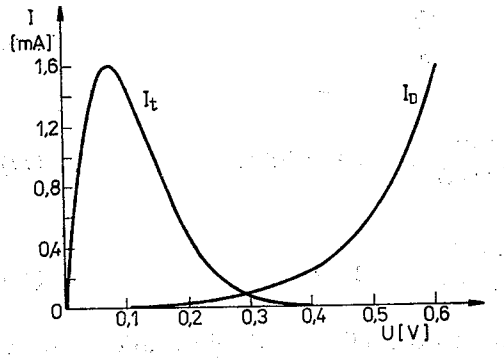
$$k = \frac{0{,}6589}{U_p} = \frac{0{,}6589}{0{,}075} = 8{,}79, \tag{66}$$

a z (51)

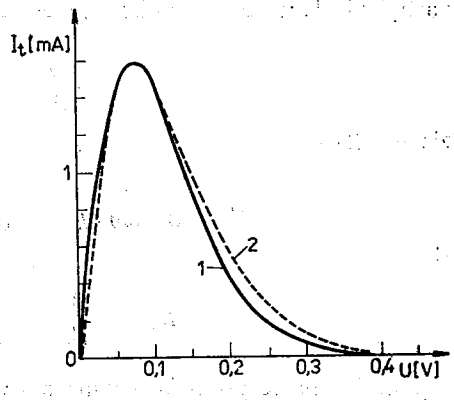
$$I_n = \frac{4I_p}{0{,}3849} = \frac{4 \cdot 1{,}6 \cdot 10^{-3}}{0{,}3849} = 16{,}63 \cdot 10^{-3}\ \text{A}. \tag{67}$$

Równanie składowej tunelowej prądu diody ma postać

$$I_t = 16{,}63 \cdot 10^{-3} \frac{\exp(-17{,}58U)[1 - \exp(-17{,}58U)]}{[1 + \exp(-17{,}58U)]^3}\ \text{A}. \tag{68}$$



Rys. 9. Wykresy składowych prądów diody tunelowej



Rys. 10. Wykresy funkcji aproksymujących składową tunelową prądu diody: krzywa 1 wg (44), krzywa 2 wg (16)

Wykres składowej tunelowej $I_t = I_t(U)$ jest przedstawiony na rys. 9 (krzywa I_t). Ten sam wykres jest powtórzony na rysunku 10 (krzywa 1), na który, dla porównania, naniesiono również wykres funkcji aproksymującej składową tunelową prądu diody według zależności (16) — (krzywa 2). Parametry składowej dyfuzyjnej obliczamy z (57), (64) i (65). Mamy

$$I_D(U_v) = I_v - I_t(U_v) = \frac{I_p}{9} - I_t(U_v) = \frac{1{,}6}{9} - 0{,}035 = 0{,}143 \text{ mA}, \tag{69}$$

$$\beta = \frac{\ln\left[\dfrac{I_p}{I_D(U_v)}\right]}{U_{pp} - U_v} = \frac{\ln \dfrac{1{,}6}{0{,}143}}{0{,}60 - 0{,}35} = 9{,}66 \tag{70}$$

oraz

$$I_0 = \frac{I_p}{\exp(\beta U_{pp})} = \frac{1{,}6 \cdot 10^{-3}}{\exp(9{,}66 \cdot 0{,}6)} = 4{,}864 \cdot 10^{-6}\text{A}. \tag{71}$$

Równanie składowej dyfuzyjnej I_D prądu diody ma postać

$$I_D = 4{,}864 \cdot 10^{-6} \exp(9{,}66U). \tag{72}$$

Wykres składowej dyfuzyjnej $I_D = I_D(U)$ jest przedstawiony na rys. 9 (krzywa I_D). Suma składowej dyfuzyjnej i składowej tunelowej daje szukane równanie charakterystyki diody tunelowej

$$I = 16{,}63 \cdot 10^{-3} \frac{\exp(-17{,}58U)[1 - \exp(-17{,}58U)]}{[1 + \exp(-17{,}58 \cdot U)]^3} + 8{,}464 \cdot 10^{-6} \exp(9{,}66 \cdot U). \tag{73}$$

Punkt przegięcia krzywej ma wg (38) i (40) współrzędne

$$U_s = \frac{1{,}1462}{k} = \frac{1{,}1462}{8{,}79} = 0{,}130 \text{ V}, \tag{74}$$

$$I_s = \frac{I_p}{\sqrt{2}} = \frac{1{,}6}{\sqrt{2}} = 1{,}131 \text{ mA}, \tag{75}$$

a nachylenie krzywej w punkcie przegięcia ma wg (53) wartość

$$\frac{dI}{dU} = \frac{1}{R_N} = \frac{Ak^3}{3} \tag{76}$$

gdzie wg (50)

$$A = \frac{I_p}{0{,}3849 \cdot k^2} = \frac{1{,}6 \cdot 10^{-3}}{0{,}3849 \cdot 8{,}79^2} = 53{,}8 \cdot 10^{-6}, \tag{77}$$

stąd

$$R_N = \frac{3}{Ak^3} = \frac{3 \cdot 10^6}{53{,}8 \cdot 8{,}79^3} = 82{,}1\Omega, \tag{78}$$

a więc obliczona rezystancja ujemna ma wartość bardzo bliską średniej wartości z granic przedziału $R_N = 60 \div 110\Omega$.

Przypadek 2: Dane I_p, $R_N = 85\Omega$, U_{pp}, I_p/I_v.

Według (54) mamy

$$k = \frac{1{,}155}{R_N \cdot I_p} = \frac{1{,}155}{85 \cdot 1{,}6 \cdot 10^{-3}} = 8{,}49, \tag{79}$$

a z (51)

$$I_n = \frac{4I_p}{0{,}3849} = \frac{4 \cdot 1{,}6 \cdot 10^{-3}}{0{,}3849} = 16{,}63 \cdot 10^{-3}\ \text{A}. \tag{80}$$

Równanie składowej tunelowej prądu diody ma więc postać

$$I_t = 16{,}63 \cdot 10^{-3} \frac{\exp(-16{,}98 \cdot U)[1 - \exp(-16{,}98 \cdot U)]}{[1 + \exp(-16{,}98 \cdot U)]^3}. \tag{81}$$

Podstawiając do (81) $U = U_v$ obliczamy wartość $I_t(U_v)$

$$I_t(U_v) = 0{,}043 \cdot 10^{-3}\ \text{A}, \tag{82}$$

a następnie obliczamy $I_D(U_v)$

$$I_D(U_v) = I_v - I_t(U_v) = \frac{I_p}{9} - I_t(U_v) = \frac{1{,}6}{9} - 0{,}043 = 0{,}136\ \text{mA} \tag{83}$$

oraz z (64)

$$\beta = \frac{\ln\left[\dfrac{I_p}{I_D(U_v)}\right]}{U_{pp} - U_v} = \frac{\ln \dfrac{1{,}6}{0{,}136}}{0{,}6 - 0{,}35} = 9{,}86 \tag{84}$$

i z (65)

$$I_0 = \frac{I_p}{\exp(\beta U_{pp})} = \frac{1{,}6 \cdot 10^{-3}}{\exp(9{,}86 \cdot 0{,}6)} = 4{,}314 \cdot 10^{-6}\text{A}. \tag{85}$$

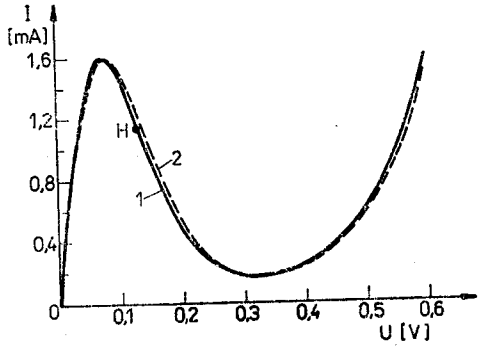
Równanie składowej dyfuzyjnej ma postać

$$I_D = 4{,}314 \cdot 10^{-6} \cdot \exp(9{,}86 \cdot U), \tag{86}$$

a równanie charakterystyki diody tunelowej

$$I = 16{,}63 \cdot 10^{-3} \frac{\exp(-16{,}98 \cdot U)[1-\exp(-16{,}98 \cdot U)]}{[1+\exp(-16{,}98 \cdot U)]^3} + 4{,}314 \cdot 10^{-6}\exp(9{,}86 \cdot U). \tag{87}$$

Wykres funkcji aproksymującej (87) jest przedstawiony na rysunku 11, (krzywa 2). Jak to



Rys. 11. Wykresy funkcji aproksymujących charakterystykę statyczną diody AEY 30 A (krzywa 1 — dla przypadku 1, krzywa 2 — dla przypadku 2)

wynika z rysunku 11, wykresy funkcji aproksymujących (73) oraz (87) różnią się tylko bardzo nieznacznie. Szczytowy punkt krzywej, opisanej równaniem (87), przypada dla napięcia

$$U_p = \frac{0{,}6589}{k} = \frac{0{,}6589}{8{,}49} = 77{,}6 \cdot 10^{-3}\ \text{V} \tag{88}$$

a więc jest przesunięty w stronę większych napięć w stosunku do wartości podanej w katalogu.

Dla obu funkcji aproksymujących napięcie, dla którego wartość prądu diody osiąga minimum, jest nieco mniejsze od katalogowego napięcia $U_p = 350$ mV.

Przykład 2

Na rysunku 12 pokazany jest wykres charakterystyki diody tunelowej typu ND 4173, zdjętej przy użyciu oscyloskopu z pamięcią firmy Tektronix. Należy wyznaczyć równanie krzywej przybliżającej tę charakterystykę.

Z wykresu odczytujemy następujące wartości charakterystyczne: $I_p = 2{,}4$ mA, $U_p = 75 \pm 3$ mV, $I_v = 0{,}25$ mA, $U_v = 350 \pm 20$ mV, $U_{pp} = 580$ mV.

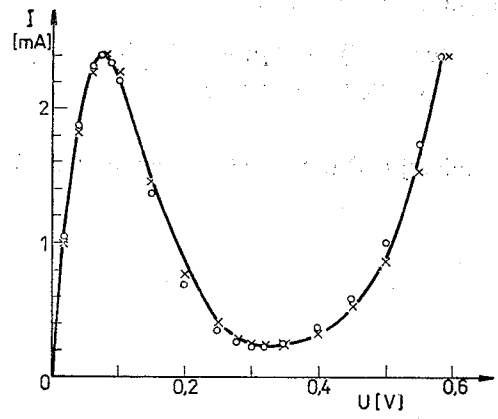
Przyjmując $U_p = 75$ mV, $U_v = 350$ mV mamy:

według (48)

$$k = \frac{0{,}6589}{U_p} = \frac{0{,}6589}{0{,}075} = 8{,}79 \tag{89}$$

według (51)

$$I_n = \frac{4 \cdot I_p}{0{,}3849} = \frac{4 \cdot 2{,}4 \cdot 10^{-3}}{0{,}3849} = 24{,}94 \cdot 10^{-3}\ \text{A} \tag{90}$$



Rys. 12. Wykres charakterystyki diody ND 4173 zdjętej z zastosowaniem oscyloskopu

oraz równanie składowej tunelowej prądu diody

$$I_t = 24{,}94 \cdot 10^{-3} \frac{\exp(-17{,}58 \cdot U)[1-\exp(-17{,}58 \cdot U)]}{[1+\exp(-17{,}58 \cdot U)]^3} \text{ A}. \tag{91}$$

Obliczona wg (91) wartość składowej tunelowej dla $U = U_v$ wynosi

$$I_t(U_v) = 0{,}0528 \cdot 10^{-3} \text{ A} \tag{92}$$

a wartość składowej dyfuzyjnej dla $U = U_v$

$$I_D(U_v) = I_v - I_t(U_v) = (0{,}2500 - 0{,}0528) \cdot 10^{-3} = 0{,}1972 \cdot 10^{-3} \text{ A}. \tag{93}$$

Według (64) i (65) otrzymujemy

$$\beta = \frac{\ln\left[\dfrac{I_p}{I_D(U_v)}\right]}{U_{pp} - U_v} = \frac{\ln\left[\dfrac{2{,}4}{0{,}1972}\right]}{0{,}58 - 0{,}35} = 10{,}865 \tag{94}$$

oraz

$$I_0 = \frac{I_p}{\exp(\beta U_{pp})} = \frac{2{,}4 \cdot 10^{-3}}{\exp(10{,}865 \cdot 0{,}58)} = 4{,}40 \cdot 10^{-6} A \tag{95}$$

Równanie przybliżonej charakterystyki diody ma postać

$$I = 24{,}94 \cdot 10^{-3} \frac{\exp(-17{,}58 \cdot U) \cdot [1-\exp(-17{,}58 \cdot U)]}{[1+\exp(-17{,}58 \cdot U)]^3} + 4{,}40 \cdot 10^{-6} \exp(10{,}865 \cdot U) \text{ A}. \tag{96}$$

Wyznaczone wg (96) wartości prądu zaznaczono na wykresie kółkami.

Ponieważ odczytane z wykresu wartości napięć są obarczone błędem, powtórzono obliczenia dla zmienionej wartości napięcia równej $U_p = 78$ mV. Dla uzyskania lepszego przybliżenia łuku krzywej przy $U > U_v$, do obliczeń przyjęto $U_{pp} = 0{,}59$ V. Przy tych danych uzyskano

$$k = 8{,}447 \tag{97}$$

$$I_n = 24{,}94 \cdot 10^{-3} \text{ A} \tag{98}$$

$$I_D(U_v) = 0{,}1832 \cdot 10^{-3} \text{ A} \tag{99}$$

$$\beta = 11{,}185 \tag{100}$$

$$I_0 = 3{,}266 \cdot 10^{-6}\text{A} \tag{101}$$

oraz równanie przybliżonej charakterystyki diody

$$I = 24{,}94 \cdot 10^{-3} \frac{\exp(-16{,}89 \cdot U) \cdot [1-\exp(-16{,}89 \cdot U)]}{[1+\exp(-16{,}89 \cdot U)]^3} + 3{,}266 \cdot 10^{-6}\exp(11{,}185 \cdot U) \text{ A}. \tag{102}$$

Wyznaczone wg (102) wartości prądu zaznaczono na wykresie krzyżykami. Z porównania wartości prądu uzskanych wg (96) i (102) wynika, że równanie (102) lepiej przybliża eksperymentalnie zdjętą charakterystykę diody.

UWAGI KOŃCOWE

W przedstawionym sposobie aproksymacji statycznej charakterystyki diody tunelowej funkcja aproksymująca stanowi sumę dwu funkcji wykładniczych, z których pierwsza przybliża składową tunelową, a druga w znany sposób składową dyfuzyjną tego prądu. Funkcja ta w zasadzie służy do przybliżania rzeczywistej charakterystyki diody tunelowej w zakresie napięć od $U = 0$ do $U = U_{pp}$. Może być również wykorzystana do przybliżania charakterystyki diody w zakresie napięć od ok. $-0.5U_p$ do napięć wyraźnie przekraczających U_{pp}.

Do wyznaczenia parametrów funkcji aproksymującej wystarcza znajomość podawanych w katalogach danych jak U_p, I_p, U_p/U_v oraz U_{pp} lub I_p, R_N, U_{pp} i U_p/U_v.

Prezentowany sposób aproksymacji nie uwzględnia występującego w obszarze doliny charakterystyki zwiększenia prądu (excess current [8]) spowodowanego występowaniem dodatkowych poziomów energetycznych w pasmie zabronionym. Pominięcie wpływu tego prądu powoduje, że minimum funkcji aproksymującej występuje przy napięciu nieco niższym od U_v. Jednakże nie ma to większego znaczenia przy analizowaniu pracy układów zawierających diody tunelowe.

Zaletą przedstawionej metody aproksymacji jest, że charakterystykę diody tunelowej w całym zakresie napięć od 0 do U_{pp} aproksymuje się jedną funkcją, dzięki czemu unika się niedogodności związanych z aproksymacją odcinkową.

BIBLIOGRAFIA

1. W. F. Chow: *Principles of Tunnel Diode Circuits*. New York: J. Wiley, 1964 (tłum. rosyjskie 1966 r.)
2. E. A. Akczurin, W. W. Rud, W. J. Spirin: *Tunnelnyje diody w tiechnikie swjazi*. Moskwa: Swjaź, 1971
3. A. S. Sidorow: *Tieorija i projektirowanije TD-schiem*. Moskwa: Sowietskoje Radio, 1971
4. S. A. Gariainow, I. D. Abiezgauz: *Połuprowodnikowyje pribory s otricatielnym soprotiwlenijem*. Moskwa: Energija, 1966
5. L. J. Weiss, T. Wysocki: *Przybliżenia wykładnicze rzeczywistych przebiegów i rozkładów o charakterze skokowym lub impulsowym*. Rozprawy Elektrotechniczne, 1986, t. 32, z. 2

6. A. Moschwitzer, K. Lunze: *Halbleiterelektronik*. Berlin: Verlag Technik, 1977
7. C. M. Sze: *Physics of Semiconductor Devices*. New York: J. Wiley, 1969 (tłum. rosyjskie 1973 r.)
8. E. Philippow (wyd): *Taschenbuch Elektrotechnik*. Berlin, Verlag Technik, 1969, Band 3

L. J. WEISS, F. WYSOCKA, T. WYSOCKI, T. WYSOCKI JR

APPROXIMATION OF STATIONARY CHARACTERISTICS OF A TUNNEL DIODE BY AN EXPONENTIAL FUNCTION

Summary

An exponential function liable to approximate the stationary characteristics of a tunnel diode is presented. The procedure of determining the parameters of this function from the known parameters of the diode is displayed. Two numerical examples are quoted.

L. J. WEISS, F. WYSOCKA, T. WYSOCKI, T. WYSOCKI, JR.

APPROXIMATION DE LA COURBE CARACTÉRISTIQUE STATIQUE DE DIODE TUNNEL À L'AIDE DE LA FONCTION EXPONENTIELLE

Résumé

Dans l'article on a présenté le mode de résolution du problème d'approximation de la courbe caractéristique statique de diode tunnel à l'aide de la fonction exponentielle et aussi le mode dedétermination des paramètres de cette fonction, à la base de connaissance des données de la diode. L'article contient aussi deux examples de l'application numérique.

L. J. WEISS, F. WYSOCKA, T. WYSOCKI, T. WYSOCKI JR.

ANWENDUNG DER EXPONENTIALFUNKTION ZUR APPROXIMATION DER STATISCHEN KENNLINIE EINER TUNNELDIODE

Zusammenfassung

In diesem Aufsatz wurde die Approximation der statischen Kennlinie einer Tunneldiode durch eine Exponentialfunktion, sowie ein Verfahren zur Bestimmung der Parameter dieser Funktion bei gegebenen Diodendaten dargestellt. Die Anwendung der vorgeschlagenen Methode wurde anhand von zwei Rechenbeispielen erläutert.

Л. Я. ВАЙС, Ф. ВЫСОЦКА, Т. ВЫСОЦКИ, Т. ВЫСОЦКИ МЛ.

АППРОКСИМАЦИЯ ВОЛЬТАМПЕРНОЙ ХАРАКТЕРИСТИКИ ТУННЕЛЬНОГО ДИОДА С ПОМОЩЬЮ ЭКСПОНЕНЦИАЛЬНОЙ ФУНКЦИИ

Резюме

Рассматривается приближение вольтамперной характеристики туннельного диода с помощью экспоненциальной функции. Описан способ определения параметров этой функции с использованием данных диода. Статья содержит два численные примеры.

ROZPRAWY ELEKTROTECHNICZNE, 1989, 35, z. 1, ss. 175–191

0.25.4.036:681.3.

Dobór optymalnej wielkości pakietów w sieci z integracją usług

MARIUSZ PIONTAS, DOMINIK RUTKOWSKI

Instytut Telekomunikacji Politechniki Gdańskiej

Otrzymano 1987.09.03

Autoryzowano do druku 1988.03.25

Przedstawiono analizę sieci z komutacją pakietów, w której przesyłane są sygnały mowy ludzkiej oraz krótkie i długie informacje komputerowe. Dla przedstawionego modelu węzła i sieci obliczono długości pakietu optymalizującego wykorzystanie przepustowości kanału.

1. PRZESYŁANIE SYGNAŁÓW MOWY W KANALE CYFROWYM

W drugiej połowie XX wieku obserwujemy niezwykle dynamiczny rozwój różnego rodzaju sieci przesyłania informacji. Wprawdzie od dawna znana jest sieć telefoniczna czy też sieć dalekopisowa, jednak od kilkunastu lat uruchamiane są w wielu krajach sieci nowego rodzaju zwane sieciami teleinformatycznymi, które umożliwiają przesyłanie informacji między końcówkami komputerowymi i komputerami ogólnego przeznaczenia. Sieci te, jak się okazuje, mogą być również wykorzystane do przesyłania innych rodzajów informacji w tym w szczególności sygnałów mowy. Wymaga to nie tylko przetworzenia tych sygnałów na postać cyfrową, lecz również formowania z nich odpowiednich bloków informacji cyfrowych, jednakże korzyści płynące z zastosowania sieci teleinformatycznych do przesyłania sygnałów mowy są znaczne. Po pierwsze wzrasta ogromnie jakość transmisji, a po drugie — znacznie maleją jej koszty.

Jest oczywiste, że majątek trwały sieci telefonicznej jest w wielu krajach ogromny, a jej zasoby, jak się okazuje, są w dużej mierze nieefektywnie wykorzystane, wobec czego przekształcenie jej w sieć telekomunikacyjną nowego rodzaju, w której stosowane będą zasady pracy sieci teleinformatycznej i która dzięki temu będzie mogła służyć również przesyłaniu innych rodzajów strumieni informacji, poza strumieniem telefonicznym, jest celowe i korzystne.

Przyczyną niskiej efektywności wykorzystania zasobów sieci telefonicznej jest niski stopień wykorzystania jej kanałów. Chodzi tu bowiem o fakt wyłączności użytkowania kanału łączącego parę komunikujących się abonentów oraz istnienie naturalnych przerw w strumieniu sygnałów mowy i naprzemienny w czasie kierunek ich przesyłania, a więc wykorzystanie na ogół w każdym momencie rozmowy tylko jednego z dwóch kanałów simpleksowych łączących abonentów. Wszystkie te uwarunkowania sprzyjają wzrostowi

zainteresowania sieciami teleinformatycznymi, zwłaszcza że integracja różnych usług telekomunikacyjnych w takich sieciach jest bardzo korzystna pod względem ekonomicznym i stosunkowo łatwa w realizacji.

Z tego względu w pracy omówione zostały niektóre aspekty wspólnego użytkowania sieci telekomunikacyjnej przez ucyfrowione sygnały mowy ludzkiej i informacje komputerowe, a w szczególności przedstawiony został model sieci, umożliwiający dobór długości pakietów przenoszących sygnały mowy.

1.1. SPECYFICZNE CECHY TRANSMISJI UCYFROWIONYCH SYGNAŁÓW MOWY

Przetworzenie sygnałów mowy na postać cyfrową nie nastręcza obecnie poważniejszych problemów technicznych. Istotna jest jednak znajomość niezbędnej szybkości transmisji v_m cyfrowych sygnałów mowy, która rzutuje na wymaganą przepustowość kanału. Szybkość ta może się zmieniać w szerokich granicach, w zależności od stosowanej metody ucyfrowienia i wymaganej jakości odtwarzanych po stronie odbiorczej sygnałów mowy.

Jak wiadomo, przy pewnych ogólnych założeniach maksymalna ilość informacji, która może być przesyłana przez kanał, dana jest wzorem Shannona

$$I = W\log_2\left(1+\frac{S}{N}\right), \tag{1}$$

gdzie:

I — przepustowość kanału (b/s),

W — szerokość pasma kanału (Hz),

S/N — stosunek mocy średnich sygnału do szumu.

Wiadomo też, że szerokość pasma sygnałów mowy rzadko przekracza 8 kHz. Ponieważ maksymalna wartość stosunku mocy średniej sygnału do szumu wynosi około 50 dB, więc $I = 106$ kb/s [1].

W celu określenia wymaganej prędkości przesyłania sygnałów mowy przy zachowaniu ich zrozumiałości, prowadzone są od dawna badania, których celem jest stworzenie prostych i jednocześnie możliwie wiernych modeli układów syntetyzujących sygnały mowy. Rezultaty tych badań wskazują, że prędkość niezbędna do przesyłania pewnych parametrów charakteryzujących właściwości sygnałów mowy może być wielokrotnie mniejsza od 106 kb/s. Wyniki te stanowią podstawę dostosowywania szybkości przekazywania sygnałów mowy do potrzeb użytkowników, a więc umożliwiają dokonywanie wyboru między jakością jej odbioru a kosztami stosowanych kanałów i końcówek. Z tego względu urządzenia przetwarzające sygnały mowy na postać cyfrową można z grubsza podzielić na dwie grupy: [2]

— urządzenia, w których bezpośrednio lub pośrednio wykorzystywane jest twierdzenie o próbkowaniu. Najczęściej informacja o wartości poszczególnych próbek (lub o różnicy między wartościami kolejnych próbek) zostaje przedstawiona w postaci ciągu informacji binarnych,

— urządzenia[1)], które wyznaczają na bieżąco wartości pewnych parametrów opisu-

[1)] urządzenia te zwane są wokoderami

jących właściwości określonego modelu sygnałów mowy i przekazują je w formie cyfrowej do miejsca ich odbioru.

W urządzeniach drugiej grupy dokonuje się analizy wycinków sygnałów mowy w kolejnych ramkach czasowych, których długość T_r jest w praktyce rzędu kilkudziesięciu milisekund. Jest to sposodowane faktem, że sygnał mowy jest procesem stochastycznym niestacjonarnym i w dłuższym przedziale czasu zachodzą znaczące zmiany wartości opisujących go parametrów, a więc analiza dłuższych wycinków wpłynęłaby na pogorszenie zrozumiałości dźwięków mowy odtwarzanych po stronie odbiorczej, natomiast analiza dokonywana w krótszych przedziałach czasu prowadziłaby do znacznego wzrostu niezbędnej szybkości przekazywania informacji o tych parametrach.

1.2. WSPÓLNE UŻYTKOWANIE SIECI PRZEZ UCYFROWIONE SYGNAŁY MOWY I INFORMACJE KOMPUTEROWE

W sieci telekomunikacyjnej wspólnie użytkowanej przez abonentów końcówek dźwiękowych i informacje komputerowe spotykamy kilka rodzajów informacji, które generowane są przez trzy rodzaje źródeł, a mianowicie: końcówki dźwiękowe, komputery i końcówki teleinformatyczne. Możemy też wyróżnić trzy rodzaje informacji: [3],

(1) krótkie informacje przesyłane w konwersacyjnym trybie współpracy użytkownika końcówki teleinformatycznej z komputerem,

(2) długie informacje (programy, informacje do/z baz danych itp.),

(3) informacje przenoszące sygnały mowy.

Informacje typu (1) są stosunkowo krótkie i jak wykazują badania 90% z nich ma długość nie przekraczającą 20 kb. Aby uzyskać konwersacyjny tryb współpracy z komputerem, należy zapewnić dostatecznie małe opóźnienie przesyłania takich informacji; nie powinno ono przekraczać 200 ms. Warto przy tym zaznaczyć, że podczas trwania seansu łączności średnia aktywność końcówki przekazującej informacje do komputera wynosi zaledwie $1 \div 5\%$ czasu trwania seansu, natomiast średnia aktywność komputera przesyłającego informacje do określonej końcówki wynosi $13 \div 35\%$ czasu trwania seansu łączności [4].

Informacje typu (2) są długie (rzędu wielu kilobitów). Dopuszczalne opóźnienia są w tym przypadku większe i zależą od przeznaczenia informacji, przy czym na ogół przekraczają 600 ms. Cechą szczególną końcówek generujących takie informacje jest ich wysoka średnia aktywność.

Warto dodać, że pomiary wykazały, iż długości informacji obu typów można aproksymować za pomocą rozkładu wykładniczego [4].

Jeśli weźmiemy pod uwagę rozmowę telefoniczną, to wiadomo, że poszczególne sygnały mowy przedzielone są przerwami. Pomiary wykazały, że czasy trwania sygnałów mowy oraz przerw między nimi są od siebie niezależne i mają rozkład zbliżony do wykładniczego o wartości średniej (dla języka angielskiego) $1{,}2 \div 1{,}5$ s [3, 4, 9].

Dopuszczalne opóźnienie tego rodzaju sygnałów jest zbliżone do opóźnienia dopuszczalnego dla informacji (1).

1.3. RODZAJE KOMUTACJI SPOTYKANE W SIECIACH PRZESYŁANIA INFORMACJI CYFROWYCH

Jednym z podstawowych parametrów, decydujących o przydatności określonej sieci do przesyłania informacji cyfrowych, jest wnoszone przez nią średnie opóźnienie. Zależy ono w decydującej mierze od stosowanego w niej rodzaju komutacji. Najczęściej spotykaną dotychczas metodą komutacji jest komutacja kanałów. Polega ona na tym, że na czas przesyłania informacji zostaje zestawiony przez odpowiednie węzły sieci (urządzenia komutacyjne) łańcuch kanałów między komunikującymi się końcówkami. Zaletami tej metody jest stosunkowo krótki czas przesyłania informacji (w systemach cyfrowych czas ten zależy głównie od prędkości transmisji w kanale) oraz łatwości implementacji. Do wad należy stosunkowo długi czas zestawiania trasy oraz niewielki stopień wykorzystania przepustowości kanałów. Pierwsza z tych wad może być prawie całkowicie wyeliminowana przez zastosowanie tzw. szybkiej komutacji kanałów. Dzięki wykorzystaniu wydzielonego kanału służącego wyłącznie celom komutacji udaje się skrócić czas zestawiania trasy[2)] z 5 s do 140 ms [3]. Stanowi to dziesiątą część średniego czasu trwania pojedynczego sygnału mowy, a więc z punktu widzenia efektywności wykorzystania przepustowości kanału opłacalne staje się zestawianie trasy po pojawieniu się kolejnego sygnału mowy i rozłączanie po zamilknięciu rozmówcy. Podobna metoda, o nazwie TASI (Time Assigned Speech Interpolation) stosowana jest w kablach podmorskich. Jest ona przystosowana do sygnałów analogowych i polega na tym, że każdemu z n rozmówców przydziela się jeden z k kanałów ($k < n$) wtedy, gdy detektor wykryje obecność sygnałów mowy. Ponieważ sygnały mowy trwają przeciętnie 40% czasu trwania rozmowy, więc stosunek n/k charakteryzujący możliwą do przeprowadzania koncentrację może sięgać wartości $1/0{,}4 = 2{,}5$. Wada tego rozwiązania polega na tym, że odcinane są początkowe fragmenty sygnałów mowy (jednak bez istotnej utraty zrozumiałości), a ponadto omawiana metoda staje się skuteczna dopiero przy dużej liczbie kanałów, rzędu 80 [3, 4, 10].

Przez długie lata komutacja kanałów była najbardziej powszechną metodą realizacji połączeń w sieciach telekomunikacyjnych. W ostatnich piętnastu latach obserwujemy jednak rozwój konkurencyjnej w stosunku do komutacji kanałów formy komutacji zwanej komutacją informacji. Polega ona na tym, że poszczególne informacje przesyłane są z końcówek źródłowych poprzez odpowiednie węzły sieci do końcówek docelowych zgdnie z określoną regułą doboru trasy, przy czym w pamięci węzłów sieci informacje te oczekują przez pewien czas na obsługę, a więc w metodzie tej nie wchodzi w grę zestawianie trasy przed rozpoczęciem transmisji określonej informacji, gdyż trasa dla każdej informacji jest ustalana na bieżąco przez węzeł, do którego informacja już dotarła. Ponadto tymi samymi kanałami międzywęzłowymi mogą być przesyłane kolejno po sobie informacje pochodzące z różnych źródeł; ma więc miejsce wspólne użytkowanie kanałów. Opóźnienia przesyłania każdej informacji wiążą się teraz głównie z czasem jej oczekiwania w pamięci węzłów, które znalazły się na jej trasie i czasem nadawania, jednakże wiadomo, że może być ono wielokrotnie mniejsze niż w przypadku komutacji kanałów. Często spotykaną odmianą komutacji informacji jest komutacja pakietów, która polega na tym, że informacja przed wprowadzeniem do sieci zostaje podzielona na bloki, które po uzupełnieniu o odpowiednie

[2)] chodzi tu o średni czas

informacje sterująco-kontrolne[3)] tworzą pakiety, te zaś przesyłane są w sieci niezależnie od siebie. Wady i zalety tej metody są obszernie omawiane w literaturze i tu wspomnimy tylko o niektórych aspektach wiążących się z przesyłaniem pakietyzowanych informacji. Wielkość i sposób formowania pakietów przenoszących sygnały mowy jest bowiem sprawą niezmiernie istotną dla sprawnej transmisji sygnałów mowy.

2. DOBÓR OPTYMALNEJ WIELKOŚCI PAKIETU PRZENOSZĄCEGO SYGNAŁY MOWY

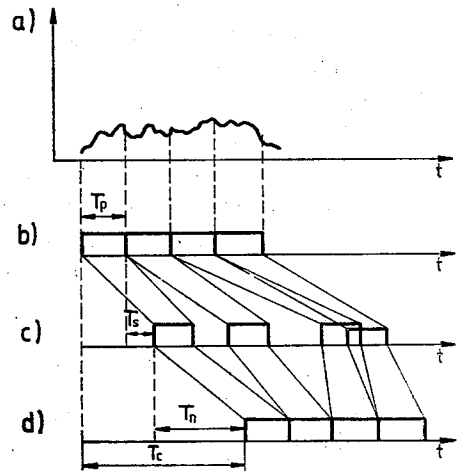
2.1. ZALEŻNOŚCI CZASOWE WYSTĘPUJĄCE PRZY PRZESYŁANIU PAKIETYZOWANYCH SYGNAŁÓW MOWY

Dobór wielkości pakietu przenoszącego sygnały mowy jest kompromisem między dwoma sprzecznymi wymaganiami, według których:

— długość l informacji właściwej w pakiecie powinna być na tyle duża, by przy ustalonej długości nagłówka l_H iloraz $\left(\frac{l}{l+l_H}\right)$, będący wskaźnikiem wykorzystania przepustowości kanałów przez informacje właściwe, był możliwie duży,

— długość pakietu $l+l_H$ powinna być na tyle mała, aby zapewnić krótki czas przesyłania sygnałów mowy i zachować naturalny charakter wypowiedzi.

Na rys. 1 zilustrowano podział wycinka sygnału mowy na segmenty i ich przesyłanie



Rys. 1. Przesyłanie sygnałów mowy za pomocą pakietów; a) wycinek sygnału mowy, b) pakiety przenoszące segmenty sygnału mowy w końcówce źródłowej, c) pakiety po przybyciu do końcówki docelowej, d) pakiety po opuszczeniu bufora nadajnika

Oznaczenia: T_p — długość segmentu mowy, T_s — czas trwania transmisji przez sieć, T_0 — opóźnienie wprowadzane w odbiorniku, T_c — całkowite opóźnienie sygnału mowy

po ucyfrowieniu w formie oddzielnych pakietów do odbiornika. Nie zaznaczono na nim części protokolarnej w poszczególnych pakietach i przyjęto, że czas konwersji sygnałów mowy na postać cyfrową jest pomijalnie krótki. Jak wiadomo, gęstość prawdopodobieństwa długości sygnałów mowy ma rozkład zbliżony do wykładniczego o wartości średniej około 1,5 s, natomiast czas trwania T_p segmentu zależy m.in. od stosowanej

[3)] chodzi tu np. o adres końcówki docelowej, typ i numer pakietu itp.

metody przekształcania sygnałów na postać cyfrową i dla wokoderów jest rzędu 10 ms. Zatem ciąg informacji binarnych uzyskany z ucyfrowienia pojedynczego sygnału mowy wypełnia przeciętnie 150 pakietów. Jak widać z rysunku, pakiet jest przesyłany przez sieć w czasie $\underset{\sim}{T}_s$. Po przybyciu do końcówki docelowej część zawierająca ucyfrowiony wycinek sygnału mowy zostaje dodatkowo opóźniona o T_0, co ma na celu przywrócenie właściwych (naturalnych) relacji czasowych między odtworzonymi dźwiękami mowy, po czym zostaje przekazana na wejście przetwornika C/A. Należy zaznaczyć, że konwersacyjny charakter rozmowy zostanie zachowany, gdy spełniona będzie zależność:

$$\underset{\sim}{T}_c = T_p + \underset{\sim}{T}_s + T_0 \leqslant T_a, \tag{2}$$

gdzie T_a jest dopuszczalnym, akceptowanym przez odbiorcę opóźnieniem rzędu 200 ms. Czas trwania segmentu T_p najwygodniej jest wybrać tak, aby był on równy całkowitej wielokrotności długości ramki

$$T_p = k \cdot T_r, \quad k = 1, 2, \ldots \tag{3}$$

przy czym k powinno być z jednej strony możliwie małe, aby spełniona została nierówność (2), z drugiej zaś od wartości k zależy efektywność wykorzystania pakietu zdefiniowana jako stosunek długości części pakietu zawierającej informację właściwą do całkowitej długości pakietu

$$\xi = \frac{k \cdot T_r \cdot v_m}{k \cdot T_r \cdot v_m + l_{il}}. \tag{4}$$

Dla wokodera z predykcją liniową v_m wynosi 9600 b/s, a $T_r = 20$ ms. Jeżeli przyjmiemy, że długośc nagłówka (części sterująco-kontrolnej) pakietu wynosi $l_H = 152$ bity, to dla pakietu przenoszącego informację z jednej ramki wartość współczynnika ξ wynosi 56%. Dla k równego 2 i 3 wartość ξ wynosi odpowiednio 72% i 80%. Dalsze zwiększanie k powoduje już niewielkie przyrosty ξ. Widać stąd, że dla $k = 1$ korzyść płynąca z wykorzystania statystycznego charakteru sygnałów mowy jest stosunkowo nieduża. Dobór wielkości k należy przeprowadzić tak, aby dla stosowanego typu przetwornika sygnału mowy zmaksymalizować wartość wyrażenia (4) przy spełnieniu nierówności (2).

2.2. ROZKŁADY DŁUGOŚCI ROZWAŻANYCH BLOKÓW INFORMACJI

Zakładać będziemy, że długości każdej z trzech wspomnianych wcześniej rodzajów informacji mają rozkład wykładniczy o różnej wartości średniej, tzn. gęstość prawdopodobieństwa długości v jest określona wzorem

$$p_i(v) = \mu_i \cdot e^{-\mu_i v}, \tag{5}$$

gdzie μ_i^{-1} jest średnią długością informacji typu (i).
Zakładamy ponadto, że informacje typu (3) dzielone są na bloki o długości danej wzorem

$$l_B^{(3)} = k \cdot v_m \cdot T_r. \tag{6}$$

Przyjmujemy tu, że informacje typu (1) i (2) (patrz p.1.2) dzielone są na bloki o maksymalnej długości równej $l_B^{(i)}$. Bloki te tworzą następnie informację właściwą w pakiecie ty-

pu (i). Długość informacji właściwej w pakiecie przenoszącym ostatni blok danej informacji równa jest długości tego bloku, a długość nagłówka[4] $l_H^{(i)}$ w każdym pakiecie jest stała.

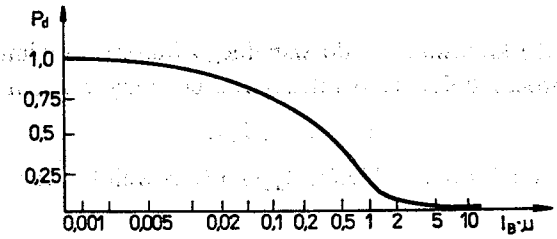
Dzieląc każdą wiadomość na bloki otrzymujemy $(n-1)$ bloków o długości $l_B^{(i)}$ i 1 blok o długości mniejszej lub równej $l_B^{(i)}$. Prawdopodobieństwo, że wiadomość składa się z n bloków wyraża się wzorem

$$P(\underset{\sim}{N}_b = n) = \int_{(n-1)l_B^{(i)}}^{nl_B^{(i)}} p_i(v)\mathrm{d}v = (e^{\mu_i l_B^{(i)}} - 1) \cdot e^{-\mu_i l_B^{(i)} \cdot n} \tag{7}$$

Możemy też obliczyć prawdopodobieństwo zdarzenia, że losowo wybrany blok powstały z podziału dowolnej informacji typu (i) ma długość równą $l_B^{(i)}$

$$P_d^{(i)} = \sum_{n=1}^{\infty} \frac{n-1}{n} P(\underset{\sim}{N}_b = n) = 1 - \left(l_B^{(i)} \cdot \mu_i - \ln(e^{l_B^{(i)}\mu_i} - 1)\right)(e^{\mu_i l_B^{(i)}} - 1) \tag{8}$$

Wykres wartości $P_d^{(i)}$ w funkcji iloczynu $l_B^{(i)} \cdot \mu_i$ znajduje się na rys. 2.



Rys. 2. Prawdopodobieństwo zapełnienia bloku w funkcji iloczynu $l_B \cdot \mu$

Korzystając z powyższych zależności, można wyprowadzić wzór na gęstość prawdopodobieństwa długości bloku dla informacji o wykładniczym rozkładzie długości z wartością średnią μ_i^{-1}. Otrzymamy więc

$$p_b^{(i)}(v) = \frac{\mathrm{dB}_i(v)}{\mathrm{d}v} = P_d^{(i)} \cdot \delta(v - l_B^{(i)}) + (1 - P_d^{(i)}) \cdot \frac{\mu_i e^{-\mu_i v}}{1 - e^{-\mu_i l_B^{(i)}}}, \quad v \in (0, l_B^{(i)}\rangle \tag{9}$$

gdzie $B_i(v)$ jest dystrybuantą rozkładu długości bloku typu (i).

Wzór na gęstość prawdopodobieństwa długości pakietu można otrzymać ze wzoru (8) zastępując $l_B^{(i)}$ przez $l_B^{(i)} + l_H^{(i)}$.

Interesująca nas średnia długość bloku wynosi więc

$$E\underset{\sim}{l}_b^{(i)} = \int_0^{l_B^{(i)}} v\,\mathrm{dB}_i(v) = P_d^{(i)} \cdot l_B^{(i)} + (1 - P_d^{(i)}) \cdot \frac{1}{\mu_i} \cdot \frac{1}{1 - e^{-\mu_i l_B^{(i)}}} \cdot \left(1 - e^{-\mu_i l_B^{(i)}}(1 + \mu_i l_B^{(i)})\right), \tag{10}$$

a średnia długość pakietu

$$E\underset{\sim}{l}_p^{(i)} = E\underset{\sim}{l}_b^{(i)} + l_H^{(i)}. \tag{11}$$

[4] w pracy tej nagłówek jest synonimem informacji sterująco-kontrolnej

Obliczone powyżej średnie długości bloków i pakietów są niezbędne do obliczenia całkowitego natężenia strumienia wejściowego węzła oraz opóźnień transmisji.

2.3. NATĘŻENIA STRUMIENI WEJŚCIOWYCH W WĘŹLE

W dalszym ciągu zakładać będziemy, że w jednostce czasu wpływa do węzła średnio R_{0i} bitów informacji i-tego rodzaju ($i = 1, 2, 3$). Oznacza to, że do węzła w jednostce czasu dociera przeciętnie $\mu_i R_{0i}$ informacji i-tego rodzaju. Oznaczamy przez λ_{0i} liczbę pakietów wysyłanych w jednostce czasu do bufora przez dystrybutor węzła. Wynosi ona

$$\lambda_{0i} = \frac{R_{0i}}{E l_b^{(i)}}. \tag{12}$$

Liczba ta zostaje powiększona o liczbę pakietów retransmitowanych, a więc całkowite natężenie strumienia informacji i-tego rodzaju wynosi

$$\lambda_i = a_i \cdot \lambda_{0i} \tag{13}$$

gdzie a_i jest średnią liczbą transmisji pakietu typu (i) niezbędną do jego bezbłędnego przesłania.

Pakiety typu (1) i (3) kierowane są do wspólnego bufora. Przyjmujemy, że momenty ich zgłoszeń tworzą proces Poissona o intensywności danej wzorem

$$\lambda_1 = \lambda_{11} + \lambda_{13}. \tag{14}$$

Momenty zgłoszeń do bufora pakietów typu (2) również tworzą proces Poissona o intensywności

$$\lambda_2 = \lambda_{12} \tag{15}$$

Zatem współczynnik obciążenia kanału przez pakiety typu (1) i (3) wynosi

$$\varrho_1 = \frac{(\lambda_{11} E l_p^{(1)} + \lambda_{13} E l_p^{(3)})}{C}, \tag{16}$$

gdzie C oznacza przepustowość kanału.

Analogicznie współczynnik obciążenia kanału przez pakiety typu (2) przyjmuje wartość

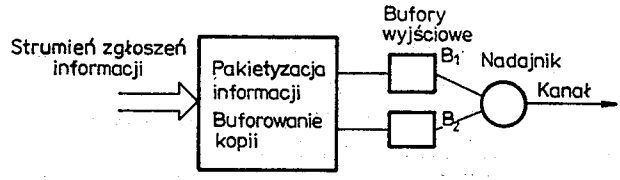
$$\varrho_2 = \frac{\lambda_{12} E l_p^{(2)}}{C}. \tag{17}$$

Zakładamy również, że liczba informacji typu (1) pozostaje w stałym stosunku do informacji typu (2), tzn.

$$\mu_1 R_{01} = \omega \mu_2 R_{02}. \tag{18}$$

Wzór ten oznacza, że na jedną informację typu (2) („długą") wpływającą do węzła przypada ω informacji typu (1) („krótkich"). W celu sprecyzowania dalszych rozważań konieczne jest przyjęcie modelu węzła sieci. W pracy tej przyjmiemy prosty model węzła (patrz rys. 3) składający się z:

— bloku, do którego napływają zarówno informacje wchodzące do sieci jak i pakiety z sąsiednich węzłów; dokonywana jest w nim pakietyzacja informacji oraz przechowywane są kopie pakietów przekazywanych do buforów wyjściowych,



Rys. 3. Schemat modelu węzła sieci

— dwóch buforów wyjściowych, w tym:
buforu B_1 przechowującego pakiety typu (1) i (3),
buforu B_2 przechowującego pakiety typu (2)
— stanowiska obsługi (nadajnika).

Pakiety trafiające do bufora (1) posiadają wyższy priorytet niż pakiety typu (2), z uwagi na wymagane mniejsze opóźnienie dopuszczalne. Przyjmujemy ponadto, że powiadomienia o dyskwalifikacji pakietów przez węzeł odbiorczy przekazywane są w pakietach wysyłanych w kierunku zwrotnym (decyzyjne sprzężenie zwrotne). Zakładamy też bezpamięciowość kanału, tj. niezależność zniekształceń elementarnych. Wobec tego prawdopodobieństwo wystąpienia co najmniej jednego błędu w pakiecie długości l_p wynosi

$$P = 1-(1-P_0)^{l_p}, \tag{19}$$

gdzie P_0 jest prawdopodobieństwem wystąpienia błędu elementarnego. Przy założeniu, że $P_0 l_p \ll 1$, możemy prawą stronę wzoru aproksymować wyrażeniem

$$P \simeq P_0 \cdot l_p \tag{20}$$

Średnia liczba transmisji niezbędnych do bezbłędnego przesłania pojedynczego pakietu o długości $l_B^{(i)}+l_H^{(i)}$ jest więc dana wzorem

$$a_i \cong \frac{1}{1-P_0(l_B^{(i)}+l_H^{(i)})} \tag{21}$$

Wzór (21) określa górną granicę wartości a_i, która jest osiągalna dla pakietu o długości maksymalnej $l_B^{(i)}+l_H^{(i)}$. Oczywiście prawdopodobieństwo osiągnięcia przez pakiet maksymalnej długości jest tym większe, im mniejszy jest iloczyn $\mu_i l_H^{(i)}$.

2.4. OPÓŹNIENIA W KANALE MIĘDZYWĘZŁOWYM

Zajmiemy się teraz analizą opóźnienia, jakiego doznaje pakiet w węźle. W opisanym modelu na czas przebywania pakietu w węźle składa się:
— czas oczekiwania w buforze,
— czas obsługi (czas trwania sygnału przenoszącego informacje zawarte w pakiecie).

Jeśli węzeł odbiorczy wykryje błąd w pakiecie, to następuje jego retransmisja, co oczywiście rzutuje na średni czas przebywania pakietu w węźle. Przyjmując te założenia, możemy określić średni czas konieczny do bezbłędnego przesłania pakietu o priorytecie k ($k = 1, 2$). Mamy więc

$$T_k = (T_{bk}+T_{tk})+(T_{bk}+T_{tk}+T_m)P_k+(T_{bk}+T_{tk}+T_m)P_k^2+\ \ldots\ =$$
$$= (T_{bk}+T_{tk}+T_m)\frac{1}{1-P_k}-T_m, \tag{22}$$

gdzie:

P_k — prawdopodobieństwo wystąpienia błędu w pakiecie o priorytecie k,

T_k — średni czas niezbędny do bezbłędnego przesłania pakietu o priorytecie k,

T_{bk} — średni czas przebywania w buforze pakietu o priorytecie k,

T_{tk} — średni czas transmisji pakietu o priorytecie k,

T_m — średni czas transmisji informacji pomocniczej o wykryciu błędu w pakiecie.

Ponieważ pakiety obsługiwane są w węźle ze stałymi priorytetami, więc wzory na średnie czasy przebywania w buforach przybierają postać

$$T_{b1} = \frac{M}{1-\varrho_1}; \qquad (23) \qquad T_{b2} = \frac{M}{(1-\varrho_1)(1-\varrho_1-\varrho_2)}, \qquad (24)$$

gdzie ϱ_1 i ϱ_2 dane są wzorami (16) i (17), a M wynosi

$$M = \lambda_1(m_2^{(1)}+m_2^{(3)})+\lambda_2 m_2^{(2)}. \qquad (25)$$

Wartości λ_1 i λ_2 określone są przez zależności (14), (15), natomiast występująca we wzorze (25) wielkość $m_2^{(i)}$ jest drugim momentem rozkładu długości pakietu, danym wzorem

$$m_2^{(i)} = \int_{l_H^{(i)}}^{l_H^{(i)}+l_B^{(i)}} \cdot v^2 \mathrm{dB}_p^{(i)}(v). \qquad (26)$$

Po podstawieniu odpowiednich wzorów otrzymujemy

$$m_2^{(3)} = \frac{1}{C^2}(l_B^{(3)}+l_H^{(3)})^2, \qquad (27)$$

$$m_2^{(i)} = \frac{1}{C_2}\left\{P_d^{(i)}(l_B^{(i)}+l_H^{(i)})^2+(1-P_d^{(i)})\cdot\mu_i(e^{-\mu_i l_H^{(i)}}-e^{-\mu_i(l_B^{(i)}+l_H^{(i)})})^{-1}\cdot\left[e^{-\mu_i l_H^{(i)}}\left(\frac{l_H^{(i)}}{\lambda_i}+\right.\right.\right.$$

$$\left.\left.\left.+\frac{2l_H^{(i)}}{\lambda_i^2}+\frac{2}{\lambda_i^3}\right)-e^{-\mu_i(l_B^{(i)}+l_H^{(i)})}\cdot\left(\frac{(l_B^{(i)}+l_H^{(i)})^2}{\lambda_i}+\frac{2(l_B^{(i)}+l_H^{(i)})}{\lambda_i^2}+\frac{2}{\lambda_i^3}\right)\right]\right\};$$

$$\text{dla} \quad i = 1, 2. \qquad (28)$$

W celu obliczenia T_k musimy jeszcze określić średni czas transmisji pakietu. W przypadku pakietów o priorytecie (1) czas ten wynosi

$$T_{t1} = \frac{\lambda_{11}E\underset{\sim}{l}_p^{(1)}+\lambda_{13}E\underset{\sim}{l}_p^{(3)}}{\lambda_{11}+\lambda_{13}}\cdot\frac{1}{C}, \qquad (29)$$

a dla pakietów o priorytecie (2) mamy

$$T_{t2} = \frac{E\underset{\sim}{l}_p^{(2)}}{C}. \qquad (30)$$

Przyjmujemy, że natężenie strumieni w obydwu kierunkach jest jednakowe, a czas niezbędny do wykrycia obecności błędów w pakiecie, oraz czas oczekiwania na pakiet przesyłany w kierunku zwrotnym są pomijalnie małe. Zakładamy, że powiadomienie o dyskwalifikacji określonego pakietu jest dołączone do innego pakietu przesyłanego

w kierunku zwrotnym lub w przypadku braku takiego pakietu wysyłane jest ono za pośrednictwem oddzielnego pakietu protokolarnego. Wartość średnia czasu transmisji powiadomienia o dyskwalifikacji pakietu równa jest sumie średnich czasów transmisji pakietów ważonej udziałem danego typu pakietów w całkowitym strumieniu informacji, a więc

$$T_m = \frac{1}{C} \frac{\sum_{i=1}^{3} \lambda_{1i} E\underset{\sim}{l}_p^{(i)}}{\sum_{i=1}^{3} \lambda_{1i}}. \tag{31}$$

Jeśli podstawimy wzory (23)—(31) do (22), otrzymamy

$$T_1 = \left(\frac{M}{1-\varrho_1} + \frac{1}{C} \left(\frac{\lambda_{11} E\underset{\sim}{l}_p^{(1)} + \lambda_{13} E\underset{\sim}{l}_p^{(3)}}{\lambda_{11}+\lambda_{13}} + \frac{\sum_{i=1}^{3} \lambda_{1i} E\underset{\sim}{l}_p^{(i)}}{\Lambda_1} \right) \right) \cdot \frac{1}{1-P_1} - \frac{\sum_{i=1}^{5} \lambda_{1i} E\underset{\sim}{l}_p^{(i)}}{\Lambda_1} \cdot \frac{1}{C}; \tag{32}$$

$$T_2 = \left(\frac{M}{(1-\varrho_1)\cdot(1-\varrho_1-\varrho_2)} + \frac{1}{C} \left(\frac{\lambda_{11} E\underset{\sim}{l}_p^{(1)} + \lambda_{13} E\underset{\sim}{l}_p^{(3)}}{\lambda_{11}+\lambda_{13}} + \frac{\sum_{i=1}^{3} \lambda_{1i} E\underset{\sim}{l}_p^{(i)}}{\Lambda_1} \right) \right) \frac{1}{1-P_2} - \frac{\sum_{i=1}^{3} \lambda_{1i} E\underset{\sim}{l}_p^{(i)}}{\Lambda_i} \cdot \frac{1}{C}; \tag{33}$$

gdzie:

$$P_1 = P_0(l_B^{(1)} + l_H^{(1)}); \quad P_2 = P_0(l_B^{(2)} + l_H^{(2)}); \quad \Lambda_1 = \sum_{i=1}^{3} \lambda_{1i}.$$

W rozważaniach przyjęto, że decyzja o transmisji danego pakietu nie zależy od odbioru pakietów wysyłanych poprzednio (tzn. rozważany jest tzw. niezależny system transmisji pakietów). Ponadto ustalono maksymalną długość trasy. Przy wyborze kierowano się wynikami badań, które mówią, że trasa między węzłem źródłowym a docelowym liczy nie więcej niż 5 kanałów międzywęzłowych [8]. Oznacza to, że dopuszczalne opóźnienie transmisji między dwoma sąsiednimi węzłami równe jest średnio 1/5 całkowitego opóźnienia dla danego typu pakietów.

Pewnym niedostatkiem omawianej analizy jest ograniczenie się wyłącznie do obliczenia opóźnień średnich, bowiem dla niektórych użytkowników bardziej interesujący jest czas opóźnienia zadanego procentu (np. 90%) przesyłanych pakietów.
Stosowanie priorytetów utrudnia jednak wyznaczenie tego rodzaju parametru i nie będziemy się nim zajmować.

Ważnym zagadnieniem w naszych rozważaniach jest określenie długości nagłówków. W oparciu o dostępną literaturę [3, 5, 6] można przyjąć, że dla pakietów typu (1) i (2) wynosić ona będzie 236 bitów, zaś dla pakietów typu (3) — 200 bitów.

Długość części pakietu typu (3) przenoszącej informację właściwą możemy obliczyć ze wzoru (6). Jeżeli przyjmiemy, że $v_m = 9600$ b/s, $k = 1$, a $T_p = 20$ ms, to długość pakietu typu (3) jest stała i równa

$$E\underset{\sim}{l}_p^{(3)} = E\underset{\sim}{l}_H^{(3)} + E\underset{\sim}{l}_B^{(3)} = 200 + 9600 \cdot 0{,}02 = 392 \text{ bity}. \tag{34}$$

Korzystając ze wzoru (8) możemy więc obliczyć prawdopodobieństwo zdarzenia, że pakiet jest w pełni wykorzystany do przenoszenia ucyfrowionych sygnałów mowy. Wynosi ono 0,942. Zatem mniej niż 6% pakietów typu (3) przenosi oprócz sygnałów mowy również tzw. „dummy information".

Odrębnym zagadnieniem, jakie musi być rozstrzygnięte przy budowie sieci, jest dobór optymalnych wielkości $l_B^{(1)}$ i $l_B^{(2)}$. Krótkie informacje właściwe w pakietach powodują bowiem, że we wzorze na T_k najbardziej znaczący staje się pierwszy składnik (T_{bk}), natomiast przy dużych wartościach $l_B^{(1)}$ i $l_B^{(2)}$ decydujący wpływ na wartość T_k mają dwa pozostałe składniki wzoru, mianowicie T_{tk} i T_m oraz czynnik $\frac{1}{1-P_0}$.

2.5. OBLICZANIE OPTYMALNYCH DŁUGOŚCI PAKIETÓW

Celem przedstawionej niżej analizy jest znalezienie takich wartości $l_B^{(1)}$ i $l_B^{(2)}$, aby przy ustalonym udziale strumienia ucyfrowionych sygnałów mowy w całkowitym strumieniu przenoszonych informacji, przepustowość kanału została jak najlepiej wykorzystana.

Jako miarę wykorzystania przepustowości kanału przyjęto stosunek natężenia strumienia informacji właściwych wprowadzanych do węzła do przepustowości kanału

$$\eta = \frac{\sum_{i=1}^{3} R_{0i}}{C}. \tag{35}$$

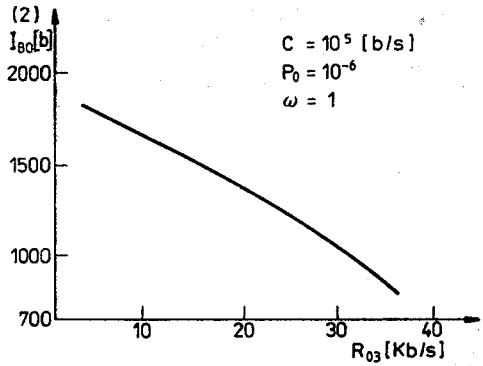
Interesuje nas również zależność optymalnej długości pakietów od udziału strumienia ucyfrowionych sygnałów mowy w całkowitym strumieniu przenoszonych informacji, a więc zależność

$$l_{B0}^{(k)} = f_k\left(\frac{R_{03}}{\sum_{i=1}^{3} R_{0i}}\right), \quad k = 1, 2. \tag{36}$$

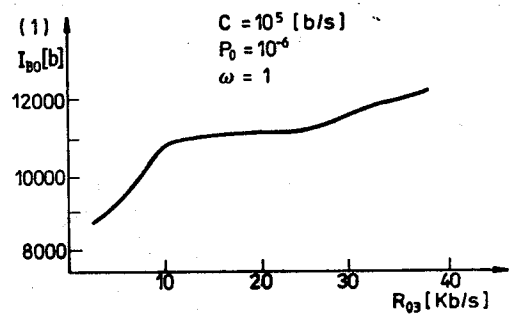
Znalezienie na drodze analitycznej długości pakietów $l_{B0}^{(1)}$ i $l_{B0}^{(2)}$, maksymalizującej natężenie strumienia informacji przenoszonych przez kanał przy zadanych opóźnieniach, nie jest zadaniem łatwym. Powodem tego jest skomplikowana postać wzorów (32), (33) oraz nieliniowa postać warunków ograniczających. Rozwiązanie tego zagadnienia możliwe jest natomiast na drodze numerycznej. W związku z tym napisany został program znajdujący wartości $l_{B0}^{(1)}$ i $l_{B0}^{(2)}$, a do obliczeń wykorzystano algorytm Gaussa-Seidela, z uwagi na łatwość jego implementacji.

Na rys. 4, 5, 6 przedstawiono typowe krzywe otrzymane dzięki rozwiązaniu zadania optymalizacyjnego dla zadanej wartości przepustowości kanału C, prawdopodobieństwa błędu elementarnego P_0, oraz stosunku liczby informacji typu (1) do liczby informacji typu (2) oznaczonego przez ω.

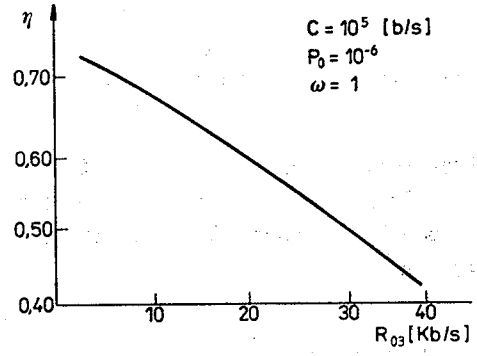
Obliczenia wartości $l_{B0}^{(1)}$ i $l_{B0}^{(2)}$ przeprowadzono dla C zmieniającego się w granicach od 100 kb/s do 0,5 Mb/s. Wybór tak dużych przepustowości podyktowany został tendencjami dominującymi w projektowanych współcześnie sieciach. O ile we wcześniejszych rozwiązaniach sieci takich jak ARPA, CYKLADY czy CYBERNET wykorzystywane były ka-



Rys. 4. Zależność optymalnej długości pakietu typu (2) od natężenia strumienia ucyfrowionych sygnałów mowy



Rys. 5. Zależność optymalnej długości pakietu typu (1) od natężenia strumienia ucyfrowionych sygnałów mowy



Rys. 6. Zależność wskaźnika wykorzystania przepustowości kanału od natężenia strumienia ucyfrowionych sygnałów mowy

nały o przepustowościach nie większych niż 50 kb/s, to w sieciach obecnie oddawanych do eksploatacji używane są kanały o przepustowościach sięgających megabitów na sekundę. Założono też, że P_0 wynosi 10^{-6}, co odpowiada kanałom o stosunkowo dobrej jakości.

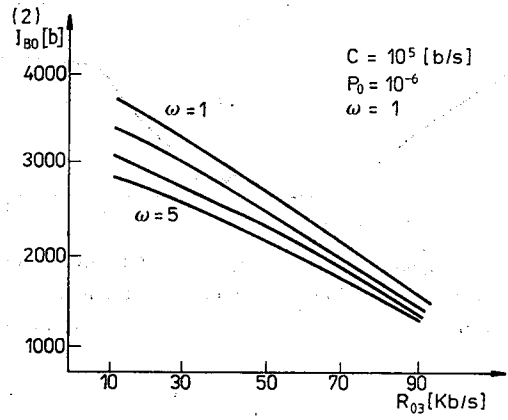
Pewnym problemem w obliczeniach był dobór warości ω. Zależy ona bowiem od sposobu użytkowania sieci. Jeśli sieć służy głównie do przesyłania długich informacji, to wartość ω bliska będzie jedności, gdyż każdemu żądaniu przesłania informacji z bazy danych (tj. informacji typu (1)) towarzyszyć będzie przesłanie informacji typu (2). Jeśli jednak przesyłane są głównie krótkie informacje między końcówkami i komputerami, a więc dominuje konwersacyjny tryb pracy, to wartość tego współczynnika jest większa od jedności. Dlatego przyjęto, że wartość ω może zmieniać się w granicach od 1 do 5.

2.6. OMÓWIENIE WYNIKÓW

W wyniku przeprowadzonych obliczeń okazało się, że wszystkie rozpatrywane przypadki posiadają pewne wspólne cechy. Zostaną one niżej omówione.

1) Zależność $l_{BO}^{(2)}$ od R_{03} jest prawie liniowa, przy czym wzrostowi R_{03} towarzyszy spadek wartości $l_{BO}^{(2)}$. Zjawisko to spowodowane jest zmniejszeniem się efektywnej przepustowości kanału dostępnej dla pakietów o niższym priorytecie. Na skutek tego przy

ustalonej długości $l_B^{(2)}$ rośnie czas transmisji T_t oraz czas buforowania T_b. Dlatego też musi ulec zmniejszeniu długość pakietu typu (1), aby mógł być spełniony warunek (2). Wartość $l_B^{(2)}$ maleje również ze wzrostem ω przy ustalonym R_{03} (patrz rys. 7).



Rys. 7. Zależność optymalnej długości pakietu typu (2) od natężenia strumienia ucyfrowionych sygnałów mowy przy zmianach parametru ω

Przyczyną tego zjawiska jest wzrost natężenia strumienia pakietów typu (1), co dodatkowo zmniejsza przepustowość kanału dostępną dla pakietów typu (2).

Prawie liniowa zależność między $l_{B0}^{(2)}$ a R_{03} pozwala aproksymować długość pakietu typu (2) wzorem

$$\hat{l}_{B0}^{(2)} = a - b \cdot R_{03}. \tag{37}$$

Wartości parametrów a i b dla wybranych przypadków podano w tabeli 1. W tabeli 2 znajdują się natomiast wartości $l_{B0}^{(2)}$ uzyskane numerycznie, wartości otrzymane za pomocą

Tabela 1

Przykładowe wartości współczynników a i b

Przepustowość kanału [b/s]	Parametr ω	Współczynnik a [b]	Współczynnik b [s^{-1}]
100 000	1	1950	28,75
200 000	1	3400	27,85
200 000	2	3150	26,0
250 000	1	4050	27,8
250 000	5	3100	20,0
500 000	1	8300	31,7

aproksymacji wzorem (37) oraz względny błąd aproksymacji. Warto zauważyć, że wartość $l_{B0}^{(2)}$ zwiększa się wraz ze wzrostem C przy ustalonym R_{03}. Wartość ta zwiększa się również wtedy, gdy R_{03} zostaje zwiększone proporcjonalnie do wzrostu C. Jest to spowodowane faktem, że przy proporcjonalnym zwiększaniu R_{03} rośnie efektywna przepustowość kanału dostępna dla pakietów typu (2).

Tabela 2

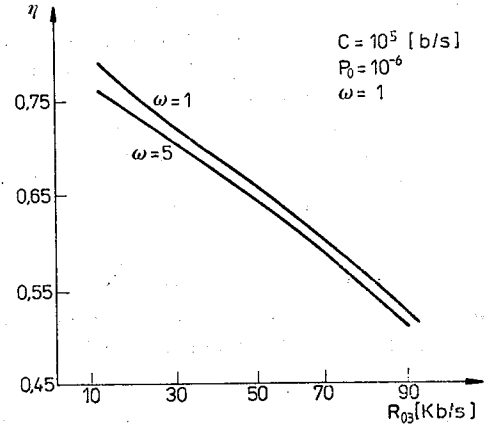
Porównanie wartości parametru $l_{B0}^{(2)}$ z wartością aproksymowaną $\hat{l}_{B0}^{(2)}$

R_{03}[kb/s]	10	20	30	40	50	60	70	80
$l_{B0}^{(2)}$[b]	2811	2574	2387	2161	1924	1674	1385	1049
$\hat{l}_{B0}^{(2)}$[b]	2860	2630	2370	2110	1850	1590	1330	1070
$\frac{\lvert l_{B0}^{(2)} - \hat{l}_{B0}^{(2)} \rvert}{l_{B0}^{(2)}}$ [%]	1,7	2,2	0,7	2,4	3,8	5,0	4,0	2,0

2) Zależność wartości $l_{B0}^{(1)}$ od C, R_{03} oraz ω jest bardziej złożona niż w przypadku $l_{B0}^{(2)}$. We wszystkich badanych przypadkach zaobserwowano wzrost wartości $l_{B0}^{(1)}$ towarzyszący zwiększaniu R_{03}, przy czym wartość $l_{B0}^{(1)}$ jest zawsze większa od 8 kb. Podstawiając do wzoru (8) wartości $N = 8$ kb oraz $\mu^{-1} = 1$ kb znajdujemy prawdopodobieństwo zdarzenia, że informacja typu (1) wypełni pakiet typu (1). Wynosi ono zaledwie $2{,}2 \cdot 10^{-4}$. Oznacza to, że prawie wszystkie informacje typu (1) przenoszone są przez pojedyncze pakiety.

W związku z tym spostrzeżeniem narzuca się metoda obsługi informacji typu (1). Nie należy ich dzielić, lecz w całości przesyłać w pojedynczych pakietach. Ułatwi to pakietyzację niewiele przy tym zmniejszając wykorzystanie przepustowości kanału.

3) Zgodnie z przewidywaniami wartość wskaźnika wykorzystania przepustowości kanału η zmniejsza się wraz ze wzrostem R_{03}. Jest to spowodowane stosunkowo długą częścią sterująco-kontrolną pakietu typu (3). Ponadto widzimy, że zwiększeniu przepustowości C towarzyszy wzrost η, gdyż wzrasta wtedy długość informacji właściwej w pakietach (patrz p.1°), co powoduje spadek względnego udziału informacji sterująco-kontrolnych w strumieniu przenoszonych informacji. Wartość η jest natomiast mało wrażliwa na zmiany ω, co obrazują wykresy z rys. 8.



Rys. 8. Zależność wskaźnika wykorzystania przepustowości kanału od natężenia strumienia ucyfrowionych sygnałów mowy przy zmianach parametru ω

Interesujący jest również wpływ, jaki na wartość η mają odchylenia od wartości optymalnych $l_{B0}^{(1)}$ i $l_{B0}^{(2)}$. Dokładne rozwiązanie tego zagadnienia jest trudne. Pozostają więc jedynie obliczenia numeryczne. Generalnie rzecz biorąc wartość η jest mało wrażliwa na odchylenia $l_B^{(1)}$ i $l_B^{(2)}$ od swoich wartości optymalnych. Wartość η jest szczególnie mało czuła na zmiany $l_B^{(1)}$. Nawet kilkudziesięcioprocentowe zmiany $l_B^{(1)}$ powodują spadek η o ułamki procenta. Jest to spowodowane dużą wartością iloczynu $l_B^{(1)} \cdot \mu_1$ przewyższającego liczbę 8. Zmiany $l_B^{(1)}$ nie mają wówczas wpływu na średnią długość pakietu (patrz wzór (10)) i stosunek długości części sterująco-kontrolnej pakietu do długości informacji właściwej pozostaje bez zmian.

Nieco większy wpływ na η mają zmiany $l_B^{(2)}$. Obliczona wartość iloczynu $l_B^{(2)} \mu_2$ waha się w granicach od 0,1 do 0,75. Oczywiście większość pakietów typu (2) wypełniona jest do swej maksymalnej długości. Dlatego też zmiany $l_B^{(2)}$ wpływają poprzez średnią długość pakietu na ilość wprowadzanej do kanału informacji sterująco-kontrolnej. Zmiana $l_{B0}^{(2)}$ o kilkadziesiąt procent powoduje spadek wartości η o kilka procent. Zauważmy jednak, że spadek wartości η o 1% oznacza zmniejszenie natężenia strumienia informacji przenoszonych przez sieć o tysiące bitów na sekundę. Dlatego też problem określenia właściwej wielkości pakietów, mimo wszystkich związanych z tym trudności, jest ważny przy projektowaniu nowoczesnych sieci przesyłania informacji, w których w grę wchodzi integracja usług.

BIBLIOGRAFIA

1. J. Peciak: *O utajnianiu mowy bez tajemnic.* Warszawa, MON 1980
2. B. Gold: *Digital speech networks.* IEEE Trans. on Comm., 1980, Vol. COM-28, pp. 1636—1658
3. I. Gitman, H. Frank: *Economic analysis of integrated voice and data networks; a case study.* Proc. of the IEEE, 1978, Vol. 66, pp. 1549—1570
4. G. Harrington: *Voice/data integration using circuit switched networks.* IEEE Trans. on Comm., 1979, Vol. COM-27, pp. 1478—1490M-27, pp. 1478—1490
5. A. Mowafi, J.W. Kelly: *Integrated voice/data packet switching techniques for future military networks.* IEEE Trans. on Comm., 1980, Vol. COM-28, pp. 1655—1662
6. J. Rossi, A. Mowafi: *Performance analysis of hybrid switching concept for integrated voice/data communication.* IEEE Trans. on Comm., 1982, Vol. COM-30, pp. 818—823
7. D. Rutkowski: *Dobór wielkości pakietu w sieci komunikacyjnej, (W:) Własności i funkcje sieci komputerowych, Część II.* Biblioteka WACS, skrypt PWr, Wrocław 1980
8. N. Abramson, F. Kuo: *Sieci telekomunikacyjne komputerów.* Warszawa, WNT, 1978
9. C. Weinstein: *Fractional Speech Loss and Talker Activity Model for TASI and for Packed-Switched Speech.* IEEE Trans. on Comm., 1978, Vol. COM-26, No. 8
10. T. Żmidziński: *Analiza wspólnego użytkowania kanału sieci telekomunikacyjnej przez strumienie: telefoniczny i informacji komputerowych.* Rozprawa doktorska, Gdańsk, 1986

M. PIONTAS, D. RUTKOWSKI

CHOICE OF THE OPTIMUM PACKET LENGHT IN INTEGRATED SERVICE NETWORK

Summary

The paper presents an analysis of packed-switching network in which three kinds of data are sent: human speech, short and long computer data. For the given model of the node and of the network, the lengths of packs optimizing the throughput of a channel, are calculated.

M. PIONTAS, D. RUTKOWSKI

CHOIX DE LA GRANDEUR DU PAQUET DANS LE RÉSEAU TÉLÉINFORMATIQUE

Résumé

On a présenté l'analyse du réseau avec commutation par paquets, par lequel sont transmis les signaux du parler humain ainsi que les courtes ou longues informations de l'ordinateur. On a calculé les longueurs du paquet optimisant le débit du modèle du noeud et du réseau présenté.

M. PIONTAS, D. RUTKOWSKI

WAHL DER OPTIMALEN PAKETGRÖßE IM NETZ MIT INTEGRIERTEN DIENSTLEISTUNGEN

Zusammenfassung

Es wird ein Netz mit Paketveränderung analysiert, worin Sprachsignale sowie kurze und lange Computerinformationen übertragen werden. Für das dargestellte Modell des Netzknotens und des Netzes werden Paketlängen berechnet, bei denen die Ausnutzung der Kanaldurchlaßfähigkeit optimal ist.

М. ПИОНТАС, Д. РУТКОВСКИ

ВЫБОР ДЛИНЫ ПАКЕТА В СЕТИ ИНТЕГРАЛЬНОГО ОБСЛУЖИВАНИЯ

Резюме

Представлен анализ сети коммутации пакетов, в которой передаются три рода данных: цифровая речь, короткие и длинные компьютерные информации. Для принятой модели узла и сети вычислена длина пакетов, оптимизующая использование скорости передачи канала.

ROZPRAWY ELEKTROTECHNICZNE, 1989, 35, z. 1, ss. 193–210

621.365.414:621.06

Closed loop identification of electric resistance furnaces

DOMINIK SANKOWSKI

Instytut Elektroenergetyki, Politechnika Łódzka

Received 1988.01.24

Authorized 1988.04.22

Theoretical and practical principles of on-line identification of the dynamics of elecrtic resistance furnaces in a closed loop configuration are presented. Test signals in the form of Multifrequency Binary Sequences (MBS) were applied to the closed loop arrangement.

1. INTRODUCTION

The applications of electroheat in the manufacturing and processing industries are widespread and of significant importance [1], [4]. Electric Resistance Furnaces are of particular interest. For example fabrication processes in the semiconductor industry require a precise, uniform and stable temperature control in the diffusion furnaces. The design and effective utilisation of this equipment requires detailed information about their static and dynamic properties. Description of system properties is attained using that body of knowledge commonly referred to under the general name of System Identification [5].

The purpose of this paper is to present a theoretical and practical bases of the identification of furnace dynamics in the closed loop configuration. A diffusion furnace is considered as a multivariable process within a temperature control loop. Test signals in the form of Multifrequency Binary Sequences (MBS) are applied to the closed loop structure. The software used in the signal processing and data reduction has been developed for specific application in the on-line non-parametric closed loop identification problem. Results using the classical sinusoidal method of identifying the same system form the basis for a comparison between this method and the MBS method.

2. DIGITAL CONTROL MICROPROCESSORS AND IDENTIFICATION IN ELECTRICAL RESISTANCE FURNACES

2.1. THE IMPACT OF MICROELECTRONICS AND SOFTWARE IN DIGITAL CONTROL

Various types of Direct Digital Control (DDC) have occored in the last few years. These developments have enabled the application of automatic control techniques in the

manufacturing and industrial processes. This is of significant importance in Electrical Resistance Furnaces which are used in many branches of industry for curing and heat treatment of materials. They also form an important stage in the processes of producing semiconductor devices in diffusion furnaces which may be regarded as multivariable systems.

The DDC systems utilize microprocessors or microcomputers as basic elements in the information loops of the process. Control algorithms used in process systems still heavily depend on PID controllers for a number of reasons [15]. A typical tuning method is based upon that proposed more than thirty years ago [3]. However, its principal disadvantage is the difficulty in precise tuning in real processes since these processes are generaly not very well known. It is of prime importance as the plant or process must usually operate with unconditional stability.

Automatic tuning based upon on-line identification is now a practical possibility in small computer environment. However, the identification must be performed in closed loop so that:

— no production is lost
— product quality must be maintained and
— operator, plant and environment must be protected.

The microcomputer offers the process engineer some other benefits which are worth mentioning. Portability of both hardware and software means that the benefit of many software functions from one hardware structure are readily available on the site of the process plant which normally is fixed in location. Hence the information loop functions of acquisition including measurement, communication, computation processing and distribution including control become more useful and more easily implemented. Versatility of these instruments is embodied in their ability to permit more precise control than simpler forms of control, including manual, since these assist in implementing more complex algorithms.

Other benefits which are worth mentioning are the lower cost and improved reliability. However, software can be expensive. Nevertheless, it can be made to suit individual requirements which may vary from time to time and from process to process. This can be a considerable benefit in a research and development environment.

2.2. THE IDENTIFICATION AIMS

It is possible to summarise identification by considering both the reason and solution techniques of the problem [5]. System design may be improved provided that data can be acquired. This is relevant to the synthesis of new process equipment or the modification of the existing plant. Generally speaking, understanding the process will usually lead to confidence in the product. Improved understanding may be acquired through knowledge of process behaviour under different conditions. Process identification allows all these benefits. Additionally it also allows the validation or checking of theoretically proposed models which can lead to mathematical descriptions of reduced complexity or simulation algorithms for further study and investigation.

2.3. SPECIAL PROBLEMS RELATED TO RESISTANCE FURNACES [13]

There are three identifiable groups of difficulties associated with resistance furnaces. Thus it is possible to enumerate special complications due to:

— the furance structure
— wide range of process functions
— influence of temperature sensors.

The furnace structure exerts significant influence for a number of reasons. Furnaces are normally of irregular shape and geometry. These influence the furnace as a dynamic system by having larger thermal inertias. Moreover, the structure has the characteristic of a distributed parameter system. Also of importance is the irregular temperature field which results from all geometrical factors. All materials suffer change with age, such as creep and deterioration of thermo-physical properties so that the dynamics of the system are strictly speaking non-stationary although the changes occur slowly with time.

There are many different process functions based upon the different technologies required for the large assortment of the treatments. Thus the operating temperature ranges ensure operation which is essentially non-linear. Charge insertion and removal is, strictly speaking, an essential disturbing operation in electrothermal process so that the overall process dynamics are changing. Finally, the effect of the measurement sensor must be taken into account. In the closed-loop identification method, the sensor influences both the control of the system and the on-line identification since it is a common element in both. The special problems with sensors are essentially associated with measurement precision which may deteriorate with ageing.

2.4. IDENTIFICATION CLASSES [6]

Identification methods may be grouped according to any of several ways, i.e.:

(i) parametric and non-parametric methods
(ii) deterministic, random, or pseudorandom methods
(iii) „*off-line*" or „*on-line*" methods
(iiii) open-loop, or closed-loop methods.

This is schematically ilustrated in Fig. 1.

Ad (i) Parametric methods generally mean estimating the parameters of a transfer function or differential difference equation based on the relationships between the experimental observations of the physical system.

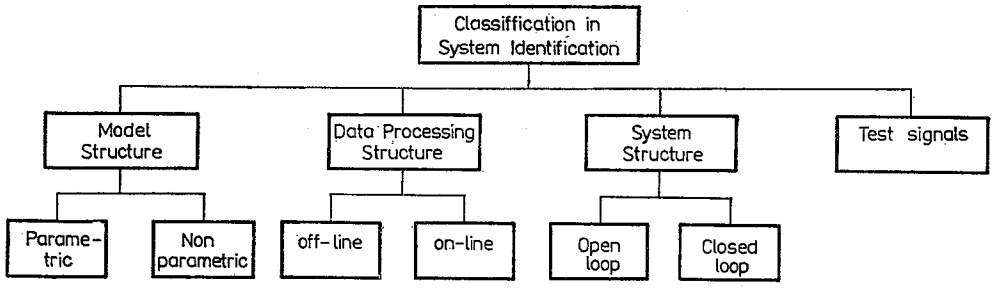


Fig. 1. Identification classes

These parametric methods require assumptions concering the choice of a specific dynamic model. Non-parametric methods, on the other hand, can be used in a very general way with linear systems because no assumption as to the order of the model need to be made.

Ad (ii) This classification is according to the type of test signals used for identification, i.e.: deterministic, random and pseudorandom [8].

Deterministic signals are those which can be described by a mathematical relationship, and can be categorized as being aperiodic, periodic, or quasi periodic. Step function and delta Dirac function can be regarded as aperiodic. Periodic signals can be further categorised as sinusoidal, composite sinusoidal, and binary sequence. Binary sequences are to be specially chosen as test signals because they simplify generation, storing and data processing. Several types of binary sequences are known; these include: Pseudo Random Binary Sequences (PRBS), Multifrequency Binary Sequences (MBS) and optimal binary sequences (optimal in the sense that the variance of the final parameter estimates is minimized). Pseudo Random Binary Sequences have a uniform energy spectrum and have been much employed in correlation analysis important in *on-line* identification. In many cases, identification of the process is sufficient at only a discrete number of frequencies, hence, the uniform energy spectrum obtained from PRBS is not appropriate. In such cases, a Multifrequency Binary Sequences (MBS) can be designed to concentrate its energy at the discrete frequencies of interest [2], [9].

Random signals can not be described by on explicit mathematical relationship and may be categorized as stationary or nonstationary. Stationary random processes may be further categorized as being either ergodic, or nonergodic.

Separate group of signal are pseudorandom included binary sequences. Although binary sequences are deterministic from the point of view of generation, it is also proper to regard them as random from the point of observers who are not aware of the code in which the information is contained.

Ad (iii) Parameter identification may be performed *off-line*, where all measurements are made first, before computations or *on-line*, where data is measured in real time, and all necessary computations are done during the intervals between measurements.

The *off-line* method is attractive, because it allows a relatively large sampling bandwidth and quite complex algorithms can be used in shaping and estimates. The *on line* approach however yields results as soon as measurements are done, and is useful in a small digital computer environment.

Ad (iiii) The input test signal in open loop experiments can be chosen by the experiment designer, or determined partly from output feedback by a controller of a given structure in the closed loop experiments.

Most processes operate as a part of a control configuration, and the inputs to the process are partly determined as feedback from other signals. In many cases security or production reasons do not permit that the controllers are removed during an identification experiment. Normally the reason for feedback control is to decrease the variance of the output. Then naturally also the information content of the meassured data desecrases, and the estimates obtained have worse accuracy. For this reason, closed loop experiment may be inferior to an open loop one, but has important advantage in avoiding the

special operation of opening the loop. It should be emphasized that the above classification is not means to be strictly applied to exclusive classes mentioned above.
It is possible to combine the classes as for example:

(i) parametric, deterministic, *off-line* open loop identification, or
(ii) nonparametric, deterministic *on-line*, closed loop identification.

3. THE BASIS OF IDENTIFICATION USING MBS TEST SIGNALS

3.1. THE DESIGN AND PROPERTIES OF MBS [2], [16]

Binary sequence inputs can be designed in such a way that they are persistently excited, periodic, and efficient, i.e. for a signal containing N_F frequencies the goal is to produce a signal with zero mean value and with all the power concentrated in the N_F dominant harmonics, and with $1/N_F$ part of the total power at each of the N_F frequencies. A number of useful MBS sequences are known to exist [2], [8], [9].

The MBS test signals used in the experimental part of this paper are proposed by Van den Bos [2]. This binary signal $u(t)$ is taken „+1" or „−1" depending on the polarity function $f(t)$ expresed by:

$$f(t) = \sum_{\substack{k=1,2,4,8 \\ 16,32}} \delta_k \cos\left(\frac{2\pi k}{N\lambda} \cdot t\right) \tag{1}$$

in discreate time intervals:

$$n\lambda \leqslant t \leqslant (n+1)\lambda \tag{2}$$

for $n = 0, 1, 2, \ldots (N-1)$
where

$$\delta_k = \begin{cases} +1 & \text{for} \quad k = 1, 2, 8, 32 \\ -1 & \text{for} \quad k = 4, 16 \end{cases} \tag{3}$$

λ — duration time of one impulse equal to sampling time
N — number of samples in one sequence

$$N = 128, 256, 512\ldots$$

The actual power $P_{\omega k}$ concentrated at frequency k is equal to $\frac{1}{2} a_{k_u}^2$ where a_{ku} is the Fourier coefficient of the k-th harmonic expressed by:

$$a_{ku} = \frac{4}{\pi k} \sum_{n=0}^{\frac{N}{2}-1} u_n \cdot \sin \frac{\pi k}{N} \cos \frac{\pi k}{N} (2n+1). \tag{4}$$

These MBS test signals have the following worthwhile advantages:
— they may be generated easily because they are binary,
— they are acurately known and periodic hence, it is not necessary to measure them

— they are even, periodic functions of time since this MBS possess only cosinusoidal harmonics. If a sequence with either time characteristic is to be generated, it is only necessary to store half the sequence,

— about 70% of the total power is concentrated at six frequencies $k\omega_0 k = 1, 2, 4, 8, 16, 32$ where $\omega_0 = \frac{2\pi}{N\lambda}$ is a basic angular frequency,

— this MBS has a mean value of zero, so that it introduces very small disturbances to the process set point.

An example of this MBS signal for $N = 128$ (first half) is shown on fig. 2. Figure 3 shows power distribution of the MBS signal.

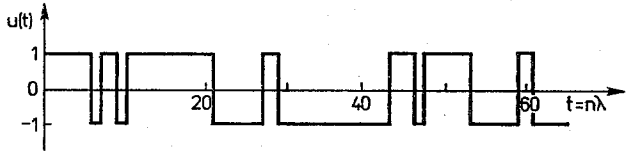


Fig. 2. MBS test signal

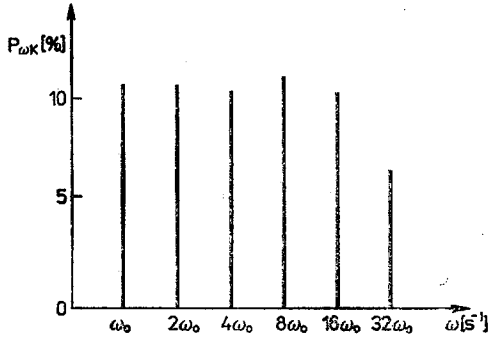


Fig. 3. Power distribution of the MBS test signal

3.2. FREQUENCY RESPONSE ESTIMATION BY MBS

Figure 4 shows a single input — output linear stationary process (electrical reistance furnace).

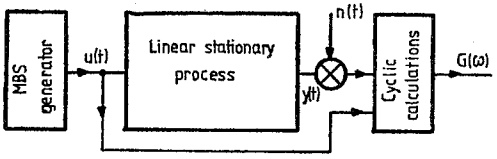


Fig. 4. Single input/output linear stationary process

For estimation, the observed output, $y(t)$ (temperature) is regarded as being composed of the true response $y_0(t)$ plus random disturbances $n(t)$, whose source may be the process itself or the data collection equipment. This $n(t)$ disturbances are of a Gaussion type with

the zero mean value. The dynamics of this single input/output linear stationary process can be expressed by its frequency response, as follows:

$$G(j\omega) = \frac{Y(j\omega)}{U(j\omega)} = |G(\omega)|e^{-j\varphi(\omega)} = P(\omega)+jQ(\omega) \tag{5}$$

where: $U(j\omega)$, $Y(j\omega)$ are Fourier transforms of the $u(t)$ and $y(t)$ respectively:

$$U(j\omega) = \int_{-\infty}^{\infty} u(t)e^{-j\omega t}\mathrm{d}t \tag{6}$$

$$Y(j\omega) = \int_{-\infty}^{\infty} y(t)e^{-j\omega t}\mathrm{d}t$$

$|G(\omega)|$ — amplitude ratio of $G(j\omega)$

$\varphi(\omega)$ — phase angle

$P(\omega)$, $Q(\omega)$ — real and imaginary part of the $G(j\omega)$ respectively.

Let $u(t)$ be the MBS signal periodic even function of time. Signals $u(t)$ and $y(t)$ can be expressed as Fourier series:

$$u(t) = \bar{u} + \sum_{k=1}^{\infty} a_{ku}\cos(k\omega_0 t)$$

$$y(t) = \bar{y} + \sum_{k=1}^{\infty} [a_{ky}\cos(k\omega_0 t)+b_{ky}\sin(k\omega_0 t)], \tag{7}$$

where: $\bar{u}$, $\bar{y}$ — are d.c. input and output components respectively. a_{ku}, a_{ky}, b_{ky} are the $u(t)$ and $y(t)$ Fourier coefficients of the k-th harmonic respectively. For the observation time T they are expressed by:

$$a_{ku} = \frac{2}{T}\int_0^T u(t)\cos(k\omega_0 t)\mathrm{d}t,$$

$$a_{ky} = \frac{2}{T}\int_0^T y(t)\cos(k\omega_0 t)\mathrm{d}t, \tag{8}$$

$$b_{ky} = \frac{2}{T}\int_0^T y(t)\sin(k\omega_0 t)\mathrm{d}t.$$

Regarding the relationships [5], [6] and [8] one obtains:

$$G(j\omega) = \frac{\int_0^T \sum_{k=1}^{\infty} [a_{ky}\cos(k\omega_0 t)+b_{ky}\sin(k\omega_0 t)][\cos\omega t - j\sin\omega t]\mathrm{d}t}{\int_0^T \sum_{k=1}^{\infty} a_{ku}\cos(k\omega_0 t)[\cos\omega t - j\sin\omega t]\mathrm{d}t}. \tag{9}$$

By using orthogonal properties of the sine and cosine function and assuming that all angular frequencies $\omega_k = k\omega_0$ are a multiple of basic angular frequency ω_0 (properties of MBS) one obtains:

$$G(k\omega_0) = \frac{a_{ky} - jb_{ky}}{a_{ku}} \tag{10}$$

Because the measurements at the output are contaminated by the Gaussian noise $n(t)$, the coefficients a_{ky} and b_{ky} respectively are estimated by $\hat{a}_{ky}$ and $\hat{b}_{ky}$:

$$\hat{a}_{ky} = \frac{2}{M \cdot T} \int_0^{M \cdot T} [y_0(t) + n(t)] \cos(k\omega_0 t) \mathrm{d}t,$$

$$\hat{b}_{ky} = \frac{2}{M \cdot T} \int_0^{M \cdot T} [y_0(t) + n(t)] \sin(k\omega_0 t) \mathrm{d}t \tag{11}$$

M — is the number of sequence periods
$M \cdot T$ — is the total experiment time.
It may be shown [10] that for this MBS signal, the relative variances of the estimates are inversely proportional to the product of total experiment time and the power present in the input signal at the corresponding frequency. For computer processing, the signal $y(t)$ is sampled with the discrete time interval:

$$y(t) = y(n\lambda) \quad n = 0, 1, \ldots (M \cdot N - 1)$$

Replacing the integrals (11) by sums, using (10) and (7) and taking the amplitude of MBS input u_0 one gets:

$$\hat{P}(k\omega_0) = \frac{2}{M \cdot N \cdot u_0 \cdot a_{ku}} \sum_{n=1}^{MN-1} [y(n\lambda) - \bar{y}] \cos(k\omega_0 n\lambda)$$

$$\hat{Q}(k\omega_0) = \frac{-2}{M \cdot N \cdot u_0 \cdot a_{ku}} \sum_{n=1}^{MN-1} [y(n\lambda) - \bar{y}] \sin(k\omega_0 n\lambda) \tag{12}$$

$k = 1, 2, 4, 8, 16, 32.$
The usage of well-known relationships:

$$G(k\omega_0) = \sqrt{P^2(k\omega_0) + Q^2(k\omega_0)}$$

$$\varphi(k\omega_0) = \frac{180}{\pi} \operatorname{arctg} \frac{Q(k\omega_0)}{P(k\omega_0)} \tag{13}$$

$k = 1, 2, 4, 8, 16, 32.$ will give six points of the complex transfer function in the range of angular frequencies $\omega_0 \div 32\omega_0$.

The computer which is used, not only to collect the data, but for control purposes as well, frequently has an idle period between samplings. The usage of this algorithm rather than the FFT algorithm, can potentially save time by utilizing these idle periods to carry out the calculations following eq. (12). These calculations can be done on-line as each sample in