POLSKA AKADEMIA NAUK
KOMITET ELEKTRONIKI I TELEKOMUNIKACJI

Nr Indeksu 363189
PL ISSN 0035-9386

KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJI

ELECTRONICS AND TELECOMMUNICATIONS QUARTERLY

TOM XXXIX — ZESZYT 1

WYDAWNICTWO NAUKOWE PWN
WARSZAWA 1993

POLSKA AKADEMIA NAUK
KOMITET ELEKTRONIKI I TELEKOMUNIKACJI

KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJI

ELECTRONICS AND TELECOMMUNICATIONS QUARTERLY

TOM XXXIX — ZESZYT 1

WYDAWNICTWO NAUKOWE PWN
WARSZAWA 1993

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

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

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

WYDAWNICTWO NAUKOWE PWN
Warszawa, ul. Miodowa 10

Ark. wyd. 17,25 Ark. druk. 14,5	Podpisano do druku w sierpniu 1993 r.
Papier offsetowy kl. III 80 g. B-1	Druk ukończono we wrześniu 1993 r.

Skład: Wyd. **GP-BIT** W-wa ul. Marymoncka 34.
Druk i oprawa:
DRUKARNIA BRACI GRODZICKICH

Z okazji wspaniałego Jubileuszu 90-lecia urodzin

Profesora Witolda Nowickiego

założyciela i wieloletniego Redaktora Naczelnego naszego kwartalnika
życzenia zdrowia, pomyślności oraz dalszego patronowania sprawom polskiej telekomunikacji

składają

Rada i Zespół Redakcyjny
KWARTALNIKA ELEKTRONIKI I TELEKOMUNKACJI

KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJI, 1993, 39, z. 1, ss. 5-18

621.317.328:621.3.017
512.86

Dyfrakcja na asymetrycznej półpłaszczyźnie impedancyjnej. Struktura rozwiązania w otoczeniu krawędzi

HALINA KUDREWICZ

Zakład Teorii Fal Elektromagnetycznych
Instytut Podstawowych Problemów Techniki PAN, Warszawa

Otrzymano 1992.04.03

Autoryzowano do druku 1992.12.11

W pracy zbadano strukturę rozwiązania zagadnienia dyfrakcji na asymetrycznej półpłaszczyźnie impedancyjnej w otoczeniu krawędzi. Analizowano rozwiązanie otrzymane przez Hudra (1976) metodą Wienera–Hopfera–Hilberta. Znaleziono asymptotyczne rozwinięcie w nieskończoności jednostronnych transformat Fouriera biorących udział w konstrukcji rozwiązania. W oparciu o te rozwinięcia skonstruowano rozwinięcie funkcji wyrażających gęstości prądów elektrycznego i magnetycznego na półpłaszczyźnie oraz rozwinięcie składowej wzdłużnej pola elektrycznego i składowej poprzecznej pola magnetycznego w aparaturze przy krawędzi.

1. WSTĘP

W literaturze są znane dwie metody rozwiązania problemu dyfrakcji na asymetrycznej półpłaszczyźnie. Pierwsza Maliużyńca (1958) i druga Hurda (1976) nazywana przez autora metodą Wienera–Hopfa–Hilberta. W każdej rozwiązanie otrzymuje się w postaci całki. W pierwszym przypadku jest to całka Sommerfelda, w drugim całka typu odwrotnej transformaty Fouriera. Żadna z tych postaci nie jest dogodna do analizy własności pola w otoczeniu krawędzi, gdzie jak wiadomo jest ono osobliwe. Jednakże, metoda Wienera–Hopfa–Hilberta daje aparat dla zbadania postaci rozwiązania przy krawędzi. Wykorzystuje bowiem funkcje analityczne, będące jednostronnymi transformatami Fouriera pewnych funkcji opisujących pole. W teorii rozwinięć asymptotycznych znane jest twierdzenie wiążące asymptotyczne rozwinięcie funkcji w zerze z asymptotycznym rozwinięciem jej transformaty Laplace'a i tym samym jednostronnej transformaty Fouriera w nieskończoności. To

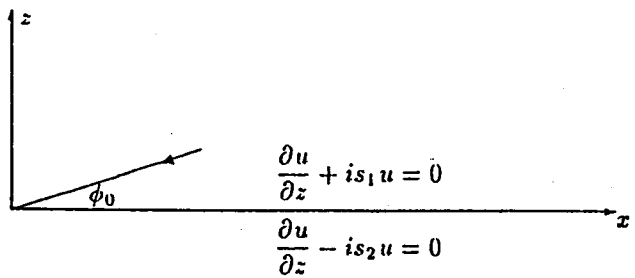
twierdzenie zostało wykorzystane w pracy. Znaleziono pierwsze wyrazy rozwinięcia asymptotycznego przy krawędzi funkcji wyrażających gęstość prądu elektrycznego i magnetycznego na asymetrycznej półpłaszczyźnie pola elektrycznego i magnetycznego w aperturze dla zagadnienia dyfrakcji fali płaskiej. Wykorzystano wyniki analizy rozwiązania zagadnienia dyfrakcji na półpłaszczyźnie impedancyjnej symetrycznej [3].

2. SFORMUŁOWANIE PROBLEMU DYFRAKCYJNEGO

Na asymetryczną półpłaszczyznę umieszczoną jak na rys. 1 pada fala płaska

$$u_i = \mathrm{e}^{-ik(x\cos\phi_0 + z\sin\phi_0)} \tag{1}$$

pod kątem ϕ_0, prostopadle do krawędzi y.



Rys. 1. Fala płaska padająca na półpłaszczyznę $z = 0$, $x \geqslant 0$

Współczynniki załamania półpłaszczyzny są różne dla każdej ze stron

$$s_1 = k \sin \phi_1, \quad s_2 = k \sin \phi_2 \tag{2}$$

i zakłada się, że są rzeczywiste oraz, że $0 < \phi_j < \pi/2$ dla $j = 1, 2$.

Należy znaleźć pole elektromagnetyczne $\boldsymbol{E}$, $\boldsymbol{H}$ (z harmoniczną zależnością od czasu $e^{-i\omega t}$, taką jak fala padająca) przy założeniu, że na półpłaszczyźnie spełnione są warunki Leontowicza:

$$\begin{aligned} \boldsymbol{n}\times\boldsymbol{E} &= \eta_1 Z[\boldsymbol{n}\times(\boldsymbol{n}\times\boldsymbol{H})] \quad \text{dla } z > 0, \\ \boldsymbol{n}\times\boldsymbol{E} &= \eta_2 Z[\boldsymbol{n}\times(\boldsymbol{n}\times\boldsymbol{H})] \quad \text{dla } z < 0, \end{aligned} \tag{3}$$

gdzie $\boldsymbol{n}$ jest wektorem normalnym zewnętrznym, $Z = \sqrt{\mu/\varepsilon}$, $\eta_j = k/s_j, j = 1, 2$.

Rozpatrzmy problem TM względem osi y. Wtedy skalarną funkcją u wyznaczającą całe pole jest składowa E_y pola $\boldsymbol{E}$. Mamy

$$\boldsymbol{E} = (0, u, 0), \qquad \boldsymbol{H} = \left(\frac{i}{\omega\mu}\frac{\partial u}{\partial z}, 0, -\frac{i}{\omega\mu}\frac{\partial u}{\partial x}\right). \tag{4}$$

Funkcja u spełnia równanie Helmholtza

$$\nabla^2 u + k^2 u = 0 \tag{5}$$

z warunkami brzegowymi

$$\begin{aligned} &\frac{\partial u}{\partial z} + is_1 u = 0 \quad \text{dla } x > 0,\ z = 0_+, \\ &\frac{\partial u}{\partial z} - is_2 u = 0 \quad \text{dla } x > 0,\ z = 0_- \end{aligned} \tag{6}$$

oraz z warunkami na ostrzu i w nieskończoności. Żądamy, żeby funkcja u była ograniczona, a w nieskończoności spełniała warunek Sommerfelda.

Rozwiązania poszukujemy w postaci sumy pola padającego i ugiętego:

$$u = u_i + u_s. \tag{7}$$

Funkcja u_i dana jest wzorem (1), funkcja u_s jest nieznana. Spełnia ona równanie Helmholtza (5), gdzie w miejsce u wstawiamy u_s. Z warunków brzegowych (6) po uwzględnieniu (7) i (1) otrzymujemy

$$\begin{aligned} &\frac{\partial u_s}{\partial z} + is_1 u_s = i(\gamma_0 - s_1)e^{i\alpha_0 x} \quad \text{dla } z = 0_+,\ x = 0, \\ &\frac{\partial u_s}{\partial z} - is_2 u_s = i(\gamma_0 + s_2)e^{i\alpha_0 x} \quad \text{dla } z = 0_-,\ x = 0, \end{aligned} \tag{8}$$

gdzie

$$\alpha_0 = -k\cos\phi_0, \quad \gamma_0 = k\sin\phi_0. \tag{9}$$

W przyjętej metodzie rozwiązywania będziemy również korzystali z warunku ciągłości funkcji u i jej pochodnej w aperturze

$$\begin{aligned} &(u_i + u_s)|_{z=0_-} = (u_i + u_s)|_{z=0_+} \quad \text{dla } x < 0, \\ &\left.\frac{\partial(u_i + u_s)}{\partial z}\right|_{z=0_-} = \left.\frac{\partial(u_i + u_s)}{\partial z}\right|_{z=0_+} \quad \text{dla } x < 0. \end{aligned} \tag{10}$$

3. SZKIC METODY WIENERA – HOPFA – HILBERTA

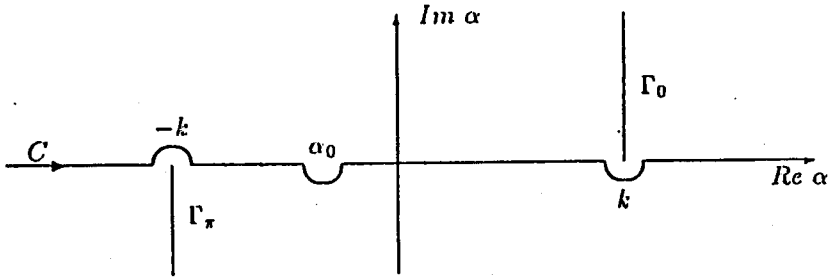
Rozwiązania szukamy w następującej postaci

$$\begin{aligned} u_s &= \int_C A(\alpha)e^{i(\alpha x + \gamma z)}d\alpha \quad \text{dla } z > 0, \\ u_s &= \int_C A(\alpha)e^{i(\alpha x - \gamma z)}d\alpha \quad \text{dla } z < 0, \end{aligned} \tag{11}$$

gdzie

$$\gamma = \sqrt{k^2 - \alpha^2}, \quad \gamma(0) = k. \tag{12}$$

Linia C jest położona na płaszczyźnie α z cięciami Γ_0, Γ_π, jak na rys. 2.



Rys. 2. Płaszczyzna zespolona α i kontur całkowania C

Do wyznaczenia pozostają amplitudy $A(\alpha)$ i $B(\alpha)$. Wstawiając funkcje (11) do warunków brzegowych (8) i warunków ciągłości (10) otrzymujemy cztery równania całkowe dla nieznanych amplitud $A(\alpha)$ i $B(\alpha)$

$$\int_C (\gamma + s_1)A(\alpha)e^{i\alpha x}d\alpha = i(\gamma_0 - s_1)e^{i\alpha_0 x} \quad \text{dla } x > 0, \tag{13}$$

$$\int_C (\gamma + s_2)B(\alpha)e^{i\alpha x}d\alpha = i(\gamma_0 + s_2)e^{i\alpha_0 x} \quad \text{dla } x > 0, \tag{14}$$

$$\int_C [A(\alpha) - B(\alpha)]e^{i\alpha x}d\alpha = 0 \quad \text{dla } x < 0, \tag{15}$$

$$\int_C \gamma[A(\alpha) + B(\alpha)]e^{i\alpha x}d\alpha = 0 \quad \text{dla } x < 0. \tag{16}$$

Oznaczmy przez Ω^+ półpłaszczyznę powyżej linii całkowania C, a przez Ω^- — półpłaszczyznę poniżej tej linii. Oznaczmy przez $U_j, j = 1, 2$ dowolną funkcję analityczną w Ω^+, ciągłą w obszarze domkniętym i znikającą dla $|\alpha| \to \infty$, a przez $L_j, j = 1, 2$ dowolną funkcję analityczną w Ω^-, ciągłą w obszarze domkniętym i znikającą dla $|\alpha| \to \infty$. Nietrudno pokazać, że jeżeli funkcje A i B spełniają następujące zależności, to spełniają również układ równań (13)—(16)

$$\begin{aligned}(s_1 + \gamma)A(\alpha) &= U_1(\alpha) + (\alpha - \alpha_0)p_1,\\ (s_2 + \gamma)B(\alpha) &= U_2(\alpha) + (\alpha - \alpha_0)p_2,\end{aligned} \tag{17}$$

$$\begin{aligned}A(\alpha) - B(\alpha) &= L_1(\alpha),\\ \gamma[A(\alpha) + B(\alpha)] &= L_2(\alpha),\end{aligned} \tag{18}$$

gdzie

$$p_1 = -(2\pi i)^{-1}(s_1 - \gamma_0),$$
$$p_2 = -(2\pi i)^{-1}(s_2 + \gamma_0). \tag{19}$$

Przez wyrugowanie funkcji $A(\alpha)$, $B(\alpha)$ z równań (17)–(18) otrzymuje się macierzowe równanie Wienera–Hopfa

$$\boldsymbol{G}(\alpha)\boldsymbol{L}(\alpha) = \boldsymbol{U}(\alpha) + (\alpha - \alpha_0)^{-1}\boldsymbol{P} \tag{20}$$

spełnione na linii C, a w ogólności, gdy liczba falowa k jest zespolona, tzn. $k = \operatorname{Re} k + i \operatorname{Im} k$, spełnione w poziomym pasie o szerokości $2 \operatorname{Im} k$.

Macierz $\boldsymbol{G}(\alpha)$, funkcje wektorowe $\boldsymbol{U}(\alpha)$ i $\boldsymbol{L}(\alpha)$ oraz wektor liczbowy $\boldsymbol{P}$ są opisane w następujący sposób

$$\boldsymbol{G}(\alpha) = \frac{1}{2}\begin{bmatrix} s_1 + \gamma & (s_1 + \gamma)/\gamma \\ -s_2 - \gamma & (s_2 + \gamma)/\gamma \end{bmatrix}, \tag{21}$$

$$\boldsymbol{L}(\alpha) = \begin{bmatrix} L_1(\alpha) \\ L_2(\alpha) \end{bmatrix}, \quad \boldsymbol{U}(\alpha) = \begin{bmatrix} U_1(\alpha) \\ U_2(\alpha) \end{bmatrix}, \quad \boldsymbol{P} = \begin{bmatrix} p_1 \\ p_2 \end{bmatrix}. \tag{22}$$

Łatwo jest pokazać, że jeżeli w otoczeniu linii C istnieje faktoryzacja

$$\boldsymbol{G}(\alpha) = \boldsymbol{G}_U(\alpha)\boldsymbol{G}_L(\alpha) \tag{23}$$

taka, że macierz $\boldsymbol{G}_U(\alpha)$ jest analityczna i nieosobliwa w Ω^+ wraz z pewnym otoczeniem linii C, a macierz $\boldsymbol{G}_L(\alpha)$ jest analityczna i nieosobliwa w Ω^- wraz z pewnym otoczeniem linii C, to rozwiązanie równania (20) w otoczeniu tej linii wyraża się wzorami

$$\boldsymbol{L}(\alpha) = \frac{\boldsymbol{G}_L^{-1}(\alpha)\boldsymbol{G}_U^{-1}(\alpha_0)\boldsymbol{P}}{\alpha - \alpha_0} + \boldsymbol{G}_L^{-1}(\alpha)\boldsymbol{W}(\alpha), \tag{24}$$

$$\boldsymbol{U}(\alpha) = \frac{\boldsymbol{G}_U(\alpha)\boldsymbol{G}_u^{-1}(\alpha_0) - \boldsymbol{I}}{\alpha - \alpha_0}\boldsymbol{P} + \boldsymbol{G}_U(\alpha)\boldsymbol{W}(\alpha), \tag{25}$$

gdzie $\boldsymbol{I}$ jest macierzą jednostkową, a $\boldsymbol{W}(\alpha)$ – wektorową funkcją całkowitą.

Po znalezieniu czynników faktoryzacji macierzy $\boldsymbol{G}(\alpha)$ i zbadaniu ich zachowania w nieskończoności okazuje się, że w klasie funkcji $\boldsymbol{U}(\alpha)$, $\boldsymbol{L}(\alpha)$ znikających w nieskończoności rozwiązanie jest jedno, dla którego $\boldsymbol{W}(\alpha) \equiv 0$. Wtedy $L_1(\alpha) = O(1/\alpha\sqrt{\alpha})$, $L_2(\alpha) = O(1/\alpha)$, $U_1(\alpha) = O(1/\sqrt{\alpha})$, $U_2(\alpha) = O(1/\sqrt{\alpha})$.

Najtrudniejszym zadaniem jest sfaktoryzowanie macierzy. Dokonał tego Hurd (1976). W podanej przez niego metodzie macierze $\boldsymbol{G}_U(\alpha)$ i $\boldsymbol{G}_L(\alpha)$ buduje się z wektorów tworzących rozwiązania wektorowego problemu Hilberta na linii Γ_π, powstałego z równania Wienera–Hopfa (23). Istota powodzenia metody leży w tym,

że dla rozpatrywanego zagadnienia, wektorowe równanie problemu Hilberta daje się rozseparować na dwa skalarne. Każde z nich rozwiązuje się poprzez całkę typu Cauchy'ego, podobnie jak skalarne równanie Wienera–Hopfa. Obliczona według tej procedury macierz $G_U(\alpha)$ ma postać:

$$G_U(\alpha) = \frac{1}{2\sqrt{K_U}} \begin{bmatrix} g & \sqrt{k+\alpha}\,g \\ g^{-1} & -\sqrt{k+\alpha}\,g^{-1} \end{bmatrix}, \tag{26}$$

gdzie funkcja $g = g(\alpha)$ jest dana wzorem

$$g^2(\alpha) = \frac{(\sqrt{k+c_1}+\sqrt{k+\alpha})(\sqrt{k-c_1}+\sqrt{k+\alpha})}{(\sqrt{k+c_2}+\sqrt{k+\alpha})(\sqrt{k-c_2}+\sqrt{k+\alpha})}, \tag{27}$$

przy czym oznaczono

$$c_j = k\cos\phi_j \quad \text{dla } j = 1, 2. \tag{28}$$

Funkcja K_U wyraża się w następujący sposób:

$$K_U(\alpha) = \exp Q(\alpha), \tag{29}$$

gdzie

$$Q(\alpha) = \frac{1}{2\pi i}\int_{\Gamma_\pi} \ln\frac{[s_1+\gamma(t)][(s_2+\gamma(t)]}{[s_1-\gamma(t)][s_2-\gamma(t)]}\,\frac{\mathrm{d}t}{t-\alpha}, \tag{30}$$

Całka ta nie daje się przedstawić w postaci funkcji elementarnych, ale można ją przedstawić jako sumę dwóch całek, z których każda zależy tylko od jednego parametru. Mamy

$$Q(\alpha) = Q(\alpha, s_1, s_2) = Q(\alpha, s_1) + Q(\alpha, s_2), \tag{31}$$

gdzie

$$Q(\alpha, s_j) = \frac{1}{2\pi i}\int_{\Gamma_\pi} \ln\frac{s_j+\gamma(t)}{s_j-\gamma(t)}\,\frac{\mathrm{d}t}{t-\alpha} \quad \text{dla } j = 1, 2. \tag{32}$$

Uwzględniając (31) w (29) możemy napisać

$$K_U(\alpha)\,K_U(\alpha, s_1, s_2) = K_U(\alpha, s_1, s_2) = K_U(\alpha, s_1)\,K_U(\alpha, s_2). \tag{33}$$

W pracy [4] pokazano, że funkcja $K_U(\alpha, s)$ jest czynnikiem faktoryzacji funkcji charakterystycznej, występującej w problemie dyfrakcji na półpłaszczyźnie symetrycznej $s_1 = s_2 = s$, przedłużonym na całą płaszczyznę α z cięciem Γ_π.

Mając czynnik $G_U(\alpha)$ faktoryzacji macierzy $G(\alpha)$, czynnik $G_L(\alpha)$, zgodnie z (23), znajdziemy ze wzoru

$$G_L(\alpha) = G_U^{-1}(\alpha) G(\alpha). \tag{34}$$

Aby obliczyć funkcje L i U wpiszemy brakujące nam macierze:

$$G_L^{-1}(\alpha) = \frac{\sqrt{K_U}}{2g(s_1+\gamma)(s_2+\gamma)} \times$$

$$\times \begin{bmatrix} (s_2+\gamma)g^2 - (s_1+\gamma) & \sqrt{k+\alpha}\,[(s_2+\gamma)g^2 + (s_1+\gamma)] \\ \gamma[(s_2+\gamma)g^2 + s_1 + \gamma] & \sqrt{k+\alpha}\,[(s_2+\gamma)g^2 - (s_1+\gamma)] \end{bmatrix}, \tag{35}$$

$$G_U^{-1}(\alpha) = \frac{1}{\sqrt{k+\alpha}} \sqrt{K_U} \begin{bmatrix} \sqrt{k+\alpha}\, g^{-1} & \sqrt{k+\alpha}\, g \\ g^{-1} & -g \end{bmatrix}. \tag{36}$$

Wstawiając (35) i (36) do (24) oraz (26) i (36) do (25) przy $W(\alpha) \equiv 0$ otrzymujemy wektorowe funkcje L i U, gdzie

$$L_1 = -\frac{g_0}{4\pi i(\alpha-\alpha_0)} \frac{\sqrt{K_U}}{\sqrt{K_{0U}}\, g(s_2+\gamma)} \times \tag{37}$$

$$\times \left\{(s_1-\gamma_0)g_0^{-2}\left(\frac{(s_2+\gamma)}{(s_1+\gamma)}g^2 - 1 + \sqrt{\frac{k+\alpha}{k+\alpha_0}}\left[\frac{(s_2+\gamma)}{(s_1+\gamma)}g^2 + 1\right]\right) + \right.$$

$$\left. + (s_2+\gamma_0)\left(\frac{(s_2+\gamma)}{(s_1+\gamma)}g^2 - 1 - \sqrt{\frac{k+\alpha}{k+\alpha_0}}\left[\frac{(s_2+\gamma)}{(s_1+\gamma)}g^2 + 1\right]\right)\right\},$$

$$L_2(\alpha) = -\frac{g_0}{4\pi i(\alpha-\alpha_0)} \frac{\sqrt{K_U}}{\sqrt{K_{0U}}\, g(s_2+\gamma)} \times \tag{38}$$

$$\times \left\{(s_1-\gamma_0)g_0^{-2}\left(\left[\frac{(s_2+\gamma)}{(s_1+\gamma)}g^2 + 1\right]\gamma + \sqrt{\frac{k+\alpha}{k+\alpha_0}}\left[\frac{(s_2+\gamma)}{(s_1+\gamma)}g^2 - 1\right]\right) + \right.$$

$$\left. + (s_2+\gamma_0)\left(\left[\frac{(s_2+\gamma)}{(s_1+\gamma)}g^2 + 1\right]\gamma - \sqrt{\frac{k+\alpha}{k+\alpha_0}}\left[\frac{(s_2+\gamma)}{(s_1+\gamma)}g^2 - 1\right]\right)\right\},$$

$$U_1(\alpha) = -\frac{1}{4\pi i(\alpha-\alpha_0)}\left\{\left[\frac{g}{g_0}\sqrt{\frac{K_U}{K_{0U}}}\left(1+\sqrt{\frac{k+\alpha}{k+\alpha_0}}\right)-2\right](s_1-\gamma_0)+\right.$$

$$\left.+\,gg_0\sqrt{\frac{K_U}{K_{0U}}}\left(1-\sqrt{\frac{k+\alpha}{k+\alpha_0}}\right)(s_2+\gamma_0)\right\}, \tag{39}$$

$$U_2(\alpha) = -\frac{1}{4\pi i(\alpha-\alpha_0)}\left\{g^{-1}g_0^{-1}\sqrt{\frac{K_U}{K_{0U}}}\left(1-\sqrt{\frac{k+\alpha}{k+\alpha_0}}\right)(s_1-\gamma_0)+\right.$$

$$\left.+\left[g^{-1}g_0\sqrt{\frac{K_U}{K_{0U}}}\left(1+\sqrt{\frac{k+\alpha}{k+\alpha_0}}\right)-2\right](s_2+\gamma_0)\right\}. \tag{40}$$

We wzorach (37)–(40) oznaczono

$$g_0 = g(\alpha_0), \qquad K_{0U} = K(\alpha_0). \tag{41}$$

Podstawiając (39) i (40) do (17) otrzymujemy

$$A(\alpha) = -\frac{g}{4\pi i(\alpha-\alpha_0)(s_1+\gamma)}\sqrt{\frac{K_U}{K_{0U}}}\left\{g_0^{-1}(s_1-\gamma_0)\left(1+\sqrt{\frac{k+\alpha}{k+\alpha_0}}\right)+\right.$$

$$\left.g_0(s_2+\gamma_0)\left(1-\sqrt{\frac{k+\alpha}{k+\alpha_0}}\right)\right\}, \tag{42}$$

$$B(\alpha) = -\frac{g^{-1}}{4\pi i(\alpha-\alpha_0)(s_2+\gamma)}\sqrt{\frac{K_U}{K_{0U}}}\left\{g_0^{-1}(s_1-\gamma_0)\left(1-\sqrt{\frac{k+\alpha}{k+\alpha_0}}\right)+\right.$$

$$\left.g_0(s_2+\gamma_0)\left(1+\sqrt{\frac{k+\alpha}{k+\alpha_0}}\right)\right\}. \tag{43}$$

Ostatecznie rozwiązanie problemu dyfrakcyjnego jest dane w postaci całkowej (11) z amplitudami $A(\alpha)$ i $B(\alpha)$ określonymi odpowiednio wzorami (42) i (43).

4. ANALIZA ROZWIĄZANIA W OTOCZENIU KRAWĘDZI

Dla zbadania postaci rozwiązania w otoczeniu krawędzi posłużymy się jednostronnymi transformatami Fouriera jakimi, jak wynika ze sposobu ich wprowadzenia, są funkcje L i U. Skorzystamy przy tym z twierdzenia wiążącego asymptotyczne rozwinięcie transformaty w nieskończoności z asymptotycznym rozwinięciem funkcji w zerze.

Wprowadzimy następujące oznaczenie dla prawostronnej transformaty Fouriera

$$\mathscr{L}_+[f(x)] = \frac{1}{2\pi}\int_0^\infty f(x)\mathrm{e}^{-i\alpha x}\mathrm{d}x. \tag{44}$$

Funkcja $\mathscr{L}_+$ jest analityczna w dolnej półpłaszczyźnie zmiennej zespolonej α.

Lewostronną transformatę Fouriera oznaczymy przez $\mathscr{L}_-$

$$\mathscr{L}_-[g(x)] = \frac{1}{2\pi}\int_{-\infty}^0 g(x)\mathrm{e}^{-i\alpha x}\mathrm{d}x. \tag{45}$$

Funkcja $\mathscr{L}_-$ jest analityczna w górnej półpłaszczyźnie zmiennej zespolonej α.

Wprowadzimy pojęcie gęstości prądu elektrycznego i magnetycznego na półpłaszczyźnie. Prądy te powstają na skutek nieciągłości na niej pola elektromagnetycznego.

Skok składowej E_y pola elektrycznego nazywamy gęstością prądu magnetycznego i oznaczamy

$$I_1(x) = E_y(x, 0_+) - E_y(x, 0_-). \tag{46}$$

Skok składowej H_x pola magnetycznego nazywamy gęstością prądu elektrycznego i oznaczamy

$$I_2(x) = H_x(x, 0_+) - H_x(x, 0_-). \tag{47}$$

Biorąc pod uwagę (11), (4) i (13) możemy napisać:

$$I_1(x) = \begin{cases} 0 & \text{dla } x < 0 \\ \displaystyle\int_C L_1(\alpha)\mathrm{e}^{i\alpha x}\mathrm{d}\alpha & \text{dla } x \geqslant 0. \end{cases} \tag{48}$$

Stąd wynika, że funkcja L_1 jest prawostronną transformatą Fouriera gęstości prądu magnetycznego

$$L_1(\alpha) = \mathscr{L}_+[I_1(x)]. \tag{49}$$

Funkcja L_2 wyraża się przez prawostronną transformatę gęstości prądu elektrycznego

$$L_2(\alpha) = \omega\mu\mathscr{L}_+[I_2(x)]. \tag{50}$$

Natomiast funkcje U_1 i U_2 są lewostronnymi transformatami

$$U_1(\alpha) = \mathscr{L}_-\left[\frac{\partial u_s}{\partial z} + is_1u_s - i(s_1 - \gamma_0)\mathrm{e}^{i\alpha_0 x}\right], \tag{51}$$

$$U_2(\alpha) = \mathscr{L}_-\left[\frac{\partial u_s}{\partial z} - is_2u_s - i(s_2 + \gamma_0)\mathrm{e}^{i\alpha_0 x}\right]. \tag{52}$$

Znajdziemy asymptotyczne rozwinięcie funkcji L_1, L_2, U_1, U_2 w nieskończoności wykorzystując rozwinięcia znanych funkcji w szeregi potęgowe oraz asymptotyczne rozwinięcie funkcji $K_U(\alpha, s)$, przedstawione w [4], na podstawie czego można napisać:

$$K_U(\alpha, s_j) = \exp\left\{\frac{1}{2\pi i}\int_{\Gamma_\pi} \ln\frac{s_j+\gamma(t)}{s_j-\gamma(t)}\,\frac{dt}{t-\alpha}\right\} = \tag{53}$$

$$= 1 - \frac{s_j}{\pi\alpha}\ln(1+\alpha/k) + O(1/\alpha) \quad \text{dla } \alpha \notin \Gamma_\pi;\ j = 1, 2.$$

Ze wzoru (27) mamy

$$g^2 = 1 - ia/\sqrt{a} + O(1/\alpha) \quad \text{dla } \alpha \to \infty,\ \alpha \notin \Gamma_\pi, \tag{54}$$

gdzie

$$a = \sqrt{k+c_1} + \sqrt{k-c_1} - \sqrt{k+c_2} - \sqrt{k-c_2}. \tag{55}$$

Stąd

$$g = 1 - ia/2\sqrt{\alpha} + O(1/\alpha) \quad \text{dla } \alpha \to \infty,\ \alpha \notin \Gamma_\pi. \tag{56}$$

Uwzględniając (53) i (33) dla funkcji $K_U(\alpha, s_1, s_2)$ mamy rozwinięcie

$$K_U = 1 - \frac{s_1+s_2}{\pi\alpha}\ln(1+\alpha/k + O(1/\alpha) \quad \text{dla } \alpha \to \infty,\ \alpha \notin \Gamma_\pi, \tag{57}$$

skąd

$$\sqrt{K_U} = 1 - \frac{s_1+s_2}{2\pi\alpha}\ln(1+\alpha/k) + O(1/\alpha) \quad \text{dla } \alpha \to \infty,\ \alpha \notin \Gamma_\pi. \tag{58}$$

Ze wzoru (37) otrzymujemy

$$L_1(\alpha) = \frac{a_1}{\alpha\sqrt{\alpha}} + \frac{a_2}{\alpha^2\sqrt{\alpha}}\ln(1-\alpha/k) + O(1/\alpha) \quad \text{dla } \alpha \to \infty,\ \alpha \notin \Gamma_0, \tag{59}$$

gdzie

$$a_1 = \frac{(s_1-\gamma_0)g_0^{-1} - (s_2+\gamma_0)g_0}{2\pi\sqrt{k+\alpha_0}\sqrt{K_{0U}}}, \tag{60}$$

$$a_2 = -\frac{[(s_1-\gamma_0)g_0^{-1} - (s_2+\gamma_0)g_0](s_1+s_2)}{4\pi^2\sqrt{K_{0U}}}. \tag{61}$$

Ze wzoru (38) otrzymujemy

$$L_2(\alpha) = \frac{b_0}{\alpha} + \frac{b_1}{\alpha^2}\ln(1 - \alpha/k) + O(1/\alpha) \quad \text{dla } \alpha \to \infty,\ \alpha \notin \Gamma_0, \tag{62}$$

gdzie

$$b_0 = -\frac{(s_1 - \gamma_0)g_0^{-1} + (s_2 + \gamma)g_0}{2\pi i\sqrt{K_0 U}}, \tag{63}$$

$$b_1 = \frac{[(s_1 - \gamma_0)g_0^{-1} + (s_2 + \gamma)g_0](s_1 + s_2)}{4\pi^2 i\sqrt{K_0 U}}. \tag{64}$$

Ze wzoru (39) otrzymujemy

$$U_1(\alpha) = \frac{c_1}{\sqrt{\alpha}} + \frac{c_2}{\alpha} + \frac{c_3}{\alpha\sqrt{\alpha}}\ln(1 + \alpha/k) + O\left(1/\alpha\sqrt{\alpha}\right) \quad \text{dla } \alpha \to \infty,\ \alpha \notin \Gamma_\pi, \tag{65}$$

gdzie

$$c_1 = -\frac{(s_1 - \gamma_0)g_0^{-1} + (s_2 + \gamma)g_0}{4\pi i\sqrt{k + \alpha_0}\ \sqrt{K_{0U}}}, \tag{66}$$

$$c_2 = -\frac{1}{4\pi i}\left\{\left[\left(1 - \frac{ia}{2\sqrt{k + \alpha_0}}\right)\frac{1}{g_0\sqrt{K_{0U}}} - 2\right](s_1 - \gamma_0) + \right.$$

$$\left. + \left(1 - \frac{ia}{2\sqrt{k + \alpha_0}}\right)\frac{g_0(s_2 + \gamma_0)}{\sqrt{K_{0U}}}\right\}, \tag{67}$$

$$c_3 = -\frac{[(s_1 - \gamma_0)g_0^{-1} + (s_2 + \gamma)g_0](s_1 + s_2)}{8\pi^2 i\sqrt{K_{0U}}\ \sqrt{k + \alpha_0}}. \tag{68}$$

Ze wzoru (40) otrzymujemy

$$U_2(\alpha) = \frac{d_1}{\sqrt{\alpha}} + \frac{d_2}{\alpha} + \frac{d_3}{\alpha\sqrt{\alpha}}\ln(1 + \alpha/k) + O\left(1/\alpha\sqrt{\alpha}\right) \quad \text{dla } \alpha \to \infty,\ \alpha \notin \Gamma_\pi, \tag{69}$$

gdzie

$$d_1 = \frac{(s_1 - \gamma_0)g_0^{-1} - (s_2 + \gamma)g_0}{4\pi i\sqrt{k + \alpha_0}\ \sqrt{K_{0U}}}, \tag{70}$$

$$d_2 = -\frac{1}{4\pi i}\left\{\left(1 - \frac{ia}{2\sqrt{k+\alpha_0}}\right)\frac{1}{g_0\sqrt{K_{0U}}}(s_1 - \gamma_0) + \right. \tag{71}$$

$$\left. + \left[\left(1 + \frac{ia}{2\sqrt{k+\alpha_0}}\right)\frac{g_0}{\sqrt{K_{0U}}} - 2\right](s_2 + \gamma_0)\right\},$$

$$d_3 = -\frac{[(s_1 - \gamma_0)g_0^{-1} - (s_2 + \gamma)g_0](s_1 + s_2)}{8\pi^2 i\sqrt{K_{0U}}\sqrt{k+\alpha_0}}. \tag{72}$$

We wzorach (56) i (59) a oznacza liczbę daną przez (55).

Wykorzystamy teraz twierdzenie wiążące asymptotyczne rozwinięcie funkcji w zerze z asymptotycznym rozwinięciem jej transformaty Laplace'a w nieskończoności [5]. Teraz tego twierdzenia brzmi:

Jeżeli funkcja $\varphi(t)$ ma rozwinięcie

$$\varphi(t) \sim \sum a_n \psi_n(t) \tag{73}$$

do N-tego wyrazu przy $t \to 0$, to jej transformata Laplace'a $\mathscr{L}[\varphi(t)] = f(s) = \int_0^\infty \varphi(t)e^{-st}dt$ ma rozwinięcie

$$f(s) \sim \sum a_n g_n(s) \tag{74}$$

do N-tego wyrazu, jednostajnie względem argumentu s przy $s \to \infty$, $s \notin S_\Delta$, $\Delta > 0$, gdzie S_Δ jest sektorem $0 < |s| < \infty$, $|\arg s| < \pi/2 - \Delta$,

$$g_n(s) = \mathscr{L}[\psi_n(t)]. \tag{75}$$

Dla jednostronnych transformat Fouriera (44) i (45) teza pozostanie prawdziwa, jeżeli współczynniki a_n we wzorze (74) podzielimy przez 2π oraz obszar S_Δ obrócimy odpowiednio o kąt $\pi/2$, zgodnie ze wskazówkami zegara dla transformaty $\mathscr{L}_+$, a przeciwnie – dla transformaty $\mathscr{L}_-$.

Z twierdzenia skorzystamy w drugą stronę. Znając rozwinięcie transformaty w nieskończoności określimy rozwinięcie funkcji w zerze, posługując się tablicami [6]. Korzystając z rozwinięcia (59) i zależności (49) otrzymamy następujące rozwinięcie funkcji $I_1(x)$ (wyrażającej gęstość prądu magnetycznego na półpłaszczyźnie) dla $x \to 0_+$:

$$I_1 = \frac{2i\sqrt{i}a_1}{\sqrt{\pi}}\sqrt{x} + \frac{2\sqrt{2i}a_2}{3}x\sqrt{x}\ln x + O(x\sqrt{x}), \tag{76}$$

gdzie a_1 i a_2 są dane odpowiednio wzorami (60) i (61).

Korzystając z rozwinięcia (62) i zależności (50) otrzymamy następujące rozwinięcie funkcji $I_2(x)$ (wyrażającej gęstość prądu elektrycznego na półpłaszczyźnie) dla $x \to 0_+$:

$$I_2(x) = \frac{ib_0}{\omega\mu} + \frac{b_1}{\omega\mu} x \ln x + O(x), \tag{77}$$

gdzie b_0 i b_1 są dane odpowiednio wzorami (63) i (64).

Rozwinięcia funkcji U_1 i U_2 w nieskończoności prowadzą do rozwinięć kombinacji liniowych składowej E_y pola elektrycznego i H_x pola magnetycznego w aperturze dla $x \to 0_-$.

Ze wzorów (65) i (51) otrzymujemy

$$E_y - \frac{\omega\mu}{s_1} H_x = \frac{c_1 i}{2s_1\sqrt{-\pi i}} \frac{1}{\sqrt{x}} - \frac{c_2}{s_1} - \frac{2c_3}{s_1\sqrt{i}} \sqrt{x} \ln x + O(\sqrt{x}), \tag{78}$$

gdzie c_1, c_2 i c_3 są dane odpowiednio wzorami (66)–(68).

Ze wzorów (69) i (52) otrzymujemy

$$E_y + \frac{\omega\mu}{s_2} H_x = \frac{d_1 i}{2s_2\sqrt{-\pi i}} \frac{1}{\sqrt{x}} - \frac{d_2}{s_2} - \frac{2d_3}{s_2\sqrt{i}} \sqrt{x} \ln x + (\sqrt{x}), \tag{79}$$

gdzie d_1, d_2 i d_3 są dane odpowiednio wzorami (70)–(72).

Jakościowy obraz rozwinięcia jest taki jak dla półpłaszczyzny impedancyjnej symetrycznej. Jest to rozwinięcie względem tych samych funkcji – innych niż w zagadnieniu dyfrakcji na półpłaszczyźnie idealnie przewodzącej. O postaci tych funkcji decyduje postać rozwinięcia w nieskończoności jednostronnych transformat Fouriera budujących rozwiązanie całego problemu dyfrakcyjnego w metodzie Hurda.

Z kombinacji liniowej równań (78) i (79) otrzymujemy

$$E_y(x) = \frac{(c_1+d_1)i}{2(s_1+s_2)\sqrt{-\pi i}} \frac{1}{\sqrt{x}} - \frac{c_2+d_2}{s_1+s_2} - \frac{2(c_3+d_3)}{(s_1+s_2)\sqrt{i}} \sqrt{x} \ln x + O(\sqrt{x}), \tag{80}$$

$$H_x(x) = \frac{(d_1 s_1 - c_1 s_2)i}{2\omega\mu(s_1+s_2)-\pi i} \frac{1}{\sqrt{x}} - \frac{d_2 s_1 - c_2 s_2}{\omega\mu(s_1+s_2)} - \frac{2(d_3 s_1 - c_3 s_2)}{\omega\mu(s_1+s_2)} \sqrt{x} \ln x + O(\sqrt{x}). \tag{81}$$

Przedstawiona metoda pozwala znaleźć pierwsze wyrazy rozwinięcia funkcji opisujących pole elektromagnetyczne w otoczeniu krawędzi, w płaszczyźnie $z = 0$. Nie daje natomiast informacji o tym jak rozwinięcie zależy od kąta obserwacji.

Jakościowy obraz rozwinięcia jest taki sam jak dla półpłaszczyzny impedancyjnej symetrycznej. W rozwinięcie wchodzą wyrazy zawierające logarytm, których nie ma w przypadku dyfrakcji na półpłaszczyźnie idealnie przewodzącej.

Otrzymano pierwsze wyrazy rozwinięcia gęstości prądu elektrycznego i magnetycznego przy krawędzi oraz składowej wzdłużnej pola elektrycznego i poprzecznej pola magnetycznego w aperturze.

BIBLIOGRAFIA

1. R.A. Hurd: *The Wiener–Hopf–Hilbert method for diffraction problems.* Can. J. Phys., (1976) **54**, 7, pp. 775–780
2. R.A. Hurd and S. Przeździecki: *Diffraction by a half-plane with different face impedances – a re-examination.* Can. J. Phys., (1981) **59**, pp. 1337–1347
3. H. Kudrewicz: *Dyfrakcja na półpłaszczyźnie impedancyjnej, struktura rozwiązania w otoczeniu krawędzi.* Prace IPPT 43/1990
4. H. Kudrewicz: *Diffraction by an impedance half-plane – dependence on the impedance parameter.* Archives of Acoustics, (1990) **15**, 1–2, pp. 151–183
5. A. Erdelyi: *Rozwinięcia asymptotyczne.* PWN, Warszawa 1967
6. M. Ryżik i Gradsztejn: *Tablice całek, sum, szeregów i iloczynów.* Państwowe Wydawnictwo Techniczno-Teoretycznej Literatury, Moskwa–Leningrad 1951 (w języku rosyjskim)
7. I.M. Braver, Kh.L. Garb, P.Sh. Fridberg, I.M. Yakover: *O zachowaniu równań Maxwella w bliskości krawędzi półpłaszczyzny, na której zadano dwustronne warunki brzegowe typu impedancyjnego* (w jęz. rosyjskim). DAN SSRR, 1986, 286, No 5, 1092–1096, Fizyka matematyczna
8. I.M. Braver, P.Sh. Fridberg, Kh.L. Garb, and I.M. Yakover: *Elektromagnetic Field Near the Edge of a Perfectly Conducting Edge and a Resistive Half-Plane.* Proc. of the 1989 URSI International Symposium on Electromagnetic Theory, Stockholm, pp. 88–90
9. I.M. Braver, P.Sh. Fridberg, Kh.L. Garb, J.M. Yakover: *The behavior of the electromagnetic field near the edge of a resistive half-plane.* JEE Trans. Antennas Propag. 1988, vol. AP-36, pp. 1760–1768

H. KUDREWICZ

DIFFRACTION BY AN ASYMMETRIC HALF–PLANE. THE STRUCTURE OF THE SOLUTION NEAR THE EDGE

Summary

For a diffraction problem on asymmetric impendance half-plane the structure of the solution near half-plane's edge has been examined. The foundation for the analysis was the Hurd's solution (1976) obtained with the Wiener–Hopf–Hilbert method. Incomplete Fourier transforms appearing in this solution were asymptotically expanded at infinity in the complex spectral plane. The expansions obtained were then used to asymptotically evaluate the functions describing electric and magnetic currents flowing in the half-plane, near the edge. They were also employed to find the asymptotic expansions of the longitudinal component of the electric field and the transversal magnetic field in the aperture near the edge.

KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJI, 1993, 39, z. 1, ss. 19-37

621.3.087

Generacja liczbowych ciągów pseudolosowych nad ciałem Galois GF (2^k) za pomocą rejestru akumulacyjnego

ANTONI ZABŁUDOWSKI, BOŻYDAR DUBALSKI

Instytut Telekomunikacji, Akademia Techniczno-Rolnicza, Bydgoszcz

Otrzymano 1992.01.09

Autoryzowano do druku 1992.05.20

W pracy dokonano analizy własności pseudolosowych sekwencji wielowartościowych (MPS) nad ciałem Galois GF (2^k) generowanych przez rejestr działający na zasadzie pracy rejestru akumulatora. Proponowana metoda jest tania w realizacji, gdyż wymaga użycia jedynie jednego rejestru zaś otrzymane sekwencje pseudolosowe cechują się dobrymi własnościami losowymi. Generator działający według proponowanej zasady wytwarza elementy ciągów liczbowych w każdym takcie zegarowym. Analiza sekwencji MPS dotyczy przebiegu funkcji autokorelacji oraz rozkładu serii generowanych liczb.

WSTĘP

W prezentowanej pracy, która stanowi kontynuację pracy [15], została zaproponowana nowa metoda generowania pseudolosowych ciągów liczbowych. Ciągi te posiadają takie same własności losowe, jak wielowartościowe sekwencje pseudolosowe (MPS) analizowane w pracy [15], lecz generatory zbudowane w oparciu o proponowaną zasadę są prostsze w realizacji układowej. Zaproponowana metoda bazuje na wykorzystaniu układu, któremu autorzy nadali nazwę rejestru akumulacyjnego, będącego transformacją rejestru liniowego (ściślej mówiąc rejestr liniowy jest szczególnym przypadkiem realizacji rejestru akumulacyjnego). W artykule opisana został idea tworzenia omawianego rejestru oraz zostały przeanalizowane parametry statystyczne sekwencji pseudolosowych uzyskiwanych z generatora MPS: równomierność rozkładu liczb pojawiających się w ciągu, przebieg funkcji autokorelacji w pełnym okresie sekwencji oraz rozkład serii generowanych liczb.

Ocena własności probabilistycznych sekwencji wielowartościowych polega na sprawdzeniu, na ile parametry losowe otrzymanych ciągów różnią się od paramet-

rów wielowartościowych sekwencji idealnych określonych postulatami R1 – R3, zdefiniowanych przez Daviesa i Golomba [3,6] dla ciągów binarnych, a które zostały rozszerzone na sekwencje wielowartościowe.

R1. Generowany ciąg zachowuje równomierność rozkładu liczb należących do zbioru $\{0, 1\ldots., (2^k - 1)\}$ lub do zbioru $\{-(2^k - 1), -(2^k - 3),\ldots -1, 1,\ldots, (2^k - 3), (2^k - 1)\}$, tzn. każda z liczb pojawia się z jednakowym prawdopodobieństwem równym $1/2^k$.

R2. Znormalizowana funkcja autokorelacji generowanej sekwencji liczb posiada następujący przebieg:

$$\frac{R_x(n)}{R_x(0)} = \begin{cases} 1 & \text{gdy } n = iL,\ i = 0,\ 1\ldots \\ A & \text{gdy } n \neq iL, \end{cases}$$

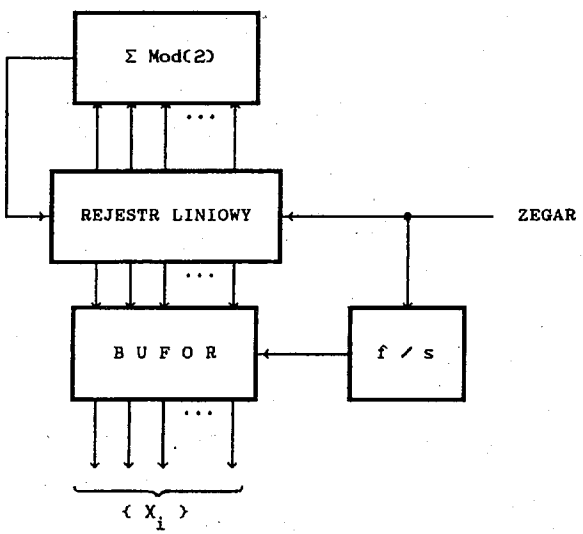
gdzie L – długość generowanego ciągu, wartość $A \cong 0$.

R3. Generowany ciąg liczbowy spełnia prawo seryjności określone w następujący sposób:

liczba serii określonej długości jest jednakowa dla każdej z liczb ze zbioru $\{0, 1,\ldots (2^k - 1)\}$ występujących w generowanej sekwencji, a liczba wszystkich serii o długości $(l - 1)$ jest 2^k-krotnie większa od liczby serii o długości l.

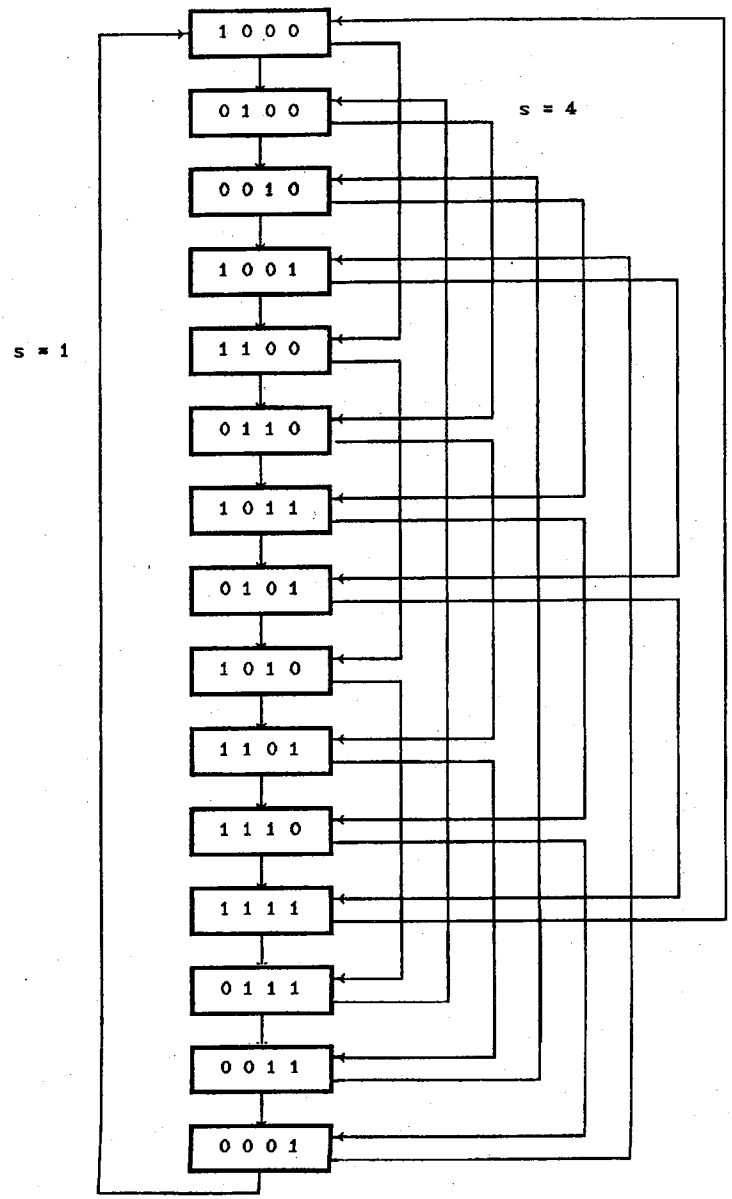
1. ZASADA GENERACJI WIELOWARTOŚCIOWYCH SEKWENCJI PSEUDOLOSOWYCH PRZY UŻYCIU REJESTRU AKUMULACYJNEGO

Przyjęta zasada generacji wielowartościowych sekwencji jest już dobrze znana i polega na zapisywaniu stanu wybranych k przerzutników tworzących rejestr liniowy, co s taktów zegarowych, do równoległego rejestru buforowego. Jeżeli sekwenc-



Rys. 1. Schemat blokowy proponowanego generatora

jom binarnym otrzymywanym z wyjść rejestru zostaną przypisane wagi 2^i, gdzie $i \in \{0,\ldots, k-1\}$, wówczas na wyjściu bufora pojawi się ciąg liczbowy, którego elementy należą do zbioru $\{0,\ldots, 2^k - 1\}$, posiadający cechy sekwencji pseudolosowych. Schemat blokowy tak zrealizowanego generatora pokazano na rys. 1. Na rys. 2 podano graf



Rys. 2. Graf stanów generatora pseudolosowej sekwencji wielowartościowej opisanego wielomianem charakterystycznym $f(x) = x^4 \oplus x \oplus 1$ dla parametrów skoku $s = 1$ oraz $s = 4$

opisujący działanie generatora wykorzystującego rejestr liniowy opisany wielomianem charakterystycznym $f(x) = x^4 \oplus x \oplus 1$, gdy elementy generowanych ciągów wielowartościowych są pobierane co $s = 1$ i $s = 4$ taktów zegarowych. Kolejność występowania stanów rejestru jest zależna od przyjętej liczby taktów zegarowych, po których stan przerzutników zapisywany jest do bufora.

Wykorzystanie omawianej metody generacji sekwencji pseudolosowych wymaga spełnienia następującego warunku:

Aby z generatora pseudolosowej sekwencji wielowartościowej otrzymać ciąg o długości $L = 2^m - 1$, który jest utworzony przez kolejne s-te binarne wektory, liczby L oraz s muszą być względnie pierwsze.

W przypadku niespełnienia tego warunku okres sekwencji wielowartościowej jest krótszy, a ciąg traci swe własności losowe [4].

Nietrudno stwierdzić, że przyjęta realizacja układowa generatora powoduje, że częstotliwość pojawienia się wektorów binarnych (binarnej reprezentacji liczb należących do ciągu wielowartościowego) na wyjściach bufora jest $l = f/s$ razy mniejsza niż częstotliwość pracy rejestru liniowego.

Niech będzie dany rejestr liniowy opisany macierzą generującą [M] generujący ciąg binarny maksymalnej długości. Załóżmy, że stan początkowy rejestru określa wektor $\langle \mathbf{X}_0 \rangle$ różny od wektora zerowego. Ponieważ sekwencję wielowartościową tworzy co s-ty wektor, można ją opisać następującym przekształceniem:

$$\begin{aligned} \langle \mathbf{X}_1 \rangle &= \langle \mathbf{X}_0 \rangle [\mathbf{M}]^s \\ \langle \mathbf{X}_2 \rangle &= \langle \mathbf{X}_0 \rangle [\mathbf{M}]^{2s} \\ \langle \mathbf{X}_3 \rangle &= \langle \mathbf{X}_0 \rangle [\mathbf{M}]^{3s} \\ \langle \dot{\mathbf{X}}_i \rangle &= \langle \dot{\mathbf{X}}_0 \rangle [\dot{\mathbf{M}}]^{is} \\ \langle \mathbf{X}_{L-1} \rangle &= \langle \dot{\mathbf{X}}_0 \rangle [\dot{\mathbf{M}}]^{(L-1)s}, \end{aligned} \tag{2}$$

przy czym iloczyn is jest obliczany MOD(L).

Z ciągu $\langle \mathbf{X}_0 \rangle$, $\langle \mathbf{X}_1 \rangle$,..., $\langle \mathbf{X}_{L-1} \rangle$ wzięto pod uwagę podwektory wymiaru $[l \times k]$ będące binarną reprezentacją sekwencji wielowartościowej. Łatwo zauważyć, że tę samą sekwencję można otrzymać nie przez wybór co s-tego wyrazu sekwencji nieprzekształconej, lecz dzięki zbudowaniu takiego rejestru, w którym zbiór wektorów binarnych jest transformowany na siebie przez macierz $[\mathbf{M}_A] = [\mathbf{M}]^s$. Nietrudno stwierdzić, że tak powstały generator działa identycznie jak automat [4], którego wejścia przerzutników określone są przez liniowe funkcje boolowskie wynikające z postaci macierzy $[\mathbf{M}_A]$. Podobieństwo idei działania zaproponowanego generatora z zasadą działania rejestru akumulatora spowodowało, że nadano mu nazwę *rejestru akumulacyjnego*. Liczbę s nazywać będziemy dalej parametrem skoku rejestru akumulacyjnego.

Dla ilustracji dotychczasowych rozważań przedstawiony zostanie prosty przykład.

Przykład 1.

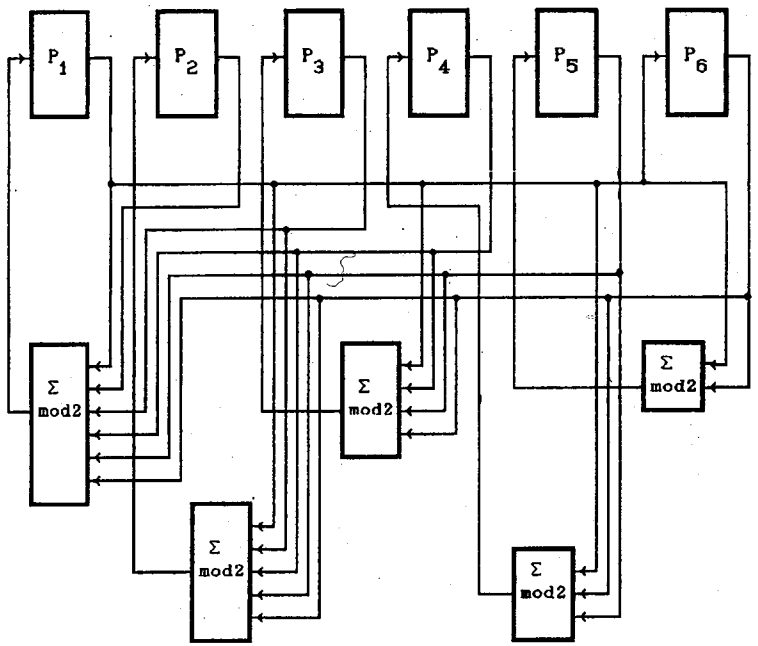
Niech będzie dany rejestr liniowy o długości $m = 6$ określony macierzą $[\mathbf{M}]$:

$$[\mathbf{M}] = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Parametr skoku s niech będzie równy 5.
Przekształcona macierz $[\mathbf{M}_A] = [\mathbf{M}]^s$ posiada postać:

$$[\mathbf{M}_A] = [\mathbf{M}]^s = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 \end{bmatrix}.$$

Wykorzystując macierz $[\mathbf{M}_A]$ zbudowano rejestr akumulacyjny, którego schemat pokazano na rys. 3.



Rys. 3. Przykład realizacji rejestru akumulacyjnego składającego się z 6-ciu przerzutników

Jeżeli wynikowe sekwencje wielowartościowe będą otrzymywane z wyjść dwóch pierwszych przerzutników tworzących rejestry, wówczas dla parametrów $s = 1$ oraz $s = 5$ wygenerowane zostaną następujące ciągi:

a) dla $s = 1$

{2 3 3 3 3 3 1 2 1 2 1 2 3 1 0 2 3 1 2 3 3 1 2 3 1 2 1 0 2 1 0 2 3 3 1 0 0 2 1
2 3 3 3 1 0 2 1 2 1 0 0 2 3 1 0 0 0 2 1 0 0 0 0}

b) dla $s = 5$

{2 3 1 2 3 2 0 0 3 2 0 0 0 3 2 3 1 2 0 3 2 3 2 3 2 0 3 2 0 3 1 1 1 2 0 0 0 0 3
1 2 3 1 1 2 0 3 1 2 0 0 3 1 1 2 3 2 3 1 1 1 1 1}.

Proponowane rozwiązanie nie zmienia przyjętej zasady generacji pseudolosowych sekwencji wielowartościowych pokazanej na rys. 2, jednakże rejestr akumulacyjny generuje liczby należące do ciągu pseudolosowego w każdym takcie zegarowym. Konstrukcja rejestru, w porównaniu z klasycznym rozwiązaniem, wymaga użycia dodatkowych bramek realizujących funkcje EXOR, lecz nie wymaga stosowania rejestru buforowego.

2. WŁASNOŚCI LOSOWE WIELOWARTOŚCIOWYCH SEKWENCJI PSEUDOLOSOWYCH OTRZYMYWANYCH Z REJESTRU AKUMULACYJNEGO

Zanim dokonana zostanie analiza własności losowych generowanych sekwencji wielowartościowych przypomnijmy, jakie parametry statystyczne sekwencji pseudolosowych będą brane pod uwagę. Podobnie jak to przedstawiono w pracach [6, 15] badanie statystycznych sekwencji przeprowadzone zostanie pod kątem sprawdzenia:

a) rozkładu liczb pojawiających się w sekwencji pseudolosowej;

b) przebiegu funkcji autokorelacji określonej dla pełnego okresu ciągu wielowartościowego;

c) rozkładu serii liczb.

Prawdopodobieństwo pojawienia się liczb ze zbioru $\{0,1,..,(2^k - 1)\}$, rozumiane tak, jak to podano w pracy [15], jest następujące:

liczba 0 pojawia się z prawdopodobieństwem

$$P_{(0)} = \frac{2^{m-k} - 1}{2^m - 1} \tag{3}$$

natomiast pozostałe liczby pojwiają się z prawdopodobieństwem

$$P_{(i)} = \frac{2^{m-k}}{2^m - 1}. \tag{4}$$

Wyznaczenie wartości funkcji autokorelacji generowanych sekwencji MPS dokonamy posługując się metodą podaną w pracy [15]. W tym celu przypomnimy ogólną zależność opisującą funkcję autokorelacji sekwencji wielowartościowej:

$$R_x(n) = \sum_{i=0}^{k-1} \sum_{j=0}^{k-1} 2^{i+j} R'_{ij(n)} - E(X)^2 . \tag{5}$$

Wartość $R'_{ij(n)}$ może być wyrażona w następujący sposób:

$$R'_{ij(n)} = \begin{cases} -1/L & \text{dla } n \neq 0 \text{ MOD(L)} \\ 1 & \text{dla } n = 0 \text{ MOD(L)} \end{cases} \tag{6.a}$$

$$R'_{ij(n)} = \begin{cases} -1/L & \text{dla } (ns \pm \delta_{ij}) \neq 0 \text{ MOD(L)} \\ 1 & \text{dla } (ns \pm \delta_{ij}) = 0 \text{ MOD(L)}, \end{cases} \tag{6b}$$

gdzie i oraz j oznaczają numery porządkowe przerzutników rejestru liniowego, z których uzyskiwane są sekwencje binarne, zaś δ_{ij} różnicę pomiędzy numerami tych przerzutników.

Wykorzystując zależności (5), (6a) i (6b) wartości funkcji autokorelacji sekwencji wielowartościowych można obliczyć korzystając z podanych poniżej zależności.

Dla $n = 0$:

$$R_x(0) = \sum_{i=0}^{k-1} \sum_{j=0}^{k-1} 2^{i+1} R'_{ij(n)} - E(X)^2 = \sum_{i=0}^{k-1} 2^{2i} - 1/L \sum_{i=0}^{k-1} \sum_{\substack{j=0 \\ j\neq i}}^{k-1} 2^{i+1} - E(X) =$$

$$= \sum_{i=0}^{k-1} 2^{2i} - 1/L \left(\sum_{i=0}^{k-1} \sum_{j=0}^{k-1} 2^{i+j} + \sum_{i=0}^{k-1} 2^{2i} \right) - E(X)^2 =$$

$$= [3(2^m - 1)]^{(-1)} [2^m(2^{2k} - 1) - 3(2^k - 1)^2] - E(X)^2 \tag{7}$$

dla $n \neq 0$ oraz $(ns \pm \delta_{ij}) \neq 0$ MOD(L), $\quad i, j = 0, 1, \ldots, k-1$:

$$R_x(n) = -1/L \sum_{i=0}^{k-1} \sum_{j=0}^{k-1} 2^{i+j} - E(X)^2 = -1/L\,(2^k - 1)^2 - E(X)^2 \tag{8}$$

dla $n \neq 0$ oraz $(ns \pm \delta_{ij}) = 0$ MOD(L) (punkty charakterystyczne):

$$R_x(n) = \sum_{\substack{i,\, j \in K \\ i+j=\delta_{ij}}} 2^{i+j} - 1/L\,(2^k - 1)^2 + 1/L \sum_{\substack{i,\, j \in K \\ i+j=\delta_{ij}}} 2^{i+j} - E(X)^2 =$$

$$= (2^m - 1)^{(-1)} [2^m \sum_{\substack{i,\, j \in K \\ i+j=\delta_{ij}}} 2^{i+j} - (2^k - 1)^2] - E(X)^2, \tag{9}$$

gdzie K jest zbiorem indeksów wszystkich tych przerzutników rejestru liniowego, z których wyjść pobierane są sekwencje binarne.

Jak łatwo się przekonać, wyrażenia (7), (8) oraz (9) mają tą samą postać jak formuły przedstawione w pracy [15], określające wartości funkcji autokorelacji sekwencji MPS otrzymywanych z k równolegle pracujących jednakowych rejestrów liniowych. Dla poparcia słuszności dotychczasowych rozważań podany zostanie przykład.

Przykład 2.

Niech będzie dany rejestr liniowy opisany wielomianem charakterystycznym:

$$f(x) = x^6 \oplus x^5 \oplus 1$$

Długość generowanej sekwencji wynosi: $L = 63$
Liczba ciągów składowych: $k = 3$
Przesunięcie wzajemne binarnych ciągów składowych:

$$\delta_{12} = 1;\ \delta_{13} = 2.$$

Wartości funkcji autokorelacji wynoszą:

$n = 0$	$R_x(0) = 1295/63$	
$n \neq 0,\ (n \pm \delta_{ij}) = 0$ MOD(L)	$R_x(-1/s) = R_x(1/s) = 591/63$	(*)
	$R_x(-2/s) = R_x(2/s) = 207/63$	(**)
$n \neq 0,\ (n \pm \delta_{ij}) \neq 0$ MOD(L)	$R_x(n) = -49/63$	

Wartości znormalizowanej funkcji autokorelacji wynoszą:

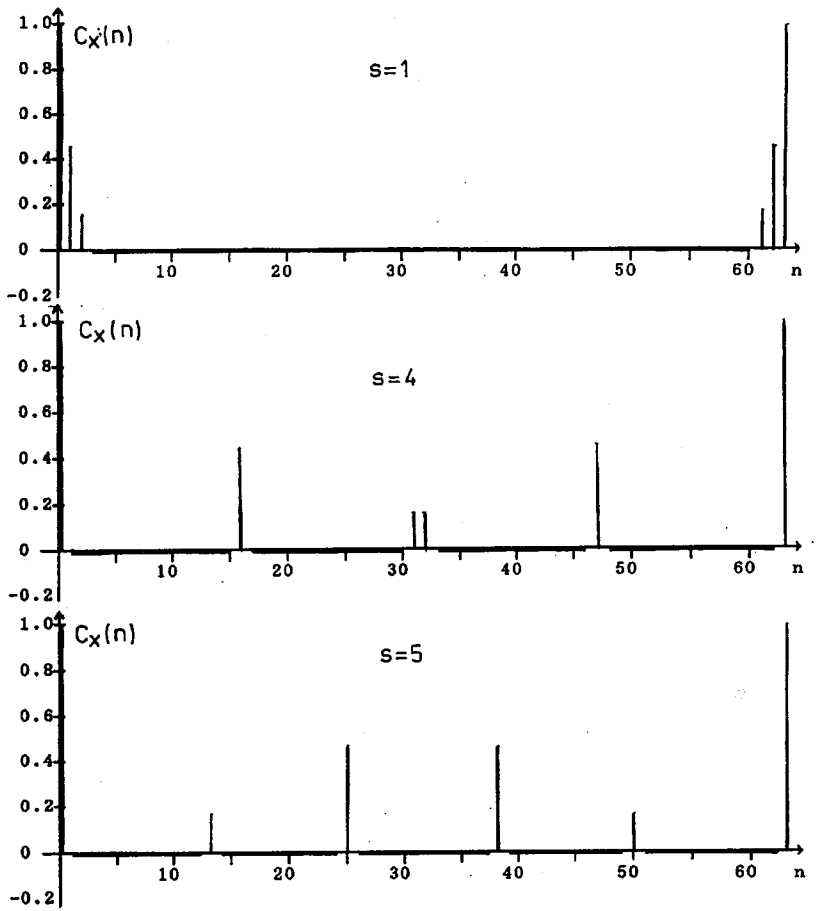
$n = 0$	$C_x(0) = 1$	
$n \neq 0,\ (n \pm \delta_{ij}) = 0$ MOD(L)	$C_x(-1/s) = C_x(1/s) = 0.45$	(*)
	$C_x(-2/s) = C_x(2/s) = 0.16$	(**)
$n \neq 0,\ (n \pm \delta_{ij}) \neq 0$ MOD(L)	$C_x(n) = -0.03$	

Tabela 1.

	Parametr skoku		
	$s=1$	$s=4$	$s=5$
0	$n=0$	$n=0$	$n=0$
$ns-1=kL$	$n=1$	$n=16$	$n=38$
$ns+1=kL$	$n=62$	$n=47$	$n=25$
$ns-2=kL$	$n=2$	$n=32$	$n=13$
$ns+2=kL$	$n=61$	$n=31$	$n=50$

W tabeli 1 podano wartości przesunięć n (punkty charakterystyczne), w których funkcja autokorelacji przybiera odpowiednio wartości (*) i (**) dla parametrów skoku $s = 1$, 4 oraz 5. Wykres przebiegu znormalizowanej funkcji autokorelacji pokazano na rys. 4.

Jeżeli sekwencja MPS jest uzyskiwana z wyjść kolejnych przerzutników tworzących rejestr liniowy ($\delta_{ij} = 1$ dla $j = i + 1$), wówczas dobór parametru skoku decyduje o przebiegu funkcji autokorelacji. I tak, jeśli wartość parametru skoku $1 \leqslant s \leqslant 2\ (k-1)$, gdzie k – długość wektora binarnego reprezentującego



Rys. 4. Wykresy znormalizowanej funkcji autokorelacji sekwencji psedolosowej otrzymywanych z przykładowego rejestru akumulacyjnego dla parametru skoku $s = 1$, $s = 4$ oraz $s = 5$

liczby ze zbioru $\{0, 1, \ldots, 2^k - 1\}$, to mogą pojawić się takie kolejne punkty n, dla których jest spełniony warunek $(n \pm \delta_{ij}) = 0$ MOD(L). Jeżeli jednak wartość parametru skoku ma wartość $s > 2\ (k-1)$, to w otoczeniu dowolnego punktu charakterystycznego n nie istnieje drugi punkt, dla którego spełniony jest warunek $(n \pm \delta_{ij}) = 0$ MOD(L).

3. ANALIZA SPEŁNIENIA WARUNKU SERYJNOŚCI PRZEZ SEKWENCJE WIELOWARTOŚCIOWE UZYSKIWANE Z REJESTRU AKUMULACYJNEGO

Dla przeprowadzenia analizy sekwencji pseudolosowych generowanych za pomocą rejestru akumulacyjnego pod kątem spełnienia postulatu seryjności takie samo założenie dotyczące liczby serii pojawiających się w sekwencji pseudolosowej, jakie uczyniono w pracy [15] dla generatorów zbudowanych z k równolegle pracujących rejestrów liniowych. Oczywiście, przedstawione we wspomnianej pracy twierdzenie 2, określające konieczny warunek na długość m użytego rejestru liniowego, w zależności od wymiaru k binarnego wektora oraz maksymalnej długości l pojwiających się serii, jest również słuszne dla sekwencji MPS uzyskiwanych z rejestru akumulacyjnego. Zatem, aby możliwy był równomierny rozkład serii liczb długość rejestru liniowego powinna być równa:

$$m = l\,k. \tag{9}$$

Uczyńmy obecnie założenie, że rejestr akumulacyjny zbudowano na bazie rejestru liniowego spełniającego zależność (9). Dla takiego przypadku określony zostanie warunek konieczny na to, by generowana sekwencja wielowartościowa spełniała postulat seryjności. W tym celu konieczne jest wprowadzenie pewnych dodatkowych, przedstawionych poniżej oznaczeń.

Niech będzie dana macierz $[\mathbf{M}]^m$ transformująca zbiór wszystkich wektorów binarnych w ten sam zbiór. Macierz tę nazywać będziemy macierzą generującą rejestru akumulującego. Przez $[\mathbf{M}_m^s]$ **rozumieć będziemy podmacierz macierzy** $[\mathbf{M}]^s$ wymiaru $[k \times m]$, która składa się z jej pierwszych k kolumn, co odpowiada k kolejnym wyjściom przerzutników rejestru akumulacyjnego, z ltórych otrzymywana jest sekwencja MPS. Zatem macierz $[\mathbf{M}]^s$ może być przedstawiona w następujący sposób:

$$[\mathbf{M}]^s = [[\mathbf{M}_k{}^s], [\mathbf{M}_{m-k}{}^s]].$$

Zdefiniujmy dalej macierz $[\mathbf{N}]$ wymiaru $[m \times m]$ składającą się z l podmacierzy $[\mathbf{M}_k{}^{si}]$, $i = 1, 2, \ldots, l$:

$$[\mathbf{N}] = [[\mathbf{M}_k{}^s], [\mathbf{M}_k{}^{2s}], \ldots, [\mathbf{M}_k{}^{ls}]]. \tag{10}$$

Analiza własności macierzy $[\mathbf{N}]$ pozwoli określić warunek konieczny i dostateczny na to, by uzyskane sekwencje wielowartościowe spełniały postulat seryjności.

Twierdzenie 1.

Jeśli macierz $[\mathbf{N}]$ zdefiniowana wyrażeniem (10), jest macierzą nieosobliwą nad ciałem Galois GF(2), wówczas sekwencja MPS generowana przez rejestr akumulujący będzie spełniała postulat seryjności.

Dowód

Zgodnie z wymogami postulatu seryjności, w całej sekwencji wielowartościowej nie istnieje podciąg o długości l złożony z samych zer. Z kolei, aby istniały serie zer

o długościach $r < l$ muszą być spełnione dla pewnego wektora początkowego $\langle X_0 \rangle$ różnego od wektora zerowego następujące warunki:

$$\begin{aligned}
&\langle X_0 \rangle \quad [M]^s = \langle X_0 \rangle \quad [[M_k{}^s], [M_{m-k}{}^s]] = \langle 0, X_1 \rangle \\
&\langle 0, X_1 \rangle \quad [M]^s = \langle 0, X_1 \rangle \quad [[M_k{}^s], [M_{m-k}{}^s]] = \langle 0, X_2 \rangle \\
&\langle 0, X_2 \rangle \quad [M]^s = \langle 0, X_2 \rangle \quad [[M_k{}^s], [M_{m-k}{}^s]] = \langle 0, X_3 \rangle \\
&\langle 0, X_{r-1} \rangle \quad [M]^s = \langle 0, X_{r-1} \rangle \quad [[M_k{}^s], [M_{m-k}{}^s]] = \langle 0, X_r \rangle \\
&\langle 0, X_r \rangle \quad [M]^s = \langle 0, X_r \rangle \quad [[M_k{}^s], [M_{m-k}{}^s]] = \langle \alpha, X_{(r-1)} \rangle
\end{aligned} \tag{11}$$

gdzie $\langle \alpha \rangle = \langle 0 \rangle$, $\langle 0 \rangle$ jest wektorem zerowym złożonym z k zer.

Warunki określone w postaci (11) można przekształcić następująco:

$$\begin{aligned}
\langle X_0 \rangle [M]^s &= \langle 0, X_1 \rangle \\
\langle X_0 \rangle [M]^{2s} &= \langle 0, X_2 \rangle \\
\langle X_0 \rangle [M]^{3s} &= \langle 0, X_3 \rangle \\
\langle X_0 \rangle [M]^{(r-1)s} &= \langle 0, X_r \rangle \\
\langle X_0 \rangle [M]^{rs} &= \langle \alpha, X_{(r+1)} \rangle
\end{aligned} \tag{12}$$

Z kolei zbiór warunków (12) można przedstawić w następujący sposób:

$$\begin{aligned}
\langle X_0 \rangle [M_k{}^s] &= \langle 0 \rangle_{[l \times k]} \\
\langle X_0 \rangle [M_k{}^{2s}] &= \langle 0 \rangle_{[l \times k]} \\
\langle X_0 \rangle [M_k{}^{3s}] &= \langle 0 \rangle_{[l \times k]} \\
\langle X_0 \rangle [M_k{}^{rs}] &= \langle 0 \rangle_{[l \times k]}.
\end{aligned} \tag{13}$$

Ostatecznie zbiór warunków przedstawiony w postaci wyrażenia (13) można zapisać w następującej formie:

$$\langle X_0 \rangle [N] = \langle \langle 0 \rangle_{[l \times kr]}, X_r \rangle \tag{14}$$

gdzie wektor $\langle 0 \rangle_{[l \times k]}$ oznacza wektor zerowy składający się z kr zer.

Wiadomo, że w ciągu wielowartościowym nie istnieje seria złożona z l zer, zatem wyrażenie (14) jest słuszne jedynie dla $r < (l-1)$, czyli nie istnieje żaden wektor wymiaru $l \times m$, różny od wektora zerowego, spełniający zależność:

$$\langle X_0 \rangle [N] = \langle 0 \rangle_{[l \times m]}. \tag{15}$$

Ponieważ jedynym wektorem $\langle X_0 \rangle$, spełniającym równanie (15) jest wektor $\langle 0 \rangle_{[l \times m]}$, a on nie należy do zbioru binarnych wektorów generowanych przez rejestr

liniowy, można zatem stwierdzić, że macierz [N] nie jest macierzą osobliwą nad GF(2), co dowodzi twierdzenia 1.

Obecnie zostanie podany przykład potwierdzający słuszność dotychczasowych rozważań.

Przykład 3.

Niech będzie dany rejestr liniowy opisany wielomianem charakterystycznym:

$$f(x) = x^6 \oplus x \oplus 1,$$

dla ktorego macierz generująca [M] ma następującą postać:

$$[\mathbf{M}] = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Analizowana sekwencja MPS jest uzyskiwana na wyjściu dwóch pierwszych przerzutników rejestru. Zatem $m = 6$, $k = 2$ oraz $l = 3$. Założono, że parametr skoku dla rejestru akumulacyjnego wynosi $s = 4$.

$$[\mathbf{M}]^4 = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix} \quad [\mathbf{M}]^8 = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \end{bmatrix} \quad [\mathbf{M}]^{12} = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

Z macierzy $[\mathbf{M}]^4$, $[\mathbf{M}]^8$, $[\mathbf{M}]^{12}$ budujemy macierz [N] wybierając dwie pierwsze kolumny każdej z wymienionych macierzy, czyli:

$$[\mathbf{N}] = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \end{bmatrix}.$$

Łatwo sprawdzić, że macierz [N] jest macierzą nieosobliwą nad ciałem GF(2). Zatem wygenerowany ciąg pseudolosowy dla parametru $s = 4$ ciąg pseudolosowy spełnia warunek seryjności. Sekwencja ta została podana poniżej.

$$\{X_i\} = 3\,2\,0\,3\,0\,3\,1\,2\,\underline{3\,3\,3}\,2\,3\,1\,\underline{0\,0}\,\underline{1\,1}\,2\,1\,0\,1\,0\,\underline{3\,3}\,\underline{1\,1\,1}\,3$$
$$0\,2\,\underline{0\,0}\,2\,3\,\underline{2\,2}\,0\,2\,\underline{1\,1}\,0\,\underline{2\,2}\,3\,0\,1\,2\,0\,1\,3\,1\,3\,2\,1\,\underline{2\,2\,2}$$
$$1\,\underline{3\,3}\,\underline{0\,0}$$

W ciągu tym występują 3 serie liczb o długości trzy (nie pojawia się seria trzech zer), 9 serii liczb o długości dwa (liczba serii zer jest większa o jeden od serii każdej z pozostalych liczb pojawiających się w otrzymanym ciągu wielowartościowym) oraz 36 serii o długości 1.

Sprawdźmy obecnie, czy postulat seryjności jest spełniony także przez sekwencję MPS uzyskaną z tego samego rejestru akumulacyjnego, dla którego parametr skoku s jest równy 5.

$$[M]^5 = \begin{bmatrix} 1&0&0&0&0&1 \\ 1&1&0&0&0&0 \\ 0&1&1&0&0&0 \\ 0&0&1&1&0&0 \\ 0&0&0&1&1&0 \\ 0&0&0&0&1&0 \end{bmatrix} \quad [M]^{10} = \begin{bmatrix} 1&0&0&0&1&1 \\ 0&1&0&0&0&1 \\ 1&0&1&0&0&0 \\ 0&1&0&1&0&0 \\ 0&0&1&0&1&0 \\ 0&0&0&1&1&0 \end{bmatrix} \quad [M]^{15} = \begin{bmatrix} 1&0&0&1&0&1 \\ 0&1&0&0&1&0 \\ 1&1&1&0&0&1 \\ 1&1&1&1&0&0 \\ 0&1&1&1&1&0 \\ 0&0&1&0&1&0 \end{bmatrix}.$$

Macierz [N] dla tego przypadku jest następująca:

$$[N] = \begin{bmatrix} 1&0&1&0&1&0 \\ 1&1&0&1&1&1 \\ 0&1&1&0&1&1 \\ 0&0&0&1&1&1 \\ 0&0&0&0&0&1 \\ 0&0&0&0&0&0 \end{bmatrix}.$$

Ponieważ macierz [N] posiada jeden wiersz zerowy zatem jest ona macierzą osobliwą nad ciałem GF(2). Stąd, zgodnie z twierdzeniem 1, rejestr akumulacyjny pracujący z parametrem skoku $s = 5$ nie powinien spełniać warunku seryjności, co w pełni potwierdzają wyniki badania sekwencji wielowartościowej uzyskanej z badanego generatora:

$$\{X_i\} = \underline{2\,2\,2\,2}\,3\,1\,3\,1\,\underline{2\,2}\,3\,\underline{0\,0}\,1\,2\,3\,0\,1\,\underline{2\,2}\,3\,1\,2\,3\,\underline{0\,0\,0\,0}$$
$$1\,\underline{2\,2\,2}\,3\,0\,1\,3\,0\,1\,3\,1\,3\,1\,3\,0\,1\,2\,3\,1\,3\,\underline{0\,0\,0}\,1\,3\,\underline{0\,0}$$
$$1\,3\,1\,2\,3\,1\,2.$$

Porównując oba przypadki analizowane w przykładzie należy zwrócić uwagę na fakt, że rozkład liczb w obu generowanych ciągach jest równomierny, również wartości funkcji autokorelacji tych sekwencji są jednakowe, różnią się jedynie miejscami występowania punktów charakterystycznych (gdy $s = 4$ punktami charakterystycznymi są $n = 16$ oraz $i = 47$, zaś gdy $s = 5$ wówczas $n = 25$ oraz $n = 38$).

PODSUMOWANIE

W przedstawionej pracy zaprezentowana została metoda generacji pseudolosowych sekwencji wielowartościowych. Do realizacji generatora sekwencji MPS wykorzystano zmodyfikowany rejestr liniowy nazwany w pracy rejestrem akumulacyjnym. Analiza sekwencji otrzymanych z takiego generatora może okazać się bardzo przydatna w praktyce. Przebieg funkcji autokorelacji sekwencji wielowartościowej jest podobny do przebiegu tej funkcji ciągu MPS otrzymywanego z k jednakowych pracujących równolegle rejestrów liniowych. Wartości funkcji autokorelacji dla punktu $n = 0$ oraz dla tych punktow n, które nie są punktami charakterystycznymi, są takie same w obu przypadkach. Tylko w punktach charakterystycznych (dla których funkcja korelacji wzajemnej binarnych ciągów składowych $R'_{ij}(n)$ jest równa 1) wartości $R_x(n)$ mogą być różne. W przypadku rejestru akumulacyjnego liczba punktów charakterystycznych (poza punktem $n = 0$) jest równa $2(k-1)$. Dla pracujących równolegle jednakowych rejestrów liniowych, liczba ta, w zależności od wybranego wzajemnego przesunięcia składowych ciągów binarnych, może się zmieniać od $2(k-1)$ do $k(k-1)$, przy czym jeżeli występuje większa liczba punktów charakterystycznych to wartości funkcji autokorelacji w tych punktach są mniejsze.

Jak to pokazano w pracy, sekwencje MPS otrzymywane z rejestru akumulacyjnego mogą spełniać również postulat seryjności, należy jednak dokonać wyboru parametru skoku poprzez analizę przekształconych macierzy generujących.

Pomimo, że w pracy skoncentrowano się na sprzętowej realizacji generatora sekwencji wielowartościowej, zdaniem autorów, zaproponowana metoda generacji jest wygodna zarówno w implementacji sprzętowej, jak i programowej przede wszystkim ze względu na prostotę realizacji oraz dobre własności losowe generowanych sekwencji pseudolosowych.

BIBLIOGRAFIA

1. N.P. Cagigal, S. Bracho: *Algorithimic determination of linear feedback in a shift register for pseudorandom binary sequence generation.* IEE Proc., 1986, vol. 133, no. 4, pp. 191–194
2. J.A. Chang: *Generation of 5 level maximal-length sequences.* Electronics Letters, 1966, vol. 2, no. 7, p. 258
3. A.C. Davies: *Properties of waveform obtained by nonrecursive digital filtering of pseudorandom binary sequences.* IEEE Trans. on Comp., 1971, vol. C-20, no. 3, pp. 270–281
4. A. Gill: *Linear sequential circuits* (przekł. z ang.). Nauka, Moskwa 1974
5. K.K. Godfrey: *Three-level m-sequences.* Electronic Letters, 1966, vol. 2, no. 7, pp. 241–243
6. S.W. Golomb: *Shift register sequences.* Holden Day, San Francisco 1967
7. T.G. Lewis, W.H. Payne: *Generalized feedback shift register pseudorandom number algorithm.* J. ACM, 1973, vol. 20, no. 3, pp. 456–468
8. S.N. Lo, Z.C. Tan: *Properties of nonrecursive digital filtered ternary maximal length pseudorandom sequences.* Proc. IEEE, 1977, vol. 65, no. 12, pp. 1719–1720
9. F.J. MacWiliams, N.J. Sloane: *Pseudorandom sequences and arrays.* Proc. IEEE, 1976, vol. 64, no. 12, 1715–1729
10. P.S. Moharir: *Generalized PN sequences.* IEEE Trans. on Inform. Theory, 1977, vol. IT-23, no. 6, pp. 782–784

11. D.R. Morgan: *Autocorrelation function of sequential m-bits word taken from n-bits shift register PN sequence.* IEEE Trans. on Comp., 1980, vol. C-29, no. 5, pp. 408–410.
12. J.H. Pangratz, H. Weinrichter: *Pseudo-random number generator based on binary and quinary maximal length sequences.* IEEE Trans. on Comp., 1979, vol. C-28, no. 9, pp. 637–642
13. G. Wustmann: *Comments on "Autocorrelation function of sequential m-bits word taken from n-bits shift register PN sequence".* IEEE Trans. on Comp., 1981, vol. C-30, no. 3, p. 241
14. G. Wustmann: *Autocorrelation function of filtered p-level maximal length sequences.* IEEE Trans. on Comp., 1982, vol. C-29, no. 1, pp. 75–76
15. A. Zabłudowski, B. Dubalski: *Generacja wielowartościowych sekwencji pseudolosowych nad ciałem Galois* (2^k) *przy wykorzystaniu jednakowych, pracujących równolegle rejestrów liniowych.* Kwartalnik Elektroniki i Telekomunikacji, 1992, t. 38 z. 4
16. A. Zabłudowski, B. Dubalski: *Generacja pseudolosowych ciągów liczbowych.* Rozprawy Elektrotechniczne, 1986, t. 32, z. 2, ss. 377–394
17. A. Zabłudowski, B. Dubalski: *The maximal PN sequences over GF(p).* Microelectron. Reliab., 1987, vol. 27, no. 4, pp. 631–637.

A. ZABŁUDOWSKI, B. DUBALSKI

GENERATION OF MULTILEVEL PSEUDORANDOM SEQUENCES OVER GALOIS FIELD GF(2^k) OBTAINED BY THE USE ACCUMULATIVE REGISTER

Summary

This paper deals with analysis of multilevel pseudorandom sequences (MPS) over GF(2^k) obtained by the use of register operating on the same principle as accumulator is doing.. The proposed method is very cheap in realization as it requires only one register and the generated MPS's possess good random properties. A generator built on the basis of the proposed method generates sequences of random numbers with every pulse of the clock. The analysis of MPS concerns autocorrelation function as well as the series of generated random numbers.

KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJI, 1993, 39, z. 1, ss. 35-54

681.3.05
681.325

Algorytmy predykcji dla koderów ADPCM

PRZEMYSŁAW DYMARSKI, ANDRZEJ CHMIELEWSKI, SŁAWOMIR KULA

Instytut Telekomunikacji Politechniki Warszawskiej

EWA ŚWIERCZ

Instytut Telekomunikacji, Politechnika Białostocka

Otrzymano 1992.05.27

Autoryzowano do druku 1992.09.04

Przeprowadzono badania porównawcze algorytmów liniowej predykcji pod kątem zastosowań w układach DPCM o szybkościach transmisji 16—32 kbit/s. Badano predyktory o strukturach transwersalnej i kratowej, oraz sekwencyjne algorytmy ich adaptacji: gradientu stochastycznego (SG) i najmniejszych kwadratów (LS). Przebadano wiele wariantów algorytmów adaptacji (np. SG z normalizacją i bez normalizacji, LS z oknem eksponencjalnym i oknem ruchomym). Wyznaczono optymalne wartości pewnych parametrów (szybkości zbieżności, liczby współczynników predykcji, długości okna). Zaproponowano metodę stabilizacji algorytmu LS z oknem ruchomym. Wskazano na metodę LS z oknem eksponencjalnym jako na najlepsze rozwiązanie problemu predykcji w modulacji DPCM.

1. WSTĘP

Sygnał mowy znajduje coraz szersze zastosowanie w wielu zagadnieniach telekomunikacji i informatyki (transmisja mowy kanałami cyfrowymi z szybkościami 1.2 – 64 kbit/s, cyfrowe systemy automatycznego udzielania informacji np. bankowej, kolejowej, lotniczej, akustyczne urządzenia wyjściowe komputerów). Konieczne staje się w związku z tym przeprowadzenie badań, które stanowiłyby podstawę do realizacji technicznej odpowiednich urządzeń cyfrowego przetwarzania mowy z szybkością mniejszą niż 32 kbit/s (zalecenia CCITT obejmują szybkości transmisji 64 i 32 kbit/s). Celem tych badań jest oszczędniejsze wykorzystanie kanałów transmisyjnych.

Niemal wszystkie stosowane obecnie metody kodowania sygnału mowy wywodzą się z zasady liniowej predykcji. Należy tu wymienić adaptacyjne kodery

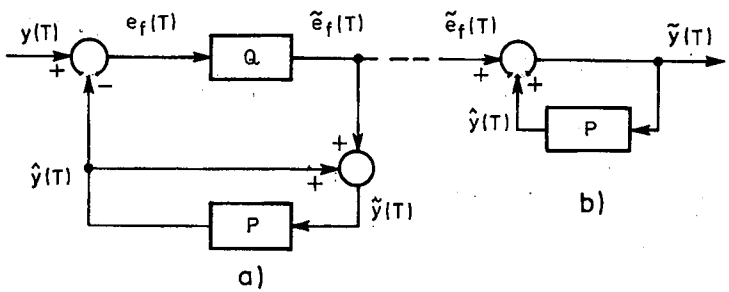
różnicowe (ADPCM) dla przepływności binarnych 16 – 32 kbit/s [2], kodery predykcyjne o pobudzeniu impulsowym (MP) dla przepływności binarnych 8 – 16 kbit/s [3], kodery predykcyjne o pobudzeniu stochastycznym (CELP) dla przepływności binarnych 4.8 – 12 kbit/s [4] i wokodery predykcyjne dla przepływności binarnych rzędu 2 kbit/s [1].

Przedmiotem pracy są algorytmy predykcji dla adaptacyjnych koderów różnicowych (ADPCM) o przepływnościach binarnych 16 – 32 kbit/s. Algorytmy liniowej predykcji można podzielić na blokowe i rekursywne (sekwencyjne). Metody blokowe są rzadko stosowane w koderach ADPCM. Ich omówienie można znaleźć w [7]. Przegląd algorytmów blokowych, gradientowych i rekursywnych LS z uwzględnieniem licznych wariantów wynikających np. z różnych kryteriów optymalizacji filtru predykcyjnego można znaleźć w [17].

W pracy dokonano przeglądu algorytmów rekursywnych liniowej predykcji, w tym algorytmów gradientu stochastycznego i najmniejszych kwadratów w wersjach autokorelacyjnej i kowariancyjnej. Uwzględniono dwie struktury predyktora: transwersalną i kratową. Opracowano programy symulacyjne dla wyminionych wyżej algorytmów liniowej predykcji i wykonano badania na 3 frazach sygnału mowy (w sumie ok. 10s materiału dźwiękowego). Przebadano wiele wariantów omawianych algorytmów (np. z normalizacją i bez normalizacji) i przeanalizowano ich właściwości w zależności od stałych określających szybkości zbieżności, rzędu filtru predykcyjnego, stałej czasowej, długości okna. Przeprowadzono także badania symulacyjne całego układu ADPCM, składającego się z adaptacyjnego predyktora i adaptacyjnego kwantyzera, dla szybkości transmisji 16, 24 i 32 kbit/s (zalecenia CCITT obejmują jedynie układ ADPCM o szybkości transmisji 32 kbit/s [4]). Projektowanie adaptacyjnych kwantyzerów nie jest przedmiotem rozważań niniejszej pracy — zostało ono przedstawione w [6]. Układ ADPCM o szybkości transmisji 24 kbit/s został zrealizowany w czasie rzeczywistym na procesorze sygnałowym TMS32010 [8].

2. PREDYKTOR W UKŁADZIE ADPCM

Układ adaptacyjnego kodera różnicowego pokazano na rys. 1.



Rys. 1. Układ ADPCM: koder (a) i dekoder (b), Q — kwantyzer, P — predyktor

Kwantowaniu podlegają próbki sygnału różnicowego $e_f(T)$, powstającego w wyniku odjęcia sygnału aproksymującego (sygnału predykcji) $\hat{y}(T)$ od sygnału mowy $y(T)$. Poza szczególnymi przypadkami (modulacja delta) próbkowanie przebiega z częstotliwością 8000 próbek/s. Sygnał aproksymujący $\hat{y}(T)$ powstaje w predyktorze P, w wyniku cyfrowego przetwarzania poprzednich próbek sygnału wyjściowego $\tilde{y}(T-1),\ldots, \tilde{y}(T-n)$. Przy braku przekłamań transmisji w koderze i dekoderze generowany jest ten sam sygnał predykcji $\hat{y}(T)$ i spełnione jest wówczas równanie:

$$\tilde{y}(T)-y(T)=(\tilde{e}_f(T)+\hat{y}(T))-(e_f(T)+\hat{y}(T))=\tilde{e}_f(T)-e_f(T)=e(T) \qquad (2.1)$$

Oznacza to, że sygnał wyjściowy dekodera różni się od sygnału wejściowego kodera jedynie o błąd kwantyzacji $e(T)$.

Miarą jakości sygnału wyjściowego $\tilde{y}(T) = y(T) + e(T)$ jest stosunek mocy składowej użytecznej $y(T)$ do mocy zniekształceń $e(T)$

$$\mathrm{SNR} = \frac{E[y^2(T)]}{E[e^2(T)]} = \frac{E[y^2(T)]}{E[e_f^2(T)]} \cdot \frac{E[e_f^2(T)]}{E[e^2(T)]} = G_p\,\mathrm{SNR}_q \qquad (2.2)$$

gdzie $G_p = \dfrac{E[y^2(T)]}{E[e_f^2(T)]}$ jest zyskiem predykcji równym stosunkowi mocy sygnału mowy do mocy sygnału różnicowego, czyli błędu predykcji, zaś $\mathrm{SNR}_q = \dfrac{E[e_f^2(T)]}{E[e^2(T)]}$ – stosunkiem mocy sygnału kwantowanego do mocy szumu kwantowania. Dla uzyskania możliwie najlepszej jakości sygnału wyjściowego, predyktor projektuje się tak, aby uzyskać maksimum zysku predykcji G_p, a kwantyzer – aby uzyskać maksimum stosunku SNR_q.

Zysk predykcji oraz jakość SNR sygnału wyjściowego określa się często w mierze segmentowej, jako średnią arytmetyczną wartości wyznaczonych w mierze logarytmicznej (dB) w segmentach o długości 10 – 30 ms.

Zadaniem kwantyzera jest zaokrąglenie wartości próbki sygnału wejściowego $e_f(T)$ do jednej z N wartości (poziomów kwantyzacji), co umożliwia zakodowanie próbki $\tilde{e}_f(T)$ w słowie binarnym o długości $B = \log_2 N$ bitów.

Aby zapewnić prawidłową pracę kwantyzera dla sygnałów o różnej mocy, stosuje się adaptację kwantyzera, uzależniając wartości poziomów kwantyzacji od czasu. Adaptacja jest niezbędna w kodowaniu mowy, gdy liczba poziomów kwantyzacji jest mała (N = 4 – 16).

Istnieją metody adaptacji kwantyzera „w przód" i „wstecz". Adaptacja „w przód" polega na określeniu optymalnych parametrów kwantyzera dla pewnego fragmentu sygnału wejściowego, a następnie skwantowaniu tego fragmentu sygnału. Tego typu adaptacją nie będziemy się zajmować, gdyż wymaga ona transmisji parametrów kwantyzera do odbiornika, a ponadto nie może być stosowana w układach DPCM, w których sygnał wejściowy kwantyzera nie może być z góry określony, ze względu na zamkniętą pętlę sprzężenia zwrotnego (rys. 1).

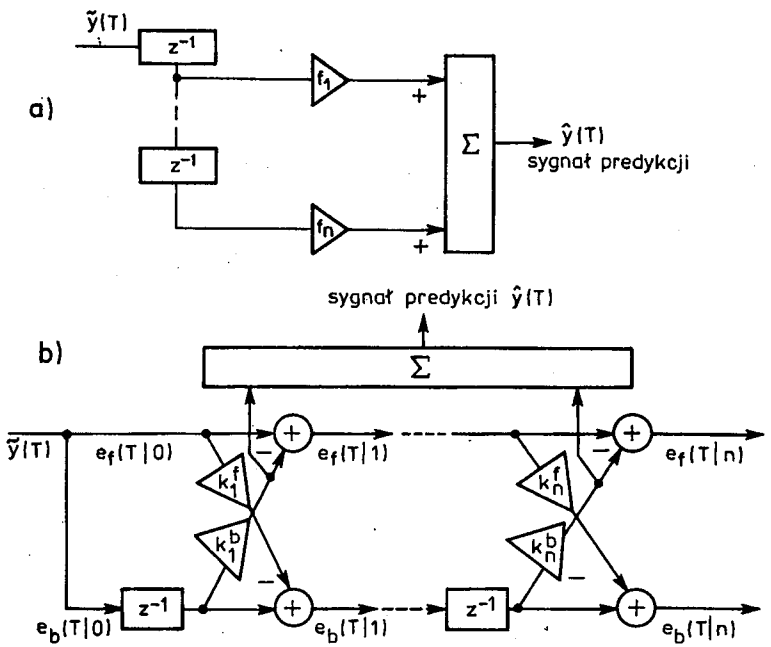
Adaptacja „wstecz" polega na korekcji parametrów kwantyzera jedynie w oparciu o wartości transmitowanych próbek $\tilde{e}_f(T)$. Większość stosowanych algorytmów adaptacji „wstecz" wywodzi się z koncepcji Jayanta [2] opracowanej dla kwantyzera równomiernego. W algorytmie Jayanta parametr Δ, określający odległości poziomów kwantyzacji, zmienia się z częstotliwością próbkowania zgodnie ze wzorem:

$$\Delta(T+1) = \Delta(T)\, M[I(T)] \tag{2.3}$$

gdzie T — numer próbki, I(T) — numer poziomu kwantowania wykorzystywany w chwili T, M(1)...M(N) — współczynniki adaptacji. Siatka poziomów kwantyzacji „rozciąga się" lub „kurczy" w zależności od wykorzystywanych poziomów kwantowania (niskie poziomy kwantowania implikują kurczenie się siatki, wysokie poziomy — jej rozciąganie.

Zagadnienie projektowania kwantyzera, tzn. wybór poziomów kwantyzacji i współczynników adaptacji M(I), jest przedstawione w pracy [6]. Zgodnie z opisanymi tam zasadami zaprojektowano kwantyzery o N = 4, 8 i 16 poziomach kwantyzacji, dla szybkości transmisji 16, 24 i 32 kbit/s. Kwantyzery te zostały zastosowane w badaniach symulacyjnych, które będą opisane w kolejnych rozdziałach pracy.

W układach ADPCM najczęściej stosowane są predyktory AR [5] o strukturze transwersalnej i kratowej (rys. 2). Innymi predyktorami (np. MA, ARMA) nie będziemy się tutaj zajmować.



Rys. 2. Predyktor liniowy AR rzędu *n* w strukturze transwersalnej (a) i kratowej (b)

Predyktor tranwersalny opisany jest wektorem współczynników predykcji $\mathbf{f} = [f_1, ..., f_n]^t$ (t oznacza transpozycję), natomiast predyktor kratowy współczynnikami odbicia „w przód" $k_1^f, ..., k_n^f$ i „wstecz" $k_1^b, ..., k_n^b$. W przypadku predyktorów o stałych współczynnikach można łatwo zapewnić równoważność obu struktur predyktora. Wystarczy spełnić równania wynikające z algorytmu Levinsona [13], podstawiając $k_i^f = k_i^b$, $i = 1, ..., n$. W przypadku predyktorów adaptacyjnych, których współczynniki są funkcjami czasu T, zapewnienie równoważności obu struktur jest sprawą bardziej skomplikowaną i jest cechą tylko nielicznych algorytmów adaptacji (np. niektórych wariantów metody najmniejszych kwadratów).

Algorytmy adaptacji predyktora można podzielić na blokowe i rekursywne (sekwencyjne). Metody blokowe utrzymują stałe wartości współczynników predykcji w obrębie ustalonego odcinka czasu – skokowe zmiany ich wartości następują co 10–20 ms [13]. Jak już wspomnieliśmy, algorytmy te są rzadko stosowane w koderach ADPCM, gdyż wymagają transmitowania współczynników predykcji do odbiornika.

Metody rekursywne adaptacji predyktora polegają na korekcji wartości współczynników predykcji w każdej chwili czasowej T. Algorytm adaptacji predyktora tranwersalnego wykorzystuje zatem rekursję:

$$\mathbf{f}(T + 1) = R(\mathbf{f}(T)] \tag{2.4}$$

gdzie $\mathbf{f}(T) = [f_1(T), ..., f_n(T)]^t$ – wektor współczynników predykcji w chwili T, oraz R – pewna funkcja. Funkcja R uwzględnia wartości sygnału wejściowego predyktora $\tilde{y}(t)$ dla $t \leqslant T$ oraz kryterium adaptacji predyktora. Podobna rekursja wykorzystywana jest do adaptacji predyktora kratowego, gdzie wyznaczane są wartości współczynników odbicia $k_1^f(T + 1), ..., k_n^f(T + 1)$ i $k_1^b(T + 1), ..., k_n^b(T + 1)$.

Metody sekwencyjne umożliwiają wyznaczenie identycznych współczynników predykcji w koderze i dekoderze ADPCM (przy braku przekłamań w transmisji), bez konieczności transmitowania dodatkowych parametrów. Przesyła się jedynie próbki skwantowanego sygnału błędu predykcji $\tilde{e}_f(T)$. Należy zwrócić uwagę na fakt, że predyktor w układzie ADPCM występuje w pętli sprzężenia zwrotnego, zarówno w koderze jak i w dekoderze. Powstaje problem kontroli stabliności takiego układu ze sprzężeniem zwrotnym. Utrzymanie stabilności jest łatwiejsze w przypadku predyktora kratowego. Jak wiadomo [1] jest ona zapewniona, gdy spełniony jest warunek:

$$|k_m^f| < 1, |k_m^b| < 1, \quad m = 1, ..., n. \tag{2.5}$$

Odpowiednie warunki stabilności dla układu z predyktorem transwersalnym nie dadzą się sformułować w tak prosty sposób.

Innym problemem jest odporność układu ADPCM na błędy w transmisji. Dotyczy to algorytmów adaptacji kwantyzera i predyktora. Wskutek przekłamania transmitowanej próbki $e_f(T)$ w odbiorniku generowane są fałszywe wartości poziomów kwantyzacji i współczynników predykcji. Wpływa to ujemnie na dalszą pracę

układu ADPCM. Dobre algorytmy adaptacji kwantyzera i predyktora charakteryzują się tym, że powstałe w ten sposób zakłócenie ma możliwie krótki czas propagacji.

3. ADAPTACJA PREDYKTORA METODAMI GRADIENTU STOCHASTYCZNEGO

3.1. PREDYKTORY O STRUKTURZE TRANSWERSALNEJ

Optymalny predyktor winien maksymalizować zysk predykcji G_p, a co za tym idzie, minimalizować wartość średniokwadratową (moc) błędu predykcji $E[e_f^2(\mathrm{T})]$. Błąd predykcji w chwili T można zapisać w postaci:

$$e_f(\mathrm{T}) = y(\mathrm{T}) - \hat{y}(\mathrm{T}) = y(\mathrm{T}) - \mathbf{f}^t(\mathrm{T})\,\mathbf{y}(\mathrm{T}-1), \tag{3.1}$$

gdzie $\mathbf{y}(\mathrm{T}-1) = [\tilde{y}(\mathrm{T}), \ldots, \tilde{y}(\mathrm{T}-n)]^t$ – wektor próbek sygnału wejściowego predyktora.

Wartości $y(\mathrm{T})$ i $e_f(\mathrm{T})$ nie mogą być wykorzystane do wyznaczania współczynników predykcji, gdyż nie są znane po stronie odbiorczej. W związku z tym zastąpimy je skwantowanymi odpowiednikami $\tilde{y}(\mathrm{T})$ i $\tilde{e}_f(\mathrm{T})$. Wielkości te związane są zależnością analogiczną do zależności (3.1):

$$\tilde{e}_f(\mathrm{T}) = \tilde{y}(\mathrm{T}) - \hat{y}(\mathrm{T}) = \tilde{y}(\mathrm{T}) - \mathbf{f}^t(\mathrm{T})\,\mathbf{y}(\mathrm{T}-1). \tag{3.2}$$

Sygnał predykcji $\hat{y}(\mathrm{T})$ został wyznaczony przy użyciu współczynników predykcji $\mathbf{f}(\mathrm{T})$, przygotowanych już w poprzedniej chwili $\mathrm{T}-1$. W chwili T przygotowuje się współczynniki dla następnej chwili, dodając poprawkę $\Delta\mathbf{f}(\mathrm{T})$:

$$\mathbf{f}(\mathrm{T}+1) = \mathbf{f}(\mathrm{T}) + \Delta\mathbf{f}(\mathrm{T}). \tag{3.3}$$

Poprawka jest skierowana przeciwnie do kierunku wzrostu minimalizowanej funkcji, czyli mocy skwantowanego błędu predykcji $\tilde{e}_f^2(\mathrm{T})$. Kierunek wzrostu wskazywany jest przez gradient tej funkcji. Wynika stąd następująca wartość poprawki:

$$\begin{aligned}\Delta\mathbf{f}(\mathrm{T}) &= -(\beta/2)\frac{\partial}{\partial\mathbf{f}}\tilde{e}_f^2(\mathrm{T}) = -(\beta/2)\,2\tilde{e}_f(\mathrm{T})\frac{\partial}{\partial\mathbf{f}}\tilde{e}_f(\mathrm{T}) \\ &= -\beta\,\tilde{e}_f(\mathrm{T})\frac{\partial}{\partial\mathbf{f}}[\tilde{y}(\mathrm{T}) - \mathbf{f}^t(\mathrm{T})\,\mathbf{y}(\mathrm{T}-1)] = \beta\,\tilde{e}_f(\mathrm{T})\,\mathbf{y}(\mathrm{T}-1).\end{aligned} \tag{3.4}$$

Rekursja

$$\mathbf{f}(\mathrm{T}+1) = \mathbf{f}(\mathrm{T}) + \beta\,\tilde{e}_f(\mathrm{T})\,\mathbf{y}(\mathrm{T}-1) \tag{3.5}$$

opisuje *metodę gradientu stochastycznego* wyznaczania współczynników predykcji [5].

Zbieżność wektora (3.5) jest zagwarantowana dla dostatecznie małego β. Parametr β reguluje jednocześnie szybkość zbieżności. Zbyt mała jego wartość oznacza małą wielkość poprawki $\beta\ \tilde{e}_f(T)\ \mathbf{y}(T-1)$ i wolną zbieżność wektora współczynników. Zbyt duża wartość β prowadzi do rozbieżności wektora współczynników predykcji.

Szybkość zbieżności zależy także od poziomu przetwarzanego sygnału $\{y(T)\}$. Wielkość poprawki zależy od kwadratu amplitudy sygnału, gdyż amplituda wpływa zarówno na wektor $\mathbf{y}(T-1)$ jak i na błąd predykcji $\tilde{e}_f(T)$. Jest to zjawisko niekorzystne, gdyż szybkość zbieżności nie powinna zależeć od poziomu sygnału. Można wyeliminować to zjawisko, uzależniając parametr β od mocy sygnału:

$$\beta(T) = \frac{L_\beta}{\sigma^2(T) + M_\beta}. \tag{3.6}$$

Stała L_β reguluje szybkość zbieżności, stała M_β zapobiega dzieleniu przez zero w przypadku braku sygnału na wejściu predyktora, natomiast $\sigma^2(T)$ jest estymatorem mocy sygnału w chwili T, który obliczać będziemy zgodnie z rekursją:

$$\sigma^2(T) = \alpha\ \sigma^2(T-1) + (1-\alpha)\ \tilde{y}^2(T). \tag{3.7}$$

Wielkość α określa szybkość reakcji estymatora na zmiany mocy chwilowej sygnału. Dla mowy spróbkowanej z częstotliwością 8 kHz przyjmuje się $\alpha = 0.9 - 0.99$. Otrzymujemy w ten sposób *algorytm gradientu stochastycznego z normalizacją.*

W systemach ADPCM ważna jest odporność algorytmu adaptacji predyktora na błędy transmisji. Pojedynczy błąd transmisji (tzn. fałyszywa wartość $\tilde{e}_f(T)$ po stronie odbiorczej) spowoduje rozbieżność ciągu wektorów $\{\mathbf{f}(T)\}$ obliczanych w nadajniku i w odbiorniku. Aby tego uniknąć, stosuje się *algorytm gradientu stochastycznego z tłumieniem:*

$$\mathbf{f}(T+1) = (1-\mu)\ \mathbf{f}(T) + \beta\ \tilde{e}_f(T)\ \mathbf{y}(T-1). \tag{3.8}$$

Parametr $\mu \ll 1$ decyduje o szybkości „tłumienia" pojawiających się błędów transmisji. W przypadku braku sygnału na wejściu predyktora wyznaczone współczynniki predykcji dążą do zera – w przypadku algorytmu gradientu stochastycznego (3.5) bez tłumienia współczynniki te pozostałyby nie zmienione.

3.2. PREDYKTORY O STRUKTURZE KRATOWEJ

Predyktor kratowy pokazano na rys. 2b. Równania m-tego ogniwa kraty mają następującą postać:

$$\begin{cases} e_f(T|m) = e_f(T|m-1) - k_m^b\ e_b(T-1|m-1) \\ e_b(T|m) = e_b(T-1)|m-1) - k_m^f\ e_f(T|\ m-1). \end{cases} \tag{3.9}$$

Zaletą struktury kratowej jest to, że optymalizację współczynników k_m^b i k_m^f można przeprowadzić niezależnie dla każdego ogniwa kraty. Współczynnik k_m^b wyznacza się

na minimum mocy błędu predykcji $e_f^2(\mathrm{T}|m)$. Do poprzedniej wartości $k_m^b(\mathrm{T})$ dodaje się poprawkę skierowaną przeciwnie do gradientu (w tym wypadku pochodnej) funkcji $e_f^2(\mathrm{T}|m)$. W chwili T gradient ma postać:

$$\frac{\partial}{\partial k_m^b}\Big|_{k_m^b(\mathrm{T})} e_f^2(\mathrm{T}|m) = 2e_f(\mathrm{T}|m)\,[-e_b(\mathrm{T}-1|m-1)] = \\ = -2e_f(\mathrm{T}|m-1)\,e_b(\mathrm{T}-1|m-1) + 2k_m^b(\mathrm{T})\,e_b^2(\mathrm{T}-1|m-1). \tag{3.10}$$

Iteracje wykonuje się według wzoru:

$$k_m^b(\mathrm{T}+1) = k_m^b(\mathrm{T}) - \frac{\beta}{2}\frac{\partial}{\partial k_m^b}\Big|_{k_m^b(\mathrm{T})} e_f^2(\mathrm{T}|m) = [1-\beta\,e_b^2(\mathrm{T}-1|m-1)]\,k_m^b(\mathrm{T}) + \\ + \beta\,e_f(\mathrm{T}|m-1)\,e_b(\mathrm{T}-1|m-1). \tag{3.11}$$

Podobne wyrażenie otrzymuje się dla współczynnika k_m^f, który wyznacza się minimalizując $e_b^2(\mathrm{T}|m)$ [5]:

$$k_m^f(\mathrm{T}+1) = [1-\beta\,e_f^2(\mathrm{T}|m-1)]\,k_m^f(\mathrm{T}) + \beta\,e_f(\mathrm{T}|m-1)\,e_b(\mathrm{T}-1|m-1). \tag{3.12}$$

Podobnie jak w predyktorze transwersalnym, szybkość adaptacji i zbieżność algorytmu zależy od wyboru parametru β. Aby uniezależnić szybkość adaptacji od poziomu sygnału, stosuje się algorytm gradientu stochastycznego z normalizacją. Parametr β we wzorach 3.11 i 3.12 zależy wówczas od mocy przetwarzanych w m-tym ogniwie kraty sygnałów wejściowych:

$$\beta_m(\mathrm{T}) = \frac{\mathrm{L}_\beta}{\sigma_m^2(\mathrm{T}) + \mathrm{M}_\beta}, \tag{3.13}$$

która jest estymowana zgodnie z rekursjami ($m = 1,\ldots, n$):

$$\sigma_m^2(\mathrm{T}) = \alpha\,\sigma_m^2(\mathrm{T}-1) + (1-\alpha)\,[e_f^2(\mathrm{T}|m-1) + e_b^2(\mathrm{T}-1|m-1)].$$

Stosowany jest także wariant algorytmu adaptacji, w którym zakłada się jednakowe wartości współczynników $k_m = k_m^b = k_m^f$ i oblicza się je w wyniku minimalizacji sumy $e_f^2(\mathrm{T}|m) + e_b^2(\mathrm{T}|m)$. Otrzymujemy wówczas następujący wzór iteracyjny:

$$k_m(\mathrm{T}+1) = \{1-\beta\,[e_b^2(\mathrm{T}-1|m-1) + e_f^2(\mathrm{T}|m-1)]\}\,k_m(\mathrm{T}) + \\ + 2\,\beta\,e_f(\mathrm{T}|m-1)\,e_b(\mathrm{T}-1|m-1). \tag{3.14}$$

Stosowane są również inne warianty algorytmu adaptacji, np. algorytm z tłumieniem (wartości 1 we wzorach (3.11), (3.12) należy zastąpić przez $1-\mu$, $\mu \ll 1$).

W układach ADPCM celowe jest kontrolowanie stabilności układu z predyktorem w pętli sprzężenia zwrotnego. W przypadku predyktora kratowego ogranicza się to do sprawdzenia, czy obliczone wartości k_m^f i k_m^b leżą w przedziale $(-1, +1)$ dla $m = 1, 2,\ldots, n$. Jeżeli ze wzorów (3.11), (3.12) lub (3.14) otrzyma się wartości spoza tego przedziału, należy sprowadzić je do wartości $+1$ lub -1. Inną możliwą reakcją

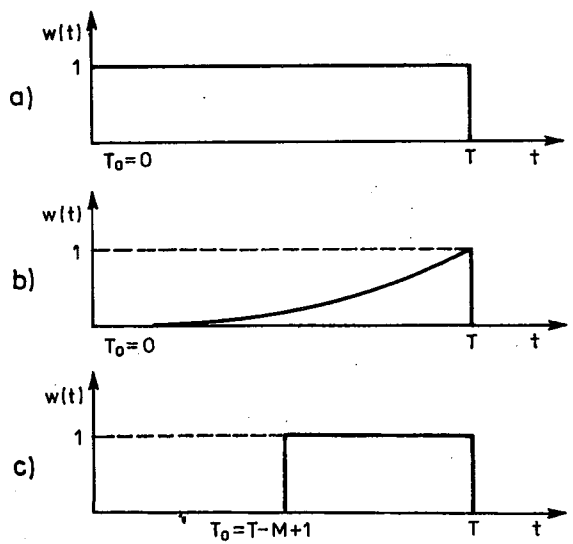
na wykrycie niestabilności jest wyzerowanie współczynników k_m^b i k_m^f w „niestabilnym" ogniwie i we wszystkich dalszych ogniwach kraty.

Należy podkreślić, że struktury transwersalna i kratowa z adaptacją współczynników metodą stochastycznego gradientu nie są sobie równoważne. Dla obu tych struktur należy przeprowadzić osobne badania symulacyjne.

4. ADAPTACJA PREDYKTORA METODĄ NAJMNIEJSZYCH KWADRATÓW

4.1. PODSTAWOWE WARIANTY ALGORYTMU NAJMNIEJSZYCH KWADRATÓW

Metody najmniejszych kwadratów (LS) obejmują algorytmy rekursywne, w których bieżące wartości współczynników predykcji $\mathbf{f}(T)$ wyznaczane są w taki sposób, aby minimalizowały energię błędu predykcji w obrębie określonego okna. Okno rozciąga się od momentu T_0 do bieżącej chwili T. Gdy $T_0 = \text{const}$ mamy do czynienia z oknem początkowym (rys. 3a). Okno takie rozszerza się wraz z upływem czasu i jest najwłaściwsze dla analizy sygnałów stacjonarnych. W przypadku sygnału mowy należy adaptować predyktor do aktualnie wypowiadanej głoski (poprzednie głoski winne być usunięte z okna). Można to osiągnąć wprowadzając okno początkowe z wagą eksponencjalną (rys. 3b) lub okno ruchome o stałej długości M (rys. 3c).



Rys. 3. Okna, w obrębie których minimalizuje się sumę kwadratów błędów predykcji: a) okno początkowe, b) okno początkowe z wagą eksponencjalną, c) okno ruchome o długości M

Algorytmy LS można podzielić na autokorelacyjne i kowariancyjne [5]. W algorytmach autokorelacyjnych dokonuje się minimalizacji energii błędu predykcji w obrębie całego okna $\tau = T_0, T_0 + 1, \ldots, T$:

$$\varepsilon_f(\mathrm{T}|n) = \sum_{\tau=\mathrm{T}_0}^{\mathrm{T}} [\tilde{e}_f(\tau)]^2 = \sum_{\tau=\mathrm{T}_0}^{\mathrm{T}} [\tilde{y}(\tau) - \mathbf{f}^t(\mathrm{T})\,\mathbf{y}(\tau-1)]^2. \tag{4.1}$$

Ze względu na konieczność synchronicznego działania algorytmu w nadajniku i odbiorniku, używa się skwantowanych wartości $\tilde{e}_f(\tau)$ i $\tilde{y}(\tau)$. Stosując okno z wagą eksponencjalną, energia $\varepsilon_f(\mathrm{T}|n)$ przyjmie postać $\sum_{\tau=\mathrm{T}_0}^{\mathrm{T}} \lambda^{\mathrm{T}-\tau} [\tilde{e}_f(\tau)]^2$.

Ponieważ $\mathbf{y}(\tau-1) = [\tilde{y}(\tau-1)\ ,\ldots,\ \tilde{y}(\tau-n)]^t$, do wyliczenia sygnału predykcji $\hat{y}(\tau) = \mathbf{f}^t(\tau)\ \mathbf{y}(\tau-1)$ dla $\tau = \mathrm{T}_0,\ \mathrm{T}_0+1\ ,\ldots,\ \mathrm{T}_0+n-1$, a co za tym idzie, do wyznaczenia energii $\varepsilon_f(\mathrm{T}|n)$, konieczne są próbki sygnału $\tilde{y}$ leżące poza analizowanym oknem ($\tau < \mathrm{T}_0$). Ponieważ próbki te mogą być nieznane, w metodzie autokorelacyjnej przyjmuje się $\tilde{y}(\tau) = 0$, $\tau < \mathrm{T}_0$.

W metodzie kowariancyjnej unika się tego założenia, minimalizując energię błędu predykcji w zawężonym zakresie $\tau = \mathrm{T}_0 + n,\ \mathrm{T}_0 + n + 1\ ,\ldots,\ \mathrm{T}$:

$$\varepsilon_f(\mathrm{T}|n) = \sum_{\tau=\mathrm{T}_0+n}^{\mathrm{T}} [\tilde{e}_f(\tau)]^2 = \sum_{\tau=\mathrm{T}_0+n}^{\mathrm{T}} [\tilde{y}(\tau) - \mathbf{f}^t(\mathrm{T})\,\mathbf{y}(\tau-1)]^2. \tag{4.2}$$

Wyrażenie to nie zawiera wartości $\tilde{y}(\tau)$, $\tau < \mathrm{T}_0$.

Wielkość $\varepsilon_f(\mathrm{T}|n)$ może być interpretowana jako kwadrat odległości euklidesowej dwóch wektorów: wektora $\mathbf{Y}(\mathrm{T})$ próbek skwantowanego sygnału mowy i wektora $\hat{\mathbf{Y}}(\mathrm{T})$ próbek sygnału predykcji. W metodzie autokorelacyjnej wektory mają postać:

$$\mathbf{Y}(\mathrm{T}) = [\hat{y}(\mathrm{T})\,\tilde{y}(\mathrm{T}-1)\ ,\ldots,\ \tilde{y}(\mathrm{T}_0)]^t. \tag{4.3}$$

$$\hat{\mathbf{Y}}(\mathrm{T}) = \begin{bmatrix} \hat{y}(t) \\ \hat{y}(\mathrm{T}-1) \\ \cdot \\ \cdot \\ \cdot \\ \hat{y}(\mathrm{T}_0) \end{bmatrix} = \begin{bmatrix} y^t(\mathrm{T}-1) \\ y^t(\mathrm{T}-2) \\ \cdot \\ \cdot \\ \cdot \\ y^t(\mathrm{T}_0-1) \end{bmatrix} \times$$

$$\times \mathbf{f}(\mathrm{T}) = \begin{bmatrix} \tilde{y}(\mathrm{T}-1) & \tilde{y}(\mathrm{T}-2) & \ldots\ \tilde{y}(\mathrm{T}-n) \\ \tilde{y}(\mathrm{T}-2) & \tilde{y}(\mathrm{T}-3) & \ldots\ \tilde{y}(\mathrm{T}-n-1) \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \tilde{y}(\mathrm{T}_0) & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \mathbf{f}(\mathrm{T}) \tag{4.4}$$

zaś w metodzie kowariancyjnej:

$$\mathbf{Y}(\mathrm{T}) = [\tilde{y}(\mathrm{T}), \ldots, \tilde{y}(\mathrm{T}_0 + n)]^{\mathrm{t}} \tag{4.5}$$

$$\hat{\mathbf{Y}}(\mathrm{T}) = \begin{bmatrix} \hat{y}(\mathrm{T}) \\ \hat{y}(\mathrm{T}-1) \\ \cdot \\ \cdot \\ \cdot \\ \hat{y}(\mathrm{T}_0+n) \end{bmatrix} = \begin{bmatrix} y^{\mathrm{t}}(\mathrm{T}-1) \\ y^{\mathrm{t}}(\mathrm{T}-2) \\ \cdot \\ \cdot \\ \cdot \\ y^{\mathrm{t}}(\mathrm{T}_0+n-1) \end{bmatrix} \times \tag{4.6}$$

$$\times \mathbf{f}(\mathrm{T}) = \begin{bmatrix} \tilde{y}(\mathrm{T}-1) & \tilde{y}(\mathrm{T}-2), \ldots, \tilde{y}(\mathrm{T}-n) \\ \tilde{y}(\mathrm{T}-2) & \tilde{y}(\mathrm{T}-3), \ldots, y(\tilde{\mathrm{T}}-n-1) \\ \cdot & \cdot \quad \cdot \\ \cdot & \cdot \quad \cdot \\ \cdot & \cdot \quad \cdot \\ \tilde{y}(\mathrm{T}_0+n-1) & \tilde{y}(\mathrm{T}_0+n-2), \ldots, \tilde{y}(\mathrm{T}_{\mathrm{T}}) \end{bmatrix} \mathbf{f}(\mathrm{T}_0)$$

W obu przypadkach wektor próbek sygnału predykcji $\hat{\mathbf{Y}}(\mathrm{T})$ jest kombinacją liniową kolumn macierzy $\mathbf{S}(\mathrm{T}|n)$ występującej we wzorach (4.4), (4.6). Jak wiadomo, minimum odległości euklidesowej między $\mathbf{Y}(\mathrm{T})$ i $\hat{\mathbf{Y}}(\mathrm{T})$ zapewnia rzut ortogonalny wektora $\mathbf{Y}(\mathrm{T})$ na przestrzeń rozpiętą na wektorach tworzących kolumny $\mathbf{S}(\mathrm{T}|n)$. Wynika stąd, że optymalnym wektorem współczynników predykcji jest wektor [5]:

$$\mathbf{f}(\mathrm{T}) = [\mathbf{S}^{\mathrm{t}}(\mathrm{T}|n)\ \mathbf{S}(\mathrm{T}|n)]^{-1}\ \mathbf{S}^{\mathrm{t}}(\mathrm{T}|n)\ \mathbf{Y}(\mathrm{T}). \tag{4.7}$$

Analogiczne rozwiązanie otrzymuje się dla okna z wagą eksponencjalną, należy tylko każdą składową wektora $\mathbf{Y}(\mathrm{T})$ i każdy wiersz macierzy $\mathbf{S}(\mathrm{T}|n)$ pomnożyć przez odpowiednią wagę $\lambda^{T-\tau}$ ($\tau = 0, 1, \ldots$).

Obliczenie współczynników $\mathbf{f}(\mathrm{T})$ byłoby znacznie prostsze gdyby kolumny macierzy $\mathbf{S}(\mathrm{T}|n) = [\mathbf{Y}(\mathrm{T}|1), \ldots, \mathbf{Y}(\mathrm{T}|n)]$ były ortogonalne. Wówczas m-ty współczynnik predykcji $f_m(\mathrm{T})$ byłby współczynnikem rzutu ortogonalnego wektora $\mathbf{Y}(\mathrm{T})$ na wektor $\mathbf{Y}(\mathrm{T}|m)$:

$$f_m(\mathrm{T}) = \frac{\mathbf{Y}^{\mathrm{t}}(\mathrm{T}|m)\ \mathbf{Y}(\mathrm{T})}{\mathbf{Y}^{\mathrm{t}}(\mathrm{T}|m)\ \mathbf{Y}(\mathrm{T}|m)}. \tag{4.8}$$

Ponieważ macierz $\mathbf{S}(\mathrm{T}|n)$ z reguły nie jest ortogonalna, można ją zortogonalizować, stosując procedurę Grama–Schmidta. Wektor $\mathbf{Y}(\mathrm{T}|1)$ pozostawia się bez zmiany, natomiast od kolejnych wektorów $\mathbf{Y}(\mathrm{T}|m)$ odejmuje się ich rzut na podprzestrzeń rozpiętą na wektorach $\mathbf{Y}(\mathrm{T}|1)$, $\mathbf{Y}(\mathrm{T}|2)$, ..., $\mathbf{Y}(\mathrm{T}|m-1)$. Otrzymuje się w ten sposób bazę ortogonalną $\mathbf{Y}_{\mathrm{ort}}(\mathrm{T}|1), \ldots, \mathbf{Y}_{\mathrm{ort}}(\mathrm{T}|n)$.

W niektórych przypadkach procedura ortogonalizacyjna może być przeprowadzona dla struktury kratowej jak na rys. 2b [16]. Metoda autokorelacyjna z oknem początkowym i metoda kowariancyjna z oknem ruchomym o stałej długości należą do tego typu przykładów. Sygnały występujące na wyjściu układów opóźniających (z^{-1}) tworzą wówczas bazę ortogonalną $\mathbf{Y}_{ort}(T|1)$,..., $\mathbf{Y}_{ort}(T|n)$. Wartości $k_1^b(T)$,..., $k_n^b(T)$ są współczynnikami rzutu ortogonalnego wektora $\mathbf{Y}(T)$ na wektory tej bazy. Wektor próbek sygnału predykcji $\hat{\mathbf{Y}}(T)$ jest sumą poszczególnych rzutów ortogonalnych. Ma on tę samą wartość jak w przypadku predyktora transwersalnego. Wynika stąd równoważność struktury transwersalnej i kratowej (dla metody autokorelacyjnej z oknem początkowym i metody kowariancyjnej z oknem ruchomym.). Ze względu na tę równoważność rozważać będziemy dalej jedynie strukturę kratową. Wybór tej właśnie struktury można uzasadnić łatwiejszą kontrolą stabilności, o czym była mowa w p. 2.

4.2. SZYBKIE ALGORYTMY REKURSYWNE

Współczynniki opisujące strukturę kratową predyktora mogą być wyznaczone drogą działań na wektorze $\mathbf{Y}(T)$ i kolumnach macierzy $\mathbf{S}(T|n)$. Wymaga to jednak wykonania wielu operacji arytmetycznych w każdej chwili T. Zauważmy jednak, że wektor $\mathbf{Y}(T)$ i macierz $\mathbf{S}(T|n)$ różnią się w niewielkim stopniu od ich odpowiedników $\mathbf{Y}(T-1)$ i $\mathbf{S}(T-1|n)$. Różnica polega na tym, że do wektora $\mathbf{Y}(T-1)$ dopisana jest dodatkowa pierwsza składowa $\tilde{y}(T)$, a do macierzy $\mathbf{S}(T-1|n)$ dodatkowy pierwszy wiersz $[\tilde{y}(T-1), \tilde{y}(T-2)$,.... $\tilde{y}(T-n)]$. Poza tym, w metodzie kowariancyjnej z oknem ruchomym, skasowana jest ostatnia składowa $\tilde{y}(T_0+n-1)$ wektora $\mathbf{Y}(T-1)$ i ostatni wiersz $[\tilde{y}(T_0+n-2), \tilde{y}(T_0+n-3)$,..., $\tilde{y}(T_0-1)]$ macierzy $\mathbf{S}(T-1|n)$. Można zatem oczekiwać, że między współczynnikami predykcji w chwili $T-1$ i T istnieje pewien związek. Wykorzystując ten związek można wyprowadzić tzw. szybki algorytm rekursywny najmniejszych kwadratów (wyprowadzenie można znaleźć w [5]). Wymaga on wykonania w chwili T obliczenia następujących rekursji (dla $m = 0$,..., $n-1$):

$$K_{m+1}(T) = \lambda K_{m+1}(T-1) + e_f(T|m)\, e_b(T-1|m) \frac{1}{1-\gamma\ (T-1|m)} \tag{4.9a}$$

$$k_{m+1}^f(T) = \frac{K_{m+1}(T)}{\varepsilon_f(T|m)}, \qquad k_{m+1}^b(T) = \frac{K_{m+1}(T)}{\varepsilon_b(T-1|m)} \tag{4.9.b.c}$$

$$e_f(T|m+1) = e_f(T|m) - k_{m+1}^b(T)\, e_b(T-1|m) \tag{4.9.d}$$

$$e_b(T|m+1) = e_b(T-1|m) - k_{m+1}^f(T)\, e_f(T|m) \tag{4.9.e}$$

$$\varepsilon_f(T|m+1) = \varepsilon_f(T|m) - k_{m+1}^b(T)\, K_{m+1}(T) \tag{4.9.f}$$

$$\varepsilon_b(T|m+1) = \varepsilon_b(T-1|m) - k_{m+1}^f(T)\, K_{m+1}(T) \tag{4.9.g}$$

$$\gamma(\mathrm{T}|m+1) = \gamma(\mathrm{T}|m) + \frac{e_b^2(\mathrm{T}|m)}{\varepsilon_b(\mathrm{T}|m)} \tag{4.9.h}$$

$$e_f^*(\mathrm{T}|m+1) = e_f^*(\mathrm{T}|m) - k_{m+1}^b(\mathrm{T})\, e_b^*(\mathrm{T}-1|m) \tag{4.9.i}$$

$$e_b^*(\mathrm{T}|m+1) = e_b^*(\mathrm{T}-1|m) - k_{m+1}^f(\mathrm{T})\, e_f^*(\mathrm{T}|m) \tag{4.9.j}$$

$$\gamma^*(\mathrm{T}|m+1) = \gamma^*(\mathrm{T}|m) + \frac{e_b^{*2}(\mathrm{T}|m)}{\varepsilon_b(\mathrm{T}|m)} \tag{4.9.k}$$

$$K_{m+1}(\mathrm{T}) \Leftarrow K_{m+1}(\mathrm{T}) - e_f^*(\mathrm{T}|m)\, e_b^*(\mathrm{T}-1|m)\, \frac{1}{1-\gamma^*(\mathrm{T}-1|m)}. \tag{4.9.l}$$

Pierwsza część wzorów 4.9 (a)–(h) opisuje metodę najmniejszych kwadratów z oknem początkowym ($\lambda \leqslant 1$ jest wagą eksponencjalną). Obie części opisują algorytm z oknem ruchomym (należy wówczas podstawić $\lambda = 1$). Znak $\Leftarrow$ oznacza podstawienie wartości $K_{m+1}(\mathrm{T})$ po zmodyfikowaniu dotychczasowej wartości występującej po prawej stronie równania.

Wyjaśnijmy znaczenie niektórych zmiennych występujących w szybkim algorytmie najmniejszych kwadratów:
Kolejne próbki sygnału $e_b(\mathrm{T}-1|m)$, $e_b(\mathrm{T}-2|m)$,..., tworzą wektor $\mathbf{E}_b(\mathrm{T}|m+1) = \mathbf{Y}_{\mathrm{ort}}(\mathrm{T}|m+1)$ tzn. $m+1$-szy wektor zortogonalizowanej bazy $\mathbf{Y}_{\mathrm{ort}}(\mathrm{T}|1)$,..., $\mathbf{Y}_{\mathrm{ort}}(\mathrm{T}|n)$. Kolejne próbki sygnału $e_f(\mathrm{T}|m+1)$ $e_f(\mathrm{T}-1|m+1)$,... tworzą współrzędne wektora błędu predykcji $\mathbf{E}_f(\mathrm{T}|m+1)$ dla predyktora o $m+1$ współczynnikach. Wektor $\mathbf{E}_f(\mathrm{T}|m+1)$ powstaje przez odjęcie od wektora $\mathbf{Y}(\mathrm{T})$ jego rzutu na podprzestrzeń rozpiętą na wektorach będących $m+1$ pierwszymi kolumnami macierzy $\mathbf{S}(\mathrm{T}|n)$. Rzut ten stanowi wektor próbek sygnału predykcji $\hat{\mathbf{Y}}(\mathrm{T}|m+1)$ uzyskany przy użyciu predyktora o $m+1$ współczynnikach. Wektor ten jest obliczany jako suma rzutów ortogonalnych wektora $\mathbf{Y}(\mathrm{T})$ na kolejne wektory $\mathbf{Y}_{\mathrm{ort}}(\mathrm{T}|1)$,..., $\mathbf{Y}_{\mathrm{ort}}(\mathrm{T}|m+1)$ czyli $\mathbf{E}_b(\mathrm{T}|1)$,..., $\mathbf{E}_b(\mathrm{T}|m+1)$:

$$\hat{\mathbf{Y}}(\mathrm{T}|m+1) = \sum_{i=1}^{m+1} k_i^b(\mathrm{T})\, \mathbf{E}_b(\mathrm{T}|i). \tag{4.10}$$

Pierwsza współrzędna tego wektora jest (dla $m = n-1$) poszukiwaną próbką sygnału predykcji $\hat{y}(\mathrm{T})$:

$$\hat{y}(\mathrm{T}) = \sum_{i=1}^{n} k_i^b(\mathrm{T})\, e_b(\mathrm{T}-1|i-1). \tag{4.11}$$

Współczynnik $k_{m+1}^b(\mathrm{T})$ jest współczynnikiem rzutu wektora $\mathbf{Y}(\mathrm{T})$ na $\mathbf{Y}_{\mathrm{ort}}(\mathrm{T}|m+1) = \mathbf{E}_b(\mathrm{T}|m+1)$

$$k_{m+1}^b(\mathrm{T}) = \frac{\mathbf{E}_b^t(\mathrm{T}|m+1)\ \mathbf{Y}(\mathrm{T})}{\mathbf{E}_b^t(\mathrm{T}|m+1)\ \mathbf{E}_b(\mathrm{T}|m+1)} = \frac{\mathbf{E}_b^t(\mathrm{T}|m+1)\ \mathbf{E}_f(\mathrm{T}|m)}{\mathbf{E}_b^t(\mathrm{T}|m+1)\ \mathbf{E}_b(\mathrm{T}|m+1)}. \tag{4.12}$$

Ostatnia równość wynika z faktu, że wektor $\mathbf{Y}(T)$ można rozłożyć na składniki ortogonalne: $\mathbf{Y}(T) = \hat{\mathbf{Y}}(T|m) + \mathbf{E}_f(T|m)$, gdzie $\hat{\mathbf{Y}}(T|m)$ leży w podprzestrzeni rozpiętej na wektorach $\mathbf{Y}_{ort}(T|1)$,..., $\mathbf{Y}_{ort}(T|m)$, a $\mathbf{E}_f(T|m)$ jest do tej podprzestrzeni ortogonalny. W związku z tym rzut wektora $\mathbf{Y}(T)$ na $\mathbf{Y}_{ort}(T|m+1)$ jest równy rzutowi wektora $\mathbf{E}_f(T|m)$ na $\mathbf{Y}_{ort}(T|m+1)$.

Porównując wzór (4.12) ze wzorem definicyjnym (4.9.c) można zinterpretować $K_{m+1}(T)$ jako iloczyn skalarny $\mathbf{E}_b^t(T|m+1)\,\mathbf{E}_f(T|m)$, oraz $\varepsilon_b(T-1|m)$ jako kwadrat normy $\|\mathbf{E}_b(T|m+1)\|^2 = \mathbf{E}_b^t(T|m+1)\,\mathbf{E}_b(T|m+1)$. Podobnie zmienna $\varepsilon_f(T|m)$ jest interpretowana jako kwadrat normy wektora $\mathbf{E}_f(T|m)$.

Z nierówności Schwarza

$$[\mathbf{E}_b^t(T|m+1)\,\mathbf{E}_f(T|m)]^2 \leqslant \|\mathbf{E}_f(T|m)\|^2\,\|\mathbf{E}_b(T|m+1)\|^2 \tag{4.13}$$

wynika, że

$$K_{m+1}^2(T) \leqslant \varepsilon_f(T|m)\,\varepsilon_b(T-1|m). \tag{4.14}$$

Naruszenie tej nierówności (np. wskutek niedokładności obliczeń) powoduje, że wartości numeryczne norm wektorów występujących we wzorach (4.9.f), (4.9.g) stają się ujemne.

Jak już wspomniano, w chwili T dokonuje się modyfikacji macierzy $\mathbf{S}(T-1|n)$ dopisując do niej nowy pierwszy wiersz i dodatkowo kasując (dla metody kowariancyjnej z oknem ruchomym) ostatni wiersz w taki sposób, aby otrzymać macierz $\mathbf{S}(T|n)$. Współczynnik $\gamma(T-1|m)$ interpretuje się jako kwadrat sinusa kąta pomiędzy podprzestrzeniami rozpiętymi na $m+1$ kolumnach macierzy $\mathbf{S}(T-1|n)$ z dopisanym nowym wierszem i bez tego wiersza. Podobnie $\gamma^*(T-1|m)$ interpretuje się w odniesieniu do usuwanego ostatniego wiersza macierzy $\mathbf{S}(T-1|n)$. Wartości tych współczynników powinny być zawarte w przedziale $<0,1)$.

W algorytmie kowariancyjnym z oknem ruchomym występują ponadto sygnały $e_f^*(T|m)$ i $e_b^*(T-1|m)$ reprezentujące „lewą" krawędź okna (rys. 3c).

W chwili T obliczenia rozpoczyna się wyliczeniem sygnału predykcji $\hat{y}(T)$ według wzoru (4.11). Po otrzymaniu sygnału skwantowanego $\tilde{y}(T)$ (rys. 1) możliwa jest inicjalizacja wartości zmiennych niezbędnych do obliczenia rekursji (4.9). Zmiennym $e_f(T|0)$ i $e_b(T|0)$ nadaje się wartości początkowe równe $\tilde{y}(T)$. Podobnie podstawia się $e_f^*(T|0) = e_b^*(T|0) = \tilde{y}(T_0+n) = \tilde{y}(T-M+n+1)$, gdzie M jest szerokością okna (rys. 3c). Wartości $\varepsilon_b(T|0) = \varepsilon_f(T|0)$ są kwadratem normy wektora $\mathbf{Y}(T)$. Oblicza się je, dodając do kwadratu normy wektora $\mathbf{Y}(T-1)$ wartość $\tilde{y}^2(T)$ i odejmując $\tilde{y}^2(T_0+n-1)$. Dla algorytmu z oknem początkowym i wagą eksponencjalną mnoży się $\|\mathbf{Y}(T-1)\|^2 = e_f(T-1|0)$ przez λ i dodaje się $\tilde{y}^2(T)$. Współczynniki γ i γ^* dla $m = 0$ są zerowe.

Inicjalizacja całego algorytmu polega na zastosowaniu w chwili T ($0 \leqslant T \leqslant n-1$) predyktora o $m = T$ współczynnikach. Dla $T \geqslant n$ stosuje się predyktor o n współczynnikach. Wszystkie sygnały dla $T < 0$ uważa się za zerowe.

5. WYNIKI BADAŃ SYMULACYJNYCH

5.1. WARUNKI I ZAKRES BADAŃ

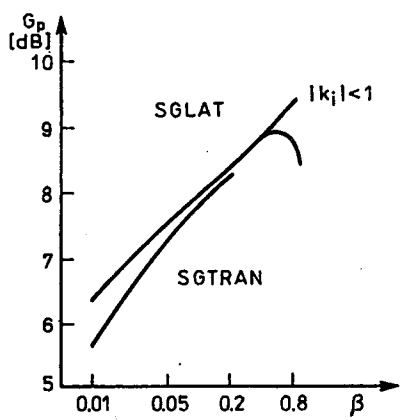
Badania symulacyjne wykonywane były na sygnale mowy. Wykorzystywano 3 frazy sygnału mowy, spróbkowanego z częstotliwością 8 kHz i przetworzonego na postać cyfrową w formacie 12-bitowym (w sumie około 10s materiału dźwiekowego). Badano następujące metody sekwencyjne liniowej predykcji:

SGTRAN: algorytm gradientu stochastycznego, predyktor transwersalny,
SGTRAN-N: algorytm gradientu stochastycznego z normalizacją, predyktor transwersalny,
SGLAT: algorytm gradientu stochastycznego, predyktor kratowy,
SGLAT-N: algorytm gradientu stochastycznego z normalizacją, predyktor kratowy,
LSLAT: algorytm najmniejszych kwadratów z oknem eksponencjalnym, predyktor kratowy,
COVLAT: algorytm kowariancyjny z oknem ruchomym, predyktor kratowy.

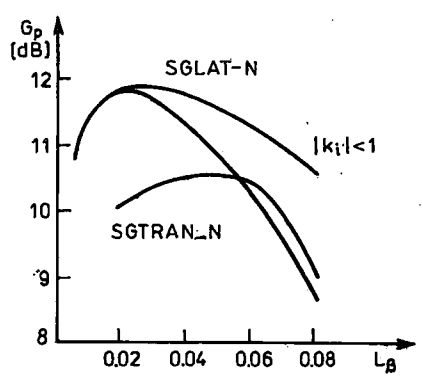
5.2 BADANIE ZYSKU PREDYKCJI

W pierwszym etapie badań rejestrowano zysk predykcji G_p w ujęciu segmentowym dla samego predyktora z adaptacją (z układu ADPCM wyeliminowano kwantyzer). Pozwoliło to na dobór optymalnych parametrów predyktora i algorytmu jego adaptacji.

W metodach gradientu stochastycznego SGTRAN i SGLAT najistotniejszą sprawą jest właściwy dobór stałej β określającej szybkość zbieżności. Dla zbyt



Rys. 4. Zysk predykcji w funkcji parametru β (n = 8)



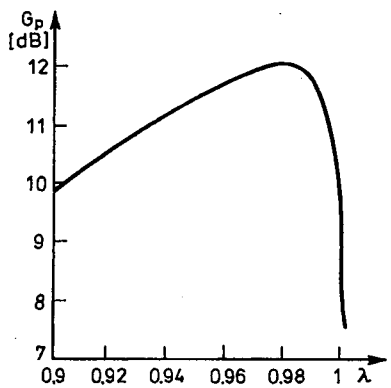
Rys. 5. Zysk predykcji w funkcji parametru L_β (n = 8)

małych wartości β szybkość zbieżności jest nazbyt powolna — objawia się to spadkiem zysku predykcji G_p (rys. 4). Dla nazbyt dużych wartości β następuje utrata stabilności algorytmu adaptacji — szczególnie w obrębie fragmentów sygnału mowy o dużym poziomie.

W metodzoe SGLAT można ograniczyć skutki utraty stabilności algorytmu adaptacji wymuszając $|k_i| < 1$, $i = 1, \ldots, n$ (jest to jednocześnie warunek stabilności układu ADPCM z predyktorem w pętli sprzężenia zwrotnego). Ostatecznie ustalono następujące wartości stałej β: dla algorytmu SGTRAN $\beta = 0.2$, dla algorytmu SGLAT $\beta = 0.4$ (zakłada się, że próbki przetwarzanego sygnału mowy mieszczą się w przedziale od -1 do $+1$).

W metodach gradientu stochastycznego z normalizacją najistotniejszy jest wybór wartości L_β — stałej określającej szybkość zbieżności i stabilność algorytmu. Działa tu podobny mechanizm, jak w przypadku metod bez normalizacji (rys. 5). Ostatecznie ustalono $L_\beta = 0.04$ dla algorytmu SGTRAN-N i $L_\beta = 0.02$ dla algorytmu SGLAT-N. Parametr M_β ma minimalny wpływ na zysk predykcji. W obliczeniach stosowano $M_\beta = 0.0001$. Parametr α, od którego zależy stała czasowa estymatora mocy sygnału (wzór 3.7), ma również niewielki wpływ na zysk predykcji — można go zmieniać w zakresie 0.8—0.98. Ostatecznie przyjęto $\alpha = 0.92$ dla algorytmu SGTRAN-N i $\alpha = 0.96$ dla algorytmu SGLAT-N.

W metodzie najmniejszych kwadratów z oknem początkowym i wagą eksponencjalną (LSLAT) należy wybrać odpowiednią wartość λ charakteryzującą efektywną długość eksponencjalnego okna. W wyniku symulacji otrzymano optymalną wartość $\lambda = 0.98$ (rys. 6).



Rys. 6. Zysk predykcji w funkcji parametru λ ($n = 8$)

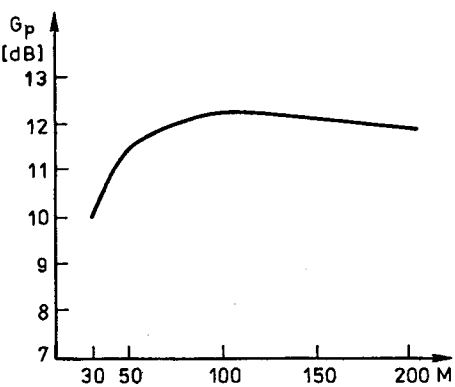
W metodzie kowariancyjnej z oknem ruchomym (COVLAT) dla niewielkich fragmentów mowy uzyskiwano wysokie wartości G_p rzędu 16 dB. Symulacje z użyciem całych fraz mowy nie dały jednak zadowalających wyników ze względu na pojawiające się przypadki utraty stabilności algorytmu adaptacji predyktora.

Szczegółowa analiza dowiodła, że występuje tu zjawisko kumulacji błędów zaokrągleń.

W algorytmie adaptacji predyktora (4.9) występują następujące zmienne, w których mogą kumulować się błędy zaokrągleń: $K_{m+1}(T)$, $m = 0,\ldots, n-1$, oraz $\varepsilon_b(T|0)$ i $\varepsilon_f(T|0)$. W przypadku braku błędów zaokrągleń spełnione są równości $K_{m+1}(T) = \mathbf{E}_b^t(T|m+1)\mathbf{E}_f(T|m)$, $m = 0,\ldots, n-1$, oraz $\varepsilon_b(T|0) = \varepsilon_f(T|0) = \|\mathbf{Y}(T)\|^2$. Po przetworzeniu dłuższego fragmentu sygnału mowy błędy zaokrągleń osiągają tak wysoki poziom, że wyżej wymienione równości nie są już spełnione. Daje się to zauważyć szczególnie w sytuacji, gdy przetwarzany jest fragment ciszy międzywyrazowej i wartości $\mathbf{E}_b^t(T|m+1)\mathbf{E}_f(T|m)$ oraz $\|\mathbf{Y}(T)\|^2$ są małe. W tych warunkach często dochodzi do naruszenia nierówności (4.14), co powoduje otrzymanie ujemnych norm wektorów w kolejnych iteracjach i utratę stabilności numerycznej całego algorytmu adaptacji. Ograniczenie wartości $K_{m+1}(T)$ tak, aby wymusić spełnienie nierówności (4.14), nie jest działaniem skutecznym, gdyż wyliczane współczynniki predykcji tracą związek z przetwarzanym sygnałem mowy (zależą od skumulowanych błędów zaokrągleń).

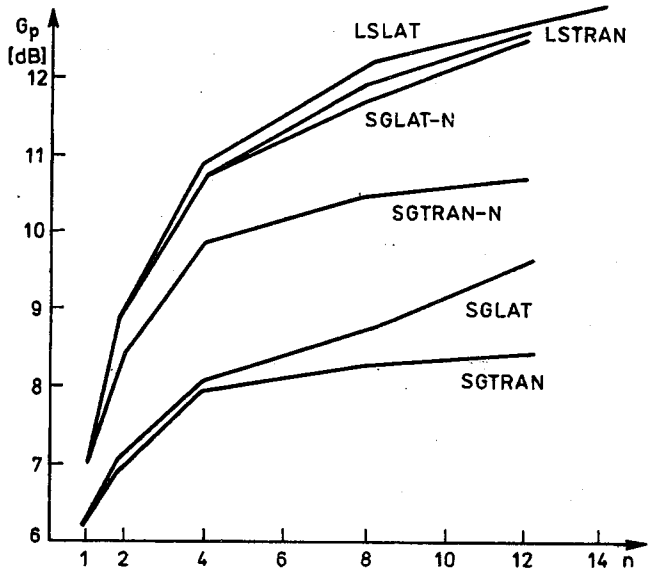
Skutecznym działaniem okazał sięł natomiast restart całego algorytmu adaptacji po każdorazowym naruszeniu nierówności (4.14). Zmienne zawierające skumulowane błędy zaokrągleń są zerowe, a cały sygnał $\tilde{y}(T)$ poprzedzający moment restartu jest traktowany jako zerowy.

Po wprowadzeniu powyższych modyfikacji zbadano wpływ długości okna M na zysk predykcji. Najlepsze wyniki osiągnięto dla okna liczącego M = 100 próbek (rys. 7).



Rys. 7. Zysk predykcji w funkcji długości okna M (n = 8)

Po ustaleniu optymalnych wartości parametrów dla każdej z wymienionych metod adaptacji predyktora, zbadano zysk predykcji w funkcji liczby współczynników predykcji n (rys. 8). Najlepszym okazał się algorytm COVLAT, jednak algorytmy LSLAT i SGLAT-N ustępują mu jedynie w niewielkim stopniu.



Rys. 8. Zysk predykcji w funkcji liczby współczynników n

Zależność zysku predykcji G_p od liczby współczynników predykcji n wykazuje szybkie nasycenie. Ostatecznie przyjęto $n = 8$ jako kompromis między jakością predyktora a liczbą operacji arytmetycznych wykonywanych w procesie adaptacji.

5.3. BADANIE UKŁADU ADPCM

Po ustaleniu optymalnych wartości parametrów dla układów adaptacji predyktora, zasymulowano układ ADPCM z kwantyzerami adaptacyjnymi o 4, 8 i 16 poziomach kwantyzacji, co odpowiada szybkości transmisji 16, 24 i 32 kbit/s. Stosowano predyktory o $n = 8$ współczynnikach.

Tablica 1

Stosunek mocy sygnału użytecznego do zniekształceń (SNR [dB]) w ujęciu segmentowym dla symulowanego układu ADPCM

ALGORYTM	Liczba poziomów kwantyzacji		
	4	8	16
bez predykcji	8.91	14.68	19.91
SGTRAN	14.59	20.98	24.28
SGLAT	14.69	21.14	24.21
SGTRAN-N	15.57	21.65	24.37
SGLAT-N	16.35	22.89	25.04
LSLAT	16.31	22.76	25.10
COVLAT	16.69	23.03	25.37

Z analizy wyników (tabl. 1) można wnioskować o opłacalności stosowania predyktorów (poprawa jakości mowy o 3–9 dB w porównaniu z układem bez predykcji). Najlepsze wyniki zapewnia metoda COVLAT, jednak metody LSLAT i SGLAT-N prawie jej dorównują. Generalnie, predyktory o strukturze kratowej zapewniają lepsze działanie układu ADPCM niż predyktory transwersalne.

6. PODSUMOWANIE

Celem pracy było dokonanie analizy porównawczej algorytmów liniowej predykcji pod kątem ich przydatności do implementacji w koderach ADPCM, szczególnie z wykorzystaniem procesorów sygnałowych. Podobne badania przeprowadzono już w [9, 10]. W niniejszej pracy problem potraktowano szerzej, badając szerszą klasę algorytmów liniowej predykcji (m.in. algorytmy kowariancyjne z oknem ruchomym) oraz układy ADPCM o różnej szybkości transmisji: 16, 24 i 32 kbit/s.

Najlepsze właściwości, z punktu widzenia zysku predykcji, posiadają algorytmy najmniejszych kwadratów. Algorytm z oknem ruchomym w swojej wersji oryginalnej [5] nie nadaje się do przetwarzania sygnału mowy ze względu na kumulację błędów zaokrągleń. Po omówionej w p. 5.2 modyfikacji algorytm ten okazał się najlepszy ze względu na zysk predykcji i stosunek synał–szum w układzie ADPCM. Algorytm z oknem początkowym i wagą eksponencjalną jest odporny na kumulację błędów zaokrągleń (w każdym momencie T skumulowane błędy maleją w stosunku λ) i niewiele ustępuje algorytmowi z oknem ruchomym.

Metody gradientowe z normalizacją, zastosowane w predyktorach o strukturze kratowej, niewiele ustępują metodom najmniejszych kwadratów. Metody gradientowe są ciągle jeszcze udoskonalane. Proponowane są warianty bardziej złożone obliczeniowo, lecz szybciej zbieżne [12].

Ostateczne decyzje co do implementacji poszczególnych algorytmów liniowej predykcji w koderach ADPCM zależą od techniki realizacji tych układów. Metody gradientu stochastycznego w wersji transwersalnej są najmniej złożone obliczeniowo (liczba mnożeń i dzieleń w każdej iteracji czasowej jest rzędu $2n$ dla algorytmu SGTRAN i $3n$ dla algorytmu SGTRAN-N). Możliwa jest ich implementacja na procesorach sygnałowych starszej generacji (np. TMS 32010). Metody gradientu stochastycznego oraz metody najmniejszych kwadratów w wersji kratowej są bardziej złożone obliczeniowo (liczba mnożeń i dzieleń na iterację czasową jest rzędu $10n$ dla algorytmu SGLAT, $13n$ dla algorytmu SGLAT-N, $11n$ dla algorytmu LSLAT i $16n$ dla algorytmu COVLAT). Możliwa jest ich implementacja na procesorach sygnałowych nowszej generacji (np. TMS 32025, TMS 32030).

Wydaje się, że predyktor kratowy z adaptacją metodą najmniejszych kwadratów z oknem początkowym i wagą eksponencjalną (LSLAT) jest szczególnie godny polecenia ze względu na wysoki zysk predykcji i jednorodną strukturę algorytmu (brak restartów).

BIBLIOGRAFIA

1. L.R. Rabiner, R.W. Schafer: *Digital processing of speech signals.* Prentice-Hall, N.J. 1978
2. N.S. Jayant, P. Noll: *Digital coding of waveforms – principles and applications to speech and video.* Prentice-Hall 1984
3. B.S. Atal, J.R. Remde: *A new model of LPC excitation for producing natural-sounding speech at low bit rates.* Proc. Int. Conf. on Acoustics, Speech and Signal Processing ICASSP'82, pp. 614–617
4. CCITT Recommendation G 721 (COM XVIII – R26) Aug. 1986
5. M.L. Honig, D.G. Messerschmitt: *Adaptive filters: structures, algorithms and applications.* Kluwer Acad. Publ. 1984
6. K. Abou-Kassem, P. Dymarski: *Kwantyzery adaptacyjne w modulacji DPCM.* Kwartalnik Elektroniki i Telekomunikacji, t. 37, 1991. z. 3–4
7. E. Świercz: *Wybrane zagadnienia blokowej filtracji adaptacyjnej sygnału mowy przy zastosowaniu autokorelacyjnej i kowariancyjnej metody LS.* KST 1990
8. P. Dymarski, A. Chmielewski, K. Abou-Kassem: *Algorytmy analizy i syntezy predykcyjnej dla procesora sygnałowego TMS 32010.* KST 1990
9. M.L. Honig, D.G. Messerschmitt: *Comparison of adaptive linear prediction algorithms in ADPCM.* IEEE Trans. COM-30, July 1982
10. R.C. Reiniger, J.D. Gibson: *Backward adaptive lattice and transversal predictors in ADPCM.* IEEE Trans. on Communications vol. 33, Jan. 1985
11. C.S. Ng, P.H. Milenkovic: *Unstable covariance LPC solutions from nonstationary speech waveforms.* IEEE Trans. ASSP-37, May 1989
12. J. Chao, H. Perez, S. Tsujii: *A fast adaptive filter algorithm using eigenvalue reciprocals as stepsizes.* IEEE Trans. ASSP-38. August 1990
13. J.D. Markel, A.H. Gray: *Linear prediction of speech.* Springer-Verlag, Berlin 1976
14. S.T. Alexander: *Adaptive signal processing – Theory and applications.* Springer Verlag, New York 1986
15. B. Friedlander: *Lattice filters for adaptive processing.* Proc. of the IEEE, vol. 70, pp. 828–867, Aug. 1982
16. P. Strobach: *Linear Prediction Theory.* Springer 1990
17. J. Szabatin, A. Wojtkiewicz: *Blokowe i rekursywne algorytmy estymacji parametrów AR szeregów czasowych.* Rozprawy Elektroniczne, 1989, z. 4, ss. 993–1046

P. DYMARSKI, A. CHMIELEWSKI, S. KULA
E. ŚWIERCZ

Linear prediction algorithms for DPCM coders of bit rates 16–32 kbit/s are compared. Transversal and lattice structures of the prediction filter are considered. The following sequential adaptation algorithms: stochastic gradient (SG) and least squares (LS) methods are analysed. Several variants of these adaptation algorithms are compared, e.g. SG with and without normalization, LS with exponential window and with sliding window. The values of some parameters are optimized e.g. speed of convergence, number of prediction coefficients, length of the window. A method of stabilization for the LS algorithm with sliding window has been proposed. The LS algorithm with exponential window has proven particularly useful for the applications in DPCM coders.

KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJI, 1993, 39, z. 1, ss. 55-89

681.3.05
621.37

On the fast algorithms of linear least square method in adaptive filtering

MARIUSZ ŻÓŁTOWSKI

Instytut Telekomunikacji, Politechnika Gdańska

Received 1992.08.07

Authorized 1992.11.20

In this study the author is sharing his interest in adaptive filtering. Namely, the different variants of fast least squares algorithms in linear case are presented in detail in review form. Additionally the use of "knee" principle to select the order of adaptive filter is explained and discussed shortly. The effect of weighting parameter on filter's performance is characterized also. The approach is aimed towards the refinement of recursive schemes in view of their importance in digital signal processing, identification and control.

I. INTRODUCTION

The idea of least squares while had already been used by Gauss to establish the orbits of celestial bodies [10] is the quite old one (Gauss–Markov theorem of Jaźwiński [5]). However the reasonable progress has been done lately to make the algorithm of least squares more suitable from the computational point of view. The step by step improvements resulted in *a priori* (L. Jung, Falconer, Morf [2, 22], Bellanger [1]) and *a posteriori* errors (Carayannis G., Manolakis D., Kaloupsidis N., [23]) based approaches to the fast algorithms of recursive least squares. New trends in the architecture of microprocessors towards high signal processing ability, including parallel processing via systolic array, wave front processors etc., while offered at low prices make chance for practical implementation of even more complex algorithms in order to improve the performance of existing systems or realization of new ones not feasible before for the reason of not meeting the requirements for the real time of performance. Therefore, such a problem while undertaken to evaluate the feasibility of existing algorithms is very suitable for

considerations. There is still a debate on the weighting factor effect on the convergence of adaptive filter algorithm and the attempt of defining the method of filter order selection. The algorithms presented in this paper are fast due to the sophisticated way of performing the calculations. This is accomplished in similar manner as Discrete Fourier Transformation is performed fast by FFT.

The reader involved in this subject should be referred to the related problems being formulated as four different approaches [21]; 1) The Method of Recursive Least Squares 2) The Filtering Theory, 3) Stochastic Approximation, 4) Model Reference Adaptive Systems [1–26]. With the extensive references in hand this paper is based on [1] mainly. It is aimed towards the refinement of recursive schemes because of their importance in digital signal processing, identification and control. To complete the derivation of the considered algorithms the reader is referred to appendices.

II. OPTIMAL LEAST SQUARES METHOD

Whenever the speed of adaptation is of importance the algorithm of the filter while performing its task should be optimal on each step of adaptation and not only asymptotically. However the constant matrix cannot be substituted for the reciprocity of autocorrelation matrix as is in the algorithms with the constant step of adaptation case [1] any more. Obviously, the arising complexity in view of real time requirements should be handled. The fast algorithms of the least squares method are to meet this challenge.

The optimal on each step of adaptation least squares algorithms are based on recurrence which updates the vector of FIR filter coefficients $D(n)_{N\times 1}$ during the subsequent discrete time instants according to [1]:

$$D(n+1) = D(n) + R_{xx}^{-1}(n)X(n+1)\{y(n+1) - X^T(n+1)D(n)\} \tag{1}$$

$$D(n) = \mathrm{col}\{D_0(n), \ldots, D_{N-1}(n)\}_{N\times 1},$$

N is the number of filter's coefficients, $y(n)$ is the reference or observation vector,

$$X(n) = \mathrm{col}\{x(n), \ldots, x(n-N+1)\}_{N\times 1}$$

is the input data vector and $R_{xx}(n)$ is the autocerralation matrix.

Obviously, the computational efforts while the least squares optimization results in (1) are not taken into account. That is why further optimization tasks may rely on the reduction of calculations or error propagation effects due to quantization. However, the minimization of number of arithmetic operations seems to be of interest most of all, especially when the reciprocity of autocorrelation matrix is of concern.

III. THE FIRST SIMPLIFICATION OF COMPUTATIONS. MATRIX INVERSION LEMMA

The autocorrelation matrix of (1) changes according to

$$R_{xx}(n+1) = wR_{xx}(n) + X(n+1)X^T(n+1) \tag{2}$$

In view of the matrix inversion lemma of Appendix 1 the reciprocity of the autocorrelation matrix is

$$R_{xx}^{-1}(n+1) = \left(I - \frac{R_{xx}^{-1}(n)X(n+1)X(n+1)^T}{w + X(n+1)^T R^{-1}(n)X(n+1)}\right)\frac{R_{xx}^{-1}(n)}{w} \tag{3a}$$

Moreover, since the gain of adaptation

$$G(n) \stackrel{\text{df}}{=} \mathrm{R}_{xx}^{-1}(n)X(n) \tag{3b}$$

has been introduced, this reciprocity is

$$R_{xx}^{-1}(n+1) = \{I - G(n+1)X(n+1)^T\}R_{xx}^{-1}(n)/w. \tag{3c}$$

That is why the filter's coefficients are updated according to

$$D(n+1) = D(n) + G(n+1)\{y(n+1) - X(n+1)^T D(n)\} \qquad n = 0, 1,... \tag{3d}$$

The knowledge of $R_{xx}^{-1}(0)$ is necesarry to start the recurrence of (3). However, the use of matrix lemma can be omitted. One can derive (3) from (2) directly. Though the number of arithmetical operations is being reduced there is still need for further simplifications of computational procedure.

IV. TOWARDS FURTHER SIMPLIFICATIONS OF COMPUTATIONAL PROCEDURE. THE INTRODUCTION OF PREDICTION ERRORS AND CONTROLLING PARAMETERS

The prediction errors can be introduced in order to look for further improvement of least squares procedure in number of arithmetical operations. Namely, the *forward* ϵ^f and *backward* ϵ^b prediction errors are defined as

$$\epsilon_1^f(n+1) \stackrel{\text{df}}{=} x(n+1) - D_f^T(n)X(n) \qquad [\text{Forward } a\ priori\,] \tag{4a}$$

$$\epsilon_2^f(n+1) \stackrel{\text{df}}{=} x(n+1) - D_f^T(n+1)X(n) \qquad [\text{Forward } a\ posteriori\,] \tag{4b}$$

$$\epsilon_1^b(n+1) \stackrel{\text{df}}{=} x(n+1-N) - D_b^T(n)X(n+1) \qquad [\text{Backward } a\ priori\,] \tag{4c}$$

$$\epsilon_2^b(n+1) \stackrel{\text{df}}{=} x(n+1-N) - D_b^T(n+1)X(n+1) \quad \text{[Backward } a \text{ posteriori]} \tag{4d}$$

Moreover the "forward" and "backward" prediction in terms of energy functionals E_f and E_b is

$$E_f(n) = \sum_{k=1}^{n} w^{n-k} \{x(k) - D_f(n)^T X(k-1)\}^2 \tag{5a}$$

$$E_b(n) = \sum_{k=1}^{n} w^{n-k} \{x(k-N) - D_b(n)^T X(k)\}^2 \tag{5b}$$

The optimal vectors of filter's coefficients resulting from the minimization of (5) are

$$D_f(n) = R_{xx}^{-1}(n-1)\, r_x^f(n) \tag{6a}$$

$$r_x^f(n) = \sum_{k=1}^{n} w^{n-k} x(k) X(k-1) \tag{6b}$$

$$D_b(n) = R_{xx}^{-1}(n)\, r_x^b(n) \tag{6c}$$

$$r_x^b(n) = \sum_{k=1}^{n} w^{n-k} x(k-N) X(k). \tag{6d}$$

In this case, on the other hand the filter's coefficients are updated according to

$$D_f(n+1) = D_f(n) + G(n)\, \epsilon_1^f(n+1) \tag{7a}$$

$$D_b(n+1) = D_b(n) + G(n+1)\, \epsilon_1^b(n+1). \tag{7b}$$

The *a priori* and *a posteriori* errors are related by

$$\epsilon_2^f(n+1) = \epsilon_1^f(n+1)\,(1 - G^T(n)X(n)) \tag{8a}$$

$$\epsilon_2^b(n+1) = \epsilon_1^b(n+1)\,(1 - G^T(n+1)X(n+1)). \tag{8b}$$

Whereas the errors's energy is given by the following formulas

$$E_f(n+1) = wE_f(n) + \epsilon_1^f(n+1)\, \epsilon_2^f(n+1) \tag{9a}$$

$$E_b(n+1) = wE_b(n) + \epsilon_1^b(n+1)\, \epsilon_2^b(n+1) \tag{9b}$$

The prediction errors ratio parameter can also be introduced

$$r(n) = \frac{\epsilon_2^f(n+1)}{\epsilon_1^f(n+1)} = \frac{\epsilon_2^b(n)}{\epsilon_1^f(n)}, \quad n = 0, 1, \ldots \tag{10}$$

This parameter can be used to control the process of convergence of adaptive filter. The derivation of recursive expressions of this chapter and the explanation of physical meaning of $r(n)$ parameter are given in Appendix 2. The results in hand are used to design fast least squares algorithms [1].

V. THE FAST LEAST SQUARES ALGORITHM BASED ON PREDICTION ERRORS

The recursive expressions concerning the gain of adaptation are of interest while looking for the acceleration of least squares algorithm. Moreover, the dependence of the autocorrelation matrix and crosscorrelation vectors on discrete time n and number N of filter coefficients should be considered:

$$\begin{aligned} R_{xx}^{N}(n) &= R_{xx}(n) \\ r_{x}^{f}(n) &= r_{x}^{f,N}(n) \\ r_{x}^{b}(n) &= r_{x}^{b,N}(n). \end{aligned} \tag{11}$$

First, from the partition of $R_{xx}^{N+1}(n+1)$ (Appendix 3)

$$R_{xx}^{N+1}(n+1)\begin{bmatrix} 0 \\ G(n) \end{bmatrix} = \begin{bmatrix} r_{x}^{f,N}(n+1)^{T} G(n) \\ X(n) \end{bmatrix} \tag{12a}$$

and

$$R_{xx}^{N+1}(n+1)\begin{bmatrix} G(n+1) \\ 0 \end{bmatrix} = \begin{bmatrix} X(n+1) \\ r_{x}^{b,N}(n+1)^{T} G(n+1) \end{bmatrix}. \tag{12b}$$

Next (Appendix 3)

$$R_{xx}^{N+1}(n+1)\begin{bmatrix} 0 \\ G(n) \end{bmatrix} = X^{N+1}(n+1) - \begin{bmatrix} \epsilon_{2}^{f}(n+1) \\ 0 \end{bmatrix} \tag{13a}$$

$$R_{xx}^{N+1}(n+1)\begin{bmatrix} G(n+1) \\ 0 \end{bmatrix} = X^{N+1} - \begin{bmatrix} 0 \\ \epsilon_{2}^{b}(n+1) \end{bmatrix} \tag{13b}$$

$$X^{N+1}(n+1)^{T} = (x(n+1), X^{N}(n)^{T}) = (X^{N}(n+1)^{T}, x(n+1-N)) \tag{14a}$$

$$R_{xx}^{N}(n)\, G^{N}(n) = X^{N}(n). \tag{14b}$$

That is why

$$R_{xx}^{N+1}(n+1)\left[G^{N+1}(n+1) - \begin{bmatrix} 0 \\ G^{N}(n) \end{bmatrix} \right] = \begin{bmatrix} \epsilon_{2}^{f}(n+1) \\ 0 \end{bmatrix} \tag{15a}$$

and

$$R_{xx}^{N+1}(n+1)\left[G^{N+1}(n+1) - \begin{bmatrix} G^{N}(n+1) \\ 0 \end{bmatrix} \right] = \begin{bmatrix} 0 \\ \epsilon_{2}^{b}(n+1) \end{bmatrix} \tag{15b}$$

$$0 = \mathrm{col}(0, \ldots, 0)_{1 \times N}$$

Also (Apendix 3)

$$R_{xx}^{N+1}(n+1)\begin{bmatrix} 1 \\ -D_f^N(n+1) \end{bmatrix} = \begin{bmatrix} E_f(n+1) \\ 0 \end{bmatrix} \tag{16a}$$

and

$$R_{xx}^{N+1}(n+1)\begin{bmatrix} -D_b^N(n+1) \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ E_b(n+1) \end{bmatrix} \tag{16b}$$

Now, while comparing (16) and (15) one obtains

$$G^{N+1}(n+1) = \begin{bmatrix} 0 \\ G^N(n) \end{bmatrix} + \frac{\epsilon_2^f(n+1)}{E_f(n+1)}\begin{bmatrix} 1 \\ -D_f^N(n+1) \end{bmatrix} \tag{17a}$$

and

$$G^{N+1}(n+1) = \begin{bmatrix} G^N(n+1) \\ 0 \end{bmatrix} + \frac{\epsilon_2^b(n+1)}{E_b(n+1)}\begin{bmatrix} -D_b^N(n+1) \\ 1 \end{bmatrix} \tag{17b}$$

Moreover the vector of adaptation can be given the form

$$G^{N+1}(n+1) \stackrel{\mathrm{df}}{=} \begin{bmatrix} G_N^{N+1}(n+1) \\ g(n+1) \end{bmatrix} \tag{18a}$$

and

$$g(n+1) \stackrel{\mathrm{df}}{=} \frac{\epsilon_2^b(n+1)}{E_b(n+1)} \tag{18b}$$

That is why

$$G^N(n+1) = G_N^{N+1}(n+1) + g(n+1)D_b^N(n+1) \tag{19a}$$

and

$$G^N(n+1) = \frac{1}{1-g(n+1)\epsilon_1^b(n+1)}\left\{G_N^{N+1}(n+1) + g(n+1)D_b(n)\right\} \tag{19b}$$

since (7b) holds true.

In addition other details are given in Appendix 3. As a matter of fact all the considerations result in the following algorithm [1]:

ALGORITHM I

At instant n:

The coefficients of adaptive filter:	$D(n)$
The coefficients of forward prediction filter:	$D_f(n)$
The coefficients of backward prediction filter:	$D_b(n)$
Data vector:	$X(n)$
Adaptation gain:	$G(n)$
The energy of forward prediction error:	$E_f(n)$
Weighting factor	w

New data at instant n

The input signal:	$x(n+1)$
The reference signal:	$y(n+1)$

$$\epsilon_1^f(n+1) \underset{(4a)}{=} x(n+1) - D_f^T(n)X(n) \tag{20a}$$

$$D_f(n+1) \underset{(7a)}{=} D_f(n) + G(n)\epsilon_1^f(n+1) \tag{20b}$$

$$\epsilon_2^f(n+1) \underset{(4b)}{=} x(n+1) - D_f^T(n+1)X(n) \tag{20c}$$

$$E_f(n+1) \underset{(9a)}{=} wE_f(n) + \epsilon_1^f(n+1)\epsilon_2^f(n+1) \tag{20d}$$

$$\begin{bmatrix} G_N^{N+1}(n+1) \\ g(n+1) \end{bmatrix} \underset{(17a)(18a)}{=} \begin{bmatrix} 0 \\ G^N(n) \end{bmatrix} + \frac{\epsilon_2^f(n+1)}{E_f(n+1)} \begin{bmatrix} 1 \\ -D_f(n+1) \end{bmatrix} \tag{20e}$$

$$\epsilon_1^b(n+1) \underset{(4c)}{=} x(n+1-N) - D_b^T(n)X(n+1) \tag{20f}$$

$$G(n+1) \underset{(19b)}{=} \frac{G_N^{N+1}(n+1) + g(n+1)D_b(n)}{1 - g(n+1)\epsilon_1^b(n+1)} \tag{20g}$$

$$D_b(n+1) \underset{(7b)}{=} D_b(n) + G(n+1)\epsilon_1^b(n+1) \tag{20h}$$

On the other hand the coefficients of the optimal adaptive filter evolve according to

$$D(n+1) \underset{(3c)}{=} D(n) + G(n+1)\epsilon_1(n+1) \tag{21a}$$

while

$$\epsilon_1(n+1) \overset{\text{df}}{=} y(n+1) - X^T(n+1)D(n). \tag{21b}$$

Moreover the following initial conditions can be chosen

$$D_f(0) = D_b(0) = G(0) = 0 \text{ and } E_f(0) = E_{f,0}. \tag{22}$$

The approximate numbers of operations and active memory cells are:
$10N + 5$ multiplications, 3 divisions and $6N$ active memory cells.

VI. THE SECOND FAST ALGORITHM OF LEAST SQUARES METHOD BASED ON ALL PREDICTION ERRORS

Also the *a posteriori* gain of adaptation G_2^N can be put into good account while the optimal least squaresfast algorithms is considered. It is defined by

$$R_{xx}(n)G_2(n+1) = X(n+1). \tag{23}$$

Next, from (A3-1) first (Appendix 4)

$$R_{xx}^{N+1}(n)\left[G_2^{N+1}(n+1) - \begin{bmatrix} G_2^N(n+1) \\ 0 \end{bmatrix}\right] = \begin{bmatrix} 0 \\ \epsilon_1^b(n+1) \end{bmatrix} \tag{24a}$$

and

$$R_{xx}^{N+1}(n)\left[G_2^{N+1}(n+1) - \begin{bmatrix} 0 \\ G_2^N(n) \end{bmatrix}\right] = \begin{bmatrix} \epsilon_1^f(n+1) \\ 0 \end{bmatrix}. \tag{24b}$$

Moreover, in view of (16)

$$G_2^{N+1}(n+1) = \begin{bmatrix} G_2^N(n+1) \\ 0 \end{bmatrix} + \frac{\epsilon_1^b(n+1)}{E_b(n)} \begin{bmatrix} -D_b(n) \\ 1 \end{bmatrix} \tag{25a}$$

and

$$G_2^{N+1}(n+1) = \begin{bmatrix} O \\ G_2^N(n) \end{bmatrix} + \frac{\epsilon_1^f(n+1)}{E_f(n)} \begin{bmatrix} 1 \\ -D_f(n) \end{bmatrix}. \tag{25b}$$

Secondly (Appendix 4)

$$D(n+1) = D(n) + w^{-1}R_{xx}^{-1}(n)X(n+1)\{y(n+1) - X^T(n+1)D(n+1)\}. \tag{26}$$

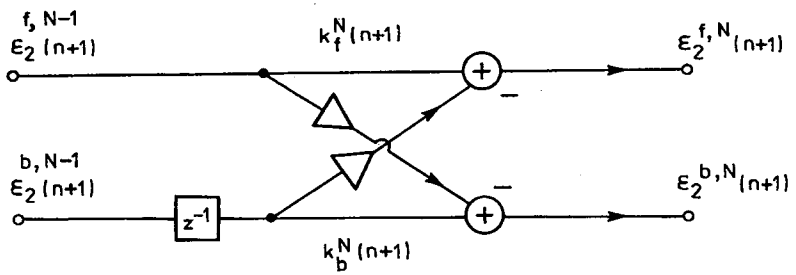


Fig. 1. The prediction error computing section of lattice filter

Next, in view of (A2-11) and (23) the following parameter can be introduced

$$r'(n+1) \stackrel{\text{df}}{=} \frac{w}{r(n+1)} = w + X^T(n+1)G_2^N(n+1) \tag{27}$$

while r is the prediction errors ratio of (10). That is why

$$\epsilon_2(n+1) + X^T(n+1)w^{-1}G_2^N(n+1)\epsilon_2(n+1) = \frac{\epsilon_2(n+1)}{r(n+1)}. \tag{28a}$$

The ϵ_2 of (28a) is the *a posteriori* error defined by

$$\epsilon_2(n+1) \stackrel{\text{df}}{=} y(n+1) - X^T(n+1)D(n+1). \tag{28b}$$

Eventually, in view of (26), (23) and (28a)

$$\epsilon_2(n+1) = r(n+1)\epsilon_1(n+1) \tag{29a}$$

while the *a priori* error ϵ_1 of (29a) is

$$\epsilon_1(n+1) \stackrel{\text{df}}{=} y(n+1) - X^T(n+1)D(n). \tag{29b}$$

On the other hand the filter's coefficients are updated in view of last expressions according to (Appendix 4)

$$D(n+1) = D(n) + G_2^N(n+1)\frac{1}{r'(n+1)}\epsilon_1(n+1). \tag{30}$$

Moreover r' can be given the recursive form in filter's order N and discrete time n (Appendix 4):

$$r'^{N+1}(n+1) = r'^N(n+1) + \frac{\epsilon_1^b(n+1)^2}{E_b(n)} = r'^N(n) + \frac{\epsilon_1^f(n+1)^2}{E_f(n)}. \tag{31}$$

Now, let

$$G_2^{N+1}(n+1) = \begin{bmatrix} G_{2,N}^{N+1}(n+1) \\ g_2(n+1) \end{bmatrix}. \tag{32}$$

Then from (25a)

$$g_2(n+1) = \frac{\epsilon_1^b(n+1)}{E_b(n)}. \tag{33}$$

Eventually, the considerations of this chapter result in the second fast least squares algorithm [3, 24, 1]

ALGORITHM 2

At instant n

The coefficients of the adaptive filter:	$D(n)$
The forward prediction coefficients:	$D_f(n)$
The backward prediction coefficients:	$D_b(n)$
Data vector:	$X(n)$
The energies of prediction errors:	$E_f(n)$, $E_b(n)$
The ratio of prediction errors:	$r'(n)$
Weighting factor:	w

New data at instant n

Input signal:	$x(n+1)$
Reference signal:	$y(n+1)$

$$\epsilon_1^f(n+1) \underset{(20a)}{=} x(n+1) - D_f^T(n)X(n) \tag{34a}$$

$$D_f(n+1) \underset{(20b),\ (A4\text{-}6)}{=} D_f(n) + \frac{G_2(n)}{r'(n)}\epsilon_1^f(n+1) \tag{34b}$$

$$E_f(n+1) \underset{(10)\,(20d)\,(27)\,(29)}{=} \{E_f(n) + \epsilon_1^f(n+1)\epsilon_1^f(n+1)/r'(n)\}w \tag{34c}$$

$$G_2^{N+1}(n+1) \underset{(32)\,(25b)}{=} \begin{bmatrix} 0 \\ G_2^N(n) \end{bmatrix} + \frac{\epsilon_1^f(n+1)}{E_f(n)} \begin{bmatrix} 1 \\ -D_f(n) \end{bmatrix} = \begin{bmatrix} G_{2,N}^{N+1}(n+1) \\ g_2(n+1) \end{bmatrix}, \tag{34d}$$

$$\epsilon_1^b(n+1) \underset{(20f)}{=} x(n+1-N) - D_b^T(n)X(n+1) \tag{34e}$$

$$G_2^N(n+1) \underset{(25a)\,(33)\,(34d)}{=} G_{2,N}^{N+1}(n+1) + g_2(n+1)D_b(n) \tag{34f}$$

$$r'^{N+1}(n+1) \underset{(31)}{=} r'^N_1(n) + \epsilon_1^f(n+1)\epsilon_1^f(n+1)/E_f(n) \tag{34g}$$

$$r'^{N}(n+1) \underset{(31)(33)}{=} r'^{N+1}(n+1) - g_2(n+1)\epsilon_1^b(n+1) \tag{34h}$$

$$E_b(n+1) \underset{(10)(9b)(27)}{=} \{E_b(n) + \epsilon_1^b(n+1)\epsilon_1^b(n+1)/r'(n+1)\}w \tag{34i}$$

$$D_b(n+1) \underset{(20h)(A4\text{-}6b)}{=} D_b(n) + \frac{G_2^N(n+1)}{r'(n+1)}\epsilon_1^b(n+1). \tag{34j}$$

Adaptive filter

$$\epsilon_1(n+1) = y(n+1) - D^T(n+1)X(n+1) \tag{34k}$$

$$D(n+1) \underset{(30)}{=} D(n) + G_2^N(n+1)\frac{1}{r'(n+1)}\epsilon_1(n+1). \tag{34l}$$

The following initial conditions can be set

$$D_f(n) = D_b(n) = G_2^N(n) = 0,\ E_f(0) = E_{f,0},\ E_b(0) = w^{-N}E_{f,0},$$
$$r'(0) = w. \tag{34m}$$

The approximate numbers of operations are:
$8N + 13$ multiplications, 7 divisions and $6N$ active memory cells.

VII. THE ALGORITHM OF LATTICE FILTER

The fast least squares algorithms being considered so far are given in terms of variable transversal filter with fixed number of coefficients. However the lattice algorithms which are recursive both in discrete time n and filter's order from 1 to N can be devised [1, 25] also. This is another challenge while looking for the unknown model of the process is of concern. The recursive expressions for

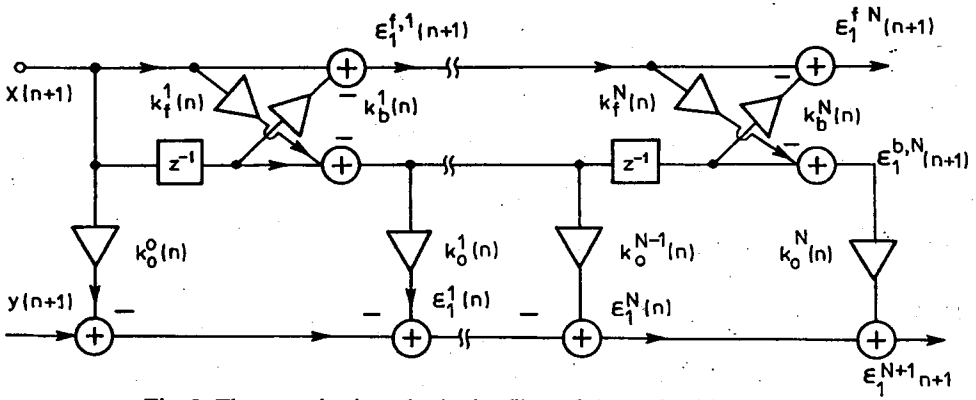


Fig. 2. The recursive in order lattice filter of the optimal least squares

prediction errors in terms of partial correlation coefficients can be obtained first. Namely, the following relations hold true (Appendix 5):

$$R_{xx}^{N+1}(n)\begin{bmatrix}\begin{bmatrix}1\\-D_f^{N-1}(n)\end{bmatrix}\dfrac{K_1^N(n)}{E_a^{N-1}(n)}\\0\end{bmatrix}=\begin{bmatrix}K_1^N(n)\\0\\\dfrac{K_1^N(n)^2}{E_f^{N-1}(n)}\end{bmatrix} \tag{35a}$$

$$R_{xx}^{N+1}(n)\begin{bmatrix}0\\\begin{bmatrix}-D_b^{N-1}(n-1)\\1\end{bmatrix}\dfrac{K_2^N(n)}{E_b^{N-1}(n-1)}\end{bmatrix}=\begin{bmatrix}\dfrac{K_2^N(n)^2}{E_b^{N-1}(n-1)}\\0\\K_2^N(n)\end{bmatrix} \tag{35b}$$

$$K_1^N(n) = r_x^{b,N}(n)^T\begin{bmatrix}1\\-D_f^{N-1}(n)\end{bmatrix} \tag{35c}$$

$$K_2^N(n) = r_x^{f,N}(n)^T\begin{bmatrix}-D_b^{N-1}(n-1)\\1\end{bmatrix} \tag{35d}$$

and

$$K_1^N(n) = K_2^N(n) = K^N(n). \tag{35e}$$

Also (see Appendix 5)

$$D_f^N(n) = \begin{bmatrix}D_f^{N-1}(n)\\0\end{bmatrix}+\frac{K^N(n)}{E_b^{N-1}(n-1)}\begin{bmatrix}-D_b^{N-1}(n-1)\\1\end{bmatrix} \tag{36a}$$

$$E_f^N(n) = E_f^{N-1}(n-1)-\frac{K^N(n)^2}{E_b^{N-1}(n)\,(n-1)} \tag{36b}$$

and

$$D_b^N(n) = \begin{bmatrix}0\\D_b^{N-1}(n-1)\end{bmatrix}+\frac{K^N(n)}{E_f^{N-1}(n)}\begin{bmatrix}1\\-D_f^{N-1}(n)\end{bmatrix} \tag{37a}$$

$$E_b^N(n) = E_b^{N-1}(n-1)-\frac{K^N(n)^2}{E_f^{N-1}(n)}. \tag{37b}$$

But the *forward* and *backward* prediction errors are given by

$$\epsilon_1^{f,N}(n+1) = x(n+1) - D_f^N(n)^T X^N(n)$$

$$\epsilon_1^{b,N}(n+1) = x(n+1-N) - D_b^N(n)^T X^N(n+1).$$

That is why (Appendix 5)

$$\epsilon_1^{f,N}(n+1) = \epsilon_1^{f,N-1}(n+1) - \frac{K^N(n)}{E_b^{N-1}(n-1)}\,\epsilon_1^{b,N-1}(n) \tag{38a}$$

$$\epsilon_1^{b,N}(n+1) = \epsilon_1^{b,N-1}(n) - \frac{K^N(n)}{E_f^{N-1}(n)}\,\epsilon_1^{f,N-1}(n+1) \tag{38b}$$

and

$$\epsilon_2^{f,N}(n+1) = \epsilon_2^{f,N-1}(n+1) - \frac{K^N(n+1)}{E_b^{N-1}(n)}\,\epsilon_2^{b,N-1}(n) \tag{39a}$$

$$\epsilon_2^{b,N}(n+1) = \epsilon_2^{b,N-1}(n) - \frac{K^N(n+1)}{E_f^{N-1}(n+1)}\,\epsilon_2^{f,N-1}(n+1). \tag{39b}$$

The last expressions are the required ones since while

$$k_f^N(n+1) \stackrel{\text{df}}{=} \frac{K^N(n+1)}{E_f^{N-1}(n+1)} \quad \text{and} \quad k_b^N(n+1) \stackrel{\text{df}}{=} \frac{K^N(n+1)}{E_b^{N-1}(n)} \tag{41}$$

they can be given the meaning of reflection coefficients or partial correlation ones. The relevant signal flow graph is shown in Fig. 1. Now one can expect that the idea of the lattice filter algorithm to be presented relies on the expressions similar to those of fast least squares of previous chapters but also recursive in adaptive filter's order N. That is why further recursive relations are of interest. Namely, the *a priori* ϵ_1 and *a posteriori* ϵ_2 errors can be considered secondly. The following relations hold true (Appendix 5)

$$R_{xx}^{N+1}(n)\left[D^{N+1}(n) - \begin{bmatrix} D^N(n) \\ 0 \end{bmatrix}\right] = \begin{bmatrix} 0 \\ K_0^N(n) \end{bmatrix} \tag{42a}$$

$$D^{N+1}(n) = \begin{bmatrix} D^N(n) \\ 0 \end{bmatrix} - \frac{K_0^N(n)}{E_b^N(n)}\begin{bmatrix} -D_b^N(n) \\ 1 \end{bmatrix} \tag{42b}$$

and

$$K_0^N(n) = \sum_{k=1}^{n} w^{n-k}\, y(k)\,\{x(k-N) - D_b^N(n)^T X^N(k)\}. \tag{42c}$$

As a matter of fact (Appendix 5)

$$\epsilon_1^{N+1}(n+1) = \epsilon_1^N(n+1) - \frac{K_0^N(n)}{E_b^N(n)}\,\epsilon_1^{b,N}(n+1) \tag{43a}$$

$$\epsilon_1^{N+1}(n+1) = \epsilon_2^N(n+1) - \frac{K_0^N(n)}{E_b^N(n)}\epsilon_2^{b,N}(n+1). \tag{43b}$$

While concerning the task of adaptation in terms of the error energy E the recursive expressions concerning this parameter can be derived next (Appendix 5)

$$E^{N+1}(n) = E^N(n) - \frac{K_0^N(n)^2}{E_b^N(n)}. \tag{44}$$

With definition of the gain of adaptation in hand

$$R_{xx}^N(n)\, G^N(n) = X^N(n) \tag{45}$$

the recursive expressions for this parameter both in discrete time n and the filter's order N can be derived either. Namely (Appendix 5)

$$G^N(n) = \begin{bmatrix} G^{N-1}(n) \\ 0 \end{bmatrix} + \frac{\epsilon_2^{b,N-1}(n)}{E_b^{N-1}(n)} \begin{bmatrix} -D_b^{N-1}(n) \\ 1 \end{bmatrix}. \tag{46}$$

Subsequently, the *a posteriori* to *a priori* errors ratio is given by (Appendix 5)

$$r^N(n) = r^{N-1}(n) - \frac{\epsilon_2^{b,N-1}(n)^2}{E_b^{N-1}(n)}. \tag{47}$$

The lattice coefficient $K^N(n)$ changes in filter's order according to (Appendix 5)

$$K^{N+1}(n) = w\,K^{N+1}(n) + \epsilon_1^{f,N}(n+1)\epsilon_2^{b,N}(n) \tag{48a}$$

or

$$K^{N+1}(n) = w\,K^{N+1}(n) + \epsilon_2^{f,N}(n+1)\epsilon_1^{b,N}(n). \tag{48b}$$

In view of (48) the meaning of the correlation coefficients of the prediction errors is well established as far as lattice coefficients $K(n)$ are of concern. Next, the energy of prediction errors can be given the recursive form (Appendix 5) either

$$E^N(n+1) = w\,E^N(n) + \epsilon_1^N(n+1)\epsilon_2^N(n+1). \tag{49}$$

Moreover, $K_0^N(n)$ coefficient and reflection coefficients $k_f^N(n)$, $k_b^N(n)$ can be considered in similar way (Appendix 5)

$$K_0^N(n+1) = w\,K_0^N(n) + \epsilon_2^N(n+1)\epsilon_1^{b,N}(n+1) \tag{50a}$$

or

$$K_0^N(n+1) = w\,K_0^N(n) + \epsilon_1^N(n+1)\epsilon_2^{b,N}(n+1). \tag{50b}$$

On the other hand (Appendix 5)

$$k_f^{N+1}(n+1) = k_f^{N+1}(n) + \frac{\epsilon_1^{b,N+1}(n+1)\epsilon_2^{f,N}(n+1)}{E_f^N(n+1)} \tag{51a}$$

$$k_b^{N+1}(n+1) = k_b^{N+1}(n) + \frac{\epsilon_2^{b,N}(n)\epsilon_1^{f,N+1}(n+1)}{E_b^N(n)}. \tag{51b}$$

Also the recursive in time form of the normalized $K_0^N(n)$ filter's coefficient is (Appendix 5)

$$k_0^N(n+1) = k_0^N(n) + \frac{\epsilon_2^{b,N}(n+1)\epsilon_1^{N+1}(n+1)}{E_b^N(n+1)} \tag{52a}$$

and

$$k_0^N(n) = K_0^N(n)/E_b^N(n). \tag{52b}$$

While the all required expressions in hand the recursive equations in time and in filter's order can be gathered to form the lattice algorithm of adaptive filter. The initial conditions can be chosen as follows:

$$\epsilon_1^{b,i}(0) = k_f^i(0) = k_b^i(0) = 0 \tag{53a}$$

$$r^i(0) = 1 \tag{53b}$$

$$E_f^i(0) = w^N E_{f,0} \tag{53c}$$

$$E_b^i(0) = w^{N-i} E_{f,0} \tag{53d}$$

$$0 \leqslant i \leqslant N-1.$$

The whole algorithm [1, 25] is:

THE ALGORITHM OF LATTICE FILTER

At time instant n

The reflection coefficients:	$k_f(n)$, $k_b(n)$
The filter's coefficients:	$k_0(n)$
The forward and backward prediction ratio:	$r(n)$
The weighting factor:	w

New data

The input signal:	$x(n+1)$
The reference signal:	$y(n+1)$

The initialization

$$\epsilon_1^{f,0}(n+1) = \epsilon_1^{b,0}(n+1) = x(n+1) \tag{54a}$$

$$\epsilon_1^{f,0}(n+1) = y(n+1) \tag{54b}$$

$$r^0(n+1) = 1 \tag{54c}$$

$$E_f^0(n+1) = E_b^0(n+1) = wE_f^0(n) + x(n+1)^2 \tag{54d}$$

The prediction

$$\epsilon^{f,i+1}(n+1) = \epsilon_1^{f,i}(n+1) - k_b^{i+1}(n)\epsilon_1^{b,i}(n) \tag{55a}$$

$$\epsilon_1^{b,i+1}(n+1) = \epsilon_1^{b,i}(n) - k_f^{i+1}(n)\epsilon^{f,i}(n+1) \tag{55b}$$

$$k_f^{i+1}(n+1) = k_f^{i+1}(n) + \epsilon_1^{f,i}(n+1)r^i(n)\epsilon_1^{b,i+1}(n+1)/E_f^i(n+1) \tag{55c}$$

$$k_b^{i+1}(n+1) = k_b^{i+1}(n) + \epsilon_1^{f,i+1}(n+1)\epsilon_1^{b,i}(n)r^i(n)/E_b^i(n) \tag{55d}$$

$$r^{r+1}(n) = r^i(n) - \frac{r^i(n)^2\epsilon_1^{b,i}(n)^2}{E_b^i(n)} \tag{55e}$$

$$E_f^{i+1}(n+1) = wE_f^{i+1}(n) + \epsilon_1^{f,i+1}(n+1)^2 r^{i+1}(n) \tag{55f}$$

$$r^{i+1}(n+1) = r^i(n+1) - \frac{r^i(n+1)^2\epsilon_1^{b,i}(n+1)^2}{E_b^i(n+1)} \tag{55g}$$

$$E_b^{i+1}(n+1) = wE_b^{i+1}(n) + \epsilon_1^{b,i+1}(n+1)^2 r^{i+1}(n+1) \tag{55h}$$

$$0 \leqslant i \leqslant N-1.$$

The filter's section

$$\epsilon_1^{i+1}(n+1) = \epsilon_1^i(n+1) - k_f^i(n)\epsilon_1^{b,i}(n+1) \tag{56a}$$

$$k_0^i(n+1) = k_0^i(n) + \frac{\epsilon_1^{b,i}(n+1)\epsilon_1^{i+1}(n+1)r^i(n+1)}{E_b^i(n+1)} \tag{56b}$$

$$0 \leqslant i \leqslant N.$$

Approximately:
$16N + 2$ multiplications, $3N$ divisions, $7N$ memmory cells.

VII. ON THE CRITERION OF FILTER'S ORDER CHOOSING

The discussion of the known criteria according to which the filter's order can be chosen has been presented by Kay and Marple [4]. However, though the energy of prediction errors is the decreasing function of the filter's order nevertheless the "knee" principle can be used to make the choice of the order of any filter only the

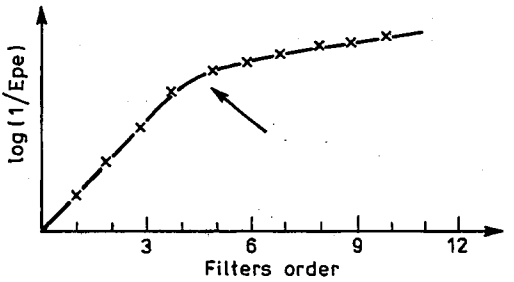


Fig. 3. The illustration of "knee principle" E_{pe} – energy of prediction error

energy of prediction errors based. The identification of "knee" (see Fig. 3) can be performed according to the place of bending where the greatest curvature of $\log(1/E_{pe})$ graph occurs.

IX. ON THE WEIGHTING PARAMETER w

The task of adaptation of the adaptive filter is to find the minimum of

$$\hat{J}_0\{D(n)\} = \sum_{k=1}^{n} w^{n-k}\{y(k) - D^T(n)X(k)\}^2 \tag{57}$$

in respect to $D(n)$.

When the effect of parameter w is of concern one can recognize (57) as the discrete time convolution of the first-order low pass filter impulse response and the sequence of the least squares fit [8]. It is well known how the smooting properties of the low pass filter depend on the parameter w [8]. If w diminishes then the filter's bandwidth is widened. As a matter of fact the recent data to the filter's input are of greater weight. It seems to be easier to get the last data based mean squared fit than the all data based one. That is why the speed of the convergence of the adaptive algorithm should be faster with w smaller ($w \cong 1$). Also, how the steady state behaviour of the adaptive algorithm depends on w is of interest. Let x and y be the stationary, discrete time sequences with R_{xx} autocorrelation matrix and r_{xy} cross-correlation vector. Then there is no effect of parameter w on the optimal steady state vector D_{opt} of filter's coefficients in ergodic case while the discrete time n tends to the infinity. Since

$$R_{xx}(n) \stackrel{\text{df}}{=} \sum_{k=1}^{n} w^{n-k}X(k)X(k)^T \tag{58}$$

and

$$r_{yx}(n) \stackrel{\text{df}}{=} \sum_{k=1}^{n} w^{n-k}y(k)X(k) \tag{59}$$

one gets

$$E\{R_{xx}(n)\} = \frac{1 - w^n}{1 - w} R_{xx} \tag{60a}$$

$$E\{r_{yx}(n)\} = \frac{1 - w^n}{1 - w} r_{yx} \tag{60b}$$

and

$$D_{opt} = R_{xx}^{-1} r_{yx}. \tag{60c}$$

Now, let consider the effect of the parameter w on the convergence of the adaptive filter in more detail. Namely, if one substitutes $E\{R_{xx}(n+1)\}$ for $R_{xx}(n+1)$ and $E\{R_{xx}^{-1}(n+1)\}$ for $R_{xx}^{-1}(n+1)$ then the coefficients of the adaptive filter are updated according to

$$D(n+1) = D(n) + \frac{1 - w}{1 - w^{n+1}} R_{xx} X(n+1)\{y(n+1) - X(n+1)^T D(n) \tag{61}$$

in view of (60a). That is why the gain of adaptation can be expected to diminish in time slower as w is smaller. This is in good accord with the previous conjecture. The acquisition of the steady state is faster on one hand while the spurious behaviour of the identified system in terms of its FIR filter model can have greater effect on the convergence on the other.

X. CONCLUSIONS AND DISCUSSION

The fast algorithms of the least squares method being optimal on each stage of adaptation can be designed to meet the requirements of real time of performance. The speeding up of calculations is due to the use of prediction errors if the autocorrelation matrix of the signal to the adaptive filter's input is partitioned. This partition may result in recursive expressions in time (Algorithm I, Algorithm II) and both in time and filter's order N (Lattice Algorithm) while the gain of adaptation, the prediction errors ratio, the energy of prediction errors etc. are of concern. These expressions in hand, outcome in the number of calculations saving procedures. Algorithm I and Algorithm II and Lattice Algorithm are *a priori* errors based ones. The algorithms which are based on *a posteriori* errors can be devised either [3, 23]. Algorithm II can be considered as more balanced one due to the use of all of prediction errors. While Lattice Algorithm is of concern the "knee" principle can be put into good account to select the order of the adaptive filter. The role of weighting parameter w can be characterized. The smaller w the faster acquisition of the steady state can be expected on one hand and greater effect of the spurious behaviour of the identified system on convergence stage may occur on the other. The problem of the least squares can be considered within the linear optimal filtering theory framework [5]. As a matter of fact the act of choice of initial conditions can be given the reasonable probabilistic meaning. Moreover there is still problem of divergence due to the filter's model imperfections. As this approach is aimed towards the refinement of recursive scheme, a new postulate concerning the task of identification can be formulated [7]. The continued interest in adaptive optimal filtering is due to the continued progress in the technology and architecture of digital signal processing systems.

REFERENCES

1. M. Bellanger: *Adaptive digital filtering*, ECCTD 1987, Paris
2. D. Falconer, L. Ljung: *Application of fast Kalman estimation to adaptive equalization.* IEEE Trans., COM-26, no 10, October 1978, pp. 1439–1446
3. G. Carayannis, D. Manolakis, N. Kalauptsidis: *A fast sequential algorithm for L.S. filtering and prediction*, IEEE Trans., vol. ASSP-31, no 6, Dec. 1983, pp. 1394–1402
4. S.M. Kay, J.R. Marple: *Spectrum analysis – A modern perspective*, Prodceedings of the IEEE, vol. 69, no 11, November 1981, pp. 1395–1396
5. A.H. Jaźwiński: *Stochastic Processes and Filtering Theory*, Academic Press, 1970
6. P. De Larminant, Y. Thomas: *Automatyka – układy liniowe*, PWN, Warszawa 1983, Polish translation from French
7. M. Żółtowski: *On the Problem of Divergence in Adaptive Filtering* (in progress)
8. A.V. Openheim, R.W. Schafer: *Cyfrowe przetwarzanie sygnałów*, WKiŁ, Warszawa 1975, Polish translation from English
9. F.C. Schweppe: *Układy dynamiczne w warunkach losowych*, WNT, Warszawa 1978, Polish translation from English
10. Gauss: *Theoria Motus Corporum Caelestium*, 1809
11. N. Wiener: *Extrapolation and Smoothing of Stationary Time Series with Engineering Applications*, New York, Technology Press and Wiley, 1949
12. R.E. Kalman: *A new approach to linear filtering and prediction problems.* Journal of Basic Engineering, vol. 82, pp. 34–35, 1960
13. N. Levinson: *The Wiener RMS (Root Mean Square) error criterion in filter design and prediction*, Journal of Mathematics and Physics, vol. 25, pp. 261–278, 1947
14. H. Robins, S. Monro: *A stochastic approximation method*, Ann. Math. Stat., vol. 25, pp. 382–386, 1954
15. Kailath: *A view on three decades of linear filtering theory*, IEEE Transaction on Information Theory, vol. IT-20, pp. 145–181, 1974
16. P. Eykhoff: *Identyfikacja w układach dynamicznych*, Warszawa 1980, PWN
17. J.G. Proakis: *Digital Communications*, New York, McGraw Hill, Inc., 1983
18. B. Widrow, S.D. Stearns: *Adaptive Digital Processing*, Prentice Hall, Englewood Cliffs, New York, USA, 1985
19. M. Bellanger: *Adaptive Digital Filtering and Signal Analysis*, Marcel Dekker Inc., New York, 1987
20. S. Haykin: *Adaptive Filtering Theory*, Prentice-Hall, Englewood Cliffs, New York, 1986
21. L. Ljung, T. Soderstrom: *Theory and Practice of Recursive Identification*, MIT Press, Cambridge 1983
22. L. Ljung, M. Morf, D.D. Falconer: *Fast calculation of gain matrices for recursive estimation schemes*, Int. J. of Control, 1978, vol. 27, no 1, pp. 1–19
23. G. Carayanis, D. Manolakis, N. Kaloupsidis: *Fast Kalman-type algorithms for sequential signal processing*, ICASSP 83, Boston, USA
24. J. Cioffi and T. Kailath: *Fast Recursive Least Squares Filters for Adaptive Filtering*, IEEE Trans., Vol. ASSP-32, no 2, April 1984, pp. 304–337
25. F. Ling, D. Manolakis, J. Proakis: *Numerically robust L.S. lattice ladder algorithms with direct updating of the coefficients*, IEEE Trans., vol. ASSP-34, no 4, August 1986, pp. 837–845
26. J.Z. Cypkin: *Podstawy teorii układów uczących się*, WNT, Warszawa 1973

APPENDIX 1

Matrix inversion lemma

The inverse matrix of

$$\mathbf{A} = \mathbf{B} + \mathbf{C}\mathbf{D}\mathbf{C}^T \tag{A1-1}$$

is given by

$$\mathbf{A}^{-1} = \mathbf{B}^{-1} + \mathbf{B}^{-1}\mathbf{C}(\mathbf{C}^T\mathbf{B}^{-1}\mathbf{C} + \mathbf{D}^{-1})\mathbf{C}^T\mathbf{B}^{-1} \tag{A1-2}$$

whenever $\mathbf{B}^{-1}$, $\mathbf{D}^{-1}$ exist.

Proof

$$\mathbf{I} = (\mathbf{I}+\mathbf{E})^{-1}(\mathbf{I}+\mathbf{E}) = (\mathbf{I}+\mathbf{E})^{-1} + (\mathbf{I}+\mathbf{E})^{-1}\mathbf{E} \tag{A1-3}$$

and

$$(\mathbf{I}+\mathbf{E})^{-1} = \mathbf{I} - (\mathbf{E}(\mathbf{E}^{-1}+\mathbf{I}))^{-1}\mathbf{E} = \mathbf{I} - (\mathbf{I}+\mathbf{E}^{-1})^{-1} \tag{A1-4}$$

for any nonsingular E. Moreover, the notion of inverse matrix $\mathbf{A}^{-1}_{n\times m}$ in general sense is [6]:

$$\mathbf{A}^{-1}_{n\times m} \text{ satisfies } \mathbf{A}\,\mathbf{A}^{-1}\,\mathbf{A} = \mathbf{A}. \tag{A1-5}$$

Let r be the rank of A. If

$$1^\circ \quad m > n, \quad r = n$$
$$\text{then} \quad \mathbf{A}^{-1} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T \tag{A1-6}$$

$$2^\circ \quad m < n, \quad r = m$$
$$\text{then} \quad \mathbf{A}^{-1} = \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}. \tag{A1-7}$$

That is why

$$(\mathbf{A}\mathbf{B})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1} \tag{A1-8}$$

in general sense also. For instance

$$\mathbf{A}\mathbf{B}(\mathbf{B}^{-1}\mathbf{A}^{-1})\mathbf{A}\mathbf{B} = \mathbf{A}\mathbf{B}\mathbf{B}^T(\mathbf{B}\mathbf{B}^T)^{-1}\mathbf{A}^{-1}\mathbf{A}\mathbf{B} = \mathbf{A}\mathbf{A}^{-1}\mathbf{A}\mathbf{B} = \mathbf{A}\mathbf{B} \quad \text{if} \quad \mathbf{B}^{-1} = \mathbf{B}^T(\mathbf{B}\mathbf{B}^T)^{-1}. \tag{A1-9}$$

Next

$$\mathbf{A}^{-1} = (\mathbf{B}+\mathbf{C}\mathbf{D}\mathbf{C}^T)^{-1} = (\mathbf{I}+\mathbf{B}^{-1}\mathbf{C}\mathbf{D}\mathbf{C}^T)^{-1}\mathbf{B}^{-1} = \mathbf{B}^{-1} + (\mathbf{I}+(\mathbf{B}^{-1}\mathbf{C}\mathbf{D}\mathbf{C}^T)^{-1})^{-1}\mathbf{B}^{-1}. \tag{A1-10}$$

This last equality follows from (A1-4) while $\mathbf{E} = \mathbf{B}^{-1}\mathbf{C}\mathbf{D}\mathbf{C}$. Subsequently

$$(\mathbf{B}^{-1}\mathbf{C}\mathbf{D}\mathbf{C}^T)^{-1} = (\mathbf{C}\mathbf{D}\mathbf{C}^T)^{-1} = \mathbf{C}^{T-1}(\mathbf{C}\mathbf{D})^{-1}\mathbf{B} = \mathbf{C}^{T-1}\mathbf{D}^{-1}\mathbf{C}^{-1}\mathbf{B}. \tag{A1-11}$$

That is why

$$\mathbf{A}^{-1} = \mathbf{B}^{-1} - (\mathbf{I}+\mathbf{C}^{T^{-1}}\mathbf{D}^{-1}\mathbf{C}^{-1}\mathbf{B})^{-1}\mathbf{B}^{-1} = \mathbf{B}^{-1} - (\mathbf{C}^{T^{-1}}(\mathbf{C}^T+\mathbf{D}^{-1}\mathbf{C}^{-1}\mathbf{B}))^{-1}\mathbf{B}^{-1}$$

$$= \ldots = \mathbf{B}^{-1} - \mathbf{B}^{-1}\mathbf{C}(\mathbf{C}^T\mathbf{B}^{-1}\mathbf{C}+\mathbf{D}^{-1})^{-1}\mathbf{C}^T\mathbf{B}^{-1}. \tag{A1-12}$$

The application of matrix inversion lemma to the autocorrelation matrix results in (3a) if the following substitutions are done:

$$\mathbf{A} = \mathbf{R}_{xx}(n+1); \quad \mathbf{B} = w\mathbf{R}_{xx}(n); \quad \mathbf{C} = \mathbf{X}(n+1); \quad \mathbf{D} = 1.$$

APPENDIX 2

1. The recursive form of the energy of prediction errors functionals

First

$$\mathbf{E}_f(n+1) = \sum_{k=1}^{n+1} w^{n+1-k} x(k)^2 - \mathbf{D}_f^T(n+1) r_x^f(n+1) \tag{A2-1}$$

and

$$r_x^f(n+1) = w \sum_{k=1}^{n} w^{n-k} x(k) X(k-1) + x(n+1) X(n) = w r_x^f(n) + x(n) X(n). \tag{A2-2}$$

Secondly

$$\mathbf{E}_f(n+1) = w \sum_{k=1}^{n} w^{n-k} x(k)^2 + x(n+1)^2 - (\mathbf{D}_f^T(n) + \mathbf{G}^T(n) \epsilon_1^f(n+1))(w r_x^f(n) + x(n+1) X(n)). \tag{A2-3}$$

Finally

$$\begin{aligned} \mathbf{E}_f(n+1) &= w\mathbf{E}_f(n) + x(n+1)^2 - \mathbf{D}_f^T(n) x(n) X(n) - \epsilon_1^f(n+1) \mathbf{G}^T(n) r_x^f(n+1) \\ &= w\mathbf{E}_f(n) + x(n)^2 - \mathbf{D}_f^T(n) x(n) X(n) - \epsilon_1^f(n+1) \mathbf{G}^T(n) \mathbf{R}_{xx}(n) \mathbf{D}_f(n+1) \\ &= w\mathbf{E}_f(n) + x(n+1) \epsilon_1^f(n+1) - \epsilon_1^f(n+1) X(n)^T \mathbf{D}_f(n+1) \\ &= w\mathbf{E}_f(n) + \epsilon_1^f(n+1)(x(n+1) - X(n)^T \mathbf{D}_f(n+1)) \\ &= w\mathbf{E}_f(n) + \epsilon_1^f(n+1) \epsilon_2^f(n+1). \end{aligned} \tag{A2-4}$$

On the other hand

$$\mathbf{E}_b(n+1) = \sum_{k=1}^{n+1} w^{n+1-k} x(k-N)^2 - \mathbf{D}_b^T(n+1) r_x^b(n+1) \tag{A2-5}$$

$$r_x^b(n+1) = \sum_{k=1}^{n+1} w^{n+1-k} x(k-N) X(k) = w r_x^b(n) + x(n+1-N) X(n+1). \tag{A2-6}$$

That is why

$$\begin{aligned} \mathbf{E}_b(n+1) &= w \sum_{k=1}^{n} w^{n-k} x(k-N)^2 + (\mathbf{D}_b(n)^T + \mathbf{G}(n+1)^T \epsilon_1^b(n+1)) \times \\ &\times (w r_x^b(n) + x(n+1-N) X(n+1) + x(n+1-N)^2 \\ &= w\mathbf{E}_b(n) - \mathbf{G}(n+1)^T \epsilon_1^b(n+1) r_x^b(n+1)) - \mathbf{D}_b(n)^T x(n+1-N) X(n+1) + \\ &+ x(n+1-N)^2 = w\mathbf{E}_b(n) + x(n+1-N) \epsilon_1^b(n+1) - \mathbf{G}(n+1)^T r_x^b(n+1) \epsilon_1^b(n+1) \\ &= w\mathbf{E}_b(n) + \epsilon_1^b(n+1)(x(n+1-N) - X(n+1)^T \mathbf{D}_b(n+1)\} \\ &= w\mathbf{E}_b(n) + \epsilon_1^b(n+1) \epsilon_2^b(n+1). \end{aligned} \tag{A2-7}$$

2. The physical meaning of parameter $r(n)$ and its estimations

The parameter $r(n)$ as *a posteriori* to *a priori* prediction errors ratio should satisfy

$$|r(n)| \leqslant 1. \tag{A2-8}$$

Moreover (8a)

$$r(n) = 1 - \mathbf{G}^T(n)X(n) = 1 - X(n)^T\mathbf{R}_{xx}^{-1}(n)X(n) \tag{A2-9}$$

and

$$X(n+1)^T\mathbf{R}_{xx}^{-1}(n+1)X(n+1) = \frac{X(n+1)^T\mathbf{R}_{xx}^{-1}(n)X(n+1)}{w + X^T(n+1)\mathbf{R}_{xx}^{-1}(n)X(n+1)} = 1 - r(n+1). \tag{A2-10}$$

Thus

$$r(n+1) = \frac{w}{w + X^T(n+1)\mathbf{R}_{xx}^{-1}(n)X(n+1)}. \tag{A2-11}$$

Now, let

$$r_0(n+1) = X^T(n+1)\mathbf{R}_{xx}^{-1}(n+1)X(n+1) \stackrel{\text{df}}{=} \|X(n+1)\|^2_{\mathbf{R}_{xx}^{-1}} \tag{A2-12}$$

where the right hand side of (A2-12) is the metric weighted by $\mathbf{R}_{xx}^{-1}$.

Suppose that the probability density function of N-dimensional Gaussian variable, necessary to use the likelihood principle is

$$p(x) = \frac{1}{\sqrt{(2\pi^N\ |\mathbf{R}_{xx}(n+1)|}} \exp(-\frac{1}{2}\|x\|^2_{\mathbf{R}_{xx}^{-1}}), \tag{A2-13}$$

where $|\mathbf{R}_{xx}|$ stands for the determinant of $\mathbf{R}_{xx}$.

The value of $r_0(n)$ is small if the input data are likely and great otherwise. The upper bound to $r_0(n)$ can be estimated as follows:

First

$$r_0(n+1) = X^T(n+1)\mathbf{R}_{xx}(n+1)^{-1}X(n+1) = X^T(n+1)\{w\mathbf{R}_{xx}(n) + X(n+1)X^T(n+1)^{-1}X(n+1)$$

$$\leqslant X^T(n+1)(X(n+1)X^T(n+1)\}^{-1}X(n+1) \tag{A2-14}$$

since $\mathbf{R}_{xx}(n)$ is positive definite (see also further remark on $\mathbf{R}_{xx}$ as information matrix).

Secondly (x vector)

$$xx^Tx = x\,\|x\|^2 = \|x\|^2x \tag{A2-15a}$$

$$(xx^T)^{-1}x = \|x\|^{2^{-1}}x. \tag{A2-15b}$$

Finally

$$0 \leqslant r_0(n) \leqslant X^T(n+1)\,\|X(n+1)\|^{2^{-1}}\,X(n+1) = 1. \tag{A2-15c}$$

The estimation of $r(n+1)$ can be also found from (2) and (A2-9). Namely

$$X(n+1) = w\mathbf{R}_{xx}(n)\mathbf{R}_{xx}^{-1}(n+1)X(n+1) + X(n+1)X^T(n+1)\mathbf{R}_{xx}^{-1}(n+1)X(n+1)$$

$$= w\mathbf{R}_{xx}(n)\mathbf{R}_{xx}^{-1}(n+1)X(n+1) + (1 - r(n+1))X(n+1) \tag{A2-15d}$$

or

$$r(n+1)X(n+1) = w\mathbf{R}_{xx}(n)\mathbf{R}_{xx}^{-1}(n+1)X(n+1). \tag{A2-15e}$$

That is why

$$r(n+1)\mathbf{I} = w\mathbf{R}_{xx}(n)\mathbf{R}_{xx}^{-1}(n+1)\mathbf{I} \tag{A2-15f}$$

and

$$r(n+1) = w\sqrt[N]{|\mathbf{R}_{xx}(n)|/|\mathbf{R}_{xx}(n+1)|} \tag{A2-15g}$$

While $\mathbf{R}_{xx}(n)$ is the information matrix in Fisher's sense [9] it is easy to notice that

$$0 \leqslant r(n+1) \leqslant 1 \tag{A2-15h}$$

holds true. This is in accord with the previous estimation of $r_0(n)$.

APPENDIX 3

1. The derivation of the (12) of Chapter V

First

$$\mathbf{R}_{xx}^{N+1}(n+1) = \sum_{k=1}^{n+1} w^{n+1-k} \begin{bmatrix} x(k) \\ X^N(k-1) \end{bmatrix} [x(k)\ X^N(k-1)]^T =$$

$$= \sum_{k=1}^{n+1} w^{n+1-k} \begin{bmatrix} X^N(k) \\ x(k-N) \end{bmatrix} [X^N(k)^T x(k-N)] =$$

$$= \begin{bmatrix} \sum_{k=1}^{n+1} w^{n+1-k} x(k)^2 & r_x^{f,N}(n+1)^T \\ r_x^{f,N}(n+1) & \mathbf{R}_{xx}^N(n) \end{bmatrix} = \begin{bmatrix} \mathbf{R}_{xx}^N(n+1) & r_x^{b,N}(n+1) \\ r_x^{b,N}(n+1)^T & \sum_{k=1}^{n+1} w^{n+1-k} x(k-N)^2 \end{bmatrix}. \tag{A3-1}$$

Secondly the expressions (12) are valid since

$$\mathbf{R}_{xx}(n)\ \mathbf{G}(n) = X(n). \tag{A3-2}$$

2. The derivation of the (13) of Chapter V

First

$$r_x^{f,N}(n+1)^T\ \mathbf{G}(n) = (\mathbf{R}_{xx}^{-1}(n)\mathbf{D}_f(n+1))^T \mathbf{G}(n) = \mathbf{D}_f^T(n+1)X(n) \tag{A3-3a}$$

and

$$r_x^{b,N}(n+1)^T\ \mathbf{G}(n+1) = (\mathbf{R}_{xx}^{-1}(n+1)\mathbf{D}_b(n+1))^T \mathbf{G}(n+1) = \mathbf{D}_b^T(n+1)X(n+1). \tag{A3-3b}$$

Secondly

$$\mathbf{R}_{xx}^{N+1}(n+1) \begin{bmatrix} 0 \\ \mathbf{G}(n) \end{bmatrix} = \begin{bmatrix} \mathbf{D}_f^T(n+1)(X(n) \\ X(n) \end{bmatrix} = X^{N+1}(n+1) - \begin{bmatrix} \epsilon_2^f(n+1) \\ 0 \end{bmatrix} \tag{A3-4a}$$

and

$$\mathbf{R}_{xx}^{N+1}(n+1) \begin{bmatrix} \mathbf{G}(n+1) \\ 0 \end{bmatrix} = \begin{bmatrix} X(n+1) \\ \mathbf{D}_b^T(n+1)X(n+1) \end{bmatrix} = X^{N+1}(n+1) - \begin{bmatrix} 0 \\ \epsilon_2^b(n+1) \end{bmatrix}. \tag{A3-4b}$$

3. The derivation of the (16) of Chapter V

First in view of (A3-1) and (A2-1)

$$\mathbf{R}_{xx}^{N+1}(n+1) \begin{bmatrix} 1 \\ -\mathbf{D}_f^N(n+1) \end{bmatrix} = \begin{bmatrix} \mathbf{E}_f(n+1) \\ r_x^{f,N}(n+1) - \mathbf{R}_{xx}^N(n)\mathbf{D}_f^N(n+1) \end{bmatrix} = \begin{bmatrix} \mathbf{E}_f(n+1) \\ 0 \end{bmatrix}. \tag{A3-5a}$$

Secondly, in view of (A3-1) and (A2-5)

$$\mathbf{R}_{xx}^{N+1}(n+1)\begin{bmatrix} -\mathbf{D}_b^N(n+1) \\ 1 \end{bmatrix}=\begin{bmatrix} -\mathbf{R}_{xx}^N(n+1)\mathbf{D}_b^N(n+1)+r_x^{b,N}(n+1) \\ \mathbf{E}_b(n+1) \end{bmatrix}=\begin{bmatrix} 0 \\ \mathrm{E}_b(n+1) \end{bmatrix}. \tag{A3-5b}$$

4. Misscellaneous

First, in view of (7b) and (19b)

$$\mathrm{D}_b(n+1)=\frac{\mathrm{D}_b(n)+\mathrm{G}_N^{N+1}(n+1)\epsilon_1^b(n+1)}{1-g(n+1)\epsilon_1^b(n+1)}. \tag{A3-6}$$

Secondly

$$1-g(n+1)\epsilon_1^b(n+1)\underset{(18b)}{=}1-\frac{\epsilon_1^b(n+1)\epsilon_2^b(n+1)}{\mathbf{E}_b(n+1)}=$$

$$\underset{(9b)}{=}1-\frac{\epsilon_1^b(n+1)\epsilon_2^b(n+1)}{w\mathbf{E}_b(n+1)+\epsilon_1^b(n+1)\epsilon_2^b(n+1)}=\frac{w\mathbf{E}_b(n)}{\mathbf{E}_b(n+1)}. \tag{A3-7}$$

That is why

$$0<1-g(n+1)\epsilon_1^b(n+1)\leqslant 1. \tag{A3-8}$$

The expression among the inequality marks approaches 1 when the error of prediction tends to zero.

APPENDIX 4

1. The derivation of (24a) and (24b)

First, from (A3-1) and (23)

$$\mathbf{R}_{xx}^{N+1}(n)\begin{bmatrix} \mathbf{G}_2^N(n) \\ 0 \end{bmatrix}=\begin{bmatrix} \mathbf{X}^N(n+1) \\ r_x^{b,N}(n)^T\mathbf{G}_2^N(n+1) \end{bmatrix}\begin{bmatrix} \mathbf{X}^N(n+1) \\ [\mathbf{R}_{xx}^N(n)\mathbf{D}_b(n)]^T\mathbf{G}_2^N(n+1) \end{bmatrix}=$$

$$=\mathbf{R}_{xx}^{N+1}(n)\mathbf{G}_2^{N+1}(n+1)-\begin{bmatrix} 0 \\ \epsilon_1^b(n+1) \end{bmatrix}. \tag{A4-1a}$$

Secondly

$$\mathbf{R}_{xx}^{N+1}(n)\begin{bmatrix} 0 \\ \mathbf{G}_2^N(n) \end{bmatrix}=\begin{bmatrix} r_x^{f,N}(n)^T\mathbf{G}_2^N(n) \\ \mathbf{X}^N(n) \end{bmatrix}=\begin{bmatrix} \mathbf{D}_f(n)^T\mathbf{X}^N(n) \\ \mathbf{X}^N(n) \end{bmatrix}=$$

$$=\mathbf{X}^{N+1}(n+1)-\begin{bmatrix} \epsilon_1^f(n+1) \\ 0 \end{bmatrix}=\mathbf{R}_{xx}^{N+1}(n)\mathbf{G}_2^{N+1}(n+1)-\begin{bmatrix} \epsilon_1^f(n+1) \\ 0 \end{bmatrix}. \tag{A4-1b}$$

Eventually (24a) and (24b) hold true.

2. The derivation of (26)

First

$$\mathbf{R}_{xx}^{N}(n+1)\mathbf{D}(n+1) = r_{yx}(n+1). \tag{A4-2}$$

Secondly

$$(w\mathbf{R}_{xx}^{N}(n) + X(n+1)X^{T}(n+1))\mathbf{D}(n+1) = wr_{yx}(n) + y(n+1)X(n+1) = \tag{A4-3}$$
$$= w\mathbf{R}_{xx}^{N}(n)\mathbf{D}(n) + y(n+1)X(n+1).$$

Thus

$$\mathbf{D}(n+1) = \mathbf{D}(n) + w^{-1}\mathbf{R}_{xx}^{N-1}(n)X(n+1)\{y(n+1) - X^{T}(n+1)\mathbf{D}(n+1)\} \tag{A4-4}$$

holds true.

3. The derivation of (30)

First from (A4-4) and (21a) ($\mathbf{G} = \mathbf{G}_1$)

$$w^{-1}\mathbf{R}_{xx}^{N-1}(n)X(n+1)\epsilon_2(n+1) = \mathbf{G}_1^{N}(n+1)\epsilon_1(n+1) = w^{-1}\mathbf{G}_2^{N}(n+1)\epsilon_2(n+1). \tag{A4-5}$$

Thus

$$\mathbf{G}_1^{N}(n) = \mathbf{G}_2^{N}(n+1)\frac{r(n+1)}{w} = \mathbf{G}_2^{N}(n+1)/r'(n+1). \tag{A4-6}$$

Eventually

$$\mathbf{D}(n+1) = \mathbf{D}(n) + \frac{\mathbf{G}_2^{N}(n+1)}{r'(n+1)}\epsilon_1(n+1) \tag{A4-7}$$

holds true.

4. The derivation of (31)

First, from (27), (25a)

$$r'^{N+1}(n+1) = w + X^{N+1}(n+1)^{T}\mathbf{G}_2^{N+1}(n+1) =$$

$$= w + X^{N+1}(n+1)^{T}\left[\begin{bmatrix}\mathbf{G}_2^{N}(n+1)\\ 0\end{bmatrix} + \frac{\epsilon_1^{b}(n+1)}{\mathbf{E}_b(n)}\begin{bmatrix}-\mathbf{D}_b(n)\\ 1\end{bmatrix}\right] =$$

$$= w + X^{N}(n+1)^{T}\mathbf{G}_2^{N}(n+1) + \frac{\epsilon_1^{b}(n+1)}{\mathbf{E}_b(n)}\{x(n+1-N) - X^{N}(n+1)^{T}\mathbf{D}_b(n)\} =$$

$$= r'^{N}(n+1) + \frac{\epsilon_1^{b}(n+1)}{\mathbf{E}_b(n)}. \tag{A4-8}$$

Secondly, in view of (25b)

$$r'^{N+1}(n+1) = w + X^{N+1}(n+1)\left[\begin{bmatrix}0\\ \mathbf{G}_2^{N}(n)\end{bmatrix} + \frac{\epsilon_1^{f}(n+1)}{\mathbf{E}_f(n)}\begin{bmatrix}1\\ -\mathbf{D}_f(n)\end{bmatrix}\right] =$$

$$= r'^{N}(n) + \frac{\epsilon_1^{f}(n+1)^2}{\mathbf{E}_f(n)}. \tag{A4-9}$$

APPENDIX 5

The algorithm of lattice filter

1. The derivation of (35)

First, from (A3-5)

$$\begin{bmatrix} \mathbf{R}_{xx}^{N}(n) & r_x^{b,N}(n) \\ r_x^{b,N}(n)^T & \sum_{k=1}^{n} w^{n-k}x(k-N)^2 \end{bmatrix}\begin{bmatrix} 1 \\ -\mathbf{D}_f(n) \\ 0 \end{bmatrix} = \begin{bmatrix} \mathbf{E}_f^{N-1}(n) \\ 0 \\ \mathbf{K}_1^{N}(n) \end{bmatrix} \tag{A5-1a}$$

$$\mathbf{R}_{xx}^{N+1}(n)$$

$$\begin{bmatrix} \sum_{k=1}^{n} w^{n-k}x(k)^2 & r_x^{f,N}(n)^T \\ r_x^{f,N}(n) & \mathbf{R}_{xx}^{N}(n+1) \end{bmatrix}\begin{bmatrix} 0 \\ -\mathbf{D}_b^{N-1}(n-1) \\ 1 \end{bmatrix} = \begin{bmatrix} \mathbf{K}_2^{N}(n) \\ 0 \\ \mathbf{E}_b^{N-1}(n-1) \end{bmatrix}. \tag{A5-1b}$$

$$\mathbf{R}_{xx}^{N+1}(n)$$

Secondly

$$\mathbf{K}_1(n) = r_x^{b,N}(n)^T\begin{bmatrix} 1 \\ -\mathbf{D}_f^{N-1}(n) \end{bmatrix} = \left[\sum_{k=1}^{n} w^{n-k}x(k-N)X(k)\right]^T\begin{bmatrix} 1 \\ -\mathbf{D}_f^{N-1}(n) \end{bmatrix} =$$

$$= \sum_{k=1}^{n} w^{n-k}x(k-N)x(k) - \mathbf{D}_f^{N-1}(n)^T\sum_{k'=0}^{n-1} w^{n-k'-1}x(k'-(N-1))X^{N-1}(k') =$$

$$= \sum_{k=1}^{n} w^{n-k}x(k-N)x(k) - \mathbf{D}_f^{N-1}(n)^T r_x^{b,N-1}(n-1) =$$

$$= \sum_{k=1}^{n} w^{n-k}x(k-N)x(k) - \mathbf{D}_f^{N-1}(n)^T\mathbf{R}_{xx}^{N-1}(n-1)\mathbf{D}_b^{N-1}(n-1) \tag{A5-2a}$$

and

$$\mathbf{K}_2^{N}(n) = r_x^{f,N}(n)^T\begin{bmatrix} -\mathbf{D}_b^{N-1}(n-1) \\ 1 \end{bmatrix} = \left[\sum_{k=1}^{n} w^{n-k}x(k)X(k-1)\right]^T\begin{bmatrix} -\mathbf{D}_b^{N-1}(n-1) \\ 1 \end{bmatrix} =$$

$$= \sum_{k=1}^{n} w^{n-k}x(k-N)x(k) - \mathbf{D}_f^{N-1}(n)^T r_x^{b,N-1}(n-1) = \sum_{k=1}^{n} w^{n-k}x(k-N)x(k) +$$

$$- \mathbf{D}_f^{N-1}(n)^T\mathbf{R}_{xx}^{N-1}(n-1)\mathbf{D}_b^{N-1}(n-1). \tag{A5-2b}$$

That is why the expressions (35) hold true.

2. The drivation of (36) and (37)

By subtracting (35b) from (A5-1a) and (35a) from (A5-1b) and equating to (A3-5a) and (A3-5b) respectively one obtains

$$\mathbf{R}_{xx}^{N+1}(n)\begin{bmatrix} 1 \\ -\mathbf{D}_f^{N-1}(n)+\mathbf{D}_b^{N-1}(n-1)\dfrac{\mathbf{K}^N(n)}{\mathbf{E}_b^{N-1}(n-1)} \\ -\mathbf{K}^N(n)/\mathbf{E}_b^{n-1}(n-1) \end{bmatrix} = \begin{bmatrix} \mathbf{E}_f^{N-1}(n)-\dfrac{\mathbf{K}^N(n)^2}{\mathbf{E}_b^{N-1}(n-1)} \\ \mathbf{0} \\ 0 \end{bmatrix}$$

$$\overset{=}{(\text{A3-5a})}\ \mathbf{R}_{xx}^{N+1}(n)\begin{bmatrix} 1 \\ -\mathbf{D}_f^N(n) \end{bmatrix} = \begin{bmatrix} \mathbf{E}_f^N(n) \\ \mathbf{0} \\ 0 \end{bmatrix} \tag{A5-3a}$$

and

$$\mathbf{R}_{xx}^{N+1}(n)\begin{bmatrix} -\mathbf{K}^N(n)/\mathbf{E}_f^{N-1}(n) \\ -\mathbf{D}_b^{N-1}(n-1)+\mathbf{D}_f^{N-1}(n)\dfrac{\mathbf{K}^N(n)}{\mathbf{E}_f^{N-1}(n)} \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{0} \\ \mathbf{E}_b^{N-1}(n-1)-\dfrac{\mathbf{K}^N(n)^2}{\mathbf{E}_f^{N-1}(n-1)} \end{bmatrix}$$

$$\overset{=}{(\text{A3-5b})}\ \mathbf{R}_{xx}^{N+1}(n)\begin{bmatrix} -\mathbf{D}_b^N(n) \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{0} \\ \mathbf{E}_b^N(n) \end{bmatrix}. \tag{A5-3b}$$

Next, one can notice that (36) and (37) follow.

3. The derivation of (38) and (39)

First, from (36a)

$$\epsilon_1^{f,N}(n+1) = x(n+1) - \left[\begin{bmatrix} \mathbf{D}_f^{N-1}(n) \\ 0 \end{bmatrix} + \frac{\mathbf{K}^N(n)}{\mathbf{E}_b^{N-1}(n-1)}\begin{bmatrix} -\mathbf{D}_b^{N-1}(n-1) \\ 1 \end{bmatrix}\right]^T X^N(n) =$$

$$= x(n+1) - \mathbf{D}_f^{N-1}(n)X^{N-1}(n) - \frac{\mathbf{K}^N(n)}{\mathbf{E}_b^{N-1}(n-1)}\left[-\mathbf{D}_b^{N-1}(n-1)^T 1\right]X^N(n) =$$

$$= \epsilon_1^{f,N-1}(n+1) - \frac{\mathbf{K}^N(n)}{\mathbf{E}_b^{N-1}(n-1)}\left[x(n-(N-1)) - \mathbf{D}_b^{N-1}(n-1)^T X^{N-1}(n)\right] =$$

$$= \epsilon_1^{f,N-1}(n+1) - \frac{\mathbf{K}^N(n)}{\mathbf{E}_b^{N-1}(n-1)}\epsilon_1^{b,N-1}(n) \tag{A5-4a}$$

and

$$\epsilon^{b,N}(n+1) = x(n+1-N) - \left[\begin{bmatrix} 0 \\ \mathbf{D}_b^{N-1}(n-1) \end{bmatrix} + \frac{\mathbf{K}^N(n)}{\mathbf{E}_f^{N-1}(n)} \begin{bmatrix} 1 \\ -\mathbf{D}_f(n) \end{bmatrix}\right]^T X^N(n+1) =$$

$$= x(n+1-N) - [0\ \mathbf{D}_b^{N-1}(n-1)^T]X^N(n+1) - \frac{\mathbf{K}^N(n)}{\mathbf{E}_f^{N-1}(n)} [1\ -\mathbf{D}_f^{N-1}(n)^T]X^{N+1}(n+1) =$$

$$= x(n-(N-1)) - \mathbf{D}_b^{N-1}(n-1)^T X^{N-1}(n) - \frac{\mathbf{K}^N(n)}{\mathbf{E}_f^{N-1}(n)} (x(n+1) - \mathbf{D}_f^{N-1}(n)^T X^{N-1}(n))$$

$$= \epsilon_1^{b,N-1}(n) - \frac{\mathbf{K}^N(n)}{\mathbf{E}_f^{N-1}(n)} \epsilon_1^{f,N-1}(n+1). \tag{A5-4b}$$

Secondly, the *a posteriori* errors are

$$\epsilon_2^{f,N}(n+1) \stackrel{\text{df}}{=} x(n+1) - \mathbf{D}_f(n+1)^T X^N(n) \tag{A5-5a}$$

$$\epsilon_2^{b,N}(n+1) \stackrel{\text{df}}{=} x(n+1-N) - \mathbf{D}_b(n+1)^T X^N(n+1). \tag{A5-5b}$$

Eventually (36a,b) hold true and (37) can be derived just the same way as shown.

4. The derivation of (42)

First

$$\epsilon_1^N(n+1) \stackrel{\text{df}}{=} y(n+1) - \mathbf{D}^N(n)^T X(n+1) \tag{A5-6a}$$

$$\mathbf{D}^N(n) = \mathbf{R}_{xx}^N(n)^{-1} r_{yx}^N(n) \tag{A5-6b}$$

$$r_{yx}^N(n) = \sum_{k=1}^{n} w^{n-k} y(k) X^N(k). \tag{A5-6c}$$

Secondly

$$\mathbf{R}_{xx}^{N+1}(n) = r_{yx}^{N+1}(n) = \begin{bmatrix} r_{yx}^N(n) \\ \sum_{k=1}^{n} w^{n-k} y(k) X(k-N) \end{bmatrix} \tag{A5-7}$$

and

$$\mathbf{R}_{xx}^{N+1}(n) \begin{bmatrix} \mathbf{D}^N(n) \\ 0 \end{bmatrix} = \begin{bmatrix} \mathbf{R}_{xx}^N(n) & r_x^{b,N}(n) \\ r_x^{b,N}(n)^T & \sum_{k=1}^{n} w^{n-k} X(k-N)^2 \end{bmatrix} = \begin{bmatrix} r_{yx}^N(n) \\ \mathbf{D}_b^N(n)^T r_{yx}^N(n) \end{bmatrix}. \tag{A5-8}$$

That is why (42a) holds true with $\mathbf{K}_0^N(n)$ given by (42c).

Also

$$\mathbf{R}_{xx}^{N+1}(n)\begin{bmatrix} -\mathbf{D}_b(n) \\ 1 \end{bmatrix} \underset{(16b)}{=} \begin{bmatrix} 0 \\ \mathbf{E}_b^N(n) \end{bmatrix} \tag{A5-9}$$

That is why (42b) holds true in view of (42a).

5. The derivation of (43)

First, in view of (A5-6a) and (42b)

$$\begin{aligned} \epsilon_1^{N+1}(n+1) &= y(n+1) - \begin{bmatrix} \mathbf{D}^N(n) \\ 0 \end{bmatrix}^T X^{N+1}(n+1) + \frac{\mathbf{K}_0^N(n)}{\mathbf{E}_b^N(n)}\begin{bmatrix} -\mathbf{D}_b^N(n) \\ 1 \end{bmatrix}^T X^{N+1}(n+1) \\ &= \epsilon_1^N(n+1) - \frac{\mathbf{K}_0^N(n)}{\mathbf{E}_b^N(n)}\left[x(n+1-N) - \mathbf{D}_b(n)^T X^{N+1}(n+1)\right] \\ &= \epsilon_1^N(n+1) - \frac{\mathbf{K}_0^N(n)}{\mathbf{E}_0^N(n)}\epsilon_1^{b,N}(n+1). \end{aligned} \tag{A5-10a}$$

Secondly

$$\begin{aligned} \epsilon_2^{N+1}(n+1) &= y(n+1) - \left[\begin{bmatrix} \mathbf{D}^N(n+1) \\ 0 \end{bmatrix} - \frac{\mathbf{K}_0^N(n+1)}{\mathbf{E}_b^N(n+1)}\begin{bmatrix} -\mathbf{D}_b^N(n+1) \\ 1 \end{bmatrix}\right]^T X^{N+1}(n+1) \\ &= \epsilon_2^N(n+1) - \frac{\mathbf{K}_0^N(n)}{\mathbf{E}_b^N(n)}\left[x(n+1-N) - \mathbf{D}_b(n+1)^T X^{N+1}(n+1)\right] \\ &= \epsilon_2^N(n+1) - \frac{\mathbf{K}_0^N(n)}{\mathbf{E}_b^N(n)}\epsilon_2^{b,N}(n+1). \end{aligned} \tag{A5-10b}$$

Next

$$\begin{aligned} \mathbf{E}(n) &= \sum_{k=1}^{n} w^{n-k}[y(n) - \mathbf{D}(n)^T X(k)]^2 = \sum_{k=1}^{n} w^{n-k} y(n)^2 - 2r_{yx}(n)^T \mathbf{D}(n) + \\ &+ \mathbf{D}(n)^T \mathbf{R}_{xx}(n)\mathbf{D}(n) = \sum_{k=1}^{n} w^{n-k} y(k)^2 - \mathbf{D}(n)^T \mathbf{R}_{xx}(n)\mathbf{D}(n). \end{aligned} \tag{A5-11}$$

The dependence on "n" is omitted in further derivation in order to avoid proliferation:

$$\begin{aligned} \mathbf{E}_{N+1}(n) &= \sum_{k=1}^{n} w^{n-k} y(k)^2 - \left[\begin{bmatrix} \mathbf{D}^N \\ 0 \end{bmatrix} - \frac{\mathbf{K}_0^N}{\mathbf{E}_b^N}\begin{bmatrix} -\mathbf{D}_b \\ 1 \end{bmatrix}\right]^T \mathbf{R}_{xx}^{N+1} \times \\ &\times \mathbf{R}_{xx}^{N+1}\left[\begin{bmatrix} \mathbf{D}^N \\ 0 \end{bmatrix} - \frac{\mathbf{K}_0^N}{\mathbf{E}_b^N}\begin{bmatrix} -\mathbf{D}_b \\ 1 \end{bmatrix}\right] \\ &= \sum_{k=1}^{n} w^{n-k} y(k)^2 - [\mathbf{D}^{N^T}\,0]\mathbf{R}_{xx}^{N+1}\begin{bmatrix} \mathbf{D}^N \\ 0 \end{bmatrix} + [\mathbf{D}^{N^T}\,0]\mathbf{R}_{xx}^{N+1}\begin{bmatrix} -\mathbf{D}_b \\ 1 \end{bmatrix}\frac{\mathbf{K}_b^N}{\mathbf{E}_b^N} + \\ &+ \frac{\mathbf{K}_0^N}{\mathbf{E}_b^N}[-\mathbf{D}_b^T\,1]\mathbf{R}_{xx}^{N+1}\begin{bmatrix} \mathbf{D}^N \\ 0 \end{bmatrix} - \frac{\mathbf{K}_0^{N^2}}{\mathbf{E}_b^{N^2}}[-\mathbf{D}_b^T\,1]\mathbf{R}_{xx}^{N+1}\begin{bmatrix} -\mathbf{D}_b \\ 1 \end{bmatrix} \end{aligned} \tag{A5-12}$$

In view of $\mathbf{R}_{xx}^{n}$ partition

$$\mathbf{R}_{xx}^{N}=\begin{bmatrix}\mathbf{R}_{xx}^{N} & r_x^{b,N}\\ (r_x^{b,N})^T & \sum_{k=1}^{n} w^{n-k}x(k-N)^2\end{bmatrix} \tag{A5-13a}$$

$$[\mathbf{D}^{N^T}\,0]\mathbf{R}_{xx}^{N+1}\begin{bmatrix}\mathbf{D}^N\\ 0\end{bmatrix}=\mathbf{D}^{N^T}\mathbf{R}_{xx}^{N}\mathbf{D}^N. \tag{A5-13b}$$

Moreover

$$[\mathbf{D}^{N^T}\,0]\mathbf{R}_{xx}^{N+1}\begin{bmatrix}-\mathbf{D}_b\\ 1\end{bmatrix}\frac{\mathbf{K}_0^N}{\mathbf{E}_b^N}=[\mathbf{D}^{N^T}\mathbf{R}_{xx'}\,\mathbf{D}^{N^T}r_x^{b,N}]\begin{bmatrix}-\mathbf{D}_b^N\\ 1\end{bmatrix}\frac{\mathbf{K}_0^N}{\mathbf{E}_b^N}=$$

$$=\mathbf{D}^{N^T}(r_x^{b,N}-\mathbf{R}_{xx}^{N}\mathbf{D}_b^N)\frac{\mathbf{K}_0^N}{\mathbf{E}_b^N}=[-\mathbf{D}_b^{N^T}\,1]\mathbf{R}_{xx}^{N+1}\begin{bmatrix}\mathbf{D}^N\\ 0\end{bmatrix}=0. \tag{A5-13c}$$

Also

$$[-\mathbf{D}_b^{N^T}\,1]\mathbf{R}_{xx}^{N+1}\begin{bmatrix}-\mathbf{D}_b^N\\ 1\end{bmatrix}=$$

$$\underset{\text{(A5-13a)}}{=}[-\mathbf{D}_b^{N^T}\mathbf{R}_{xx}^{N}+r_b^{b,N^T}-\mathbf{D}_b^{N^T}r_x^{b,N}+\sum_{k=1}^{n}w^{n-k}x^2(k-N)]\begin{bmatrix}-\mathbf{D}_b^N\\ 1\end{bmatrix}=$$

$$=\mathbf{D}_b^{N^T}\mathbf{R}_{xx}^{N}\mathbf{D}_b^N-r_x^{b,N^T}\mathbf{D}_b^N-\mathbf{D}_b^{N^T}r_x^{b,N}+\sum_{k=1}^{n}w^{n-k}x^2(k-N)=$$

$$=\sum_{k=1}^{n}w^{n-k}x^2(k-N)-\mathbf{D}_b^{N^T}\mathbf{R}_{xx}^{N}\mathbf{D}_b^N=\mathbf{E}_b^N. \tag{A5-13d}$$

That is why

$$\mathbf{E}_{N+1}(n)=\mathbf{E}_N(n)-\frac{\mathbf{K}_0^N(n)^2}{\mathbf{E}_b^N(n)}. \tag{A5-14}$$

6. The derivation of (46)

First

$$\mathbf{R}_{xx}^{N}(n)\begin{bmatrix}G^{N-1}(n)\\ 0\end{bmatrix}=\begin{bmatrix}\mathbf{R}_{xx}^{N-1}(n) & r_x^{b,N-1}\\ r_x^{b,N-1}(n)^T & \sum_{k=1}^{n}w^{n-k}x^2(k-N)\end{bmatrix}\begin{bmatrix}G^{N-1}(n)\\ 0\end{bmatrix}=$$

$$=\begin{bmatrix}X^{N-1}(n)\\ r_x^{b,N-1}(n)^TG^{N-1}(n)\end{bmatrix}=\begin{bmatrix}X^{N-1}(n)\\ \mathbf{D}_b^{N-1}(n)^T\mathbf{R}_{xx}^{N-1}(n)G^{N-1}(n)\end{bmatrix}=$$

$$=\begin{bmatrix}X^{N-1}(n)\\ \mathbf{D}_b^{N-1}(n)^TX^{N-1}(n)\end{bmatrix}=\begin{bmatrix}X^{N-1}(n)\\ x(n+1-N)-\epsilon_2^{b,N-1}(n)\end{bmatrix}=$$

$$=X^N(n)-\begin{bmatrix}0\\ \epsilon_2^{b,N-1}(n)\end{bmatrix} \tag{A5-15a}$$

and

$$\begin{bmatrix} \mathbf{G}^{N-1}(n) \\ 0 \end{bmatrix} = \mathbf{G}^N(n) - \mathbf{R}_{xx}^N(n)^{-1} \begin{bmatrix} 0 \\ \epsilon_2^{b,N-1}(n) \end{bmatrix}. \tag{A5-15b}$$

However

$$\mathbf{R}_{xx}^N(n) \begin{bmatrix} -\mathbf{D}_b^N(n) \\ 1 \end{bmatrix} \underset{(16b)}{=} \begin{bmatrix} 0 \\ \mathbf{E}_b^{N-1}(n) \end{bmatrix}. \tag{A5-16}$$

That is why

$$\mathbf{G}^N(n) = \begin{bmatrix} \mathbf{G}_{(n)}^{N-1} \\ 0 \end{bmatrix} + \frac{\epsilon_2^{b,N-1}(n)}{\mathbf{E}_b^{N-1}(n)} \begin{bmatrix} -\mathbf{D}_b^{N-1}(n) \\ 1 \end{bmatrix}. \tag{A5-17}$$

7. The derivation of (47)

First

$$\begin{aligned} \epsilon_2^N(n) &= y(n) - \mathbf{D}^N(n)^T X^N(n) = y(n) - \mathbf{D}^N(n-1)^T X^N(n) + \\ &+ (\mathbf{D}^N(n-1) - \mathbf{D}^N(n))^T X^N(n) \\ &= \epsilon_1^N(n) - \mathbf{G}^N(n)^T \epsilon_1^N(n) X^N(n). \end{aligned} \tag{A5-18}$$

Secondly, in view of (A2-9)

$$\frac{\epsilon_2^N(n)}{\epsilon_1^N(n)} = 1 - \mathbf{G}^N(n)^T X^N(n) = r^N(n). \tag{A5-19}$$

Eventually

$$r^N(n) - r^{N-1}(n) = \mathbf{G}^N(n)^T X^{N-1}(n) + G^{N-1}(n)^T X^{N-1}(n) =$$

$$= -\begin{bmatrix} \mathbf{G}^{N-1}(n) \\ 0 \end{bmatrix}^T X^N(n) - \frac{\epsilon_2^{b,N-1}(n)}{\mathbf{E}_b^{N-1}(n)} \begin{bmatrix} -\mathbf{D}_b^{N-1}(n) \\ 1 \end{bmatrix} X^N(n) +$$

$$+ \mathbf{G}^{N-1}(n)^T X^{N-1}(n) = -\frac{\epsilon_2^{b,N-1}(n)}{\mathbf{E}_b^{N-1}(n)} [x(n+1-N) - \mathbf{D}_b^{N-1}(n)^T X^{N-1}(n)] =$$

$$= -\frac{\epsilon_2^{b,N-1}(n)^2}{\mathbf{E}_b^{N-1}(n)}. \tag{A5-20a}$$

Moreover the initial condition is

$$r^1(n) = r^0(n) - \frac{\epsilon_2^b(n)^2}{\mathbf{E}_b^0(n)} \underset{r^0(n)=1}{=} 1 - \frac{x(n)^2}{\sum_{k=1}^{n} w^{n-k} x^2(k)}. \tag{A5-20b}$$

8. The derivation of (49)

From the definition of lattice coefficient $\mathbf{K}^{N+1}(n+1)$

$$\mathbf{K}^{N+1}(n+1) = \sum_{k=1}^{n+1} w^{n+1-k}x(k-N-1)x(k) - \mathbf{D}_f^N(n+1)^T\mathbf{R}_{xx}^N(n)\mathbf{D}_b^N(n) =$$

$$= w\sum_{k=1}^{n} w^{n-k}x(k)x(k-N-1) + x(n+1)x(n_f-N) - \mathbf{D}_f^N,\ \mathbf{R}_{xx}^N,\ \mathbf{D}_b^N(n) =$$

$$= w\mathbf{K}^{N+1}(n) + w\mathbf{D}_f^N(n)^T\mathbf{R}_{xx}^N(n-1)\mathbf{D}_b^N(n-1) + x(n+1)x(n-N) +$$

$$- \mathbf{D}_f^N(n+1)^T\mathbf{R}_{xx}^N(n)\mathbf{D}_b^N(n).$$

Next, in view of already known relations

$$\mathbf{D}_b(n) = \mathbf{D}_b(n-1) + \mathbf{G}(n)\epsilon_1^b(n) \tag{A5-22a}$$

$$\mathbf{D}_f(n+1) = \mathbf{D}_f(n) + \mathbf{G}(n)\epsilon_1^f(n) \tag{A5-22b}$$

$$\mathbf{R}_{xx}^N(n) = w\mathbf{R}_{xx}^N(n-1) + X(n)X(n)^T \tag{A5-22c}$$

$$\epsilon_2^b(n) = \epsilon_1^b(n)\{1 - X(n)^T\mathbf{G}(n)\}. \tag{A5-22d}$$

First

$$\mathbf{D}_f^N(n+1)^T\mathbf{R}_{xx}^N(n)\mathbf{D}_b^N(n) = \mathbf{D}_f^N(n)^T\mathbf{R}_{xx}^N(n-1)\mathbf{D}_b^N(n)w + \mathbf{D}_f^N(n)^TX(n)X(n)^T\mathbf{D}_b^N(n) +$$
$$+ \mathbf{G}(n)^T\epsilon_1^{f,N}(n+1)\mathbf{R}_{xx}^N(n)\mathbf{D}_b^N(n) =$$
$$= \mathbf{D}_f^N(n)^T\mathbf{R}_{xx}^N(n-1)\mathbf{D}_b^N(n-1)w + \mathbf{D}_f^N(n)^Tw\mathbf{R}_{xx}^N(n-1)\mathbf{G}(n)\epsilon_1^b(n) +$$
$$+ \mathbf{D}_f^N(n)^TX(n)X(n)^T\mathbf{D}_b^N(n) + \epsilon_1^{f,N}(n+1)X^N(n)^T\mathbf{D}_b^N(n). \tag{A5-23a}$$

Secondly

$$\mathbf{D}_f^N(n)^Tw\mathbf{R}_{xx}^N(n-1)\mathbf{G}(n)\epsilon_1^b(n) = \mathbf{D}_f^N(n)^T\{\mathbf{R}_{xx}^N(n) - X(n)X(n)^T\}\mathbf{G}(n)\epsilon_1^b(n) =$$
$$= D_f^N(n)^T\{X(n)\epsilon_1^b(n) - X(n)X(n)^T\mathbf{G}(n)\epsilon_1^b(n)\} =$$
$$= \mathbf{D}_f^N(n)^TX(n)\epsilon_1^b(n)\{1 - X(n)^T\mathbf{G}(n)\} = \mathbf{D}_f^N(n)^TX(n)\epsilon_2^{b,N}(n). \tag{A5-23b}$$

On the other hand

$$\mathbf{D}_f^N(n)^TX(n)X(n)^T\mathbf{D}_b^N(n) + \epsilon_1^{f,N}(n+1)X(n)^T\mathbf{D}_b^N(n) = x(n+1)X(n)^T\mathbf{D}_b^N(n). \tag{A5-23c}$$

That is why

$$\mathbf{K}^{N+1}(n) = w\mathbf{K}^{N+1}(n) + x(n+1)x(n-N) - \mathbf{D}_f^N(n)^TX(n)\epsilon_2^{b,N}(n) +$$
$$- x(n+1)X(n+1)^T\mathbf{D}_b^N(n) = w\mathbf{K}^{N+1}(n) + \epsilon_1^{f,N}(n+1)\epsilon_2^{b,N}(n). \tag{A5-23d}$$

However

$$\epsilon_2^f(n+1) = \epsilon_1^f(n+1)\{1 - X(n)^T\mathbf{G}(n)\}. \tag{A5-23e}$$

As a matter of fact

$$\epsilon_1^{f,N}(n+1)\epsilon_2^{b,N}(n) = \epsilon_2^{f,N}(n+1)\epsilon_1^{b,N}(n). \tag{A5-23f}$$

Obviously (48) hold true.

9. The derivation of (49)

First

$$\mathbf{E}^N(n+1) = \sum_{k=1}^{n} w^{n+1-k}y(k)^2 - \mathbf{D}^N(n+1)^T\mathbf{R}_{xx}^{n+1}(n+1)\mathbf{D}^N(n+1)$$

$$= w\sum_{k=1}^{n} w^{n-k}y(k)^2 + y(n+1)^2 - \mathbf{D}^N(n+1)r_{yx}^N(n+1)$$

$$= w\sum_{k=1}^{n} w^{n-k}y(k)^2 + y(n+1)^2 - \{\mathbf{D}^N(n)^T + \mathbf{G}^N(n+1)^T\epsilon_1^N(n+1)\} \times$$

$$\times \{wr_{yx}^N(n) + y(n+1)X^N(n+1)\}$$

$$= w\mathbf{E}_N(n) + y^2(n+1) - \mathbf{D}^N(n)^T y(n+1)X^N(n+1) +$$

$$-\mathbf{G}^N(n+1)^T\epsilon_1(n+1)r_{yx}^N(n+1)$$

$$= w\mathbf{E}_N(n) + y(n+1)\epsilon_1^N(n+1) - \epsilon_1^N(n+1)X(n+1)^T\mathbf{D}(n+1)$$

$$= w\mathbf{E}_N(n) + \epsilon_1^N(n+1)\epsilon_2^N(n+1) \qquad \text{(A5-24)}$$

since

$$\mathbf{G}(n) = \mathbf{R}_{xx}^{-1}(n)X(n) \qquad \text{(A5-25a)}$$

$$\mathbf{R}_{xx}(n)\mathbf{D}(n) = r_{yx}(n) \qquad \text{(A5-25b)}$$

$$\mathbf{D}^N(n+1) = \mathbf{D}^N(n) + \mathbf{G}^N(n+1)\epsilon_1^N(n+1). \qquad \text{(A5-25c)}$$

Eventually (49) holds true.

10. The derivation of (50)

First

$$\mathbf{K}_0^N(n+1) = \sum_{k=1}^{n+1} w^{n+1-k}y(k)(x(k-N) - \mathbf{D}_b^N(n+1)^T X^N(k)) =$$

$$= w\sum_{k=1}^{n+1} w^{n-k}y(k)\{x(k-N) - [\mathbf{D}_b^N(n)^T + \mathbf{G}^T(n+1)\epsilon_1^{b,N}(n+1)]X^N(k)\} =$$

$$= w\mathbf{K}_0^N(n) - \mathbf{G}^T(n+1)r_{yx}(n+1)\epsilon_1^{b,N}(n+1) + y(n+1)\epsilon_2^{b,N}(n+1) +$$

$$+ y(n+1)\mathbf{G}^T(n+1)X^N(n+1)\epsilon_1^{b,N}(n+1) =$$

$$= w\mathbf{K}_0^N(n) - X^N(n+1)^T\mathbf{D}^N(n+1)\epsilon_1^{b,N}(n+1) + y(n+1)\epsilon_2^{b,N}(n+1) +$$

$$+ y(n+1)X^N(n+1)^T\mathbf{G}(n+1)\epsilon_1^{b,N}(n+1) =$$

$$= w\mathbf{K}_0^N(n) + y(n+1)\{1 - X^N(n+1)^T\mathbf{G}(n+1)\}\epsilon_1^b(n+1) +$$

$$+ y(n+1)X^N(n+1)^T\mathbf{G}(n+1)\epsilon_1^{b,N}(n+1) - X^N(n+1)^T\mathbf{D}^N(n+1)\epsilon_1^{b,N}(n+1) =$$

$$= w\mathbf{K}_0^N(n) + \epsilon_2^N(n+1)\epsilon_1^{b,N}(n+1). \qquad \text{(A5-26)}$$

Secondly

$$\epsilon_2^N(n+1)\epsilon_1^{b,N}(n+1) = \epsilon_1^N(n+1)\epsilon_2^{b,N}(n+1) \tag{A5-27}$$

11. The derivation of (51)

According to definition

$$k_f^{N+1}(n+1) = \mathbf{K}^{N+1}(n+1)/\mathbf{E}_f^N(n+1) \tag{A5-28a}$$

$$k_b^{N+1}(n+1) = \mathbf{K}^{N+1}(n+1)/\mathbf{E}_b^N(n). \tag{A5-28b}$$

Moreover

$$\mathbf{K}^{N+1}(n+1) = w\mathbf{K}^{N+1}(n) + \epsilon_2^{f,N}(n+1)\epsilon_1^{b,N}(n) \tag{A5-29a}$$

$$\mathbf{E}_f^N(n+1) = w\mathbf{E}_f^N(n) + \epsilon_1^{f,N}(n+1)\epsilon_2^{f,N}(n+1) \tag{A5-29b}$$

$$\mathbf{E}_b^N(n+1) = w\mathbf{E}_b^N(n) + \epsilon_1^{b,N}(n+1)\epsilon_2^{b,N}(n+1) \tag{A5-29c}$$

$$\epsilon_1^{b,N+1}(n+1) = \epsilon_1^{b,N}(n) - k_f^{N+1}(n)\epsilon_1^{f,N}(n+1) \tag{A5-29d}$$

$$\epsilon_1^{f,N+1}(n+1) = \epsilon_1^{f,N}(n+1) - k_b^{N+1}(n)\epsilon_1^{b,N}(n). \tag{A5-29e}$$

That is why

$$\{\mathbf{E}_f^N(n+1) - \epsilon_1^{f,N}(n+1)\epsilon_2^{f,N}(n+1)\}k_f^{N+1}(n) = \\ = K^{N+1}(n+1) - \epsilon_2^{f,N}(n+1)\epsilon_1^{b,N}(n) \tag{A5-30a}$$

and

$$\{\mathbf{E}_b^N(n) - \epsilon_1^{b,N}(n)\epsilon_2^{b,N}(n)\}k_b^{N+1}(n) = \mathbf{K}^{N+1}(n+1) - \epsilon_1^{b,N}(n)\epsilon_2^{f,N}(n+1). \tag{A5-30b}$$

Eventually

$$k_f^{N+1}(n+1) = k_f^{N+1}(n) + \frac{\epsilon_1^{b,N+1}(n+1)\epsilon_2^{f,N}(n+1)}{\mathbf{E}_f^N(n+1)} \tag{A5-31a}$$

and

$$k_b^{N+1}(n+1) = k_b^{N+1}(n) + \frac{\epsilon_2^{b,N}(n)\epsilon_1^{f,N+1}(n+1)}{\mathbf{E}_b^N(n)}. \tag{A5-31b}$$

12. The derivation of (52)

First

$$k_0^N(n)\mathbf{E}_b^N(n) = \mathbf{K}_0^N(n) = k_0^N(n)\{\mathbf{E}_b^N(n+1) - \epsilon_1^{N,b}(n+1)\epsilon_2^{N,b}(n+1)\}/w = \\ = \{k_0^N(n+1) - \epsilon_1^N(n+1)\epsilon_2^{b,N}(n+1)\}/w. \tag{A5-32a}$$

Secondly

$$k_0^N(n+1) = k_0^N(n) + \frac{\epsilon_2^{b,N}(n+1)}{\mathbf{E}_b^N(n+1)}\{\epsilon_1^N(n+1) - k_0^N(n)\epsilon_1^{N,b}(n+1)\} = \\ = k_0^N(n) + \frac{\epsilon_2^{b,N}(n+1)}{\mathbf{E}_b^N(n+1)}\epsilon_1^{N+1}(n+1). \tag{A5-32b}$$

M. ŻÓŁTOWSKI

O SZYBKICH ALGORYTMACH METODY NAJMNIEJSZYCH KWADRATÓW W LINIOWEJ FILTRACJI ADAPTACYJNEJ

W artykule przedstawiono czytelnikom zainteresowania autora dotyczące adaptacyjnej filtracji. Przedstawiono różne warianty metody najmniejszych kwadratów w szczegółach w przypadku liniowym w formie przeglądowej. Ponadto krótko wyjaśniono zasadę „kolana" do wyboru rzędu filtru adaptacyjnego. Scharakteryzowano również wpływ parametru wagowego *w* na działanie filtru.

KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJI, 1993, 39, z. 1, ss. 91-118

681.32
621.372.54

Application of quantile filtering to sampling time jitter correction in digital signal measurement systems

TOMASZ ADAMSKI

Instytut Podstaw Elektroniki, Politechnika Warszawska

Received 1992.05.20

Authorized 1992.09.30

The presence of time jitter between the trigger signal and the sampling pulse in equivalent-time digital oscilloscopes can cause appreciable distortion of the recorded waveform. In the paper a method of correction for such distortions is proposed. The method is based on applying median or quantile filtering to measured data (i.e. to a sequence of samples). Theoretical background of this median/quantile method of sampling time jitter correction is thoroughly discussed. A hardware solution of the method, using systolic sorting arrays was suggested to implement proposed algorithms. The input signal of these circuits is a sequence of N real numbers, the output signal is nondecreasing ordered (i.e. sorted) input sequence. Along with ordering, each sorting algorithm chooses k-th (in the value) element of the ordered sequence then proposed arrays can be used to median and quantiles computations.

1. INTRODUCTION

In electronics measurement systems with sampling (like digital oscilloscopes, fast data acquisition systems or picosecond waveform recorders) user cannot neglect random unstability of sampling, called sampling time jitter. This unstability is due to inherent noise phenomena and to some extent also to the method of sampling. In equivalent time digital oscilloscopes the jitter causes troublesome noisy scattering of the measured signal image and appreciable distortions of the recorded waveform. The contribution deals with sampling jitter digital correction of equivalent time sampling systems. Simple and fast method for jitter correction by means of median or quantile filtering is proposed. The accuracy and other theoretical aspects of the method are analyzed in the sequel and were practically verified in new constructed, computer based (IBM PC/AT) instrumentation. This instrumentation picosecond waveform recorder PZ 1079-1 is the equivalent time digital oscilloscope with 2 GHz bandwidth and was constructed in the Picosecond Laboratory of Institute of

Electronics Fundamentals at Warsaw University of Technology. At present only software implementation of the method (on an IBM PC/AT computer) is completely elaborated but some hardware solutions (systolic arrays), significantly accelerating the correction, are proposed in the section 7 of the paper. Presented systolic arrays are very simple and can be easily implemented as ASICs in VLSI technology. The comparison of the median and quantile filtering method with simplest arithmetic averaging (used mostly by sampling oscilloscopes producers) and very sophisticated deconvolution method is included to the paper and shortly presented in the section 4.

2. THEORETICAL PRELIMINARIES

If $0 < p < 1$, then an order p quantile of a real random variable X is defined as such a number $a_p \in \mathrm{R}$, that

$$P(X \leqslant a_p) \geqslant p, \quad P(X \geqslant a_p) \geqslant 1 - p. \tag{2.1}$$

A quantile of the order $p = 1/2$ is called the median. For given random variable X and fixed $p \in (0, 1)$ a quantile a_p always exists. It is possible that the quantile a_p is unique but the set of a_p satisfying the conditions (2.1) can be also infinite. If the random variable X has a continuous distrbution function F then each number $a_p \in R$ satisfying the equation $F(a_p) = p$ is a quantile of the order p.

The proposed in the sequel method of jitter correction is based on the following, easy to verify, two theorems.

Theorem 2.1. Let x_p be a quantile of the order p (where $0 < p < 1$), of the real random variable X and f: $\mathrm{R} \to \mathrm{R}$ the real (B(R), B(R)) measurable function

1) If the function f satisfies the condition (2.2)

$$\text{for each } t_1, t_2 \in \mathrm{R};\ t_2 \leqslant x_p \leqslant t_2 \to f(t_1) \leqslant f(x_p) \leqslant f(t_2) \tag{2.2}$$

then $f(x_p)$ is the quantile of the order p of the random variable $f(X)$. If the function f satisfies the condition (2.3)

$$\text{for each } t_1, t_2 \in \mathrm{R};\ t_1 \leqslant x_p \leqslant t_2 \to f(t_1) \geqslant f(x_p) \geqslant f(t_2) \tag{2.3}$$

then $f(x_p)$ is the quantile of the order $1-p$ of the random variable $f(X)$.

2) If the distribution function of the random variable X is strictly increasing in a neighbourhood of x_p and f is continuous in a neighbourhood of x_p then (2.2) implies, that $f(x_p)$ is the unique quantile of the order p of the random variable $f(X)$ and similary (2.3) implies, that $f(x_p)$ is the unique quantile of the order $1-p$ of the random variable $f(X)$.

Theorem 2.2. Assume x_p is a quantile of the order p (where $0 < p < 1$) of a real random variable X, there exists such a fixed $\varepsilon > 0$ that $|X - x_p| \leqslant \varepsilon$, and the real function f: $[x_p - \varepsilon, x_p + \varepsilon] \to \mathrm{R}$ is $(\mathrm{B}([x_p - \varepsilon, x_p + \varepsilon]), \mathrm{B}(\mathrm{R}))$ measurable.

1) If the function f satisfies the condition (2.4)

$$\text{for each } t_1, t_2 \in \mathrm{R};\ x_p-\varepsilon \leqslant t_1 \leqslant x_p \leqslant t_2 \leqslant x_p+\varepsilon \rightarrow f(t_1) \leqslant f(x_p) \leqslant f(t_2) \tag{2.4}$$

then $f(x_p)$ is the quantile of the order p of the random variable $f(X)$. If the function f satisfies the condition (2.5)

$$\text{for each } t_1, t_2 \in \mathrm{R};\ x_p-\varepsilon \leqslant t_1 \leqslant x_p \leqslant t_2 \leqslant x_p+\varepsilon \rightarrow f(t_2) \leqslant f(x_p) \leqslant f(t_1) \tag{2.5}$$

then $f(x_p)$ is the quantile of the order $1-p$ of the random variable $f(X)$.

2) If the distribution function of the random variable X is strictly increasing in a neighbourhood of x_p and f is continuous in a neighbourhood of x_p then (2.4) implies, that $f(x_p)$ is the unique quantile of the order p of the random variable $f(X)$ and similary (2.5) implies, that $f(x_p)$ is the unique quantile of the order $1-p$ of the random variable $f(X)$.

Corollary 2.3. Let the input signal f: $\mathrm{R} \rightarrow \mathrm{R}$ be measurable and assume that t_k, is the quantile of the order $p \in (0, 1)$ of the random variable $t_k + T(t_k)$. If the condition (2.2) is satisfied (or $|T(t_k)| < \varepsilon$ and the condition (2.4) is satisfied) then $f(t_k)$ is the quantile of the order p of the random variable $f(t_k + T(t_k))$. If the condition (2.3) is satisfied (or $|T(t_k)| < \varepsilon$ and is satisfied the condition (2.5)) then $f(t_k)$ is the quantile of the order $1-p$ of the random variable $f(t_k + T(t_k))$.

If the distribution function of the random variable $T(t_k)$ is strictly increasing in a neighbourhood of 0 and f is continuous in a neighbourhood of the point t_k then

1) if the condition (2.2) is satisfied (or $|T(t_k)| < \varepsilon$ and the condition is satisfied (2.4)) then $f(t_k)$ is the unique quantile of the order p of the random variable $f(t_k + T(t_k))$,

2) if the condition (2.3) is satisfied (or $|T(t_k)| < \varepsilon$ and is satisfied the condition (2.5)) then $f(t_k)$ is the unique quantile of the order $1-p$ of the random variable $f(t_k + T(t_k))$.

Definition 2.1. Let $(X_1, X_2, \ldots, X_n)$ be an n-dimensional random vector defined on the probabilistic space $(\Omega, \mathcal{M}, \mathrm{P})$ and $k \in \{1, 2, \ldots, n\}$. Define function $\xi_k^{(n)}$: $\Omega \rightarrow \mathrm{R}$ in the following way. Let $\omega \in \Omega$ and assume that α is such a permutation of the set $\{1, 2, \ldots, n\}$ (in general α is dependent on ω) that $X_{\alpha(1)}(\omega) \leqslant X_{\alpha(2)}(\omega) \leqslant \ldots \leqslant X_{\alpha(n)}(\omega)$. Then we admit

$$\xi_k^{(n)}(\omega) = X_{\alpha(k)}(\omega). \tag{2.6}$$

The function $\xi_k^{(n)}$ is $(\mathcal{M}, \mathrm{B(R)})$ measurable (see theorem 2.4) then $\xi_k^{(n)}$ is a random variable. The random variable $\xi_k^{(n)}$ is called the order k statistics. The number k called an order of the order statistics $\xi_k^{(n)}$.

Theorem 2.4. The function $\xi_k^{(n)}$ (from the definition 2.1) defined for a random vector $(X_1, X_2, \ldots, X_n)$ is a random variable.

Proof. Let S_p be a set of all permutations of the set $\{1, 2, \ldots, n\}$. If $\alpha \in S_n$, then $\mathrm{A}_\alpha = \{X_{\alpha(1)} \leqslant X_{\alpha(2)} \leqslant \ldots \leqslant X_{\alpha(n)}\} \in \mathcal{M}$. Hence the set

$$\mathrm{B}_i = \bigcup_{\substack{\alpha(k)=i \\ \alpha \in S_n}} \mathrm{A}_\alpha \in \mathcal{M}, \text{ but } \xi_k^{(n)} = \sum_{i=1}^{n} X_i \chi_{B_i}, \text{ then } \xi_k^{(n)} \text{ is } (\mathcal{M}, \mathrm{B(R)}) \text{ measurable. QED.}$$

Let $(X_i)_{i=1}^{\infty}$ be the sequence of independent random variables with the same probability distribution as a random variable X. Assume that X and $(X_i)_{i=1}^{\infty}$ are defined on the same probabilistic space $(\Omega, \mathcal{M}, P)$. The random variable $\xi_{k(n)}^{(n)}$, where $k(n) = [n\lambda]+1$ for fixed $\lambda \in (0, 1)$ is so called quantile statistics of order λ for the n-dimensional random vector $(X_1, X_2, \ldots, X_n)$. Basic properties of the random variable $\xi_{k(n)}^{(n)}$ can be summarized the following theorem.

Theorem 2.5. Assume $(X_i)_{i=1}^{\infty}$ is a sequence of independent real random variables defined on the probabilistic space $(\Omega, \mathcal{M}, P)$ and with the same probability distribution F as the random variable X. Let λ be arbitrary number from (0, 1) and $\xi_{k(n)}^{(n)}$ order λ quantile statistics (where $k(n) = [n\lambda] + 1$) for the vector $(X_1, X_2, \ldots, X_n)$.

1) If there is the unique quantile a_λ of order λ for the random variable X, then $\xi_{k(n)}^{(n)} \to a_\lambda$ (P almost everywhere), when $n \to \infty$.

2) If the random variable X has a continuous density function f and a_λ is an order λ quantile, for the random variable X, the density function f is continuous and positive in the point a_λ then introducing the following notation

$$Y^{(n)} = \sqrt{\frac{n}{\lambda(1-\lambda)}} f(a_\lambda)(\xi_{k(n)}^{(n)} - a_\lambda) \tag{2.7}$$

we have for each $y_1, y_2 \in R$, $y_1 < y_2$

$$\lim_{n\to\infty} P(y_1 < Y^{(n)} < y_2) = \frac{1}{\sqrt{2\pi}} \int_{y_1}^{y_2} e^{-y^2/2}\, dy. \tag{2.8}$$

Proof of the above theorem can be found in [13], [14], [15].

The part 1) of the theorem 2.5 says that quantile statistics $\xi_{k(n)}^{(n)}$ is a consistent estimator (in the sense of convergence P almost everywhere) of the order λ quantile a_λ of the random variable X.

If $X_1, X_2, \ldots, X_n$ are independent real random variables with the same probability distribution with the distribution function $F(x)$ then the distribution function of the random variable $\xi_k^{(n)}$ defined for the random vector $(X_2, X_2, \ldots, X_n)$ has the distribution function Φ_{kn} given by the following formula (2.9) (see [13], [14], [15])

$$\Phi_{kn}(x) = \sum_{m=k}^{n} \frac{n!}{m!(n-m)!} (F(x))^m (1 - F(x))^{n-m} \tag{2.9}$$

or equivalently by the formula (2.10)

$$\Phi_{kn}(x) = \frac{n!}{(k-1)!(n-k)!} \int_0^{F(x)} t^{k-1}(1-t)^{n-k}\, dt. \tag{2.10}$$

If the distribution function F has a probability density f then there is a probability density of the random variable $\xi_k^{(n)}$ given by (2.11)

$$f_{kn}(x) = \frac{n!}{(k-1)!(n-k)!} (\mathrm{F}(x))^{k-1}(1-\mathrm{F}(x))^{n-k} f(x). \tag{2.11}$$

Proof of the formula (2.9) is based on the formula on frequency in the Bernoulli's scheme with probability of success even $\mathrm{F}(x) = p$. The formula (2.10) is obtained by integration by parts. The formula (2.11) can be obtained from (2.10) by differentiation.

If we take into account specific probability distributions, we can see much slower convergence rate of $\xi_{k(n)}^{(n)} \to a_\lambda$ (P almost everywhere) where $n \to \infty$, in comparison with the convergence $(X_1 + X_2 + \ldots + X_n)/n \to \mathrm{E}(X)$ (P almost everywhere). This result is a consequence of the probability distribution of the statistics $\xi_k^{(n)}$.

3. PROBABILISTIC MODEL OF JITTERED SAMPLING

To explain the principles of the median/quantile method of jitter correction, we have to introduce at first a mathematical model of sampling with jitter. In this section such a mathematical model of jittered sampling in equivalent time digital oscilloscopes is shortly discussed. Assume the analog input signal f: $\mathrm{R} \to \mathrm{R}$ is sampled in equvalent time points $t_1, t_2, \ldots, t_\mathrm{M}$ (particularly we can have $t_k = k \cdot \Delta t$ but it is not important in the sequel). Sampling time points t_k, $k \in \{1, 2, \ldots, \mathrm{M}\}$ are not exactly real sampling points. Because of noise phenomena present in sampling circuit real sampling points are randomly shifted on the time axis. This shifting is called sampling time jitter. In relatively slow systems these shifts are not critical and errors caused by jitter can be neglected. In fast systems, like wide band equivalent-time waveform recorders (ETWR) or sampling oscilloscopes random shifting is in the range 10–200 ps and causes very troublesome distortion of the measured signal. The aim of the sampling time jitter correctors is to reduce distortions introduced by jitter.

Let $\mathrm{T}(t_k)$ be the random variable describing the random shift of the real sampling point $t_k + \mathrm{T}(t_k)$ related to assumed, deterministic one t_k. Because, we repeat the sampling at the each point t_k (where $k = 1, 2, \ldots, \mathrm{M}$ and M is number of sampling points) in independent manner, we can describe the all sampling process by M random sequences $(f(t_k + T_n(t_k)))_{n=1}^{\infty}$ for $k = 1, 2, \ldots, \mathrm{M}$, where $(\mathrm{T}_n(t_k))_{n=1}^{\infty}$ is a sequence of independent random variables with distributions equal to the distribution of $\mathrm{T}(t_k)$. In practice the number of samples for each k is limited to N_0, then the input information for the median/quantile algorithm consist of M random vectors of N_0 dimensions or in other words, of M finite random sequences $(f(t_k + \mathrm{T}_n(t_k)))_{n=1}^{\mathrm{N}_0}$ for $k = 1, 2, \ldots, \mathrm{M}$. In the sequel we assume that the input analog signal is not corrupted by the additive noise, the jitter is limited (i.e. there is such

a number $\varepsilon > 0$ that $|T_n(t_k)| \leqslant \varepsilon$ for $k = 1, 2, \ldots, M$ and $n \in N$) and distribution functions of all r.v. $T_n(t_k)$ are continuous and strictly monotone a neighbourhood of 0.

4. ERRORS INTRODUCED BY SAMPLING TIME JITTER

In this section we assume that the input signal is a (B(R), B(R)) measurable function $f: R \to R$ and is sampled in time points $t_i + T(t_i)$, where $t_1, t_2, \ldots, t_M \in [a, b]$ are deterministic sampling points. We would like to assess measurement errors introduced by sampling time jitter. In other words, we want to assess the value $er(t) \underline{\underline{df}} f(t+T(t)) - f(t)$ for a fixed sampling point t particularly for $t_1, t_2, \ldots, t_M \in [a, b]$.

If the signal f is a (B(R), B(R)) measurable function (for instance f is continuous) then $er(t)$ is for each sampling point t a random variable but its probability distribution depends on f and is difficult to exact calculations.

Assume that jitter is limited i.e. for each t, we have $|T(t)| \leqslant \varepsilon$. If the signal $f \in C^{(n)}(R)$ then from Taylor'sformula we obtain

$$|er(t)| \leqslant \sum_{k=1}^{n-1} \frac{|f^{(k)}(t)|}{k!} \varepsilon^k + \frac{\varepsilon^n}{n!} \sup_{h \in [-\varepsilon, \varepsilon]} |f^{(n)}(t+h)|. \tag{4.1}$$

If $f^{(n+1)} \in C(R), f^{(k)} \in L^1(R, \mathscr{L}, l_1)$ for $k = 0, 1, \ldots, n+1$ and $\mathscr{F}$ denotes the Fourier transform then $\mathscr{F}(f^{(n)})(\omega) = (j\omega)^n \mathscr{F}(f)(\omega)$. Hence

$|f^{(n)}(t)| \leqslant \frac{1}{2\pi} \int_R |\omega|^n |F(\omega)| l_1(d\omega)$, where F is a spectrum of the signal f. Then from (4.1), we have the following enequality (4.2)

$$\begin{aligned} |er(t)| &\leqslant \sum_{k=1}^{n-1} \frac{|f^{(k)}(t)|}{k!} \varepsilon^k + \frac{\varepsilon^n}{2\pi n!} \int_R |\omega|^n |F(\omega)| l_1(d\omega) \leqslant \\ &\leqslant \frac{1}{2\pi} \sum_{k=1}^{n} \frac{\int_R |\omega|^k |F(\omega)| l_1(d\omega) \varepsilon^k}{k!} \leqslant \frac{1}{\pi} \int_{R^+} (e^{\varepsilon|\omega|} - 1) |F(\omega)| l_1(d\omega). \end{aligned} \tag{4.2}$$

If the spectrum F of the input signal f is limited i.e. there is such a ω_g, that supp $F \subset [-\omega_g, \omega_g]$ then the integral on the right side of (4.2) is finite and we obtain

$$|er(t)| \leqslant \frac{1}{\pi} \int_{[0, \omega_g]} (e^{\varepsilon\omega} - 1) |F(\omega)| l_1(d\omega) \tag{4.3}$$

The simplest but rather pesimistic assessment of $er(t)$, can be also obtained, when limited by ω_g spectrum of the signal f and limited jitter are assumed. In this case, specifically for $f(t) = A \cdot \sin(\omega_g t)$, where $A > 0$, we have $|er(t)| \leqslant A \cdot \omega_g \varepsilon$.

In general, we do not assume, that the sampling time jitter is limited.

In such a case, if the signal $f \in C^{(1)}(R)$, $T(t) \in L^2(\Omega, \mathcal{M}, P)$ and $\sup_{t \in R} |f'(t)| < +\infty$ then from the Lagrange theorem about the mean value we have

$$D^2(er(t)) \leqslant (\sup_{t \in R} |f'(t)|)^2 \cdot D^2(T(t)) \tag{4.4}$$

and for limited jitter

$$D^2(er(t)) \leqslant \varepsilon^2 (\sup_{t \in [a-\varepsilon, b+\varepsilon]} |f'(t)|)^2 . \tag{4.5}$$

If $f \in C^{(2)}(R)$ and $f, f', f'' \in L^1(R, \mathcal{L}, l_1)$ then (denoting spectrum of the signal f by F) the formulas (4.4) and (4.5) can be rewritten as (4.6) and (4.7)

$$D^2(er(t)) \leqslant 2 \cdot \int_{R^+} \omega |F(\omega)| \, d\omega \cdot D^2(T(t)) \tag{4.6}$$

$$D^2(er(t)) \leqslant 2\varepsilon^2 \int_{R^+} \omega |F(\omega)| \, d\omega . \tag{4.7}$$

If $f(t) = A \sin(\omega_g t)$ then from (4.4) we have $D^2(er(t)) \leqslant A^2 \omega_g^2 D^2(T(t))$.

Assume that the analyzed in the sequel sampling system is an equivalent time waveform recorder. In such a system for each $k = 1, 2, \ldots, M$ a set of samples can be gathered creating input data for correction algorithm of errors introduced by sampling time jitter. In general these statistics taken for every $k = 1, 2, \ldots, M$ allow to correct distortions caused by jitter, but the restoration needs deconvolution algorithm (see [3]) or in the case when jitter statistics depends on k, solution of the integral Fredholm equation of the first kind. Then correction leads to the ill posed problem and needs complicated numerical regularization techniques (see [3], [4]). In this context, simple methods, like mean value or quantile statistics computations, working quite correctly only under certain assumptions become attractive. The simplest method to correct scattered signal image is the mentioned mean value computation i.e. averaging. In this method for each $k = 1, 2, \ldots, M$ and sampling point t_k the number

$$MN(t_k, N_0) \stackrel{df}{=} \frac{1}{N_0} \sum_{i=1}^{N_0} f(t_k + T_i(t_k)) \tag{4.8}$$

is computed. The method is quite correct for linear signal $f(t) = At+B$, $A, B \in R$ i.e. $MN(t, N_0) \to f(t)$ with probability 1 but in general (for example for pulse signals) it introduces some errors. These errors can be easily assesed in the following way. Let $h(t) \stackrel{df}{=} E(f(t+T(t)))$. We would like to find $|h(t)-f(t)|$ and $\sup_{t \in R} |h(t)-f(t)|$ for fixed sampling time t. Assume that the input signal $f \in C^{(n)}(R)$ and for each $t \in R$, $T(t) \in L^n(\Omega, \mathcal{M}, P)$, $f(t+T(t)) \in L^1(\Omega, \mathcal{M}, P)$. Using Taylor's formula we obtain

$$|h(t)-f(t)| \leqslant \sum_{k=1}^{n-1} \frac{|f^{(k)}(f)}{k!} |E((T(t))^k)| + \frac{1}{n!} (\sup_{x \in A} |f^{(n)}(t+x)| |E|(T(t))^n|| \tag{4.9}$$

where $A = [-\varepsilon, \varepsilon]$ if the sampling time jitter is limited (i.e. exists such a number $\varepsilon > 0$, that for each sampling time t, $|T_n(t)| < \varepsilon$) or $A = R$ if the sampling time jitter is not limited.

If $f^{(n+1)} \in C(R)$, $f^{(k)} \in L^1(R, \mathscr{L}, l_1)$ for $k = 0, 1, \ldots, n+1$ then denoting by F the Fourier transform of the signal f, we can (4.9) rewrite as

$$|h(t) - f(t)| \leqslant \sum_{k=1}^{n-1} \frac{|f^{(k)}(t)|}{k!} |E((T(t))^k)| + \frac{E|((T(t))^n)|}{2\pi n!} \cdot$$

$$\cdot \int_R |\omega|^n |F(\omega)| l_1(d\omega) \leqslant \frac{1}{2\pi} \sum_{k=1}^{n-1} \frac{E|(T(t))^k)| \int_R |\omega|^n |F(\omega)| l_1(d\omega)}{k!} +$$

$$+ \frac{1}{2\pi} \frac{E|(T(t))^n|}{n!} \int_R |\omega|^n |F(w)| l_1 d\omega. \tag{4.10}$$

In particular from the formula (4.9) for $f \in C^{(1)}(R)$ and $f \in C^{(2)}(R)$ we have

$$|h(t)-f(t)| \leqslant \sup_{x \in A} |f'(t+x)| \cdot E|T(t)| \tag{4.11}$$

$$|h(t)-f(t)| \leqslant |f'(t)||E(T(t))| + \frac{1}{2} \sup_{x \in A} |f''(t+x)| \cdot E|T(t)|^2. \tag{4.12}$$

If additionally sampling time jitter is limited then two enequalities (4.13), (4.14) follow from the formulas (4.11) and (4.12)

$$|h(t) - f(t)| \leqslant \varepsilon \sup_{x \in [-\varepsilon, \varepsilon]} |f'(t+x)| \text{ for } f \in C^{(1)}(R) \tag{4.13}$$

$$|h(t)-f(t)| \leqslant \varepsilon |f'(t)| + \frac{\varepsilon^2}{2} \sup_{x \in [-\varepsilon, \varepsilon]} |f''(t+x)| \text{ for } f \in C^{(2)}(R). \tag{4.14}$$

If the probability distribution of the random variables $T(t)$ is for each $t \in R$ symmetric (i.e. jitter is symmetric), then formulas (4.9), (4.10) and (4.12) become simpler because odd moments $E((T(t))^k)$ vanish. In particular for $f \in C^{(2)}(R)$ from (4.12) we obtain

$$|h(t)-f(t)| \leqslant \frac{1}{2} \sup_{x \in A} |f''(t+x)| \cdot D^2(T(t)). \tag{4.15}$$

If $f \in C^{(3)}(R)$ and $f^{(k)} \in L^1(R, \mathscr{L}, l_1)$ for $k = 1, 2, 3$ then from (4.10) we obtain

$$|h(t)-f(t)| \leqslant \frac{D^2(T(t))}{2\pi} \int_{R^+} \omega^2 |F(\omega)| \, l_1(d\omega). \tag{4.16}$$

If sampling time jitter is limited, the signal f is analytical on R (i.e. $f \in C^{(\infty)}(R)$) and $\sup_{k \in N} |f^{(k)}(t)| = G(t) < +\infty$ then because $E((T(t))^k) \leqslant \varepsilon^k$ we obtain

$$|h(t)-f(t)| \leqslant \sum_{k=1}^{\infty} \frac{|f^{(k)}(t)|}{k!} \leqslant \sum_{k=1}^{\infty} \frac{G(t)}{k!} \varepsilon^k = G(t)(e^{\varepsilon}-1). \tag{4.17}$$

If additionally the random variable $T(t)$ has a symmetric probability distribution then $E(T(t)^k) = 0$ for k odd and (4.17) can be rewritten as (4.18)

$$|h(t)-f(t)| \leqslant \sum_{k=1}^{\infty} \frac{f^{(2k)}(t)}{(2k)!} |E((T(t)^{2k})| \leqslant G(t)(\cosh(\varepsilon)-1). \tag{4.18}$$

From assessments (4.9)–(4.18) we can easily obtain error assessments (of the signal f measurement) in the supremum norm.

For example, if $f \in C^{(3)}(R), f^{(k)} \in L^1(R, \mathscr{L}, l_1)$ for $k = 1, 2, 3$ and $E(T(t)) = 0$ for each $t \in R$ then for ε limited sampling time jitter we obtain

$$\sup_{t\in R} |h(t)-f(t)| \leqslant \frac{\varepsilon^2}{2} \sup_{t\in R} |f''(t)| \leqslant \frac{\varepsilon^2}{2\pi} \int_{R^+} \omega |F(\omega)| l_1(d\omega). \tag{4.19}$$

In the sequel we give several examples of distortions introduced by averaging when sampling time jitter is present.

Example 1. If for each $t \in [a, b]$ sampling time jitter is limited i.e. $|T(t)| \leqslant \varepsilon$ and a probability distribution $P(t, \cdot)$ of the random variable $T(t)$ is symmetric related to zero or $E(T(t)) = 0$, then averaging for arbitrary fixed $t \in [a, b]$ does not introduce errors in the case of linear signals, i.e. for signals defined by the formula f: $[a-\varepsilon, b+\varepsilon] \ni t \to At+B \in R$, where $A, B \in R$. Indeed, $E(A(Tt)+(t)+B) = At+B$ for $t \in [a, b]$.

Example 2. Let $f(t) = A\sin(\omega_g t)$; where $A, \omega_g \in R^+$ and assume that $P(t, \cdot)$, the probability distribution of the random variable $T(t)$ is symmetric,

$$h(t) = E(A\cdot\sin(\omega_g(t+T(t))) = A(\sin(\omega_g t)E(\cos(\omega_g T(t)))+\cos(\omega_g t)\cdot$$

$$\cdot E(\sin(\omega_g T(t))) = A\sin(\omega_g t)\cdot E(\cos(\omega_g T(t))) = A\cdot\sin(\omega_g t)\cdot \mathrm{Re}\,\varphi(t, \omega_g)$$

where $\varphi(t, \cdot): R \ni \omega \to E(e^{i\omega T(t)}) \in C$ is a characteristic function od the probability distribution $P(t, \cdot)$. Similarly for a consinusoidal signal $f(t) = A\cos(\omega_g t)$, we have $h(t) = A\cdot\cos(\omega_g t)\cdot \mathrm{Re}\,\varphi(t, \omega_g)$.

For each $t, \omega \in R$, $|\mathrm{Re}\,\varphi(t, \omega)| \leqslant |\varphi(t, \omega)| \leqslant |\varphi(t, 0)| \leqslant 1$. Therefore, if the probability distribution of the random variable $T(t)$ does not depend on t, then *"symmetric jitter"* modifies an amplitude of the averaged sinusoid and introduces a phase shift but the shape of h remains sinusoidal (in particular may be $h(t) \equiv 0$ or the phase shift equal to π may occur). In the case where the random variable $T(t)$ has for each $t \in R$ the uniform probability distribution on the interval $[-\varepsilon, \varepsilon]$ we obtain

$$h(t) = A\sin(\omega_g t)\, E(\cos\omega_g(t))) = \frac{A}{2\varepsilon} \sin(\omega_g t) \int_{\varepsilon}^{\varepsilon} \cos(\omega_g x)dx = f(t)\cdot Sa(\omega_g \varepsilon), \tag{4.20}$$

where Sa: $R \ni x \to Sa(x) = \begin{cases} \sin x/x & \text{for } x \in R/\{0\} \\ 1 & \text{for } x = 0 \end{cases}$.

In the case, when the random variable $T(t)$ have for each $t \in R$ the Gaussian probability distribution $N(0, \sigma)$, we obtain $\varphi(t, \omega) = \exp(-\sigma^2\omega_g^2/2)$ then

$$h(t) = A \cdot \sin(\omega_g t) \operatorname{Re} \varphi(t, \omega_g) = A \cdot \sin(\omega_g t) \cdot \exp\left(-\frac{\sigma^2\omega_g^2}{2}\right). \tag{4.21}$$

If the probability distribution of the random variable $T(t)$ does not depend on t but is not symmetric then the jitter can both change amplitude and introduce a phase shift φ (h remains a sinusoid) according to the formula

$$E(A \cdot \sin(\omega_g(T(t)+t))) = A_1 \sin(\omega_g t) + A_2 \cos(\omega_g t) = \tilde{A} \cdot \sin(\omega_g t + \varphi)$$

where $A_1 = A \cdot E(\cos(\omega_g T(t)))$, $A_2 = A \cdot E(\sin(\omega_g T(t)))$, $\tilde{A}^2 = A_1^2 + A_2^2$, and φ is a solution of equations $\sin\varphi = A_2/\tilde{A}$, $\cos\varphi = A_1/\tilde{A}$.

If the signal $f \in C^{(1)}(R)$ is periodic with a period $T = 2\pi/\omega_g$ then its Fourier series $\sum_{n=-\infty}^{\infty} c_n e^{j\omega_g n t}$ is uniform convergent for $t \in R$ to $f(t)$. Hence

$$h(t) = E(f(t+T(t))) = \sum_{n=-\infty}^{\infty} c_n E(e^{j\omega_g n(k+T(t))}) = \sum_{n=-\infty}^{\infty} c_n e^{j\omega_g n t} \cdot \varphi(t, \omega_g n). \tag{4.22}$$

Example 3. Assume the random variable $T(t)$ have for each $t \in R$ the same uniform probability distribution on $[-\varepsilon, \varepsilon]$ and the input signal f is the step function $f(t) = 1(t)$. Then we have

$$h(t) = E(1(t+T(t)) = \frac{1}{2\varepsilon} \int_{-\varepsilon}^{\varepsilon} 1(t+x)dx = \begin{cases} 1 \text{ for } t \geqslant \varepsilon \\ 1/2 + \dfrac{1}{2\varepsilon} t \text{ for } -\varepsilon < t < \varepsilon \\ 0 \text{ for } t \leqslant -\varepsilon. \end{cases} \tag{4.23}$$

If $T(t)$ have for each $t \in R$ the Gaussian probability distribution $N(0, \sigma)$ then we obtain

$$h(t) = E(1(t+T(t)) = \frac{1}{\sigma\sqrt{2}} \int_R 1(t+x)\exp\left(-\frac{x^2}{2\sigma^2}\right)dx = \frac{1}{2} + \frac{1}{2}\operatorname{erf}\left(\frac{\sqrt{2}}{2\sigma}t\right), \tag{4.24}$$

where $\operatorname{erf}(x) \overset{df}{=} \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2)\,dt$ is the error function.

For both probability distribution (uniform and Gaussian) obtained rise time of measured step function is significantly augmented by the sampling time jitter (for example up to 1.6 ε in the case of uniform probability distribution see fig. 4.1).

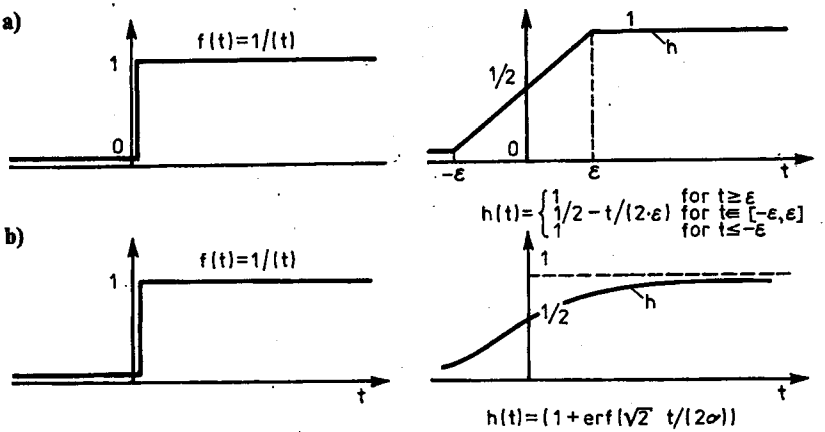


Fig. 4.1. Step function distortion caused by sampling time jitter when averaging algorithm is applied a) for each $t \in R$ the random variable $T(t)$ has the uniform distribution on $[-\varepsilon, \varepsilon]$ b) for each $t \in R$, $T(t)$ has the Gaussian distribution $N(0, \sigma)$, erf denotes the error function.

5. QUANTILE AND MEDIAN FILTERING METHODS OF SAMPLING TIME JITTER CORRECTION

In this section we analyze applications of median and quantile filters to the correction of errors introduced by sampling time jitter. Using as a basis corollary 2.3 and theorem 2.5, we can apply to the estimation of $f(t_k)$, the quantile statistics of the order p_k, where

$$P(T(t_k) \leqslant 0) \geqslant p_k;\ P(T(t_k) \geqslant 0) \geqslant 1-p_k. \tag{5.1}$$

More strictly, the proposed algorithm is the following:

1. Input data and assumptions. Let like in the section 3 the sampling be described by the random sequence (5.2)

$$(f(t_k + T_n(t_k)) + Q_n(t_k))_{n=0}^{\infty}. \tag{5.2}$$

Assume, that the additive noise $(Q_n(t_k))_{t \in R}$ is not present. Hence description of the sampling becomes simpler and (5.2) can be written with the formula (5.3)

$$(f(t_k + T_n(t_k))_{n=0}^{\infty}. \tag{5.3}$$

According to assumptions from the section 3 about the random sequence (5.2) we suppose that for each fixed $k \in \{1, 2, \ldots, M\}$, the sequence $(T_n(t_k)_{n=0}^{\infty}$ is the sequence of independent random variables with the same probability distribution (but in general this distribution can depend on k).

For each $k \in \{1,2, \ldots, M\}$ we take only finite number N_0 of samples, then n input data are described by M finite sequences of random variables $(f(t_k + T_n(t_k)))_{n=0}^{N_0}$ for $k = 1, 2, \ldots, M$ or for fixed $\{\omega_n\} \in \Omega$ by M real sequences $(f(t_k + T_n(t_k))(\omega_n))_{n=0}^{N_0}$.

In the sequel we assume that the distribution function of random variables $T_n(t_k)$ are strictly increasing in a neighbourhood of zero and sampling time jitter is limited i.e. there is such a number $\varepsilon > 0$ that $|T_n(t_k)| \leqslant \varepsilon$ for each $k \in \{1, 2, ..., M\}$. When the above assumptions are fulfilled, zero is a unique quantile of the order p_k of the random variable $T(t_k)$.

In the described below algorithm for $f(t_k)$ estimation by quantile statistics we do not assume that the probability distributions of random variables $T_n(t_k)$ for $k = 1, 2,..., M$ are identical or for fixed k symmetrical related to 0. But we have to know, whether zero is a unique quantile of the random variable $T_n(t_k)$ and to know the order of this quantile.

2. Assume, we know M positive numbers $p_1, p_2,...p_M$ satisfying the condition (5.1) for each $k \in \{1, 2, ..., M\}$. Hence, for each $k \in \{1, 2, ..., M\}$ zero is the quantile of the order p_k of the random variable $T_n(t_k)$. If random variables $T(t_k)$ for $k = 1,...M$ are symmetrical related to 0 or have the medians equal to 0 then $p_1 = p_2 = ... = p_M = 1/2$.

3. Let $\xi_j^{(N_0)}(k)$ be the order statistics for the random vector (5.4). Denote $j(k) = [N_0 \cdot p_k]+1$ and $r(k) = [N_0 \cdot (1-p_k)]+1$ then under admitted assumptions $\xi_{j(k)}^{(N_0)}(k)$ and $\xi_{r(k)}^{(N_0)}(k)$ are quantile statistics appropriately of the order p_k and $1-p_k$ defined for the random vector (5.4) (i.e. the sequence of samples)

$$(f(t_k + T_1(t_k)), f(t_k + T_2(t_k)), ..., f(t_k + T_{N_0}(t_k))). \tag{5.4}$$

For each $k \in \{1, ...,M\}$, we have a realization of the random vector (5.4) and compute $\xi_{j(k)}^{(N_0)}(k)(\omega)$ and $\xi_{r(k)}^{(N_0)}(k)(\omega)$ i.e. values of the quantile statistics of the order p_k and $1-p_k$ for (5.4).

4. For each $k \in \{1, ..., M\}$ we now have number $\xi_{j(k)}^{(N_0)}(k)(\omega)$, if N_0 is large enough, we can admit $f(t_k) \cong \xi_{j(k)}^{(N_0)}(k)$, if the condition (2.4) is satisfied and $f(t_k) \cong \xi_{r(k)}^{(N_0)}(k)$, if the condition (2.5) is satisfied.

The above described algorithm becomes simpler if zero is for each $k = 1, 2, ..., M$ a median of the random variable $T_n(t_k)$. We have not to choose in such a case between quantile statistics $\xi_{j(k)}^{(N_0)}(k)$ and $\xi_{r(k)}^{(N_0)}(k)$, when estimation of the value $f(t_k)$ is done because both values are equal. It follows from experimental data gathered for wide band waveform recorder PZ-1079, that the assumption of symmetry of the random variable $T_n(t_k)$ can be considered as fulfilled.

Assume a function f (i.e. input signal) is continuous and monotonic by intervals and has compact support. Let I be the interval $[\tilde{t}_0, \tilde{t}_K]$, where $\tilde{t}_0 = \inf(\text{supp } f)$, $\tilde{t}_K = \sup(\text{supp } f)$ and $\tilde{t}_1, \tilde{t}_2, ..., \tilde{t}_{K-1} \in I$ such points that the function f is on the intervals $\{(\tilde{t}_0, \tilde{t}_1), (\tilde{t}_1, \tilde{t}_2), ...,\} (\tilde{t}_{K-1}, \tilde{t}_K)$ monotonic. Denote $A \overset{df}{=} \{t \in I; f \text{ is not continuous in the point } t\}$ and let $U = I - \left(\bigcup_{i=0}^{K} [\tilde{t}_i-\varepsilon, \tilde{t}_i+\varepsilon] \cup A \right)$. If follows from

corollary 2.3 and theorem 2.5 that for each $k \in \{1, 2, \ldots, M\}$, if the condition (2.4) is satisfied then

$$t_k \in U \Rightarrow \xi^{(N_0)}_{j(k)}(k) \xrightarrow[N_0 \to \infty]{} f(t_k) \quad \text{P almost everywhere} \tag{5.5}$$

where $j(N_0) = [N_0 P_k] + 1$. Similarly if the condition (2.5) is satisfied then for $r(k) = [N_0 \cdot (1 - p_k)] + 1$ we have

$$t_k \in U \Rightarrow \xi^{(N_0)}_{r(k)}(k) \xrightarrow[N_0 \to \infty]{} f(t_k) \quad \text{P almost everywhere.} \tag{5.5'}$$

Hence for $t_k \in U$ random variable $\xi^{(N_0)}_{j(k)}(k)$ or $\xi^{(N_0)}_{r(k)}$ is a consistent estimator of the value $f(t_k)$ (in the sence P almost everywhere convergence). Then proposed algorithm is correct for $t_k \in U$. Of course for $p_k = 1/2$ there is no difference between (5.5) and (5.5').

6. ACCURACY AND ERRORS OF THE QUANTILE/MEDIAN METHOD

The proposed above median/quantile method of sampling jitter correction is from computational point of view simple, effective and completely correct in the case of monotone signals. When the input signal is increasing and decreasing on an interval, errors can occurs in the neighbourhood of extremes but results of computer simulation of the median/quantile algorithm prove that presented method stays very useful in jitter correction. Other applied in practice methods also introduce their own errors. For example the most sophisticated deconvolution method (see [2]) is very sensitive to input data, then we have to use complex regularization techniques. On the other hand the simplest averaging method mostly used by digital oscilloscopes producers is correct only for linear signals.

A question arises: how should be N_0 (assuming that $t_k \in U$) so that approximation $f(t_k)$ by an appropriate quantile statistics defined for the random vector (5.4) would be sufficiently exact. The analysis will be done for simplicity for the case when $p_k = 1/2$ and a continuous, strictly increasing distribution function F_k of the random variable $T_n(t_k)$. The convergence $\xi^{(N_0)}_{j(k)}(k) \xrightarrow[N_0 \to \infty]{} f(t_k)$ P almost everywhere implies the convergence in probability i.e. for arbitrary small $\delta_1 > 0$, $\delta_2 > 0$ there exists such a integer $\tilde{N}_0$, that for $N_0 > \tilde{N}_0$ we have

$$P(|\xi^{(N_0)}_{j(k)}(k) - f(t_k)| \geqslant \delta_1) < \delta_2, \tag{6.1}$$

or equivalently

$$P(\xi^{(N_0)}_{j(k)}(k) \geqslant f(t_k) + \delta_1) + P(\xi^{(N_0)}_{j(k)}(k) \leqslant f(t_k) - \delta_1) < \delta_2. \tag{6.2}$$

If $\Phi^{(k)}_{j,N_0}$ denotes a distribution function of the random variable $\xi^{(N_0)}_j(k)$, then the unequality (6.2) can be written in the following way

$$1 - \Phi^{(k)}_{j(k),N_0}(f(t_k) + \delta_1) + \Phi^{(k)}_{j(k),N_0}(f(t_k) - \delta_1) < \delta_2 \tag{6.3}$$

and then using the formula (5.10) we obtain

$$1 - \frac{N_0!}{([N_0 \cdot p_k])!(N_0 - N_0 \cdot p_k] - 1)!} \int_{\tilde{F}_k(f(t_k)-\delta_1)}^{\tilde{F}_k(f(t_k)+\delta_1)} t^{[N_0 \cdot p_k]} \cdot (1-t)^{N_0 - [N_0 \cdot p_k] - 1} dt < \delta_2, \quad (6.4)$$

where $\tilde{F}_k$ is a distribution function of the random variable $f(t_k + T_n(t_k))$.

If inside intervals where the function f is continuous, this function has the Lipshitz property i.e. there is $L > 0$, such that if f is continuous on an interval (x_1, x_2) then for $y_1, y_2 \in (x_1, x_2)$ we have $|f(y_1) - f(y_2)| \leqslant L|y_1 - y_2|$. In such a case, we can replace the condition (6.4) about N_0 by new one (6.5) stronger but simpler to verify and independent from the input signal f. If the condition (2.4) is satisfied then

$$1 - \frac{N_0!}{([N_0 \cdot p_k])!(N_0 - [N_0 \cdot p_k] - 1)!} \int_{F_k(-\delta_1')}^{F_k(\delta_1')} t^{[N_0 p_k]} (1-t)^{N_0 - [N_0 \cdot p_k]^{-1}} dt < \delta_2, \quad (6.5)$$

where $\delta_1' = \delta_1/L$ and F_k is a distribution of the random variable $T_n(t_k)$.

Similarly, if the condition (2.5) is satisfied then (6.4) can be replaced by the following condition (6.6)

$$1 - \frac{N_0!}{([N_0 \cdot p_k])!(N_0 - [N_0 \cdot p_k] - 1)!} \int_{1-F_k(\delta_1')}^{1-F_k(-\delta_1')} t^{[N_0 \cdot p_k]} (1-t)^{N_0 - [N_0 \cdot p_k]^{-1}} dt < \delta_2. \quad (6.6)$$

The following formula (6.7) holds for all integers N_0 satysfying conditions (6.5) and (6.6)

$$P(|\xi_{j(k)}^{(N_0)}(k) - f(t_k)| > \delta_1) < \delta_2. \quad (6.7)$$

Then, we obtain *"global error assessemnt"* for sampling points t_k from the set U. For arbitrary fixed $\delta_1, \delta_2 > 0$, if N_0 satisfies the conditions (6.5) and (6.6) then

$$P(\exists_k |\xi_{j(k)}^{(N_0)}(k) - f(t_k)| > \delta_1) < M\delta_2. \quad (6.8)$$

Of course, if N_0 is large enough then we can (according to the theorem 2.5) replace in the formula (6.3) the distribution of the random variable $\xi_{j(k)}^{(N_0)}(k)$ by a Gaussian distribution ($\xi_{j(k)}^{(N_0)}(n)$ has the probability distribution asymptotically Gaussian).

Systematic errors outside the set of correctness U, i.e. values $|f(t_k) - a_k|$ (where $\xi_{j(k)}^{(N_0)}(k) \to a_k$ P almost everywhere) for $t_k \notin U$, may be significant, but can be easily assessed. For example if the input signal f is a diferentiable function on R then the error does not exceed the value $\varepsilon \cdot \sup_{t \notin R} |f'(t)|$. Besides, if $t_k \notin U$ then for some (sufficiently vicious) input signals f, a small shift of extremes of the function $\{1, 2, \ldots, M\} \ni k \to a_k \in R$ (in comparison with the function $\{1, 2, \ldots, M\} \ni k \to f(t_k) \notin R$) can occur. Mostly, this local *"phase shift"* introduced by the algorithm is not important in practical measurements. The convergence (5.5) of the quantile statistics sequence is much slower than the convergence of the mean values sequence. It

follows from comparison of variances of random variables $\xi_{j(k)}^{(N_0)}(t)$ and $MN(N_0, t_k)$ for fixed N_0, because these variances can be considered as a kind of measure of convergence velocity. It is obvious that the variance of the random variable $\xi_{j(k)}^{(N_0)}(k)$ depends on a solpe of the function f in a ε-neighbourhood of t_k. As a consequence, we obtain "larger scattering" of random variable $\xi_{j(k)}^{(N_0)}(k)$ realizations for function with greater slopes.

Example 1. Assume the input signal is the step function $1(t - t_0)$ and sampling time jitter satisfies assumptions of the corollary 2.3. Then median/quantile algorithm restores exactly in the limit (i.e. when $N_0 \to \infty$) input signal outside the point t_0 (see fig. 6.1). Output data of the algorithm obtained for finite values N_0 ($N_0 = 8, 16, 32, 64, 110$) by computer simulation are shown in the fig. 6.2. Gaussian distribution of sampling time jitter was assumed with standard deviation $\sqrt{D^2(T(t_k))} = 20$ ps and the mean value $E(T(t_k)) = 0$ for each $k \in \{1, ..., M\}$. In practice, for ideal restoration of the step function $N = 100$ is sufficient.

Described above estimation algorithm for $f(t_k)$ (where $k = 1, 2, ..., M$ and $t_k \in U$) is correct when the signal f is not additively noised. If the sequence of samples $(f(t_k + T_n(t_k)))_{n=0}^{\infty}$ is additively noised then the sampling process for each $k \in \{1, ..., M\}$ is described by the sequence $(f(t_k + T_n(t_k)) + Q_n(t_k))_{n=0}^{\infty}$, where for each

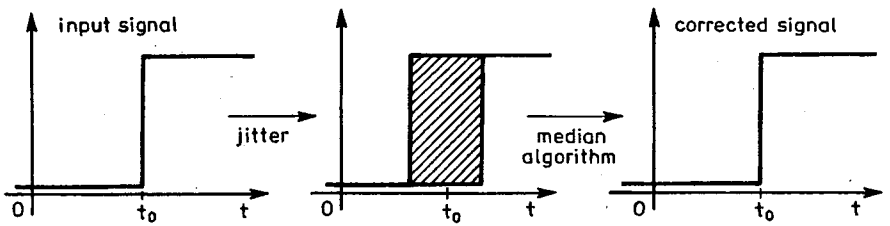


Fig. 6.1. Signal restoration with the median/quantile algorithm for the step function

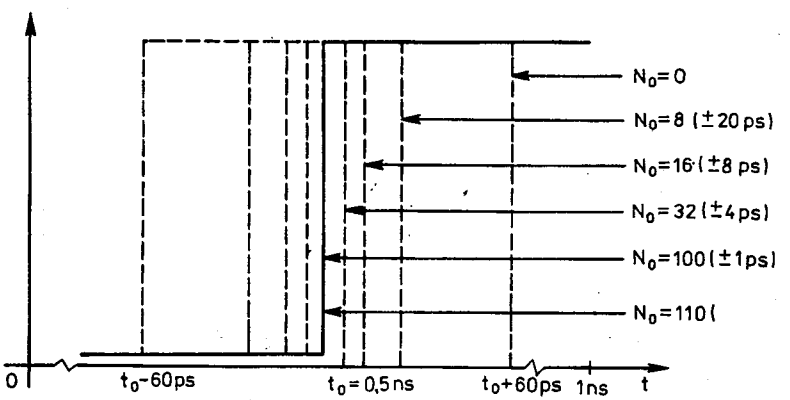


Fig. 6.2. Intervals of erroneous restoration of the step function (computer simulation results)

fixed n i k random variables $T_n(t_k)$ and $Q_n(t_k)$ are independent. The sequence $(Q_n(t_k))_{n=0}^{\infty}$ describing the noise is for each $k \in \{1, 2, ..., M\}$, a random sequence of independent random variables with the same probability distribution. The following fact holds. If we have two real, independent random variables X and Y and a_λ denotes a quantile of the order λ of the random variable X then (also if the random variable Y has a symmetric probability distribution related to 0) the number a_λ have not to be a quantile of the order λ of the random variable X + Y. It is not difficult to give appropriate examples. Therefore, in general, if $f(t_k)$ is a quantile of the order λ of the random variable $f(t_k + T_n(t_k))$ then $f(t_k)$ have not to be a quantile of the order λ of the random variable $f(t_k + T_n(t_k)) + Q_n(t_k)$.

Using proposed quantile statistics algorithm in the case when the input signal is additively noised, we can introduce a systematic error. If, for example, beside admitted assumptions about f, t_k, $T_n(t_k)$ and $Q_n(t_k)$ we additionaly assume, that random variables $T_n(t_k)$ and $Q_n(t_k)$ are for each n, k symmetric related to 0 (0 is in such a case median of the random variables $T_n(t_k)$ and $Q_n(t_k)$) and for each $k \in \{1, 2, ..., M\}$ the sequence $(f(t_k + T_n(t_k)) + Q_n(t_k))_{n=0}^{\infty}$ is a sequence of independent random variables with the same probability distribution and the same strictly monotonic distribution function then the value of above mentioned systematic error is equal to $|f(t_k) - a'_k|$, where a'_k is a median of the random variable $f(t_k + T_n(t_k)) + Q_n(t_k)$.

Under above assumptions, if the function f can be expanded in power series (in particular f can be a polynome), a distribution function of the random variables $Q_n(t_k)$ is strictly monotonic then the systematic error $|f(t_k) - a'_k|$ for $t_k \in U$ is mainly influenced by even order derivatives $f^{(i)}(t_k)$ and if these derivatives are equal to 0 then the systematic error is also equal to 0. It follows from the following fact. The sum of two symmetric independent random variables with unique medians equal to zero has the unique median equal to zero too. In particular if f is a straight line then $f(t_k)$ remains unique median of the random variables sum $f(t_k + T_n(t_k)) + Q_n(t_k)$ and the systematic error do not occur.

Described above jitter correction algorithm based on quantile statistics was tested by computer simulation for different input signals, different values N_0, quantizations and jitter distributions proving its usefulness. Advantage of the algorithm is its simplicity (in both software and hardware implementation) and robustness. We do not assume specific jitter distributions. Other important advantage of the method is possibility of its parallelization. For each point t_1, t_2, ..., t_M computations are independent and implemented with the same very simple procedure. Disadvantage of the algorithm is fast increasing with N_0 time of computations, proportional (depending on sorting algorithm) to $N_0^2 \cdot M$ or $M \cdot N_0 \log N_0$. Also convergence $\xi_{j(k)}^{(N_0)}(k) \to f(t_k)$ when $N_0 \to +\infty$, observed in the computer simulation is relatively slow. For example, for linear signal with the slope 1 V/ns, Gaussian jitter distribution with $\sigma = 20$ ps, without quantization and $N_0 = 8$, error do not exeed 0.01 V and for $N_0 = 1024$, 0.001 V.

7. IMPLEMENTATION OF THE MEDIAN/QUANTILE METHOD

Implementation of the above described sampling jitter correction alorithm can be software or hardware one. Software implementation is practically, despite of the simplicity of the method, very time consuming, particularly for large N_0. Then using specialized systolic arrays for rank statistics computation (i.e. systolic sorting circuits) seems to be very useful, when we want to have real time correction. A systolic circuit (or systolic array) is a digital network which satisfy the following three conditions:

– regularity and modularity of the architecture,
– locality of connections,
– pipelining and parallel processing.

The first two qualities facilitate VLSI implementation, the third one determines the throughput of the system.

All proposed, in the sequel, systolic sorting circuits are based on specific sorting algorithms (see for instance [3], [4], [6]). The simplest sorting procedure called in [4] *"bubllesort procedure"* is the following:

```
procedure bubblesort (a):
    i, j: integer;
    a: array[1..N0] of real;
    for i: = 1 to N0 do                                        (7.1)
    for j: = 1 to N0−i (or to N0−1) do
    if a[j] > a[j+1] then change positions a[j] and a[j+1],
```

but much more useful as a start point for hardware systolic solution is so called modified bubblesort procedure which (assuming $N_0 \geqslant 3$), is the following:

```
procedure modified bubblesort (a):
    i, j: integer;
    a: array[1..N0] of real;
    for i: = 1 to N0 do                                        (7.2)
    begin
if not 2|i then for j: = 1 to N0−1 step 2 do
if a[j] > a[j+1] then change positions a[j] and a[j+1];
if 2|i then for j: = 1 to N0−1−2·[N0+1]2 step do
if a[j+1] > a[j+2] then change positions a[j+1] and a[j+2];
        end;
```

Computational complexity of this algorithm is proportional to N_0^2. Described below sorting circuits implement a parallelized version of this algorithm.

Main questions about sorting systolic arrays are the following:

– Is a proposed systolic array really the sorting one?
– How are the throughput and delay of the array?

– Is the circuit well suited to VLSI implementation?
– Is the circuit flexible i.e. whether we can easily obtain a large sorting circuit from small ones?

The problem posed in the first question is in general much more complicated than it seems (see for example [3] and [4]) and is beyond the scope of the paper. Three last questions are discussed below.

It is well known (see [3]) that the median and arbitrary quantile for the N_0 dimensional random vector can be computed in the mean time proportional to $N_0 \log N_0$ (quicksort procedure). But the same procedure must be repeated for each sample number $k = 1, 2, \ldots, M$ then the overall computational complexity of the proposed median/quantile algorithm is proportional to $M \cdot N_0 \log N_0$. In practice the integer numbers M and N_0 are rather large for example $M = 1024$, $N_0 = 4096$, hence for real time capabilities of the jitter corrector, we need a hardware

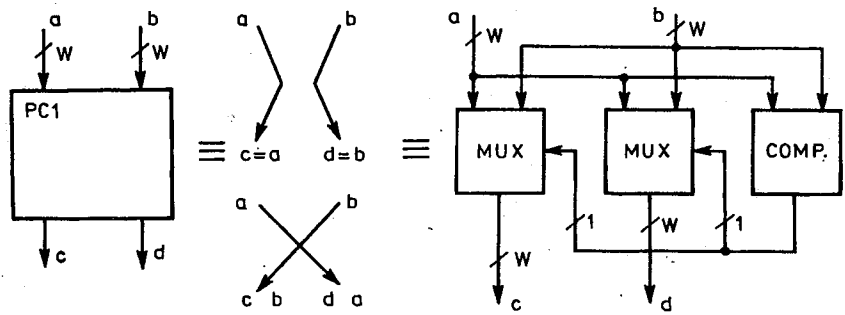


Fig. 7.1. The PC1 elementary cell, parallel implementation, W – the lenth of computer word

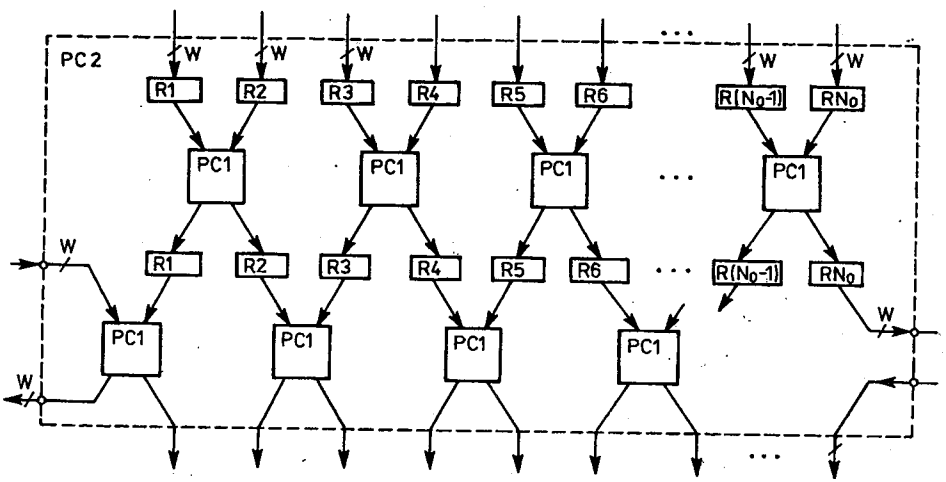


Fig. 7.2. The PC2 elementary cell (one level of the sorting systolic array from the fig. 7.7 a)) with parallel word processing

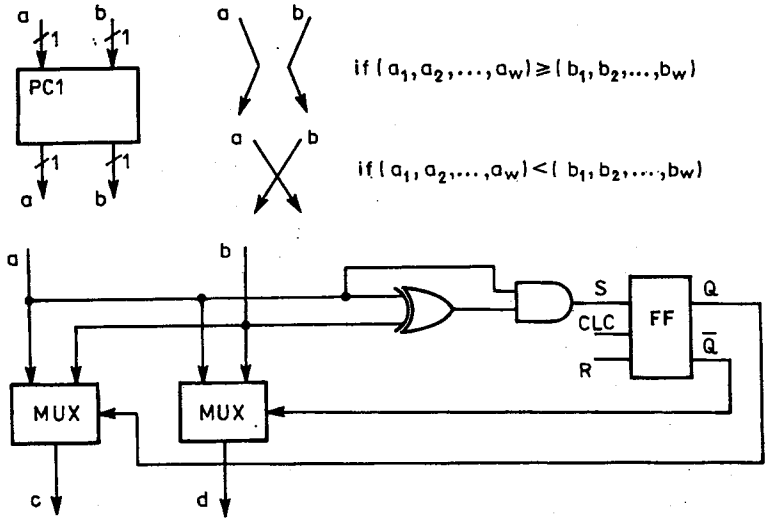


Fig. 7.3. The PC1 elementary cell with serial word processing, where $a = (a_1, a_2, \ldots, a_w)$, $b = (b_1, b_2, \ldots, b_w)$, $c = (c_1, c_2, \ldots, c_w)$, $d = (d_1, d_2, \ldots, d_w)$, $a_i, b_i, c_i, d_i \in \{0, 1\}$ and $\leqslant$ is a lexicographic order in the set $(0, 1)^w$

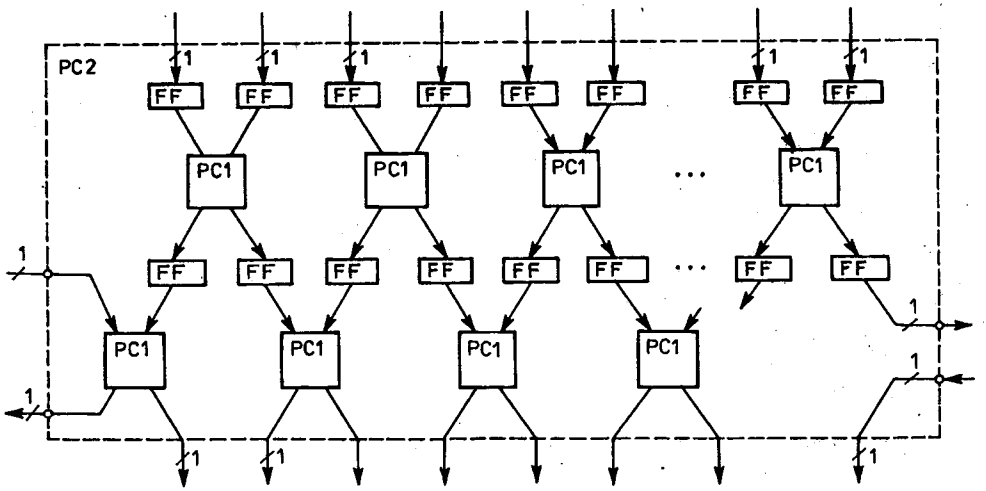


Fig. 7.4. Serial implementation of the elementary processing cell PC2

implementation of the sorting algorithm. Such a universal implementation well suited to VLSI techniques is shown in figs. 7.6 and 7.7. Delays denoted by points are the D-type flip-flops (or parallel-in parallel-out registers). Compare/exchange circuits PC1 implement delayed operation: if $a > b$ then ($c: = b$; $d: = a$) else ($c: = a$; $d: = b$). More information about circuits for quantile computation can be found in [1].

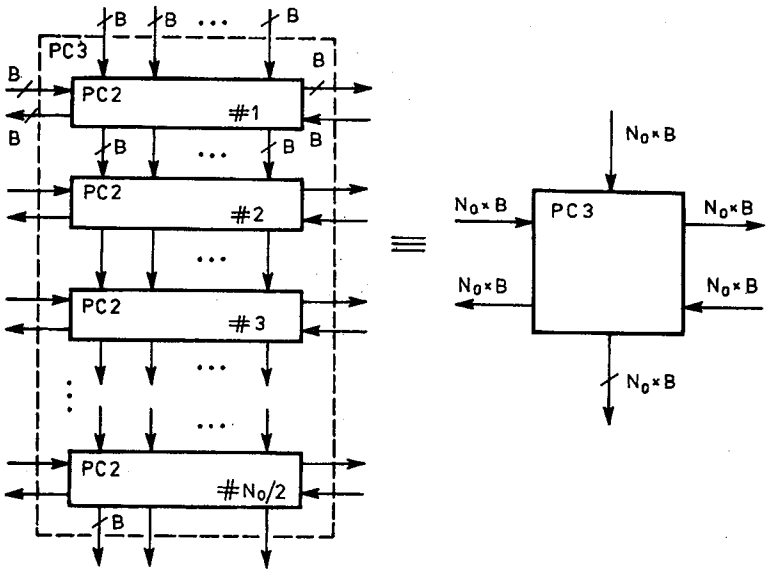


Fig. 7.5. The processing cell PC3 composed of $N_0/2$ levels (one level = PC2 cell). This connection gives as a result sorting systolic array (B = W for parallel word processing and B = 1 for serial implementation)

Assume, $N_0 \in N$ denotes the length of the finite sequence $(a_n)_{n=1}^{N_0}$, where $\{a_n; 1 \leqslant n \leqslant N_0\} \subseteq A$ and $(A \cdot \leqslant)$ is and arbitrary totally ordered set. Sorting (i.e. ordering) from formal point of view is equivalent to computing such a permutation $\pi: \{1, 2, ..., N_0\} \to \{1, 2, ..., N_0\}$ that $a_{\pi(1)} \leqslant a_{\pi(2)} \leqslant ... \leqslant a_{\pi(N_0)}$. Algorithm which performs sorting (i.e. computes the permutation π) is called a sorting algorithm. The sequences $(a_n)_{n=1}^{N_0}$ and $(a_{\pi(n)})_{n=1}^{N_0}$ are respectively the input and output data of the sorting algorithm. The set A can be an arbitrary totally ordered set for example R^n or the set C of complex numbers with the lexicographic order but as a rule in statistical applications A is the set R of real numbers.

Assume that $2|N_0$ (i.e. N_0 is even) and for simplicity a_n are integers from the interval $[0, m]$, where $m \in N$. Sorting algorithm chooses k-th (in the value) element $a_{\pi(k)}$ of the ordered sequence for arbitrary $k \in \{1, 2, ..., N_0\}$ then can be considered as a procedure for median (quantile or, in general, rank statistics) computing. For fixed k algorithms of $a_{\pi(k)}$ computation are a little simpler than full sorting algorithms. A variety of sorting algorithms and their properties are described in details in monographs [3] and [4]. Modified bubblesort procedure is less effective in sequential implementation (when compared for example with quicksort procedure) but can be easily parallelized.

We can assume that N_0 is even, without loss of generality because for N_0 odd, we can use a sorting circuit with N_0+1 inputs applying maximal number of the representation to one of them. On the other hand N_0 even gives natural symmetry and modularity to the circuit.

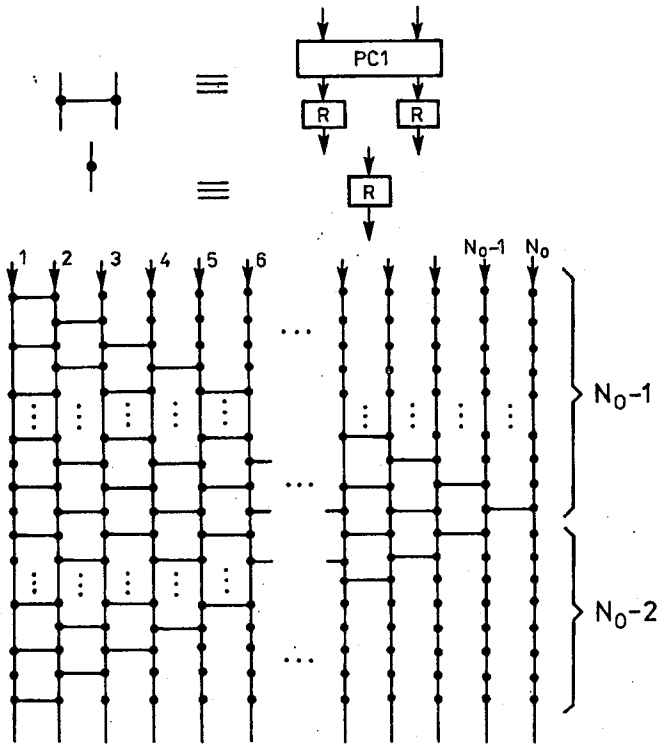


Fig. 7.6. Implementation of the classical *"bubblesort"* algorithm as a systolic array composed of $m = 2N_0 - 3$ levels

Systolic arrays presented in the sequel consist of three kinds of elementary processing cells PC1, PC2, and PC3. The processing cell PC1 is a pair of multiplexers driven by a comparator and is described in fig. 7.1. The PC1 cell implements a compare/exchange operation. The PC2 cell, shown in fig. 7.2 is one level of all proposed systolic circuits with parallel word processing and consist of N_0 PC1 cells and $2N_0$ registers. All registers have W-bits length equal to the length of processed words representing numbers. The processing cell PC3 (see fig. 7.5) is a set of $N_0/2$ cascaded PC2 cells. The solution of PC1, PC2 and PC3 as circuits with parallel word processing is fast but rather expensive. Besides, large number of pins (for large N_0) makes difficult on chip realization. More convenient for on chip implementation seems to be solution of the elementary processing cells as a serial word processing devices. The idea is similar to the distributed arithmetic concept used widely in DSP. The PC1, PC2 and PC3 cells with serial processing are shown respectively in fig. 7.3, 7.4 and 7.5.

The serial PC1 has a flip-flop FF reset every W cycles of the clock. The first inequality in bit comparison (indicating that the left side input is greater then the

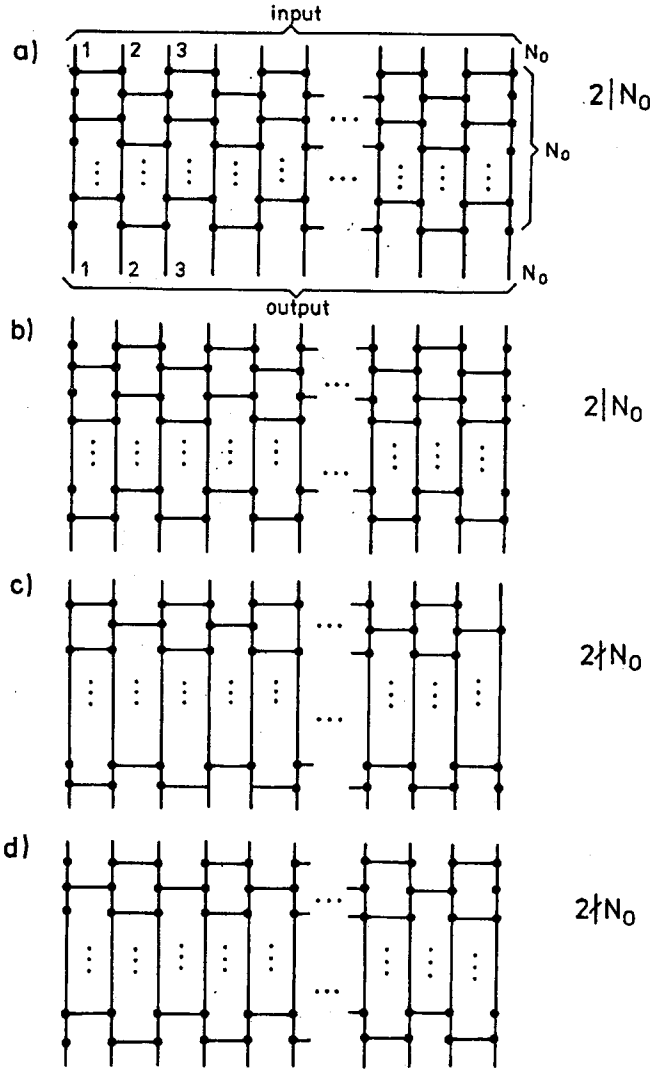


Fig. 7.7. Systolic sorting arrays with N_0 levels

right one) set FF to 1. Two input multiplexers MUX choose right side input if controlled by Q = 0 and left side input if Q = 1. Circuits from fig. 7.8 can be treated as modifications of the PC3 cell shown in fig. 7.5.

Systolic circuits with serial word processing are not so fast as systolic arrays with parallel word processing (serial method is exactly W times slower) but for the same N_0 have W times less transistors and connecting pins.

Presented sorting circuits are systolic arrays using described basic processing cells PC1, PC2 and PC3. We need $N_0/2$ cascaded PC2 cells or one PC3 appropriately

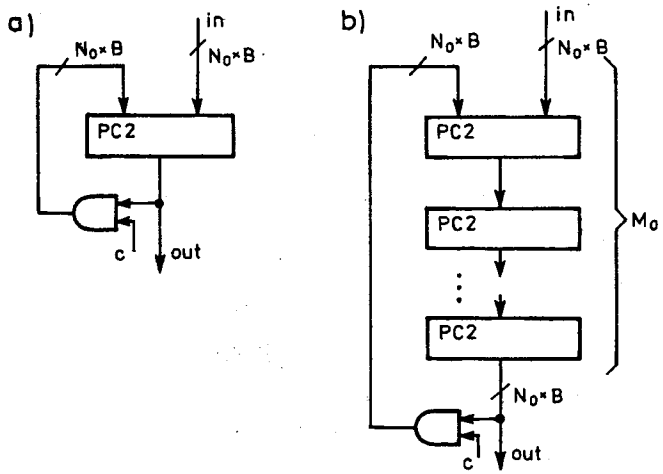


Fig. 7.8. One level a), and M_0 level connection b) of PC2 cells giving as a results sorting circuits (where $M_0|(N_0/2)$, c is a control signal, B=1 for serial word procesing and B=W for parallel one)

connected (see circuits in fig. 7.5, 7.6 and 7.7) to sort N_0 binary numbers. From theoretical point of view these circuits implement lexicographic sorting of binary vectors. Circuits shown in figs. 7.5 and 7.7 a), b), c), d) compute the output result (i.e. sorted sequence of numbers) in N_0 clock cycles in parallel case and WN_0 in serial solution. The input/output delay of the circuits depends on number of levels. Of course we seek sorting circuits with minimal number of levels and minimal number of PC1 cells. If the throughput of the circuit from fig. 7.8 a) is A then it is AM_0 and $AN_0/2$ appropriately for circuits shown in fig. 7.8 b) and 7.7. The number M_0 of levels (or more precisely cascaded PC2 cells) must be divisor of $N_0/2$. When $M_0 = N_0/2$ circuits from figs. 7.7 and 7.8 b) are equivalent. Shortcoming of the circuit from fig. 7.8 is more complicated synchronization of the data flow in comparison with the circuit from fig. 7.7 but in the second concept we can easily exchange the throughput and complexity of the system. A method of extending the number N_0 of inputs is explained in the figs. 7.9 and 7.10.

Described above circuits sort the all input sequence $(a_n)_{n=1}^{N_0}$. When we need only k-th element of the sorted sequence i.e. $a_{\pi(k)}$ (for example when median or other rank statistics is computed), we can significantly simplify sorting systolic array, cutting surplus parts of the circuits, using the principle shown in fig. 7.11. For example when the median is computed, we need only 3/4 part of the area of the full sorting circuit but number of level (then the input/output delay of the sorting circuit) remains the same.

An interface between the systolic sorting array and host computer is shown in fig. 7.12.

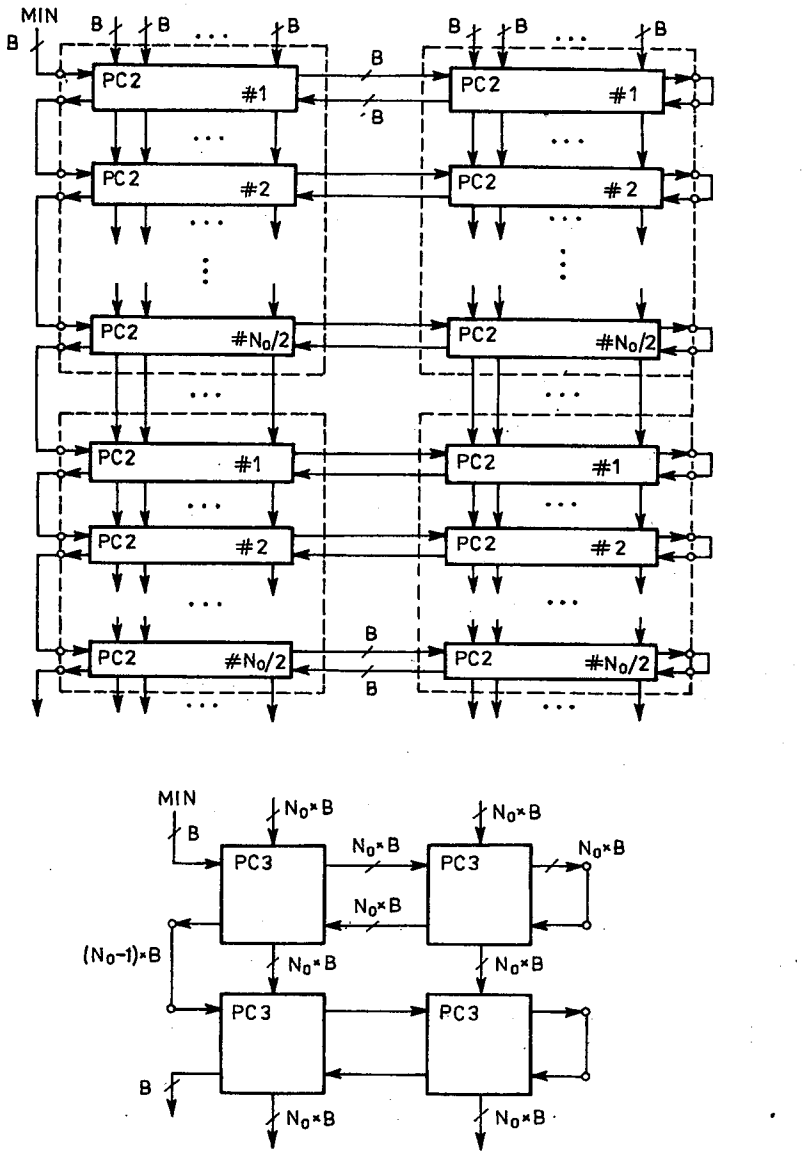


Fig. 7.9. The connection of four PC3 cells extending the input two times (B = 1 for serial word processing and B = W for parallel one)

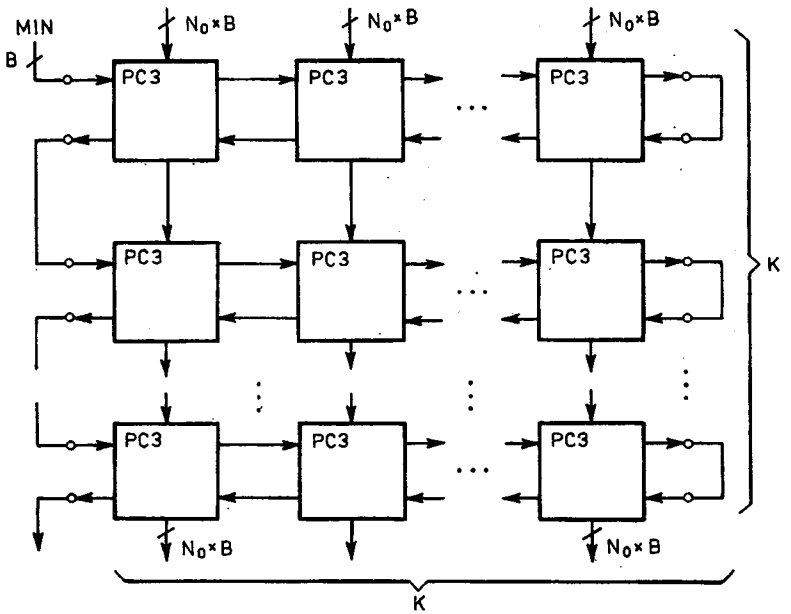


Fig. 7.10. The connection of K^2 PC3 cells extending the input K times

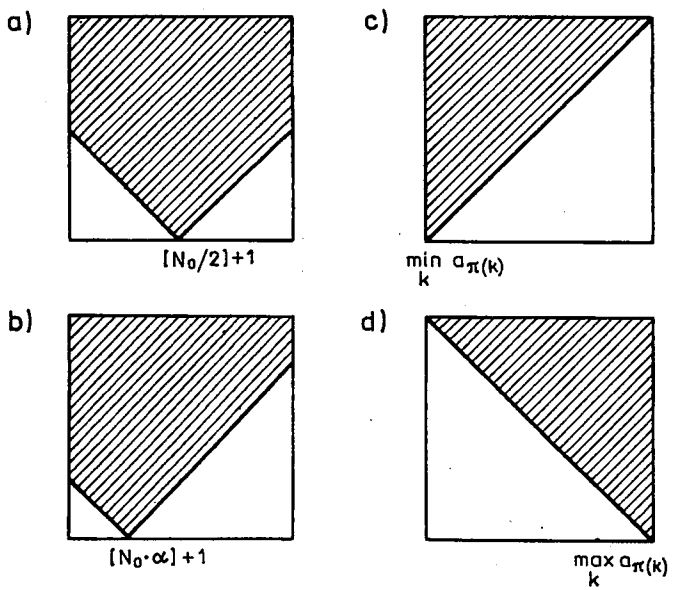


Fig. 7.11. How to modify a sorting systolic array for a) median computation (choosing the element number $[N_0/2]+1$ from a sequence $(a_n)_{n=1}^{N_0}$) b) quantile of the order α computation (choosing the element number $[N_0\alpha]+1$) c) the minimal value computation d) the maximal value computation. Grey area is important part of the circuit, white surplus one

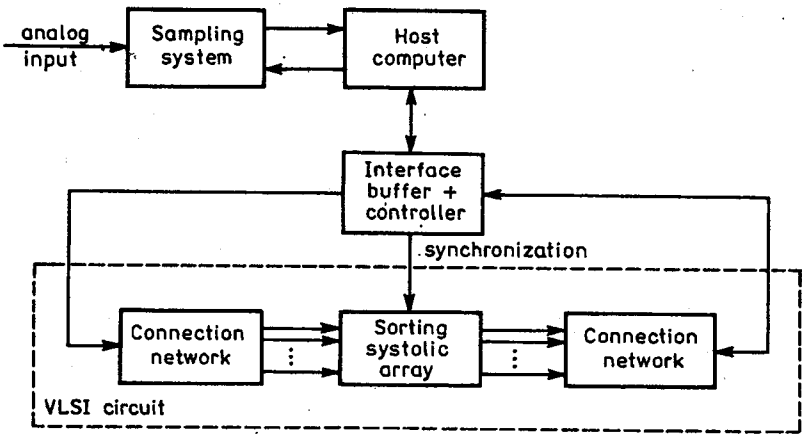


Fig. 7.12. Interface between systolic sorting array, host computer and sampling system

8. CONCLUSIONS

1. Presented in the paper a median/quantile filtering method is relatively simple in both software and hardware implementation (when compared for instance with deconvolution method) and useful in jitter correction particularly for pulse signals. The method is quite correct for a large class of signals but in general can introduce some errors. The signal image is not longer scattered on the screen. We can say the sampling time jitter is filtered. It is also important that for proposed algorithms the input information about random variables $T(t_k)$ is simple to identify.

2. A variety of fast systolic arrays for sorting procedures were suggested. The high degree of modularity and locality of connections in described solutions significantly facilitate their on chip implementation. It seems that the most convenient sorting algorithm for hardware purposes is the modified bubble sort procedure. The designer can easily exchange complexity of the system and its throughput. Presented circuits can be of course used to fast median/quantile computations. This feature give possibility of design sampling time jitter correctors in digital oscilloscopes based on quantiles computation and working in real time. The discussed circuits can be also used for median or quantile filter implementation in image processing (see for instance [12]). It is worth to underline a remarkable connection flexibility in described sorting circuits. From N_0 input circuits we can (by simple connection in appropriate two dimensional matrix) obtain the $n \cdot N_0$ input sorting circuit.

3. In modern electronics electronic (digital and analog) circuits can be considered as *"silicon implementations"* of algorithms. Such an general idea seems to be a contemporary approach to electronic circuits & systems and is similar to ASIC concept.

List of Symbols

$m \mid n$	– m is a divisor of n
Z	– set of integers
N	– set of natural numbers
R	– set of real numbers
$\langle m, n \rangle$	– set $(i \in Z;\ m \leqslant i \leqslant n)$
B(R)	– σ-field of Borel sets of R
B(X)	– σ-field of Borel sets of metric space X
C	– set of complex numbers
R^n	– n-dimensional Euclidean space
$\leqslant$	– relation of linear ordering
$[x]$	– integer part of a real number x
$[n]_m$	– n modulo m i.e. the residue modulo m
$(\Omega, \mathcal{M}, P)$	– probabilistic space

REFERENCES

1. T. Adamski: *Systolic Circuits for Fast Median Computation with Application to Sampling Time Jitter Correctors.* Electronics and Telecommunication Quarterly; 1991, No. 4
2. W. Gans: *The Measurements and Deconvolution of Time Jitter in Equivalent Time Waveform Samplers.* IEEE Proceedings; March 1983
3. A.V. Aho, J.E. Hopcroft, J.D. Ullman: *The Design and Analysis of Computer Algorithms.* Addison-Wesley, 1974
4. D. Knuth: *The Art of Computer Programming.* vol. 3. *Sorting and Searching.* Addison-Wesley 1973
5. C. Groetsch: *The Theory of Tikhonov Regularization for Fredholm Equations of First Kind.* Pitman 1984
6. E.M. Reigold, J. Nivergelt, N. Deo: *Combinatorial Algorithms. Theory and Practice.* Warszawa: PWN 1985
7. S.Y. Kung, H.J. Whitehouse, T. Kailath: *VLSI and Modern Signal Processing.* Prentice-Hall 1985
8. S.Y. Kung: *VLSI Array Processors.* IEEE ASSP Magazine, July 1985
9. W. Moore, A. McCabe, R. Urquhart: *Systolic Arrays.* Adam Hilger, Bristol 1987
10. C. Mead, L. Conway: *Introduction to VLSI Systems.* Addison-Wesley, Bristol 1987
11. D.P. Agrawal: *Advanced Computer Architecture.* IEEE Computer Society Press, 1985
12. T.S. Huang: *Two-dimensional digital signal processing.* Springer, Berlin, Heidelberg, New York 1981
13. J. Bartoszewicz: *Wyklady ze statystyki matematycznej.* Warszawa: PWN 1989
14. M. Fisz: *Rachunek prawdopodobieństwa i statystyka matematyczna.* Warszawa: PWN 1967
15. R.J. Serfling: *Aproximations Theorems of Mathematical Statistics.* John Wiley, 1980

T. ADAMSKI

ZASTOSOWANIE FILTRACJI KWANTYLOWEJ DO KOREKCJI CHWILOWEJ NIESTAŁOŚCI MOMENTU PRÓBKOWANIA W CYFROWYCH SYSTEMACH POMIAROWYCH

Streszczenie

Niestałość chwilowa momentu próbkowania występująca w szerokopasmowych oscyloskopach cyfrowych może powodować (a zależy to od kształtu sygnału mierzonego) znaczne zniekształcenia pomiaru. W pracy została przeprowadzona analiza błędów wprowadzanych przez chwilową niestałość momentu próbkowania oraz zaproponowano metodę korekcji tych błędów wykorzystującą filtrację kwantylową. Podano również szereg rozwiązań układowych służących do obliczania statystyk kwantylowych. Proponowane układy są szybkimi układami systolicznymi dobrze nadającymi się do wykonania jako układy scalone typu ASIC.

KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJI, 1993, 39, z. 1, ss. 119-130

519.68
621.396.6

The data exchanges and the messages flows in the wireless robotics microprocessor DELTA computer networks

KAZIMIERZ BIEŃKOWSKI, MARWAN GHABALLY

Instytut Informatyki, Politechnika Warszawska

Received 1992.06.02

Authorized 1992.09.09

In the paper the analysis of the new microcomputer DELTA network is given. The DELTA network is oriented towards the applications of the computer tools in the heavy machines area (e.g. excavators, bulldozers, self out-loading trucks, etc.). The paper continues and expands the results achieved during the period 1986—1990 (now there is the continuation of this under the grant from the Polish Scientific Research Committee), when the two institutes from the Warsaw Institute of Technology (WIT) were realizing the No. CPBP 02.13 Scientific topic, named " The applications of the artificial intelligence methods in the heavy machinery and in the moving vehicles". These were: the Institute for Computer Science (ICS) and the Institute for Heavy Working Machinery (IHWM) from WIT, but also several other Polish technical institutes were involved in this (these were the technical universities from Gdańsk, Łódź, Poznań, Wrocław and some other non-university institutes). There are given: the definitions, the analysis of the state graphs, the lists of messages, the lists of the statuses and the lists of commands for the three types of the data stations (the host computers) that are used in the DELTA networks. These are: the On-board Data Stations (ODSs), the Remote-control Data Stations (RDSs) and the Ground Data Stations (GDSs). In DELTA the mechanisms for the co-operation of DELTA with any general purpose computer networks are provided. This necessity may arise, e.g. when the use of maps, the geological and other huge technical information should be provided in DELTA. All the results given in the paper may be used for the simulations and for the system software generation, as well as for the hardware efficiency analysis in DELTA.

1. INTRODUCTION

In the last time we observe that the designers of the heavy machines are very interested to include microprocessors on them [1], and to make the heavy machine as

This paper was prepared under the auspices of doc. dr hab eng K. Bieńkowski, and it's one of the M. Ghabally's researches for Ph.D. work.

full automatic. Also these designers are interested in some remote control devices to control the heavy machines. This paper explains some portion of the so called DELTA system and the mutual messages between its stations, especially, related to the machine actions.

The DELTA system is a computer, robotics, intelligence and multimicroprocessor network [2], consisting at least of the three Data Stations (DSs), every of each having a microprocessor, or microcomputer [3,4], termed here as the „Data Station". Each DS in DELTA works accordingly to its design and its program, with the possibility to communicate and having interrupts from the other DSs, so this computer system can be looking at as the „Open System Computer Network".

The DELTA system applications are specially useful in the heavy robotics machine systems that work within unnatural climate, the human being can't endure it (example: poisonous gas zones, nuclear radiation files, rugged zones, etc.) or in some environments that force the worker to leave the machine's board, and to control the work from many sides on the ground, [6]. There are the following types of these DSs in DELTA:

The On-board Data Station (ODS):
The On-board Data Station (ODS), equipped with the microprocessor, is placed on the Robot Machine's (RM), e.g., excavator, board. The ODS contains all the control software and hardware necessary to work with, for instance for action and stabilization protecting the RM from falling down at the action time [5].

The Remote Data Station:
The Remote Data Station (RDS), in DELTA is a carried (portable) steering device used by a supervising worker who is forced to be aside of the working area, or from the RM for the safety or other reasons. The RDS contains the microprocessor, the monitor screen, the joysticks, the indicators and all the necessary keys, to control remotely the RM in the full extent.

The Ground Data Station:
The Ground Data Station (GDS) is a fixed computer station that stays out of the side of the working zone, because of having not any bind on the size, the weight, the energetical power, the processing power and any of the environment dangers. So, this GDS may have the very large capacity memory, the full content of complete information about the work environment and all the necessary information about the system. From the processing power point of view this GDS is the "*master*" of the DELTA system. The GDS has the possibility to communicate all DELTA stations with another system or networks by GDS's communicating possibilities. The general idea of DELTA was sketched in 1986–1988 period ([1], [2], [3], [4], [5], [6]). The first possibility for DELTA's implementation arised during the 1987–1990 period, under the auspices of the CPBP 02.13 Polish Basic Scientific Research Program (the exact name for this was: "The Artificial Intelligence Systems in the Working and Moving

Machines"). During the CPBP 02.13 realization only the problems of the ODS design were resolved ([7], [8]).

This paper is destined to make a further progress of the full (ODS-RDS-GDS) DELTA research. Some points of this were made in [9] and in [10], but the problem of the full (ODS-RDS-GDS) DELTA remains. In this paper we claim for the originality (priority) especially for the full DELTA state graph synthesis, considering the wireless method of the interstation information messages exchanges.

2. THE FULL DATA DESCRIPTION IN DELTA SYSTEM

2.1. THE SERVICE PRIMITIVES

A service is formally specified by a set of primitives available to a user or other entity to access the service [11]. These primitives tell the service to do some action or report on an action taken by a peer entity. In the ISO/OSI model, the Service Primitives (SP) can be divided into four classes as follows ([11], [12], [13]):

Request Primitive (Rq) is using to invoke a function or to get the work to be done (for example: to establish a connection).

Indication Primitive (In) is using to:

a. invoke a function or,

b. show a function (work) has been invoked at a Service Access Point (SAP), Fig. 1a [14].

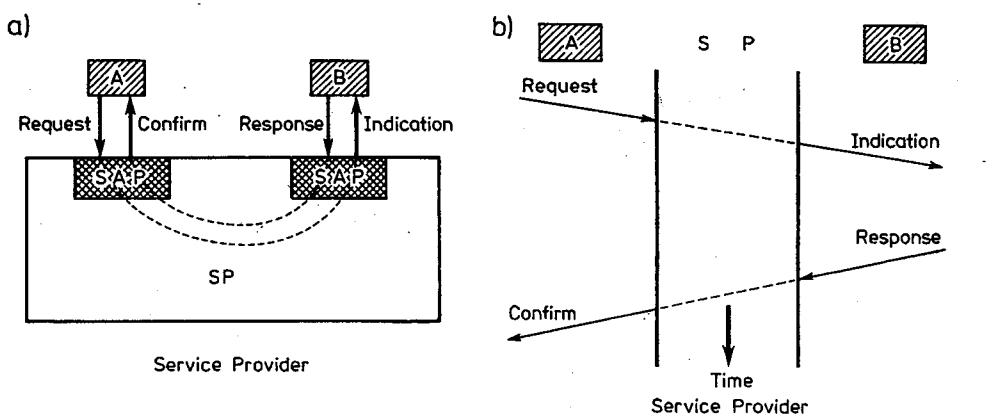


Fig. 1 Services Primitives

Response Primitive (Rs). The peer entity (side B in Fig. 1a) uses the Rs to inform another entity, it wants, about accepting or rejecting the proposed connection.

Confirm Primitive (Cf). The originating protocol entity creates an Cf and passes this up to the user, to be informed about its request [12]. The SPs can have parameters, as about the maximum message size to be used on the connection, and about the type of service desired.

Services can be either confirmed or unconfirmed. In confirmed service, there are all that four SPs, but in an unconfirmed service, there are just a *Rq (request)* and an *In (indication)*. The OSI/ISO CONNECT message is always a confirmed service, because the remote peer must agree to establish a connection, [11]. The ISO/OSI is abbreviated from the term "International Standard Organization/Open System Interconnection".

2.2. THE SERVICES PRIMITIVES IN ODS

The ODS receives the action commands from each other station (GDS, RDS), so SP in ODS contains: a CONNECT to establish connection with the other stations. When the ODS wants to send a message, its SP contains a DATA. The ODS DATA message consists of the addresses of: the target station (TS), the source station (SS), then also of the proper message (M) and the relevant data. the DATA message is unconfirmed. The DISCONNECT message is for disconnect a communication.

The list of the discussed ODS SPs is as follows:

CONNECT.request-TS-ODS
CONNECT.indication-TS-ODS
CONNECT.response-ODS-SS
CONNECT.confirm-ODS-SS
DATA.request-TS-ODS-M_data
DATA.indication-TS-ODS-M_data
—
—
DISCONNECT.request-TS-ODS
DISCONNECT.indication-TS-ODS

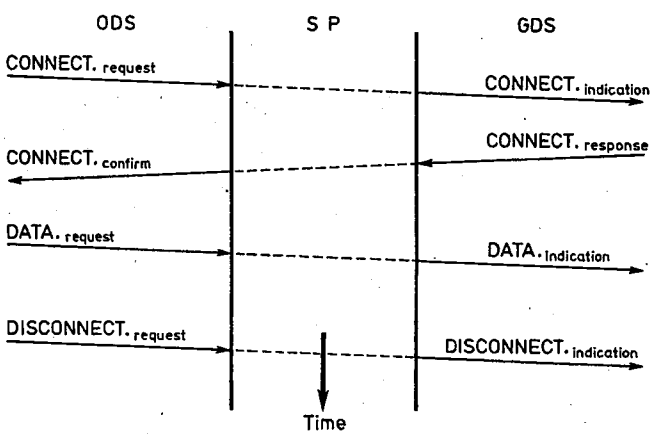


Fig. 2

The transfer of the RM arm 1 angle (∝o1) data, from ODS to GDS inside the ASD message, is a good example of SP in ODS, as is illustrated in Fig. 2, as follows below:

CONNECT.request-GDS-ODS
CONNECT.indication-GDS-ODS
CONNECT.response-ODS-GDS
CONNECT.confirm-ODS-GDS
DATA.request-GDS-ODS-ASD-∝o1
DATA.indication-GDS-ODS-ASD-∝o1
DISCONNECT.request-GDS-ODS
DISCONNECT.indication-GDS-ODS

2.3. THE SERVICES PRIMITIVES IN RDS

The RDS sends the action commands to ODS only inside Control Command (CC), CCr message, so the target station in RDS-PS always is the ODS. For instance, the data in some CCr, that is an action command, as to move the arm 1 to high (A1h), is indicating **only** the **direction** of the angle changes in the incremental (the "delta") form, not in the absolute angle value form. The list of exemplified RDS DELTA messages for this RDS CCr command is as follows:

CONNECT.request-ODS-RDS
CONNECT.indication-ODS-RDS
CONNECT.response-RDS-ODS
CONNECT.confirm-RDS-ODS
DATA.request-ODS-RDS-CCr-data
DATA.indication-ODS-RDS-CCr-data
—
—
DISCONNECT.request-ODS-RDS
DISCONNECT.indication-ODS-RDS

2.4. THE SERVICES PRIMITIVES IN GDS

The GDS sends a number of messages to each station. The common form of SP is as follows: M is a message that may be one of GDS's messages. -CCgo, Mgr, etc. Below there is given the exemplified sequence of the following GDS messages:

CONNECT.request-TS-GDS
CONNECT.indication-TS-GDS
CONNECT.response-GDS-SS

[1] The second "small" letter in symbols like ∝o is used to define the type of the station, here "o" is for the ODS, that is "On-board" DS. In the following text "g" is for the GDS, and "r" is for RDS, and so on. For instance pair "go" means the message index sent from GDS to ODS etc.

CONNECT.confirm-GDS-SS
DATA.request-TS-GDS-M-data
DATA.indication-TS-GDS-M-data
—
—
DISCONNECT.request-TS-GDS
DISCONNECT.indication-TS-GDS

For example, when the GDS sends to ODS the command as how to *move the RM to go forward 'Mfg'*, the list of the sequential GDS messages is as follows:

CONNECT.request-ODS-GDS
CONNECT.indication-ODS-GDS
CONNECT.response-GDS-ODS
CONNECT.confirm-GDS-ODS
DATA.request-ODS-GDS-CCg-Mfg
DATA.indication-ODS-GDS-CCg-Mfg
DISCONNECT.request-ODS-GDS
DISCONNECT.indication-ODS-GDS

3. THE DATA EXCHANGES AND THE MESSAGES IN DELTA

The ODS, RDS and GDS stations of the DELTA system, exchange between themselves, several types of messages. In this paper we will explain some of them, as follows.

3.1. THE MESSAGES AND THE STATE GRAPH IN THE ODS

The ODS does not send any action commands to other stastions, but its only sent messages are, so called the Action State Data (ASD), it sends to the GDS. The subset 'Action' of the ASD consists of the ODS parameters that define exactly the ODS site and its movements as follows:

ASD: = {(<Action>:	<Direction angle>,	$'\beta'$
	<arm_1_angle>,	$'\propto o1'$
	<arm_2_angle>,	$'\propto o2'$
	<arm_3_angle>,	$'\propto o3'$
	<x_site_RM>,	'Xs'
	<y_site_RM>,	'Ys'
	<z_site_RM>,	'Zs'
	<inclination angle>,	$'\delta s'$
	<cab_angle>,)	$'\sigma o'$

The state graph of the ODS is illustrated in Fig.3. Also the state graph of RM arms action is illustrated in Fig. 4.

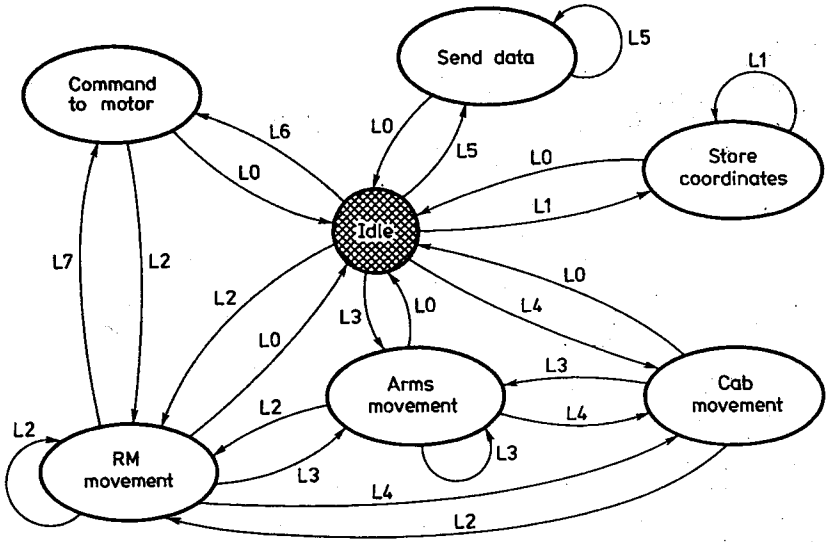


Fig. 3 The main state graph of ODS

L0 : {Out time}
L1 : {(Xw, Yw, Zw)¦(Xd, Yd, Zd)}
L2 : {Mfg¦Mfr, Mbg¦Mbr, Mlg¦Mlr, Mrg¦Mrr}
L3 : {αg1¦A1h¦A1l, αg2¦A2h¦A2l, αg3¦A3h¦A3l}
L4 : {σg¦Clr, σg¦Crr}
L5 : {Rgo¦Command_control_from_ODS-CPU}
L6 : {Stg¦Str, Spg¦Spr}
L7 : {Sgd}

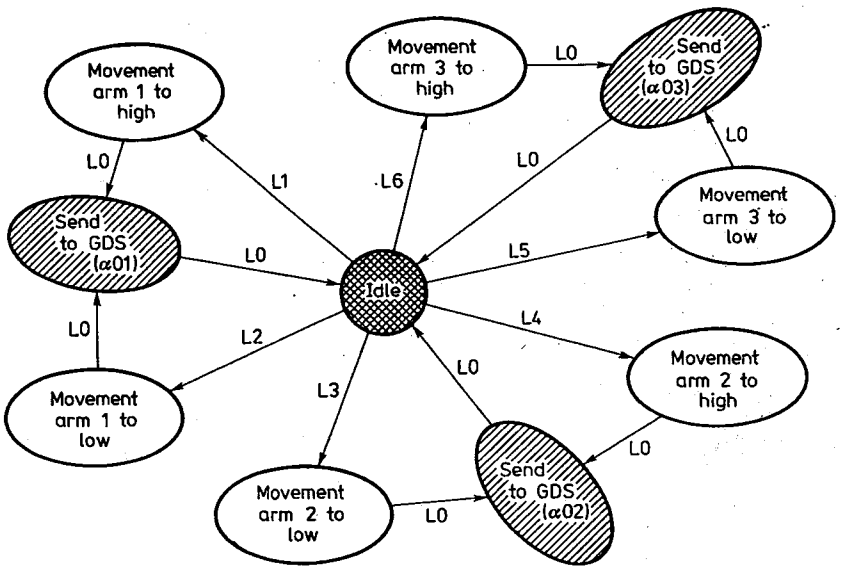


Fig. 4 The state graph of arms movement in ODS

L0 : {Out time}
L1 : {αg1¦A1h}
L2 : {αg1¦A1l}
L3 : {αg2¦A2l}
L4 : {αg2¦A2h}
L5 : {αg3¦A3l}
L6 : {αg3¦A3h}

3.2. THE MESSAGES AND THE STATE GRAPH IN THE RDS

The RDS sends its control command (CCr) to the ODS only. The CCr contains its subset 'Action', that is a set of increments or decrements of appropriate RM's movements. These are: the three arms movements, as to the low or to the high, the overall RM progressive movement: to the forward (Mfr), to the back (Mbr), to the left (Mlr) and to the right (Mrr). Also CCr contains a cab turn incremental ("delta") command to the left (Clr) or to the right (Crr). All these movement commands are done by the RM remote operator (using the portable RDS DELTA station) by means of the joysticks at the RDS. The list of the appropriate RDS's sequential messages is as follows:

CCr: = {(<action_command>:	<arm_1_to_high>,	'A1l'
	<arm_1_to_low>,	'A1l'
	<arm_2_to_high>,	'A2h'
	<arm_2_to_low>,	'A2l'
	<arm_3_to_high>,	'A3h'
	<arm_3_to_low>,	'A3l'
	<RM_forward_movement>,	'Mfr'
	<RM_back_movement>,	'Mbr'
	<RM_turn_left>,	'Mlr'
	<RM_turn_right>,	'Mrr'
	<cab_turn_left>,	'Clr'

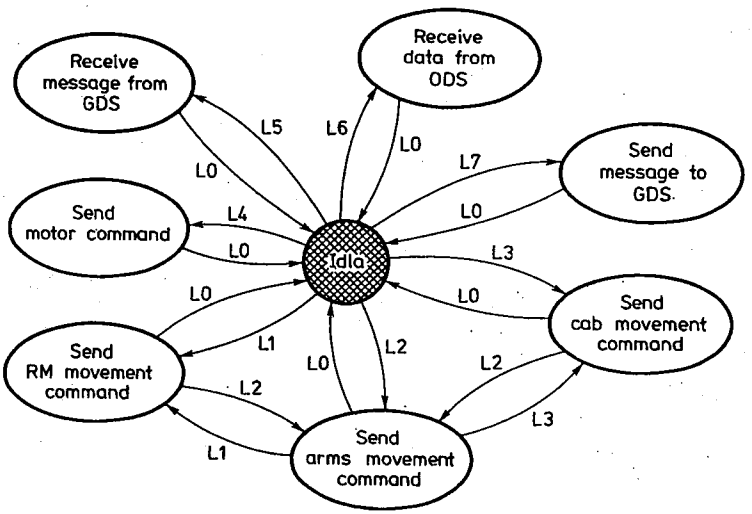


Fig. 5 The main state graph RDS

L0 : {Out time}
L1 : {J3_back ¦ J3_forward ¦ J4_left ¦ J4_right}
L2 : {J1_B_left ¦ J1_B_right ¦ J2_A1_left ¦ J2_A2_right}
L3 : {J1_R_left ¦ J1_R_right}
L4 : {Key1 ¦ Key2}
L5 : Nog ¦ Kog ¦ Sgr ¦ Cwr ¦ Tgr_data}
L6 : {Dor}
L7 : {No_ODS ¦ OK_ODS ¦ Request_data}

The action in the RDS is controlled by the joysticks. The joystick (two-plane control) J1 control is for a cab movement and for a bucket movement. The joystick J2 controls each RM's arms: 1 and 2. The stick (one-plane control) J3 controls the RM's movement to the forward and to the back. The stick J4 controls the MR's movement to the left and to the right. The state graph of the RDS is shown in Fig. 5.

3.3. THE MESSAGES AND THE STATE GRAPH IN GDS

The GDS sends to the ODS its Control Command (CCgo). The CCgo contains its Action Command (AC) field. The AC of the CCgo consists of the sub-commands of the three angles: $\alpha g1$, $\alpha g2$, $\alpha g3$, changed for each three excavator's arms, that all

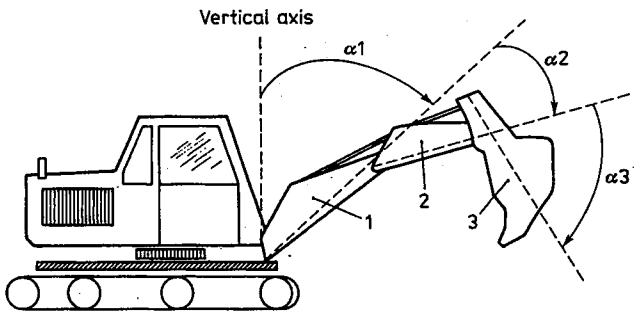


Fig. 6 The excavator arms angles

are illustrated in the Fig. 6. The CCgo contains also the four following subsets for movement of the all bulk RM: to the forward (Mfg), to the back (Mbg), turn to the left (Mlg) and turn to the right (Mrg), as follows below:

CCgo: = {(<Action_Command>:
<movement_arm_1_angle>, 'αg1'
<movement_arm_2_angle>, 'αg2'
<movement_arm_3_angle>, 'αg3'
<RM_forward_movement>, 'Mfg'
<RM_back_movement>, 'Mbg'
<RM_turn_left>, 'Mlg'
<RM_turn_right>, 'Mrg'
<cab_turn_angle> 'αg'

The state graph of the GDS is shown in Fig. 7.

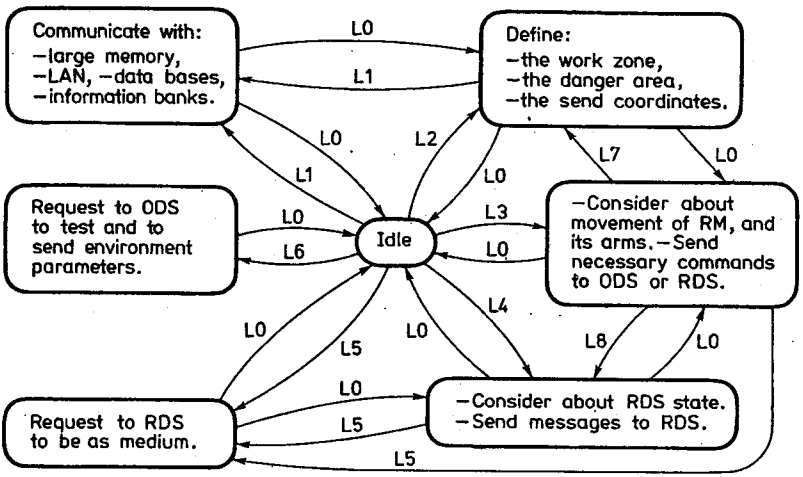


Fig. 7 The main state graph of GDS

L0 : {Out time}
L1 : {CPU-GDS_do_communicate_with_'address'}
L2 : {Start_of_work | CPU-GDS_do_define_work_area | CPU-GDS_do_define_work_danger | CPU-GDS_send_(Xw, Yw, Zw)_to_ODS | CPU-GDS_send_(Xd, Yd, Zd)_to ODS}
L3 : {(α01 | α02 | α03 | σo | β | δ | Sdo | Spo | Ab) | CPU-GDS_do_'command' | CPU-GDS_send_to_ODS_data'}
L4 : {CPU-GDS_send_to_RDS_'command' | Nor | Rd | Kor}
L5 : {CPU-GDS_send_to_RDS_'Nog' | CPU-GDS_send_to_RDS_'Kog'}
L6 : {Te | Hu | Ap | Pc | CPU-GDS_send_to_ODS_'Rgo'}
L7 : {CPU-GDS_read_work_area | CPU-GDS_read_danger_area}
L8 : {CPU-GDS_send_to_RDS_'command'}

CONCLUSION

The three stations in the DELTA system, exchange the messages between themselves, in the wireless method. These messages are limited, according to each function, also according to the relations between the stations. We observe from the state graphs, that each station has its own number of limited states.

The poposed DELTA system, is a possible to be built system, that will be able to execute its functions, according to its design. The DELTA system, in its prototype version was, on the level of main concepts, developed (especially in the level of its abstract ergonomical ideas and its functional possibilities), and partly implemented (only the ODS part of the system) in the 1986–1989 period, during the realization of the CPBP 02.13 Polish scientific program. This was realized by the two departments of the Warsaw Institute of Technology (WIT): the Institute for Heavy Working Machines (IHWM/WIT – that was responsible for the heavy machine part of the system) and the Institute for Computer Science (ICS/WIT – that was responsible for the computer part of the system). In the implementation of the

DELTA, realized to the moment, the prototype ODS station was designed, named the DELTA/1/88. This was installed on the board of the 611 type hydraulic excavator (now in operation and under investigation in the IHWM/WIT).

Still, the other parts of DELTA, that is the full ODS-RDS-GDS network, are only under the present scientific research. This is done, mainly in the IHWM-ICS/WIT, but also in the other Polish technical universities (placed in Gdańsk, Łódź, Poznań and Wrocław) by means of the funds granted from the Polish Scientific Researches Committee.

In the next future the simulation model of the full DELTA (similiar to that described in [15]) will be under the extensive investigation.

The other work, now in progress, is oriented towards the wireless data link controller synthesis. This requires the application of the special two-channel Zilog Z8530 HDLC controllers in every DS (with the additional Intel 8085AH microprocessor, 32K ROM, 32K RAM, simulator programs, debugger programs etc.).

The perspectives of the full DELTA practical uses are also good. There exists now a good couple of systems, that are very similiar to our DELTA ([16], [17]), some of them are using the wireless RDS idea too ([18], [19]). The example of the partly GDS-ODS system is extensively described in [16].

BIBLIOGRAPHY

1. K. Bieńkowski: *DELTA – system wspomagania decyzji operatora autonomicznej maszyny roboczej.* Materiały Konferencji Infogryf 88. 18–21. 10.1988 str.6, Kołobrzeg
2. K. Bieńkowski: *Architektura systemu komputerowego DELTA inteligentnej maszyny roboczej.* II PW. 1987, pp.6
3. K. Bieńkowski: *Standardy interfejsów układowych i programowych systemu komputerowego DELTA inteligentnej maszyny roboczej.* II PW. 1988, pp.6
4. K. Bieńkowski: *Hierarchiczny system informatyczny DELTA robotów i inteligentnych maszyn roboczych.* Prace Naukowe ICT Politechniki Wrocławskiej. Nr 78. pp. 23–29, seria: Konferencje, 1988
5. K. Bieńkowski: *System komputerowy DELTA inteligentnej maszyny roboczej.* II PW, 1987, pp. 7
6. K. Bieńkowski: *Wielomikrokomputerowy system precyzyjnej orientacji przestrzennej maszyny roboczej klasy DELTA/OP.* II PW, 1987, pp. 6
7. H. Dobrowolski: *Oprogramowanie mikrokomputera pokładowego (blok BMP) systemu DELTA/1 koparki hydraulicznej 611.* II PW. 1988, pp. 27
8. H. Dobrowolski: *Oprogramowanie komputera pokładowego systemu monitorująco-ostrzegawczego koparki 611 (system DELTA/1) – opis dokumentacyjny.* text: program listings: 13+4+3+1+18=39pp. II PW, 1989, pp. 14
9. H. Dobrowolski: *Koncepcja rozwoju systemu komputera pokładowego maszyny roboczej jako modularnego systemu wieloprocesorowego o elastycznej architekturze.* II PW, 1990, pp. 13
10. K. Bieńkowski: *Protokoły wymiany informacji w sieciach komputerowych robotyki.* pp. 20–25, III-a Krajowa Konferencja Robotyki Wrocław 19–21.09.1990. Prace ICT Politechniki Wrocławskiej 83. Konferencje 38
11. A.S. Tanenbaum: *Computer Networks.* Prentice-Hall International. USA. 2nd edition. 1989

12. F. Halsall: *Data Commuunications Computer Networks and OST.* Addison-Wesley Pbulishing. Great Britain, 2nd edition, 1988
13. M. Ghabally: *Layered Protocols of ISO/OSI and CCITT/X.25.* Institute of Computer Science Research Report No. 4/91. Warsaw, 1991
14. U. Black: *Computer Networks Protocols Standards and Interfaces.* Prentice-Hall International, New Jersey, 1987
15. H. Dobrowolski: *Program symulacyjny bloku BMP komputera pokładowego systemu DELTA/1 — dokumentacja użytkowa (wersja 1.0).* pp. 27, II PW, 1988
16. Alfa-Lavel Sattcontrol: *SattMobile. Computerised Materials Handling System for Manual Trucks and Cranes in Warehouses and Production Areas.* pp. 12. Sattcontrol AB S-205 Malmo Sweden, 1990
17. Komatsu Ltd.: Pay Load Meter Nr 311 SY10E/M. Tokyo, Japan, 1987, pp. 6
18. Hetronic Stauersysteme:
    — *Funksteuerungen für den Maschineneinsatz GA 609 TG,*
    — *Radio-Telecommandes Pour L'Exploitation Forestiere,*
    — *Radio Remote Control Systems With Proportional Functions,*
    — *Radio Remote Control For Cranes,*
    — *No More Worries With Control Cables.*
    Hetronic Steuer Systeme GMBH. Adalbert Stifter Str. 2. D/W-8301 Largaguaid
19. HBC-electronic:
    — *Funksteuerungen für Kräne und Maschine. Systeme FST740. FST720*
    — *Funksteurung Fahrzeugkräne. Radiomatic 730. Radiomatic 760*
    HBC-electronic Funktechnik GMBH. Haller Str. 49—53. D-7180 Crailsheim

K. BIEŃKOWSKI, M. GHABALLY

WYMIANY DANYCH I RUCH KOMUNIKATÓW BEZPRZEWODOWYCH W MIKROPROCESOROWYCH SIECIACH KOMPUTEROWYCH KLASY DELTA

Streszczenie

W artykule dokonano analizy nowej bezprzewodowej (z radiowymi liniami transmisyjnymi) sieci komputerowej (SK) typu DELTA, przeznaczonej do zastosowań w zinformatyzowanych systemach ciężkich maszyn roboczych (koparek, spychaczy, samochodów samowyładowczych, itp.). Praca jest kontynuacją i rozwinięciem tematu 3.5 CPBP 02.13 „Elementy sztucznej inteligencji w maszynach roboczych i pojazdach", realizowanego w latach 1986—1990 (obecnie kontynuowanego w ramach grantu) przez zespoły Instytutu Informatyki oraz Instytutu Maszyn Roboczych Ciężkich z Politechniki Warszawskiej, przy współpracy innych zespołów naukowych (m.in. z Politechnik: Gdańskiej, Łódzkiej, Poznańskiej, Wroclawskiej oraz z Instytutu Podstawowych Problemow Techniki PAN z Warszawy i Kielc).

Zdefiniowano i przeanalizowano grafy stanów oraz zespoły statusów i komend dla trzech typów węzłowych stacji danych (mikrokomputerów węzłowych) występujących w SK typu DELTA: w stacjach pokładowych (ODS), w stacjach do sterowania zdalnego (RDS) i w stacjach naziemnych (GDS). Przewidywane są mechanizmy współpracy SK typu DELTA z dowolnymi innymi SK, na przykład, przy konieczności uzupełniania danych kartograficznych, geologicznych i innych. Podano listy komunikatów i statusów dla tych grafów, które mogą stanowić podstawę do wygenerowania programów symulacyjnych i systemowych SK klasy DELTA.

KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJI, 1993, 39, z. 1, ss. 131–151

621.317.42
621.3.013

On the constitutive distributed parameter modelling of the singlecomponent real processes. Part I. The partial differential constitutive state equations describing the singlecomponent real processes

WACŁAW NIEMIEC

Politechnika Śląska w Gliwicach

Received 1992.07.01

Authorized 1992.09.09

In the article some general principles of the construction of the partial differential constitutive state equation and its partial differential equation of the continuity has been presented. These principles in the article in the forms of 4 (four) rules have been caught:

RULE I – the signs of the summation of balance phenomenal effects,
RULE II – the definition of the following derivative,
RULE III – the partial differential constitutive state equation form,
RULE IV – the partial differential equation of the continuity with its invariance interpretation.

The above considerations have been interpreted by the example balances of mass/charge, thermic energy and momentum of mass/electric charge generation and transport in the electronic apparatuses type electronic lamps or electrochemic processes **without electric parameters**. For the general and example considerations space-time interpretation of the principle of the phenomenal constitutive invariance was given in the article by the variable and constant physical coefficients.

1. INTRODUCTION

The problems of the constitutive distributed parameter modelling of the real processes have their genesis in American scientific literature and for that reason many years through they had not pertinent reflection in the Polish literature of this topic. It arised from the taken direction of the literature and scientific researches in Poland as others than those of the American which appeared from the business conditions or even political respects in the science. Not numerous articles reprinted

from the American literature were rather an information than an encurage to the choose of constitutive way of modelling research works [7]. Meanwhile even in Polish science literature appeared voices of known scientists that: "obtained (from the other than constitutive modelling) results are rather the consequences of accepted assumptions than real image of the process" [13]. Simultaneously in International Federation of Automatic Control Newsletters was possible to meet the discussions over congresses, conferences and symposia in which clear is stated the task of "a bridge between control sciences and technology" [14].

For this purpose of the description of physical kinetics-source and phenomenal features of the technological processes from the point of view of yield and quality aspects of the products of these processes constitutive distributed parameter modelling has been determined. The constitutive distributed parameter modelling of the real processes contains the following operations:

– gathering of the dates for the classification of the real processes [2–6],

– constitutive features of the processes; choice of balances and state variables, physical phenomena and their sources [7], [8], [2–6],

– construction of the partial differential constitutive state equations in the forms of the following derivatives [2–6], [7],

– determination of the constitutive invariance of the constitutive distributed parameter model from the balance partial differential equations of the continuity pertinent to the balances of the model [2–6], [7],

– mathematical classification of the partial differential constitutive state equations and their partial differential equations of the continuity as the constitutive invariance [2–6], [11],

– solution of constitutive distributed parameter model by [2–6]:

A. existence of the phenomenal initial conditions,

B. existence of the phenomenal initial and boundary conditions,

– discussion of the possibility of the phenomenally distributed parameter control of the real processes by the use of the phenomenal boundary conditions [2–6], [15],

– choice of the phenomenally distributed parameter control structure for the phenomenally distributed parameter control of the real processes [2–6], [15].

In the aims of the first part of this article we concentrate our focus on the principles of the construction of the constitutive distributed parameter model but the solution and applications of its results in the second part will be considered.

2. GATHERING OF THE DATES FOR THE CLASSIFICATION OF THE REAL PROCESSES WITH RESPECT TO THE CONSTITUTIVE DISTRIBUTED PARAMETER CONTROL

For the complete description of the real technological process we need to gather the information about this process being the answers for the below placed questions:

– how many composite processes contains complete real process?

– how many physical phases contains every composite process?
– how many state variables describes given phase or given composite process?
– how many source elements being decisive for the existence of the physical phenomena act in given phase or given composite process?
– how many physical phenomena are decisive for the given balance?
– how many common physical phenomena are the links between balances?

The answers for these questions are the necessary and sufficient condition for the preparation of the constitutive distributed parameter model of the considered real process. The structure of these informations by the constitutive distributed parameter modelling fulfils the relation presented in Fig. 1.

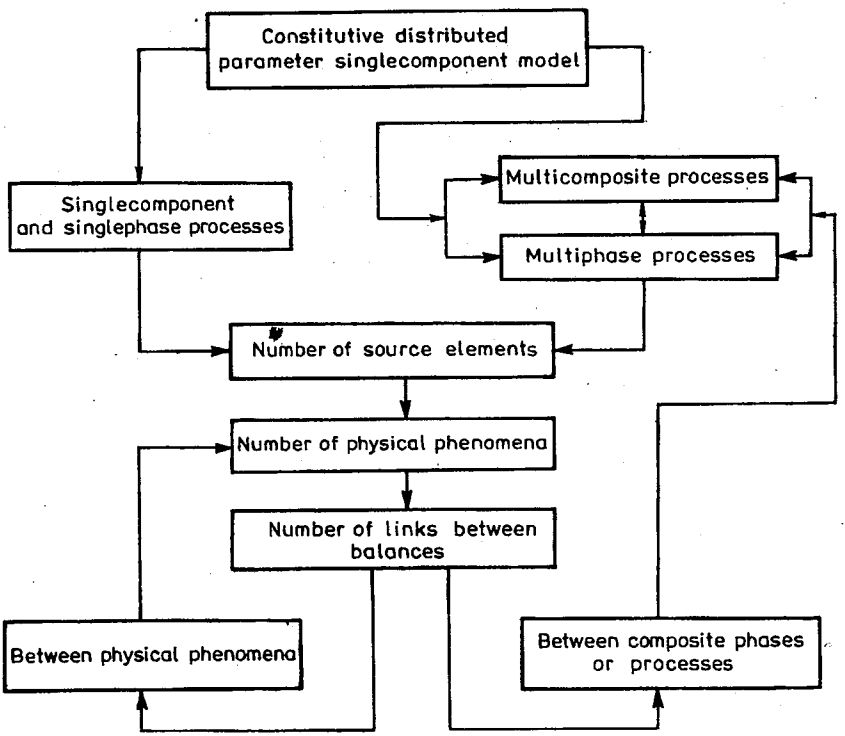


Fig. 1

The example technological processes pertinent to the structure from Fig. 1 are:

– singlecomponent and singlephase processes; industrial cleaning of the water by vapouration,
– singlecomponent and multiphase processes: continuous mass crystallization process which contains composite processes such as:
– primary nucleation process, appearance of the crystal phase inside supersaturated solution,

– crystal growth process, enlarge of the crystal phase inside the supersaturated solution,

– secondary nucleation process, mechanical generation of additional nucleons from the collisions: crystal-crystal, crystal-mixer, crystal-walls of crystallizer.

This modelling approach can be extended for the multicomponent constitutive distributed parameter modelling according to the procedure which is presented in Fig. 2.

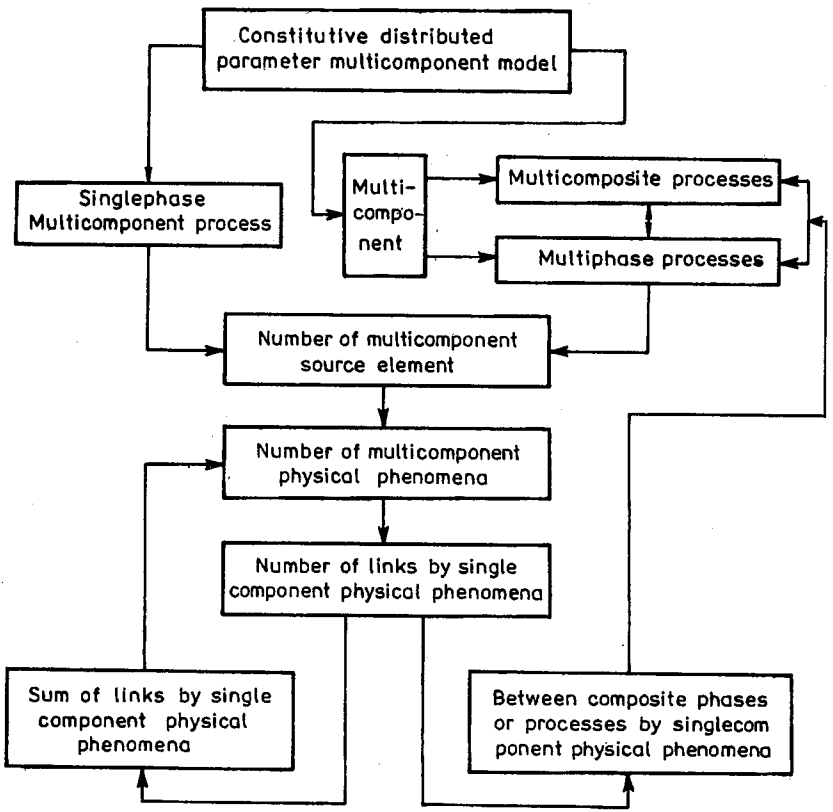


Fig. 2

The content of Fig. 2 is related to the following classified technological processes:

– singlephase multicomponent processes; for example multicomponent electrolytic processes and others,

– multiphase and multicomponent processes; for example multicomponent continuous mass crystallization process with its composite processes and others.

3. CONSTITUTIVE FEATURES OF THE REAL PROCESSES FOR THEIR CONSTITUTIVE DISTRIBUTED PARAMETER MODELLING

The following elements are the outset for the constitutive distributed parameter modelling of the real processes [1–6], [15]:

I. The local basis:
$Z(x,y,z,t)$ – the point of the processing of "working medium", constant coordinates point,
$\bar{\Omega}_{(x,y,z)t}$ – the locally selected volume-time element around the point $Z(x,y,z,t)$,
$F_{(x,y,z)t}$ – outside oriented surface of $\bar{\Omega}_{(x,y,z)t}$,
$n^{(z)}$ – the normal outside surface orientation vector of the surface $F_{(x,y,z)t}$.

II. The global basis:
$Q^R(\zeta^R,\eta^R,\beta^R,\tau)$ – the variable point of the reciprocal point $Z(x,y,z,t)$,
Ω_{Rt} – the working space volume-time containing the point $Q^R(\zeta^R,\eta^R,\beta^R,\tau)$,
F_{RT} – outside oriented surface of Ω_{RT},
$n_R^{(z)}$ – the normal outside surface orientation vector of the surface F_{Rt}.

The circulation of the normal outside surface orientation vector $n^{(z)}$ has underlying significance for the summations of surface effects of the physical phenomena of the potential fields in relation to the physical phenomena of the rotational fields in the adequate balance. Continuing it is necessary to stress here that this approach is connected to the fact of closed volume-time element $\bar{\Omega}_{(x,y,z)t}$ – so that influences of the physical phenomena in considered balances can be treated in the separate way. The circulation of the normal outside surface orientation vector $n_R^{(z)}$ is related to the phenomenal boundary control tasks on the surface F_{Rt} and by the circulation of $n^{(z)}$ can have influences on the single physical phenomena of the considered process.

III. The balance procedure [1–6]:

1. Choice of adequate balance state variables being important from the point of view of the yield and quality properties of the products of the real processes, for the coordinates "x,y,z,t" $[S_1, W_1]$,
2. Determination of the physical phenomen a for the state variables from the point (1), acting through the surface $F_{(x,y,z)t}$,
3. Determination of the physical phenomena for the state variables of the point (1) inside $\bar{\Omega}_{(x,y,z)t}$, $[S_2, W_2]$,
4. Determination of the phenomenal links between balances by their phenomenal volume effects inside $\bar{\Omega}_{(x,y,z)t}$, $[S_2, W_2]$, if not $[S_1, W_1]$,
5. Determination of the circulation of the normal outside surface orientation vector $n^{(z)}$ for $F_{(x,y,z)t}$ in relation to $n_R^{(z)}$ of F_{Rt} and its influence on the signs of summations of the potential fields (pot. fields) and rotational fields (rot. fields) [8], [12], [16]

$$\underline{\text{RULE I:}} \oiint_{F_{(x,y,z)t}} \left[\begin{pmatrix}\text{pot.}\\ \text{fields}\end{pmatrix} \pm \begin{pmatrix}\text{rot.}\\ \text{fields}\end{pmatrix}\right] dF_{(x,y,z)} \xrightarrow{\text{div}} \iiint_{\Omega_{(x,y,z)t}} \text{div}\left[\begin{pmatrix}\text{pot.}\\ \text{fields}\end{pmatrix} \mp \begin{pmatrix}\text{rot.}\\ \text{fields}\end{pmatrix}\right] d\Omega_{(x,y,z)}$$

6. Determination of the source elements for the physical phenomena of the point 2, for "t" $[S_1^d, W_1^d]$,
7. Introduction of the general balance example formula pertinent to the modelling task

$$\frac{\partial}{\partial t}\iiint_{\Omega_{(x,y,z)t}} [S_1, W_1]\, d\Omega_{(x,y,x)} = \oiint_{F_{(x,y,z)t}} \left[\begin{array}{c}\text{Geometric sum of}\\ \text{phenomenal fluxes}\\ \text{type (2) according}\\ \text{to RULE I } [S_1, W_1]\end{array}\right] dF_{(x,y,z)} +$$

$$+ \iiint_{\Omega_{(x,y,z)t}} \left[\begin{array}{c}\text{Sum of effects of}\\ \text{single physical}\\ \text{phenomena type}\\ (3) \text{ or } (4)\\ [S_2, W_2] \text{ if not } [S_1, W_1]\end{array}\right] d\Omega_{(x,y,z)} \pm \iiint_{\Omega_{(x,y,z)t}} \left[\begin{array}{c}\text{Sum of effects of}\\ \text{activity of sources (3)}\\ \text{for the physical}\\ \text{phenomena type (2)}\\ [S_1^d, W_1^d]\end{array}\right] d\Omega_{(x,y,z)} \qquad (1)$$

We consider the locally distributed parameters around the point $\mathbf{Z}(x,y,z,t)$ which is surrounded by the locally selected volume-time element $\bar{\Omega}_{(x,y,z)t}$ with its surface $\mathbf{F}_{(x,y,z)t}$ outside oriented by the normal outside surface orientation vector $\mathbf{n}^{(z)}$ having its circulation. The locally selected volume-time element $\bar{\Omega}_{(x,y,z)t}$ is phenomenally closed so that the influences of the physical phenomena on the state vector of mass/charge, energy and momentum coordinates of the point $\mathbf{Z}(x,y,z,t)$ are treated in the separate way:

$$\left[\begin{array}{c}\text{Mass/Charge}\\ \text{Energy}\\ \text{Momentum}\end{array}\right] \longleftrightarrow \left[\begin{array}{c}\text{Mass/Charge}\\ \text{Energy}\\ \text{Momentum}\end{array}\right] \qquad (2)$$

Mutual reciprocal relation

$$\left.\begin{array}{c}\mathbf{Z}(x,y,z,t)\\ \Omega_{(x,y,z)t}\end{array}\right] \longleftrightarrow \left[\begin{array}{c}Q^R(\xi^R, \eta^R, \beta^R, \tau)\\ \Omega_{Rt}\end{array}\right.$$

according to the reciprocity principle for the isotropic and anisotropic nonhomogeneous media with the space and time memories [9–10]. The locally distributed parameter approach is connected with "**v**" the field vector velocity of the movement of so called "*working medium*" for adequate balances $\frac{D}{Dt}$ [8], [2–6] consequently and the definition of the partial differential equations of the continuity for the given balance from $\frac{D}{Dt}=0$ condition [8], [2–6]. The application of Gauss law [12] for eq. (1) leads by $\mathbf{F}_{(x,y,z)t} \xrightarrow[\text{div}]{\text{RULE I}} \bar{\Omega}_{(x,y,z)t}$ to the volume effects and after equalization

of the operations under volume integrals one gets the partial differential form of the eq. (1), [1–6], [8]

$$\frac{\partial}{\partial t}[S_1,W_1] = \begin{bmatrix} u_1^1\nabla^2 S_1 & ; & u_1^2\nabla^2 W_1 \\ \text{grad}\, u_1^1\, \text{grad}\, S_1 & ; & \text{grad}\, u_1^2\, (\text{grad})W_1 \end{bmatrix} \mp \begin{bmatrix} S_1 \text{div}\, v & ; & W_1 \text{div}\, v \\ v\, \text{grad}\, S_1 & ; & (v\text{grad})W_1 \end{bmatrix} \pm$$

$$+\begin{bmatrix} u_2^1\nabla^2 S_2 & ; & u_2^2\nabla^2 W_2 \\ \text{grad}\, u_2^1\, \text{grad} S_2 & ; & \text{grad}\, u_2^2(\text{grad})\, W_2 \end{bmatrix} \pm \begin{bmatrix} S_1^d(t) & ; & W_1^d(t) \end{bmatrix}, \tag{3}$$

where:

S_1 – scalar balance state variable, $S_1^d(t)$ – scalar source element,
u_1^1 – physical coefficient of phenomenon related to the scalar state variable S_1,
W_1 – vector balance state variable, $W_1^d(t)$ – vector source element,
u_1^2 – physical coefficient of phenomenon related to the vector state variable $\mathbf{W}_1$,
v – field vector velocity of the working medium,
S_2 – scalar link state variable from the other connected balance,
u_2^1 – physical coefficient of phenomenon related to the scalar link state variable S_2,
W_2 – vector link state variable from the other connected balance,
u_2^2 – physical coefficient of phenomenon related to the vector link state variable W_2.

For the eq. (3) we introduce the definition of the following derivative with the result [2–6], [8]

$$\underline{\text{RULE II}}: \quad \frac{D[S_1;W_1]}{Dt} = \frac{\partial[S_1;W_1]}{\partial t} - [\text{grad}\, u_1^1\, \text{grad}\, S_1\ ;\ \text{grad}\, u_1^2(\text{grad})W_1] \pm$$
$$\pm[v\text{grad}\, S_1\ ;\ (v\,\text{grad})W_1] - [\text{grad}\, u_2^1\, \text{grad}\, S_2\ ;\ \text{grad}\, u_2^2\, (\text{grad})W_2] \tag{4}$$

and consequently [1–6], [8]

<u>RULE III</u>:

$$\frac{D[S_1;W_1]}{Dt} = [u_1^1\nabla^2 S_1\ ;\ u_1^2\nabla^2 W_1] \mp [S_1\, \text{div}\, v\ ;\ W_1 \text{div}\, v] \pm [S_1^d(t), W_1^d(t)] +$$
$$+[u_2^1\nabla^2 S_2\ ;\ u_2^2\nabla^2 W_2]. \tag{5}$$

According to the definition of partial differential equations of the continuity as the constitutive invariance for the continous media with the space and time memories fulfilling the condition $\frac{D[S_1;W_1]}{Dt} = 0$ [7] we have the general continuity formula [2–6]

<u>RULE IV</u>:

$$\frac{\partial[S_1;W_1]}{\partial t} = [\text{grad}\, u_1^1\, \text{grad}\, S_1\ ;\ \text{grad}\, u_1^2\, (\text{grad})W_1] \mp [v\,\text{grad}\, S_1\ ;\ (v\,\text{grad})W_1] +$$
$$+[\text{grad}\, u_2^1\, \text{grad}\, S_2\ ;\ \text{grad}\, u_2^2\, (\text{grad})W_2]^* \tag{6}$$

[]* – in many balance partial differential equations of the continuity this part is included only for the complete presentation of the phenomenal continuity from the other balance, for example the problem of diffusion enthalpy transport in energy balance.

The above presented general considerations of the constitutive distributed parameter modelling have:

– their geometrical interpretation of the space and time aspects presented in Fig. 3,
– their validity for all multibalance, multicomponent and multiphase cases of the real processes.

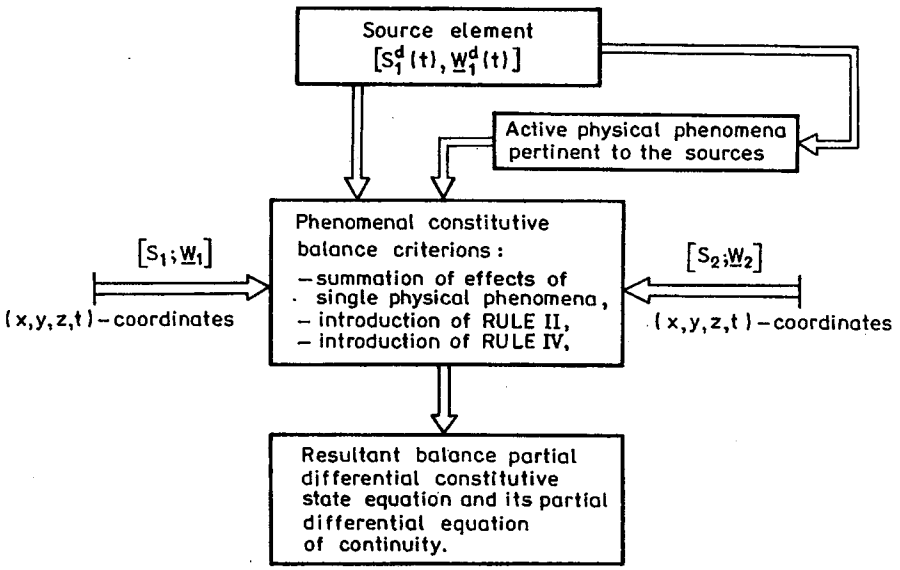


Fig. 3

The approach which has been given above has its example interpretation for the electrochemistry or electronic lamps parameters and a lot of others [16–18] as singlecomponent and singlephase problems [1–6]. We consider mass/charge, energy and momentum balances of the concentration of the singlecomponent ions or electrons electric charge with respect to the following physical phenomena: diffusion concentration and enthalpy transport, heat transfer, field vector velocity by presence of concentration, thermic energy and momentum source elements obtained from the electrical source conditions after pertinent transformation [7], [11], [1–6], [16–18].

As a consequence of this approach the below balances can be deduced:

– the mass balance of the concentration

$$\frac{\partial}{\partial t} \iiint\limits_{\Omega_{(x,y,z)t}} C \, d\Omega_{(x,y,z)} = \oiint\limits_{F_{(x,y,z)t}} [D_C \operatorname{grad} C \pm Cv] \, dF_{(x,y,z)t} \pm \iiint\limits_{\Omega_{(x,y,z)t}} (V_B G) \, d\Omega_{(x,y,z)} \qquad (7)$$

The circulation of the normal outside surface orientation vector $n^{(z)}$ with the Gauss law [12] after simple transformations of the eq. (7) leads to:

$$\frac{\partial C}{\partial t} = \operatorname{grad} D_C \operatorname{grad} C + D_C \nabla^2 C \mp C \operatorname{div} v \mp v \operatorname{grad} C \pm V_B G \,. \tag{8}$$

Introducing now the definition of the following derivative we have the ability to obtain [2–6], [8]

$$\frac{DC}{Dt} = \frac{\partial C}{\partial t} - \operatorname{grad} D_C \operatorname{grad} C \pm v \operatorname{grad} C \tag{9}$$

and consequently

$$\frac{DC}{Dt} = D_C \nabla^2 C \mp C \operatorname{div} v \pm V_B G. \tag{10}$$

From the condition for the following derivative $\frac{DC}{Dt} = 0$ [2–6], [8] we get the partial differential equations of the continuity for the following cases:

– for the variable coefficient $D_C = D_C(x,y,z,t)$

$$\frac{\partial C}{\partial t} = \operatorname{grad} D_C \operatorname{grad} C \mp v \operatorname{grad} C \,, \tag{11}$$

– for the constant coefficient $D_c \neq D_c(x,y,z,t)$

$$\frac{\partial C}{\partial t} = \mp v \operatorname{grad} C \,, \tag{12}$$

– the energy balance connected to the concentration.
With assumption that the mechanical energy influences can be neglected in the energy balance of such continuous concentration the introduction balance formula can be written as:

$$\frac{\partial}{\partial t} \iiint_{\Omega_{(x,y,z)t}} (CH) \, d\Omega_{(x,y,z)} = \oiint_{F_{(x,y,z)t}} [\lambda_C \operatorname{grad} T \pm CHv] \, dF_{(x,y,z)} \pm \iiint_{\Omega_{(x,y,z)t}} (H_B V_B G) \, d\Omega_{(x,y,z)} +$$

$$+ H_C \iiint_{\Omega_{(x,y,z)t}} \operatorname{div} (D_C \operatorname{grad} C) \, d\Omega_{(x,y,z)} \,. \tag{13}$$

Continuing, the introduction of the circulation of the normal outside surface orientation vector $\mathbf{n}^{(z)}$ influences on the summation of the potential fields and the rotational fields and the Gauss law [12] we can obtain:

$$\frac{\partial}{\partial t}(CH) = \operatorname{div}(\lambda_C \operatorname{grad} T) \mp \operatorname{div}(CHv) \pm H_B V_B G + H_C \operatorname{div}(D_C \operatorname{grad} C). \tag{14}$$

The eq. (14) is transformed to the form

$$H\frac{\partial C}{\partial t}+C\frac{\partial H}{\partial t} = \operatorname{grad}\lambda_C \operatorname{grad} T+\lambda_C\nabla^2 T\mp CH\operatorname{div} v\mp vH\operatorname{grad} C\mp \mp vC\operatorname{grad} H\pm H_B V_B G+H_C\operatorname{grad} D_C\operatorname{grad} C+H_C D_C\nabla^2 C. \tag{15}$$

Now we introduce into our considerations the partial differential equation of the concentration continuity for the constant coefficient $D_c \neq D_c(x,y,z,t)$, given by the eq. (12)

$$\frac{\partial C}{\partial t} = \mp v\operatorname{grad} C \tag{16}$$

which is subsequently multiplied both handsides by the enthalpy "H" [J/kg] to the form, as follows [2–6]

$$\frac{\partial C}{\partial t} = \mp vH\operatorname{grad} C. \tag{17}$$

Subsequently we subtract both handsides the eq. (17) from the eq. (15) with the result

$$C\frac{\partial H}{\partial t} = \operatorname{grad}\lambda_C\operatorname{grad} T+\lambda_C\nabla^2\mp CH\operatorname{div} v\mp vC\operatorname{grad} H\pm H_B V_B G+ +H_C\operatorname{grad} D_C\operatorname{grad} C+H_C D_C\nabla^2 C. \tag{18}$$

Now we recall the partial differential equation of the continuity for the variable coefficient $D_C=D_C(x,y,z,t)$ in the form (11) as below

$$\frac{\partial C}{\partial t} = \operatorname{grad} D_C\operatorname{grad} C\mp vH_C\operatorname{grad} C \tag{19}$$

which is both handsides multiplied by the specific enthalpy "H_C" for the concentration "C"

$$H_C\frac{\partial C}{\partial t} = H_C\operatorname{grad} D_C\operatorname{grad} C\mp vH_C\operatorname{grad} C. \tag{20}$$

For the continuation of our considerations we introduce the element "CH" $\left[\frac{J}{m^3}\right]$ so that there are [2–6]

– time derivative

$$C\frac{\partial H}{\partial t} = Cc_p\frac{\partial T}{\partial t}+H_C\frac{\partial C}{\partial t} \tag{21}$$

and

– space derivative

$$C\operatorname{grad} H = Cc_p\operatorname{grad} T+H_C\operatorname{grad} C \tag{22}$$

where:

$$c_p = \frac{\partial H}{\partial t}|_C \left[\frac{J}{kg\,°K}\right] \quad \text{and} \quad H_C = \frac{\partial (CH)}{\partial C}|_T \left[\frac{J}{kg}\right]$$

and the eq. (21) is subsequently multiplied both handsides by the field vector velocity "$\mp v$" to the form

$$\mp vC\,\mathrm{grad}\,H = \mp vCc_p\,\mathrm{grad}\,T \mp v \mp vH_C\,\mathrm{grad}\,C. \tag{23}$$

Applying the eq. (21) and the eq. (23) into the eq. (19) we have the ability to obtain

$$Cc_p\frac{\partial T}{\partial t} + H_C\frac{\partial C}{\partial t} = \mathrm{grad}\,\lambda_C\,\mathrm{grad}\,T + \nabla^2 T \mp CH\,\mathrm{div}\,v \mp c_p\,\mathrm{grad}T \mp \mathrm{grad}C \pm \pm H_B V_B G + H_C D_C \nabla^2 C. \tag{24}$$

Subtraction of the eq. (20) from the eq. (23) makes

$$Cc_p\frac{\partial T}{\partial t} = \mathrm{grad}\,\lambda_C\,\mathrm{grad}T + \lambda_C\nabla^2 T \mp CH\,\mathrm{div}\,v \mp vCc_p\,\mathrm{grad}T \pm H_B V_B G + H_C D_C \nabla^2 C. \tag{25}$$

Consequently to the concentration balance we introduce for the eq. (25) the definition of the following derivative after simple transformations [2–6], [8];

$$C\frac{DT}{Dt} = C\frac{\partial T}{\partial t} - \frac{1}{c_p}\,\mathrm{grad}\,\lambda_C\,\mathrm{grad}T \pm vCc_p\,\mathrm{grad}T \tag{26}$$

and subsequently the eq. (25) obtains its following derivative form

$$C\frac{DT}{Dt} = \frac{\lambda_C}{c_p}\nabla^2 T \mp CT\,\mathrm{div}\,v \pm \frac{H_B}{c_p}V_B G + \frac{H_c}{c_p}D_c\nabla^2 C. \tag{27}$$

The continuity condition $C\frac{DT}{Dt} = 0$ [2–6], [8] enables us to obtain the partial differential equations of the continuity for the cases of:

– the variable coefficients $\lambda_C = \lambda_C(x,y,z,t)$ and $D_C = D_C(x,y,z,t)$

$$Cc_p\frac{\partial T}{\partial t} = \mathrm{grad}\lambda_C\,\mathrm{grad}T \mp Cc_p v\,\mathrm{grad}T[+H_C\mathrm{grad}D_C\,\mathrm{grad}C]^* \tag{28}$$

[]* – the included element only for the presentation of the thermic energy continuity in the stream of diffusion,

– the constant coefficients $\lambda_C \neq \lambda_C(x,y,z,t)$ and $D_C \neq D_C(x,y,z,t)$

$$\frac{\partial T}{\partial t} = \mp v\,\mathrm{grad}T \tag{29}$$

– the momentum balance of the concentration

By the general balance assumption the outset formula of the balance forces is:

$$\frac{\partial}{\partial t}\iiint\limits_{\Omega_{(x,y,z)t}}(\mathrm{Cv})\,\mathrm{d}\Omega_{(x,y,z)}=\iiint\limits_{\Omega_{(x,y,z)t}}\left[\eta_w\nabla^2\mathbf{v}+\frac{\eta_w}{3}\operatorname{grad}\operatorname{div}\mathbf{v}\right]\mathrm{d}\Omega_{(x,y,z)}\mp\iiint\limits_{\Omega_{(x,y,z)t}}[\mathbf{v}\operatorname{div}(\mathrm{Cv})+$$

$$+(\mathrm{Cv}\operatorname{grad})\mathbf{v}]\,\mathrm{d}\Omega_{(x,y,z)}\pm\iiint\limits_{\Omega_{(x,y,z)t}}\mathbf{M}(t)\,\mathrm{d}\Omega_{(x,y,z)}. \tag{30}$$

After simple transformations the eq. (30) obtains its partial differential form

$$\mathbf{v}\frac{\partial C}{\partial t}+C\frac{\partial \mathbf{v}}{\partial t}=\eta_w\nabla^2\mathbf{v}+\frac{\eta_w}{3}\operatorname{grad}\operatorname{div}\mathbf{v}\mp\mathbf{v}\operatorname{div}(\mathrm{Cv})\mp(\mathrm{Cv}\operatorname{grad})\mathbf{v}\pm\mathbf{M}(t). \tag{31}$$

Now in our considerations it is necessary to introduce the partial differential equation of the concentration continuity by the constant coefficient $D_C\neq D_C(x,y,z,t)$ in the form of the eq. (12) multiplied both handsides by the field vector velocity "v":

$$\mathbf{v}\frac{\partial C}{\partial t}=\mp\mathbf{v}(\mathbf{v}\operatorname{grad}C) \tag{32}$$

obtaining from the eq. (31) the result

$$C\frac{\partial \mathbf{v}}{\partial t}=\eta_w\nabla^2\mathbf{v}+\frac{\eta_w}{3}\operatorname{grad}\operatorname{div}\mathbf{v}\mp C\mathbf{v}\operatorname{div}\mathbf{v}\mp(C\mathbf{v}\operatorname{grad})\mathbf{v}\pm\mathbf{M}(t). \tag{33}$$

For the eq. (33) we introduce the definition of the following derivative [2–6], [8]

$$C\frac{D\mathbf{v}}{Dt}=C\frac{\partial \mathbf{v}}{\partial t}\pm C(\mathbf{v}\operatorname{grad})\mathbf{v} \tag{34}$$

and consequently the following derivative form of the eq. (33) can be written as:

$$C\frac{D\mathbf{v}}{Dt}=\eta_w\nabla^2\mathbf{v}\mp C\mathbf{v}\operatorname{div}\mathbf{v}\pm\mathbf{M}(t)+\frac{\eta_w}{3}\operatorname{grad}\operatorname{div}\mathbf{v}. \tag{35}$$

Assuming in the definition of the following derivative that continuity condition is fulfilled $C\frac{D\mathbf{v}}{Dt}=0$ [2–6], [8] the partial differential equation of the continuity obtained in this way contains two different cases:

– for the variable coefficient $\eta_w=\eta_w(x,y,z,t)$

$$\frac{\partial \mathbf{v}}{\partial t}=\mp(\mathbf{v}\operatorname{grad})\mathbf{v} \tag{36}$$

– for the constant coefficient $\eta_w\neq\eta_w(x,y,z,t)$

$$\frac{\partial \mathbf{v}}{\partial t}=\mp(\mathbf{v}\operatorname{grad})\mathbf{v}. \tag{37}$$

4. THE DISCUSSION OF COMPLETE FORM OF THE SYSTEM OF PARTIAL DIFFERENTIAL CONSTITUTIVE STATE EQUATIONS OF THE DEDUCED MODEL

The partial differential constitutive state equations of mass/charge, energy and momentum balances deduced in the previous chapters can be presented as a system form:

$$\frac{DC}{Dt}=D_C\nabla^2C\mp C\,\mathrm{div}\,v\pm V_BG \tag{1}$$

(I)
$$C\frac{DT}{Dt}=\frac{\lambda_C}{c_p}\nabla^2T\mp CT\,\mathrm{div}\,v\pm\frac{H_B}{c_p}V_BG+\frac{H_C}{c_p}D_C\nabla^2C \tag{2}$$

$$C\frac{Dv}{Dt}=\eta_w\nabla^2v\mp Cv\,\mathrm{div}\,v\pm M(t)+\frac{\eta_w}{3}\,\mathrm{grad}\,\mathrm{div}\,v \tag{3}$$

which all partial differential constitutive state equations in the forms of the following derivatives fulfil the RULE III of the balance procedure III. The definition of the following derivative for all partial differential constitutive state equations of the system (I)

$$\frac{D}{Dt}=\frac{\partial}{\partial t}\pm v\,\mathrm{grad}\ -\ (\text{sum of adequate gradient operations})$$

contains all properties of the RULE II [2–6], [8].

For the system (I) the partial differential equations of the continuity have two cases interpretation from the condition $\frac{D}{Dt}=0$ [2–6], [8]

– <u>for the variable coefficients</u> $D_C=D_C(x,y,z,t)$, $\lambda_C=\lambda_C(x,y,z,t)$ and $\eta_w=\eta_w(x,y,z,t)$

$$\frac{\partial C}{\partial t}=\mathrm{grad}\,D_C\,\mathrm{grad}\,C\mp v\,\mathrm{grad}\,C \tag{1'}$$

(II)
$$Cc_p\frac{\partial T}{\partial t}=\mathrm{grad}\,\lambda_C\,\mathrm{grad}\,T\mp Cc_p v\,\mathrm{grad}\,T\,[+H_C\,\mathrm{grad}\,D_C\,\mathrm{grad}\,C]^* \tag{2'}$$

$$\frac{\partial v}{\partial t}=\mp(v\,\mathrm{grad})\,v \tag{3'}$$

[]* – the included element only for the presentation of the thermic energy continuity in the stream of diffusion phenomenon.

– <u>for the constant coefficients</u> $D_C\neq D_C(x,y,z,t)$, $\lambda_C\neq\lambda_C(x,y,z,t)$ and $\eta_w\neq\eta_w(x,y,z,t)$

$$\frac{\partial C}{\partial t}=\mp v\,\mathrm{grad}\,C \tag{1''}$$

(III)
$$\frac{\partial T}{\partial t}=\mp v\,\mathrm{grad}\,T \tag{2''}$$

$$\frac{\partial v}{\partial t}=\mp(v\,\mathrm{grad})\,v. \tag{3''}$$

The systems of the partial differential equations of the continuity (II) or (III) fulfil the properties of the RULE IV: for the coefficients $u_1^1 = u_1^1(x,y,z,t)$, $u_1^2 = u_1^2(x,y,z,t)$, $u_2^1 = u_2^1(x,y,z,t)$ and $u_2^2 = u_2^2(x,y,z,t)$ being variable – see the system (II) and consequently for the constant coefficients $u_1^1 \neq u_1^1(x,y,z,t)$, $u_1^2 \neq u_1^2(x,y,z,t)$, $u_2^1 \neq u_2^1(x,y,z,t)$ and $u_2^2 \neq u_2^2(x,y,z,t)$ the system (III) is valid.

From the above systems of partial differential equations of the continuity (II) or (III) we can observe that the system of the partial differential constitutive state equations, in the forms of the following derivatives (I) is invariant to its variable or constant coefficients adequate to the physical phenomena of the constructed model. For the construction of of system (I) important are the following process properties:

- the constructed model should be based on all state variables of the process related to its physical phenomena,
- should contain all source or kinetics aspects of the process with respect to pertinent balance,
- should be based on all characteristic balances which describe in complete form all technological aspects of the considered process with respect to the yield and quality problems of its products.

The other properties of the system (I) can be presented as follows:

- PROPERTY P1; The system (I) remains its validity in the processing point $Z(x,y,z,t)$ and whole working space volume-time Ω_{Rt}.
- PROPERTY P2; Mathematically system (I) can be classified as [11] a quasi-linear system of partial differential equations because its all highest space darivatives are in the first power: $i=1$; $(\nabla^2 C)^{i=1}$, $(\nabla^2 T)^{i=1}$, $(\nabla^2 v)^{i=1}$ and $(\text{grad div} v)^{i=1}$ and this system is decomposable for the LINEAR SYSTEMS of partial differential equations of potential fields and rotational field according to the PROPERTY P3 [10].
- PROPERTY P3; The system (I) is decomposable for: – the partial differential equations of the potential fields, and – the partial differential equations, in the forms of the following derivatives of the rotational field. It appears because the column: $\mp C \text{div} v$ in the eq. (I.1), $\mp CT\text{div} v$ in the eq. (I.2), $\mp C \text{div} v$ in the eq. (I.3) is related to the definition of the rotational fields: exists $\mathbf{B}$ – vector potential of the flow field so that $\mathbf{v} = \text{rot}\,\mathbf{B}$ and $\text{div rot}\,\mathbf{B} \equiv 0$ [12], [2–6].
- PROPERTY P4; For the system (I) there is the possibility of the determination of the initial and boundary conditions pertinent to the balance which should be fulfilled for all and every physical phenomenon of this balance.
- PROPERTY P5; The system (I) is invariant to its physical phenomenal coefficients according to the partial differential equations of the continuity (II) or (III) as THE CONSTITUTIVE INVARIANCE, so that THESE COEFFICIENTS ARE ASSU-

MED TO BE CONSTANT FOR THE ANALYTICAL SOLUTION OF THE SYSTEM (I) – see RULE IV.

– PROPERTY P6; The systems of partial differential equations of the continuity (II) or (III) are connected to the system (I) on the basis of common columns:

$$\text{system (I)} \Rightarrow \begin{bmatrix} \mp\, \text{Cdiv}\,\mathbf{v} \\ \mp\, \text{CTdiv}\,\mathbf{v} \\ \mp\, \text{Cvdiv}\,\mathbf{v} \end{bmatrix} \underset{\text{image [9], [10].}}{\overset{\text{Reciprocal}}{\longleftrightarrow}} \begin{bmatrix} \mp\, \text{vgradC} \\ \mp\, \text{CvgradT} \\ \mp\, \text{C(vgrad)v} \end{bmatrix} \Leftarrow \text{systems (II) or (III)}$$

according to the RULE III and RULE IV.

– PROPERTY P7; The properties P5 and P6 are the consequences of the assumptions of the RULE II for the definition of the following derivative $\frac{D}{Dt}$ for all balances of the model and the RULE IV for the definition of the partial differential equations of the continuity for all balances from the condition $\frac{D}{Dt}=0$.

– PROPERTY P8; The signs of the summations of potential and rotational fields in mass/charge, energy and momentum balances are determined by the RULE I and are a result of adequate balance problems.

– PROPERTY P9; For the constant space coordinates of the volume-time element $\bar{\Omega}_{(a,b,c)t}$ around source point $Z(x,y,z,t)$ $x=a$, $y=b$, $z=c$ for time "t" – all space partial differential operations are equal to ZERO and the system (I) gets its source form [1–6], [7].

$$(I^s) \qquad \frac{dC}{dt}=\pm V_B G \tag{1}$$

$$C\frac{dT}{dt}=\pm\frac{H_B}{c_p}V_B G \tag{2}$$

$$C\frac{dv}{dt}=\pm M(t) \tag{3}$$

– PROPERTY P10; If the source system of ordinary differential equations (I^s) bases in its two coordinates on even partially identical elements there is a possibility to combine them. For example in the system (I^s) there is a possibility to consider a combination of the eq. (I^s.1) and the eq. (I^s.2) with the result [1–6]

$$C\frac{dT}{dt}=\pm\frac{H_B}{c_p}\frac{dC}{dt} \tag{P.10.1}$$

where modulus of the left handside of the eq. (I^s.1)

$$V_B G = |\pm V_B G| \qquad \text{(P.10.2)}$$

has been used – possibility of inverse transformation.

Considering the properties of the system of partial differential constitutive state equations in the forms of the following derivatives (I) it is necessary to stress that every physical phenomenon can be treated with respect to its invariance partial differential continuity operations, so that the content of Table 1 is valid

Table 1

The single physical phenomena of the potential fields and the rotational field of the model and their constitutive aspects

Physical phenomena \ Mathematical formulae			Ground formulation	Constitutive form	Invariance formula
Potential fields		Difussion heat transfer	$\mathrm{div}(D_C \mathrm{grad} C)$ $\mathrm{div}(\lambda_C \mathrm{grad} T)$	$D_C \nabla^2 C$ $\lambda_C \nabla^2 T$	$\mathrm{grad} D_C \mathrm{grad} C$ $\mathrm{grad} \lambda_C \mathrm{grad} T$
Rotational field	Convection	to scalar	$\mathrm{div}(Sv)$	$S\mathrm{div}v$	$v\mathrm{grad}S$
		to vector	$(\mathrm{div}W)v$	$W\mathrm{div}v$	$(v\mathrm{grad})W$

5. A GEOMETRICAL INTERPRETATION OF THE CONSTITUTIVE DISTRIBUTED PARAMETER MODELLING

As a consequence of the constitutive approach to the distributed parameter modelling of the real processes as the continuous media with the space and time memories we can consider the following aspects of this scientific problem:

1^{00} – a general interpretation of the constitutive invariance based on the general formula of the partial differential equation of continuity governed by the RULE IV,

2^{00} – an example interpretation of the complete idea of the constitutive invariance for the deduced in the article system of partial differential constitutive state equations (I) in the forms of the continuity systems of partial differential equations (II) or (III).

The above presented approach makes possible on the detailed form presentation [7], [2–6]:

– for the content of the point 1

We recall the general continuity formula given by the RULE IV

$$\frac{\partial[S_1;W_1]}{\partial t}=[\text{grad}\,u_1^1\,\text{grad}\,S_1;\,\text{grad}\,u_1^2\,(\text{grad})W_1]\mp[v\,\text{grad}\,S_1;\,(v\,\text{grad})W_1]+$$
$$+[\text{grad}\,u_2^1\,\text{grad}\,S_2;\,\text{grad}\,u_2^2\,(\text{grad})W_2] \tag{1°1}$$

and separate from the above formula the following elements

– for time "t" we have

$$L_t=\frac{\partial[S_1:W_1]}{\partial t} \tag{1°2}$$

– for the scalar part related to „u_1^1, S_1"

$$L_{u_1^1 S_1}=[\text{grad}\,u_1^1\,\text{grad}\,S_1] \tag{1°3}$$

– for the vector part connected to „u_1^2, W_1"

$$L_{u_1^2,W_1}=[\text{grad}\,u_1^2\,(\text{grad})W_1] \tag{1°4}$$

– for the scalar part "S_1" in the field vector velocity "v"

$$L_{v,S_1}=[\mp v\,\text{grad}\,S_1] \tag{1°5}$$

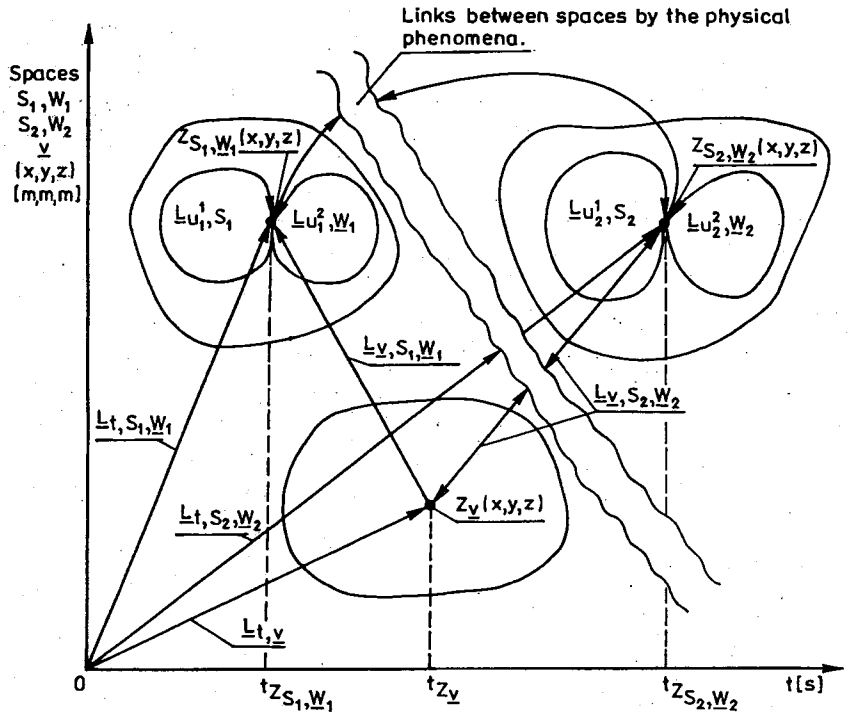


Fig. 4

– for the vector part $\mathbf{W}_1$ in the field vector velocity "v"

$$\mathbf{L}_{v,W_1}=[\mp(v\,\mathrm{grad})\mathbf{W}_1] \tag{1°6}$$

– for the scalar links based on u_2^1, S_2

$$\mathbf{L}_{u_2^1 s_2}=[\mathrm{grad}\,u_2^1\,\mathrm{grad}\,S_2] \tag{1°7}$$

– for the vector links based on u_2^2, $\mathbf{W}_2$

$$\mathbf{L}_{u_2^2 W_2}=[\mathrm{grad}\,u_2^2\,(\mathrm{grad})\mathbf{W}_2]. \tag{1°8}$$

The lead radius $\mathbf{L}_{:::}$ is very helpful by the geometric interpretation of space and time parts of the constitutive invariance for the continuous media with the space and time memories. The geometric interpretation of this idea in Fig. 4 has been presented.

For the partial differential equation of the continuity presented by the RULE IV two cases of the time interpretation from the figure Fig. 4, are valid:

– the balance phenomena appear in the identical time "t"

$$t_{Z_{s_1,W_1}}=t_{Z_v}=t_{Z_{s_2,W_2}}=t \tag{1°9}$$

– the balance phenomena times are different for all physical phenomena of the balance formula

$$t_{Z_{s_1,W_1}}\neq t_{Z_v}\neq t_{Z_{s_2,W_2}}\neq t \tag{1°10}$$

– for the modelling example idea given by the point 2°°

To our analysis is taken the system of partial differential equations of the continuity (II) possessing the form for the variable coefficients $D_C=D_C(x,y,z,t)$, $\lambda_C=\lambda_C(x,y,z,t)$ and $\eta_w=\eta_w(x,y,z,t)$

$$\frac{\partial C}{\partial t}=\mathrm{grad}\,D_C\,\mathrm{grad}\,C\mp v\,\mathrm{grad}\,C \tag{2°1}$$

$$Cc_p\frac{\partial T}{\partial t}=\mathrm{grad}\,\lambda_C\,\mathrm{grad}\,T\mp Cc_p v\,\mathrm{grad}\,T\,[+H_C\,\mathrm{grad}\,D_C\,\mathrm{grad}\,C]^* \tag{2°2}$$

$$\frac{\partial v}{\partial t}=\mp(v\,\mathrm{grad})\,v \tag{2°3}$$

– the time "t" operations

$$\mathbf{L}_{t,C}=\left[\frac{\partial C}{\partial t}\right];\quad \mathbf{L}_{t,T,C}=\left[Cc_p\frac{\partial T}{\partial t}\right];\quad \mathbf{L}_{t,v}=\left[\frac{\partial v}{\partial t}\right] \tag{2°4}$$

– the scalar part related to "D_C, C" – diffusion phenomenon in mass balance

$$\mathbf{L}_{D_C,C}=[\mathrm{grad}\,D_C\,\mathrm{grad}\,C] \tag{2°5}$$

– the scalar part connected to "D_C, C" – diffusion phenomenon in energy balance

$$\mathbf{L}_{D_C,C,T}=[H_C\,\mathrm{grad}\,D_C\,\mathrm{grad}\,C]^* \tag{2°6}$$

– the scalar part for "λ_C, T" – heat transfer phenomenon in energy balance

$$L_{\lambda_C,T}=[\text{grad}\,\lambda_C\,\text{grad}\,T] \tag{2.7}$$

– the scalar part of "C" – in the field vector velocity "v"

$$L_{v_C,C}=[\mp v\,\text{grad}\,C] \tag{2.8}$$

– the scalar part of "T" – in the field vector velocity "v"

$$L_{v,T}=[\mp Cc_p v\,\text{grad}\,T] \tag{2.9}$$

– the vector part of "v" – in the field vector velocity "v"

$$L_{v,v}=[\mp (v\,\text{grad})v]. \tag{2.10}$$

The geometric interpretation of the constitutive invariance of the system (II) as the continuous medium with the space and time aspects is given in Fig. 5.

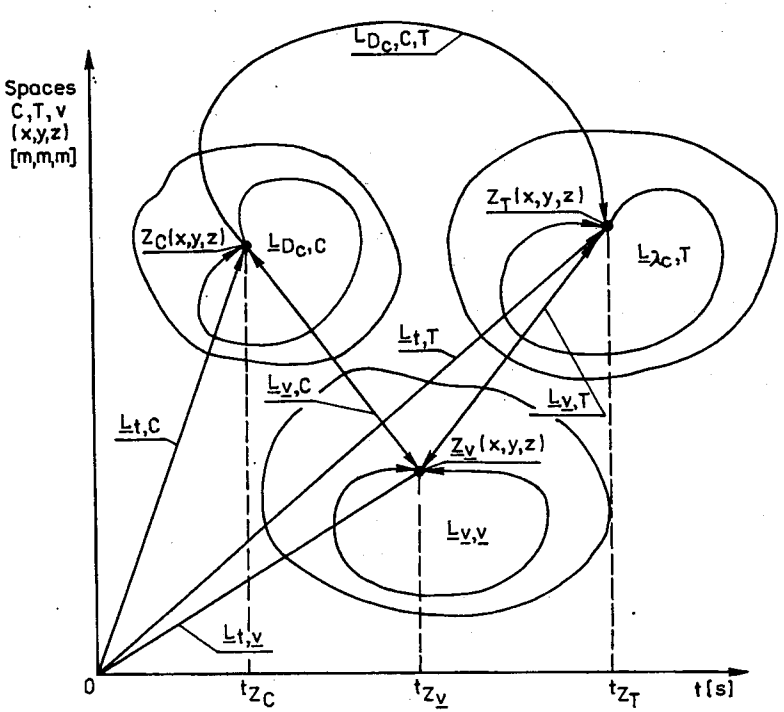


Fig. 5

Two time cases for the system (II) are possible:

– the identical balance time for all balance phenomena as given "t"

$$t_{Z_C}=t_{Z_v}=t_{Z_T}=t \tag{2.11}$$

– the different balance times for every balance phenomenon

$$t_{ZC} \neq t_{Z_v} \neq t_{Z_T} \neq t \tag{2.12}$$

Consequently to the analysis step for the system (II) we concentrate our focus on the system of the constant coefficients $D_C \neq D_C(x,y,z,t)$, $\lambda_C \neq \lambda_C(x,y,z,t)$ and $\eta_w \neq \eta_w(x,y,z,t)$ partial differential equations of the continuity (III) rewritten as follows:

$$\frac{\partial C}{\partial t} = \mp v \operatorname{grad} C \tag{2.13}$$

$$\frac{\partial T}{\partial t} = \mp v \operatorname{grad} T \tag{2.14}$$

$$\frac{\partial v}{\partial t} = \mp (v \operatorname{grad}) v \tag{2.15}$$

which can be separated for the below parts:

– the time "t" operations

$$L_{t,C} = \left[\frac{\partial C}{\partial t}\right]; \quad L_{t,T} = \left[\frac{\partial T}{\partial t}\right]; \quad L_{t,v} = \left[\frac{\partial v}{\partial t}\right] \tag{2.16}$$

– the scalar "C" in the field vector velocity "v"

$$L_{v,C} = [\mp v \operatorname{grad} C] \tag{2.17}$$

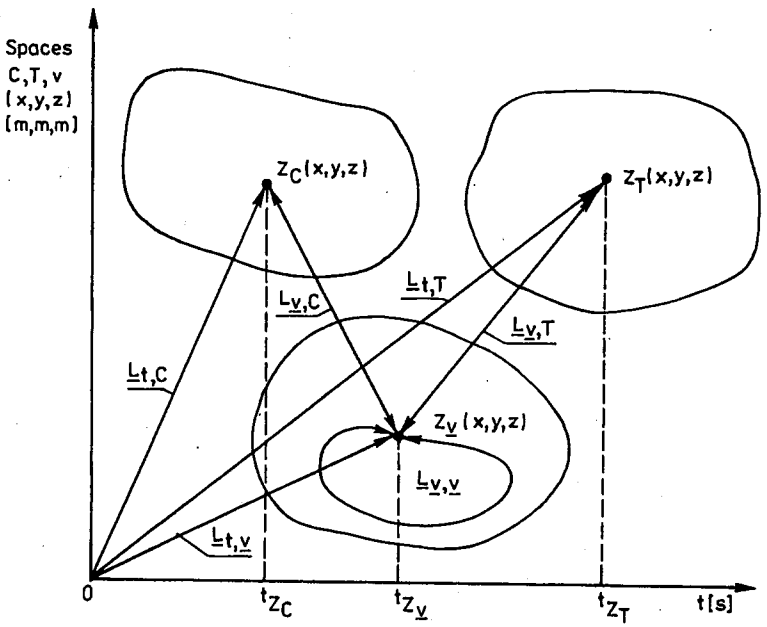


Fig. 6

– the scalar "T" in the field vector velocity "v"

$$L_{v,T}=[\mp v\,\mathrm{grad}\,T] \tag{2.18}$$

– the vector "v" in the field vector velocity "v"

$$L_{v,v}=[\mp (v\,\mathrm{grad})\,v]. \tag{2.19}$$

In Fig. 6 the above presented approach in the detailed form has been given. We consider two cases of the time conditions for the system (III) as follows:

– the identical balance time "t"

$$t_{ZC}=t_{Z_v}=t_{Z_T}=t \tag{2.20}$$

– the different balance times

$$t_{ZC}\neq t_{Z_v}\neq t_{Z_T} \tag{2.21}$$

6. CONCLUSIONS

This article is an attempt to the generalization of the constitutive distributed parameter modelling concept by the introduction of the rules: RULE I, RULE II, RULE III and RULE IV. These rules make possible on the determination of the procedure of this kind of distributed parameter modelling by the use of the physical phenomena balanced with their effects for the singlecomponent continuous media with the space and time memories. The procedure of construction of partial differential constitutive state equations in the forms of the following derivatives presented in this article by the RULE I, RULE II, RULE III and RULE IV for all balances describing the considered real process remains valid.

This general approach for the mass/charge, energy and momenum balances by the constitutive description of singlecomponent electric active media has been illustrated. All above presented RULES have been separately discussed for adequate steps of the construction of balance partial differential constitutive state equations with complete concord to the phenomenal distributed parameter modelling of the real singlecomponent processes. For the RULE IV – the partial differential formula of the continuity as the constitutive invariance a general geometric space and time interpretation has been suggested – see Fig. 4. For this idea an example illustration for the variable physical coefficients $D_C=D_C(x,y,z,t)$, $\lambda_C=\lambda_C(x,y,z,t)$ and $\eta_w=\eta_w(x,y,z,t)$ in figure Fig. 5 is given and subsequently for the constant coefficients $D_C\neq D_C(x,y,z,t)$, $\lambda\neq\lambda_C(x,y,z,t)$ and $\eta_W\neq\eta_W(x,y,z,t)$ the results are shown in Fig. 6. The above mentioned analysis enables us to formulate the general conclusion that the RULE I, RULE II, RULE III and RULE IV have underlying significance for the creation of the constitutive distributed parameter modelling theory for the singlecomponent real processes.

NOTATION

$\bar{\Omega}_{(x,y,z)t}$	– the locally selected volume-time element (cuboidal form) around the processing point $Z(x,y,z,t)$, and x=a, y=b, z=c – constant coordinates,	m^3-s
$F_{(x,y,z)t}$	– the outside oriented surface for $\bar{\Omega}_{(x,y,z,)t}$	m^2-s
$n^{(z)}$	– the normal outside surface orientation vector $F_{(x,y,z)t}$ for which circulation has underlying significance for the summation of effects of potential and rotational fields in adequate balances,	
Ω_{Rt}	– working space volume-time for the real processes	m^3-s
F_{Rt}	– the outside oriented surface for Ω_{Rt}	m^2-s
$n_R^{(z)}$	– the normal outside surface orientation vector for F_{Rt} which circulation is in relation to $n^{(z)}$ and the boundary control tasks by the phenomenal boundary conditions	
C	– the concentration of the singlecomponent processing medium	$\frac{kg}{m^3}$
T	– the temperature of the singlecomponent processing medium	°K
v	– the field vector velocity of the singlecomponent processing medium	$\frac{m}{s}$
D_C	– the diffusion coefficient for the singlecomponent processing medium	$\frac{m^2}{s}$
H_C	– the own enthalpy for the concentration "C"	$\frac{J}{kg}$
λ_C	– the heat transfer coefficient for the singlecomponent processing medium	$\frac{W}{m°K}$
c_p	– the specific heat of the singlecomponent processing medium	$\frac{J}{kg°K}$
η_w	– the coefficient of the dynamic viscosity	$\frac{Ns}{m^2}$
V_B	– the intensity of the generation of the mass/charge	$\frac{mol}{s}$
G	– the own molar mass for the mass/charge generation	$\frac{kg}{mol\,m^3}$
H_B	– the specific enthalpy for the intensity of mass/charge generation	$\frac{J}{kg}$
M(t)	– the force generation function of mass/charge	$\frac{N}{m^3}$

REFERENCES

1. W. Niemiec: *The mathematical analysis of the kinetics of the crystal growth reaction during the continuous mass crystallization process.* 171st Meeting, The Electrochemical Society, Philadelphia, Pennsylvania, USA, May 10–15, 1987, Paper No. 68
2. W. Niemiec: *On the constitutive theory of modelling and information for distributed parameter control of the continuous mass crystallization process.* Part I: *Partial differential constitutive state equations describing phenomena of the continuous mass crystallization process.* Third Int'l Conference on Liquid Metal Engineering and Technology in Energy Production, Oxford, England, 9–13 April, 1984, Paper No. 191
3. W. Niemiec: *On modelling and information construction for control of distributed parameter chemical processes in fluid phase.* 3 IFAC Symposium "Control of Distributed Parameter Systems", Toulouse, France, 29 VI–2 VII, 1982, Session 23
4. W. Niemiec: *A mathematical model of distributed parameters of the continuous mass crystallization process for the adaptive control.* PhD Thesis. Silesian Technical University 1979
5. W. Niemiec: *The constitutive theory of modelling and information for phenomenal distributed parameter control of multicomponent chemical processes in gas, fluid and solid phase.* **Part I.** *The constitutive distributed parameter model of multicomponent chemical processes in gas, fluid and solid phase.* 7th Miami Int'l Conference on Alternative Energy Sources, Miami Beach Florida, USA, 9–11 December, 1985, Session "Hydrocarbons/Energy transfer", pp. 579–588
6. W. Niemiec: *A mathematical model of the distributed parameters of the continuous mass crystallization process for the adaptive control.* Part I: *The construction of partial differential constitutive state equations for the continuous mass crystallization process.* Poznańskie Towarzystwo Przyjaciół Nauk, Wydział Nauk Technicznych, Prace Komisji Automatyki i Informatyki, Tom XV 1989, pp. 81–118
7. A. Morgan: *On the construction of constitutive equations for continuous media.* Archiwum Mechaniki Stosowanej 1965, vol. 17, No. 1, pp. 145–174
8. B. Średniawa: *Hydrodynamika i teoria sprężystości.* PWN, Warszawa 1977
9. W. Nowacki: *Dynamiczne zagadnienia termosprężystości.* PWN, Warszawa 1966
10. W. Nowacki, Z. Olesiak: *Termodyfuzja w ciałach stałych.* PWN, Warszawa 1991
11. C. Coulson, A. Jeffrey: *Fale modele matematyczne.* WNT, Warszawa 1982
12. T. Trajdos: *Matematyka dla inżynierów.* PWN, Warszawa 1974
13. J. Synowiec: *Wpływ niektórych czynników na średni wymiar kryształów soli nieorganicznych.* IChN, Gliwice–Wrocław 1971
14. *IFAC Newsletters 1980–1992*
15. S. Węgrzyn: *Podstawy automatyki.* PWN, Warszawa 1974
16. A. Tichonow, A. Samarski: *Równania fizyki matematycznej.* PWN, Warszawa 1966
17. Z. Pietrzyk, E. Leśkiewicz: *Pomiar gęstości i temperatur elektronów par rtęci w dyszy przetwornika kalorymetrycznego.* IPPT-PAN, Warszawa 1968, Nr 16/68
18. S. Moćko: *Związek między aktywnością katalityczną a własnościami elektrycznymi.* IPPT-PAN, Warszawa 1978, Nr 40/78

W. NIEMIEC

O KONSTYTUTYWNYM ROZŁOŻONYM PARAMETRYCZNIE MODELOWANIU JEDNOSKŁADNIKOWYCH RZECZYWISTYCH PROCESÓW
CZĘŚĆ I. RÓWNANIA RÓŻNICZKOWE CZĄSTKOWE KONSTYTUTYWNE STANU OPISUJĄCE JEDNOSKŁADNIKOWE RZECZYWISTE PROCESY

Streszczenie

W artykule przedstawiono ogólne zasady konstruowania równania różniczkowego cząstkowego konstytutywnego stanu oraz jego równania różniczkowego cząstkowego ciągłości. Zasady te ujęto w artykule w postaci 4 (czterech) reguł:

REGUŁA I – znaki sumowania bilansowych efektów zjawisk,
REGUŁA II – definicja pochodnej śledczej,
REGUŁA III – postać równania różniczkowego cząstkowego konstytutywnego stanu,
REGUŁA IV – postać równania różniczkowego cząstkowego ciągłości z jego inwariantną interpretacją.

Rozważania powyższe zinterpretowano przykładem bilansów masy/ładunku, energii termicznej i pędu w układach generacji i transportu ładunków elektrycznych i masy dla elektronicznych urządzeń typu lampy elektroniczne lub procesach elektrochemicznych **bez uwzględniania elektrycznych parametrów**. Dla rozważań ogólnych jak i przykładu podano przestrzenno-czasową interpretację zasady zjawiskowej inwariancji konstytutywnej przy zmiennych i stałych współczynnikach fizykalnych zjawisk.

KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJI, 1993, 39, z. 1, ss. 155-186

621.317.42
621.3.013

On the constitutive distributed parameter modelling of the singlecomponent real processes
Part II. The solution of partial differential constitutive state equations for the singlecomponent real processes

WACŁAW NIEMIEC

Politechnika Śląska w Gliwicach

Received 1992.07.01

Authorized 1992.09.09

The article is devoted to the presentation of the constitutive approach to the solution of the general formula of the partial differential constitutive state equation deduced in Part I of this article. Suggested general constitutive solution approach has consequently been applied to the solution of the deduced in Part I example system of mass/charge, thermic energy and momentum partial differential constitutive state equations. Two cases of this example solution by:
A. existence of the initial conditions,
B. existence of the initial and boundary conditions,
for all its physical phenomena have been considered.

The boundary controlability index defined on the basis of mass/charge thermic energy and momentum state vector form has been introduced and discussed.

1. INTRODUCTION

In this part of the article we concentrate our attention on the analytical solution of:

– the general partial differential constitutive state equation, deduced in [1],

– the example system of balance mass/charge, thermic energy and momentum partial differential constitutive state equations, deduced in [1],

by existence of phenomenally determined [2–6]:

A. the initial conditions,

B. the initial and boundary conditions.

For the realization of this task the important role play the following literature proved ideas:

– the reciprocity principle for the isotropic and anisotropic nonhomogeneous media with the space and time memories [7], [8]

– the properties of the potential and rotational fields [9],

– the definition of the following derivative $\frac{D}{Dt}$ and the definition of the partial differential equation of the continuity from the condition $\frac{D}{Dt}=0$ [10], [2–6],

– the properties of the phenomenal Green functions [11].

We consider the general analytical solution of the general partial differential constitutive state equation to use the obtained in this way corollary subsequently to the complete multibalance example case. According to the proved in Part I of the aticle [1] the constitutive invariance of the general and example models the physical phenomenal coefficients are assumed to be space-time constant for the analytical phenomenal solutions. Such given approach provides the complete analytical solution of mass/charge, thermic energy and momentum coordinates state vectors based on the analytical source and phenomenal solutions of the potential and rotational fields [2–6]. This makes possible the application of the obtained complete solution as follows [13–17]:

– to the identification of the technological aspects of the considered processes,

– the yield and quality of the products of the considered processes control problems by the use of the phenomenal boundary conditions.

2. SOME CONSTITUTIVE SOLUTION ASPECTS OF THE CONSTITUTIVE DISTRIBUTED PARAMETER MODELLING

According to the geometric interpretation of the constitutive problems of the general partial differential constitutive state equation [1] and the example system of balance mass/charge, thermic energy and momentum partial differential constitutive state equations [1] the solutions of both cases should possess all constitutive aspects:

I. – the system of partial differential constitutive state equations presented as a source should in all points of Ω_{Rt} and F_{Rt} be fulfilled, according to the PROPERTY P1 in [1].

II. – the system of partial differential constitutive state equations in its homogeneous part, PROPERTY P3 in [1], fulfils:

1. – the phenomenal partial differential equations of the physical phenomena of the potential fields,

2. – the phenomenal partial differential equations in the forms of the following derivatives of the rotational field.

III. – the complete solution of the constitutive distributed parameter model should contain all source and phenomenal problems in the forms of state vectors having pertinent coordinates if consequently to the balances these coordinates exist.

The general partial differential constitutive state equation deduced in [1] fulfils the problems I, II and III, too. The analytical solutions of the constitutive distributed parameter models fulfil the rule:

$$[v]=[I]+[II.1]+[II.2] \tag{2.1}$$

and $[\hat{V}]$ – the state vector of complete solution. The formula (2.1) for both cases of:
A. — the existence of the initial conditions,
B. — the existence of the initial and boundary conditions,
for every phenomenon and all physical phenomena of the considered cases remains valid.

3. THE GENERAL PARTIAL DIFFERENTIAL CONSTITUTIVE STATE EQUATION

Let us recall the general formula of the partial differential constitutive state equation in the form of the RULE III [1]

$$\frac{D[S_1;W_1]}{Dt}=[u_1^1\nabla^2 S_1;\ u_1^2\nabla^2 W_1]\mp[S_1\,\mathrm{div}v;\ W_1\,\mathrm{div}v]\pm[S_1^d(t);\ W_1^d(t)]+$$
$$+[u_2^1\nabla^2 S_2;\ u_2^2\nabla^2 W_2]. \tag{3.1}$$

As a consequence of the application of the PROPERTY P9 in [1] the source form of the eq. (3.1) for the point $Z(x,y,z,t)$, all space partial differential operations are equal to zero and we have

$$\frac{d[S_1;W_1]}{dt}=\pm[S_1^d(t);\ W_1^d(t)] \tag{3.2}$$

The integration of the eq. (3.2) leads to its source form

$$[S_1^z(t);W_1^z(t)]=[S_{10};W_{10}]\pm\int_0^t[S_1^d(t);W_1^d(t)]\,dt \tag{3.3}$$

and the homogeneous part of the eq. (3.1) is

$$\frac{D[S_1;W_1]}{Dt}=[u_1^1\nabla^2 S_1;u_1^2\nabla^2 W_1]\mp[S_1\,\mathrm{div}v;W_1\,\mathrm{div}v]+[u_2^1\nabla^2 S_2;u_2^2\nabla^2 W_2]. \tag{3.4}$$

For the solution of the eq. (3.4) the following conditions should be fulfilled:

– the partial differential constitutive state equation (3.4) pertinent to the PROPERTY P3 has its forms [1]

– for the potential fields

$$(I^{hp}) \qquad \frac{\partial[S_1;W_1]}{\partial t}=[u_1^1\nabla^2 S_1;u_1^2\nabla^2 W_1] \tag{3.5}$$

$$\frac{\partial[S_1;W_1]}{\partial t}=[u_2^1\nabla^2 S_2;u_2^2\nabla^2 W_2] \tag{3.6}$$

– for the rotational field

$$(I^{hr}) \qquad \frac{D[S_1;W_1]}{Dt}=\mp[S_1\,\mathrm{div}v;W_1\,\mathrm{div}v] \tag{3.7}$$

– the source functions state vector for the solutions of the partial differential equations of potential and rotational fields (3.5), (3.6) and (3.7) are:

$$\left.\begin{Bmatrix} S_1^z(t);W_1^z(t) \\ S_2^z(t);W_2^z(t) \\ \chi^z(t) \end{Bmatrix}\longrightarrow\left\{\begin{array}{l|l} \text{eq. (3.5)} & \text{complete solution possibility} \\ \text{eq. (3.6)} & \text{link and source supplement} \\ \text{eq. (3.7)} & \text{ground equation and source supplement} \end{array}\right.\right\} \tag{3.8}$$

being simultaneously necessary and sufficient condition of the analytical solution of the equation of the RULE III.

4. EXAMPLE INTERPRETATION OF THE CONSTITUTIVE SOLUTION OF MULTIBALANCE CONSTITUTIVE DISTRIBUTED PARAMETER MODEL

Let us recall the model partial differential constitutive state equations in the forms of the following derivatives (I) in [1]:

$$\frac{DC}{Dt}=D_C\nabla^2 C\mp C\,\mathrm{div}v\pm V_B G \tag{1}$$

(I)

$$C\frac{DT}{Dt}=\frac{\lambda_C}{c_p}\nabla^2 T\mp CT\,\mathrm{div}v\pm\frac{H_B}{c_p}V_B G+\frac{H_C}{c_p}D_C\nabla^2 C \tag{2}$$

$$C\frac{Dv}{Dt}=\eta_w\nabla^2 v\mp Cv\,\mathrm{div}v\pm M(t)+\frac{\eta_w}{3}\mathrm{grad}\,\mathrm{div}v \tag{3}$$

where: $\frac{D}{Dt}=\frac{\partial}{\partial t}\pm v\mathrm{grad}$ – (sum of adequate gradient operations) [10], [1–6].

The analysis of the system (I) for its constitutive analytical solution contains the below listed considerations.

4.1. THE SOURCE DECOMPOSITION OF THE SYSTEM OF PARTIAL DIFFERENTIAL CONSTITUTIVE STATE EQUATIONS (I)

According to the PROPERTY P9 in [1] cited as PROPERTY I in chapter 2 the following Theorem can be proved:

Theorem 1. If for the defined $\Phi(x,y,z)$ – the scalar function and $\Psi(x,y,z)$ – the vector function of the space coordinates these space coordinates are assumed to be constant, $x=a=\text{const}$, $y=b=\text{const}$, $z=c=\text{const}$, there are fulfilled relations:

$$\Phi(a,b,c)=\text{const} \quad \text{and} \quad \Psi(a,b,c)=\text{const} \tag{3.1.1}$$

and consequently

$$\nabla\Phi=0 \quad \text{and} \quad \nabla\Psi=0 \tag{3.1.2}$$

Proof. The proof is coming from the definition of the differentiation of the multivariable functions [9].

As a consequence of the utilization of this point of view for the constant cooridnates $x=a=\text{const}$, $y=b=\text{const}$ and $z=c=\text{const}$ of the locally selected volume-time element $\bar{\Omega}_{(a,b,c)t}$ around processing point Z(x,y,z,t), from the system of partial differential constitutive state equations (I) one obtains its source form

$$(\mathrm{I}^s)\qquad \begin{aligned} \frac{dC}{dt} &= \pm V_B G \quad (1)\\ C\frac{dT}{dt} &= \pm\frac{H_B}{c_p} V_B G \quad (2)\\ C\frac{dv}{dt} &= \pm M(t) \quad (3) \end{aligned} \Bigg\} \int_0^t \text{Source functions: } C^z(t),\ T^z(t),\ v^z(t).$$

Initial conditions: C_0, T_0, v_0.

Remark 1. For the integration of the eq. (I^s.2) modulus of the eq. (I^s.1) has been used to keep the possibility of inverse transformation in the point Z(x,y,z,t).

4.2. THE ANALYSIS OF THE HOMOGENEOUS PART OF THE SYSTEM OF PARTIAL DIFFERENTIAL CONSTITUTIVE STATE EQUATIONS (I)

In continuation of the previous chapters we need to consider the homogeneous part of the system (I) having the form:

$$\frac{DC}{Dt}=D_C\nabla^2 C \mp C\,\mathrm{div}\mathbf{v} \tag{1}$$

$$(\mathrm{I}^h)\qquad C\frac{DT}{Dt}=\frac{\lambda_C}{c_p}\nabla^2 T \mp CT\,\mathrm{div}\mathbf{v}+\frac{H_C}{c_p}D_C\nabla^2 C \tag{2}$$

$$C\frac{Dv}{Dt}=\eta_w\nabla^2\mathbf{v}\mp Cv\,\mathrm{div}\mathbf{v}+\frac{\eta_w}{3}\mathrm{grad}\,\mathrm{div}\mathbf{v} \tag{3}$$

where: $\frac{D}{Dt}=\frac{\partial}{\partial t}\pm$vgrad $-$ (sum of adequate gradient operations) [10], [1–6].

Introducing now the definition of the rotational fields:

Definition 1. Exists vector **B**, vector potential of the flow field such that $\mathbf{v}=\mathrm{rot}\mathbf{B}$ and $\mathrm{divrot}\mathbf{B}\equiv 0$ [9], [10], [1–6],

we have from the system (I^h):

– for the rotational fields

(I_1^{hp})

$$\frac{\partial C}{\partial t}=D_C\nabla^2 C \tag{1}$$

$$C\frac{\partial T}{\partial t}=\frac{\lambda_C}{c_p}\nabla^2 T+\frac{H_C}{c_p}D_C\nabla^2 C. \tag{2}$$

To the eq. $(I_2^{hp}.2)$ we introduce the rules [1]

$$C\frac{\partial H}{\partial t}=C\frac{\partial T}{\partial t}+\frac{H_C}{c_p}\frac{\partial C}{\partial t} \tag{3}$$

and without source [1]

$$H=c_pT. \tag{4}$$

Consequently we have

(I_2^{hp})

$$\frac{H_C}{c_p}\frac{\partial C}{\partial t}=\frac{H_C}{c_p}D_C\nabla^2 C \tag{1}$$

$$C\frac{\partial T}{\partial t}+\frac{H_C}{c_p}\frac{\partial C}{\partial t}=\frac{\lambda_C}{c_p}\nabla^2 T+\frac{H_C}{c_p}D_C\nabla^2 C. \tag{2}$$

Substraction of the eq. $(I_2^{hp}.1)$ from the eq. $(I_2^{hp}.2)$ allows to obtain at last

(I_3^{hp})

$$\frac{\partial C}{\partial t}=D_C\nabla^2 C \tag{1}$$

$$C\frac{\partial T}{\partial t}=\frac{H_C}{c_p}D_C\nabla^2 C \tag{2}$$

$$C\frac{\partial T}{\partial t}=\frac{\lambda_C}{c_p}\nabla^2 T \tag{3}$$

– for the rotational field

(I_1^{hr})

$$\frac{DC}{Dt}=\mp C\,\mathrm{div}\mathbf{v} \tag{1}$$

$$C\frac{DT}{Dt}=\mp CT\,\mathrm{div}\mathbf{v} \tag{2}$$

$$C\frac{D\mathbf{v}}{Dt}=\eta_w\nabla^2\mathbf{v}\mp C\mathbf{v}\,\mathrm{div}\mathbf{v}+\frac{\eta_w}{3}\,\mathrm{grad}\,\mathrm{div}\mathbf{v} \tag{3}$$

where:

$$\frac{D}{Dt}=\frac{\partial}{\partial t}\pm \mathrm{vgrad} \quad [10], [1-6].$$

5. ANALYTICAL SOLUTION OF THE PHENOMENAL PARTIAL DIFFERENTIAL EQUATIONS OF POTENTIAL FIELDS AND ROTATIONAL FIELD WITH RESPECT TO THEIR SOURCES

A. EXISTENCE OF THE INITIAL CONDITIONS

1. The source functions for the phenomenal solutions

The integration of the system (I^s) pertinent to the procedure described in the chapter 3.1 makes possible to obtain:

– from the eq. (I^s1) after integration we have

$$C^z(t)=C_0\pm\int_0^t V_B\mathrm{Gdt} \tag{1.1}$$

consequently,

– from the eq. (I^s2) with respect to REMARK 1 one obtains

$$T^z(t)=T_0\pm\frac{H_B}{c_p}\ln\left|\frac{C^z(t)}{C_0}\right| \tag{1.2}$$

with,

– from the eq. (I^s3) one can have

$$\mathrm{v}^z(t)=\mathrm{v}_0\pm\int_0^t M(t)\,\mathrm{dt}. \tag{1.3}$$

The formulae (1.1), (1.2) and (1.3) are the sources for the phenomenal responses of the continuous medium inside Ω_{Rt}. The source functions [$C^z(t)$, $T^z(t)$, $v^z(t)$] are the coordinates of the source state vector correlated to adequate physical phenomena of the potential fields and rotational field.

2. The phenomenal solutions of the potential fields

2.1. The diffusion concentration transport (variable point $Q(\xi,\eta,\zeta,\tau)$

Let us consider the partial differential equation of the diffusion concentration transport (I_3^{hp}1) rewritten as:

$$\frac{\partial C}{\partial t}-D_C\nabla^2 C=0. \tag{2.1.1}$$

The concentration function can be presented as follows

$$C(x,y,z,t)=U(x,u,z,t)\,\bar{C}(t) \tag{2.1.2}$$

which modifies the eq. (2.1.1) to the system form

$$\frac{\partial^2 U}{\partial x^2}+\frac{\partial^2 U}{\partial y^2}+\frac{\partial^2 U}{\partial z^2}+\lambda U=0 \tag{2.1.3}$$

and

$$\frac{d\bar{C}}{dt}=-\lambda D_C \bar{C}. \tag{2.1.4}$$

Furthering we assume that the conditions on the brim of the volume-time element $\bar{\Omega}_{(a,b,c)t}$ are equal to zero [11]

$$\begin{bmatrix} U(0,\ y,\ 0)\\ U(x,\ 0,\ 0)\\ U(0,\ 0,\ z)\end{bmatrix}=0 \qquad \text{and} \qquad \begin{bmatrix} U(a,\ y,\ 0)\\ U(x,\ b,\ 0)\\ U(0,\ y,\ c)\end{bmatrix}=0. \tag{2.1.5}$$

The introduction of the new form of the function U(x,y,z) as follows

$$U(x,y,z)=X(x)\,Y(y)\,Z(z) \tag{2.1.6}$$

into eq. (2.1.3) allows to obtain

$$\begin{matrix} X''+\vartheta X=0,\\ Y''+\mu Y=0,\\ Z''+\rho Z=0,\end{matrix} \qquad \begin{bmatrix} X(0)\\ Y(0)\\ Z(0)\end{bmatrix}=0 \qquad \text{and} \qquad \begin{bmatrix} X(a)\\ Y(b)\\ Z(c)\end{bmatrix}=0. \tag{2.1.7}$$

The space ordinary differential equations (2.1.7) have the solutions

$$\begin{aligned} X_n(x)&=\sin\frac{n\Pi x}{a}, & \vartheta&=\left(\frac{n\Pi}{a}\right)^2\\ Y_m(y)&=\sin\frac{m\Pi y}{b}, & \mu&=\left(\frac{m\Pi}{b}\right)^2\\ Z_k(z)&=\sin\frac{k\Pi z}{c}, & \rho&=\left(\frac{k\Pi}{c}\right)^2 \end{aligned} \tag{2.1.8}$$

with eigenvalues

$$\lambda_{n,m,k}=\left(\frac{n\Pi}{a}\right)^2+\left(\frac{m\Pi}{b}\right)^2+\left(\frac{k\Pi}{c}\right)^2 \tag{2.1.9}$$

and continuing our considerations the eigenfunctions

$$U_{n,m,k}=A_{n,m,k}\sin\frac{n\Pi x}{a}\sin\frac{m\Pi y}{b}\sin\frac{k\Pi z}{c}. \tag{2.1.10}$$

In the approach of this article we need to determine the amplitude $A_{n,m,k}$ respectively to the dimensions "*a,b,c*" for time "t" so that the norm of the eigenfunctions $U_{n,m,k}$ with the wieght equal to 1 (one) is equal to 1 (one) [11], [10], [2–6]

$$N_e = \int_0^a \int_0^b \int_0^c U^2_{n,m,k}\,dx\,dy\,dz = A^2_{n,m,k} \int_0^a \sin^2 \frac{n\Pi x}{a}\,dx \int_0^b \sin^2 \frac{m\Pi y}{b}\,dy \cdot \int_0^c \sin^2 \frac{k\Pi z}{c}\,dz = 1. \quad (2.1.11)$$

The integration of the eq. (2.1.11) makes possibility to obtain the value of the amplitude of the eigenfunctions

$$A_{n,m,k} = \sqrt{\frac{8}{abc}}. \quad (2.1.12)$$

The eigenfunctions belonging to the constant coordinates point $Z(x,y,z,t)$ have the form

$$U_{n,m,k}(x,y,z) = \sqrt{\frac{8}{abc}}\,W(x,y,z) \quad (2.1.13)$$

where is that

$$W(x,y,z) = \sin \frac{n\Pi x}{a} \sin \frac{m\Pi y}{b} \sin \frac{k\Pi z}{c}$$

and consequently for the variable coordinates point $Q(\xi, \eta, \beta, \tau)$ the eigenfunctions are

$$U_{n,m,k}(\xi,\eta,\beta) = \sqrt{\frac{8}{abc}}\,W(\xi,\eta,\zeta) \quad (2.1.14)$$

by the additional formula

$$W(\xi,\eta,\zeta) = \sin \frac{n\Pi\xi}{a} \sin \frac{m\Pi\eta}{b} \sin \frac{k\Pi\zeta}{c}.$$

After time integration of the eq. (2.1.4) one can obtain

$$\bar{C}(t) = D^C_{n,m,k}\, e^{\lambda_{n,m,k} D_c t}. \quad (2.1.15)$$

In the eq. (2.1.15) the integration constant possesses the form

$$D^C_{n,m,k} = \iiint_{\Omega_R} C_0 \sqrt{\frac{8}{abc}}\,W(\xi,\eta,\beta)\,d\xi\,d\eta\,d\zeta, \quad (2.1.16)$$

where exists

$$W(\xi,\eta,\beta) = \sin \frac{n\Pi\xi}{a} \sin \frac{m\Pi\eta}{b} \sin \frac{k\Pi\zeta}{c}.$$

The source function for the concentration is:

$$C^z(t) = C_0 \pm \int_0^t V_B G\,dt. \quad (2.1.17)$$

The above presented considerations enable to state the resultant solution for the diffusion concentration transport:

$$C(Z,t)=\int_0^t \iiint_{\Omega_R} \left\{ \sum_{\substack{n=1\\m=1\\k=1}}^{\infty} D^{C}_{n,m,k} e^{-\lambda_{n,m,k} D_o (t-\tau)} \frac{8}{abc} W(x,\xi;y,\eta;z,\zeta) \right\} (C_0 \pm \pm \int_0^{t\to\tau} V_B G \mathrm{d}t)\, \mathrm{d}\xi \mathrm{d}\eta \mathrm{d}\beta \mathrm{d}\tau , \tag{2.1.18}$$

where:

$$W(x,\xi;y,\eta;z,\zeta)=W(x,y,z)\ W(\xi,\eta,\beta)$$

and the part of the solution (2.1.18) inside brackets {} is the phenomenal Green function

$$G^{D}_{Ph}(Z,t;Q,\tau)=\left\{ \sum_{\substack{n=1\\m=1\\k=1}}^{\infty} D^{C}_{n,m,k} e^{-\lambda_{n,m,k} D_o (t-\tau)} \frac{8}{abc} W(x,\xi;y,\eta;z,\zeta) \right\} \tag{2.1.19}$$

which properties represent reciprocal constant image for points [7–8], [11], [2–6]

$$\underset{\in \bar{\Omega}_{(a,b,c)t}}{Z(x,y,z,t)} \xleftrightarrow[\text{constant image}]{\text{Reciprocal}} \underset{\in \Omega_{Rt}}{Q(\xi,\eta,\zeta,\tau)} \tag{2.1.20}$$

2.2. The diffusion enthalpy transport (variable point $Q(\xi,\eta,\zeta,\tau)$

The partial differential equation of the topic problem has the form

$$C\frac{\partial T}{\partial t}=\frac{H_C}{c_p} D_C \nabla^2 C. \tag{2.2.1}$$

The source formula for the partial differential equation (2.2.1) is:

$$C\frac{dT}{dt}=\frac{H_C}{c_p}\frac{dC}{dt} \tag{2.2.2}$$

and this equation is coming from the eq. (2.2.1) and the diffusion relation

$$\frac{H_C}{c_p}\frac{\partial C}{\partial t}=\frac{H_C}{c_p} D_C \nabla^2 C. \tag{2.2.3}$$

From the eq. (2.2.1) and eq. (2.2.3) we can write

$$\frac{H_C}{c_p}\frac{\partial C}{\partial t}=C\frac{\partial T}{\partial t}=\frac{H_C}{c_p} D_C \nabla^2 C \tag{2.2.4}$$

which can be modified to

$$C\frac{\partial T}{\partial t}=\frac{H_C}{c_p}\left(\frac{\partial C}{\partial t}-D_C\nabla^2 C\right) \tag{2.2.5}$$

with additionaly

$$\frac{H_C}{c_p}\frac{\partial C}{C}=\partial T. \tag{2.2.6}$$

Consequently to the solution of the chapter 2.1 the solution of the problem of the diffusion enthalpy transport due to above considerations is given by the formula:

$$T(Z,t)=T_0(Z,t)+\frac{H_C}{c_p}\left[\ln\left|\int_0^t\iiint_{\Omega_R}\left\{\sum_{\substack{n=1\\m=1\\k=1}}^{\infty}D^C_{n,m,k}e^{-\lambda_{n,m,k}D_C(t-\tau)}\frac{8}{abc}W(x,\xi;y,\eta;z,\zeta)\right\}(C_0\pm\right.\right.$$

$$\left.\left.\pm\int_0^{t\to\tau}V_B G\,dt)\,d\xi d\eta d\zeta d\tau\right|-\ln|C_0|\right]. \tag{2.2.7}$$

The solution given by the eq. (2.2.7) possesses the identical phenomenal Green function as numbered (2.1.19) for the solution (2.1.18) in the chapter 2.1.

2.3. The heat transfer temperature transport (variable point $Q(\xi,\eta,\zeta,\tau)$)

From the system of partial differential equations of the potential fields (I_3^{hp}) the heat transfer temperature transport is governed by the formula (I_3^{hp}.3); rewritten as:

$$C\frac{\partial T}{\partial t}-\frac{\lambda_C}{c_p}\nabla^2 T=0. \tag{2.3.1}$$

For the eq. (2.3.1) is introduced the temperature function

$$T(x,y,z,t)=V(x,y,z)\bar{T}(t) \tag{2.3.2}$$

which transforms the partial differential equation (2.3.1) to the system form

$$\frac{\partial^2 V}{\partial x^2}+\frac{\partial^2 V}{\partial y^2}+\frac{\partial^2 V}{\partial z^2}+\lambda V=0 \tag{2.3.3}$$

and

$$\frac{d\bar{T}}{dt}=-\lambda\frac{\lambda_C}{c_p C(t)}\bar{T}. \tag{2.3.4}$$

With the assumption of the zero brim conditions of the temperature for $\bar{\Omega}_{(a,b,c)t}$ (see chapter 2.1) we have the eigenvalues

$$\lambda_{n,m,k}=\left(\frac{n\Pi}{a}\right)^2+\left(\frac{m\Pi}{b}\right)^2+\left(\frac{k\Pi}{c}\right)^2. \tag{2.3.5}$$

The eigenfunctions for the constant coordinates point Z(x,y,z,t) are:

$$V_{n,m,k}(x,y,z)=\sqrt{\frac{8}{abc}}W(x,y,z) \tag{2.3.6}$$

for W(x,y,z) – see chapter 2.1.

The eigenfunctions for the variable coordinates point $Q(\xi,\eta,\zeta,\tau)$ have the form

$$V_{n,m,k}(\xi,\eta,\beta)=\sqrt{\frac{8}{abc}}W(\xi,\eta,\zeta). \tag{2.3.7}$$

The integration of the eq. (2.3.4) leads to the result

$$\bar{T}(t)=D^{T}_{n,m,k}\,e^{-\lambda_{n,m,k}\frac{\lambda_c}{c_p}[J(t)]}$$

where (2.3.8)

$$[J(t)]=\int\frac{dt}{C_0\pm\int_0^t V_B G\,dt}.$$

The eq. (2.3.8) contains the integration constant possessing the form

$$D^{T}_{n,m,k}=\iiint_{\Omega_R} T_0\sqrt{\frac{8}{abc}}\,W(\xi,\eta,\zeta)\,d\xi\,d\eta\,d\zeta. \tag{2.3.9}$$

With the source function of the heat transfer phenomenon eq. (1.2) rewritten here as

$$T^{z}(t)=T_0\pm\frac{H_B}{c_p}\ln\left|\frac{C^{z}(\tau)}{C_0}\right| \tag{2.3.10}$$

the heat transfer temperature transport solution has the form

$$T(Z,t)=\int_0^t\iiint_{\Omega_R}\left\{\sum_{\substack{n=1\\m=1\\k=1}}^{\infty}D^{T}_{n,m,k}\,e^{-\lambda_{n,m,k}\frac{\lambda_c}{c_p}[J(t-\tau)]}\frac{8}{abc}W(x,\xi;y,\eta;z,\zeta)\right\}\left[T_0\pm\right.$$

$$\left.\pm\frac{H_B}{c_p}\ln\left|\frac{C^{z}(\tau)}{C_0}\right|\right]d\xi\,d\eta\,d\zeta\,d\tau, \tag{2.3.11}$$

where $W(x,\xi;y,\eta;z,\zeta)=W(x,y,z)\,W(\xi,\eta,\zeta)$.

For the solution (2.3.11) the phenomenal Green function has the form:

$$G^{T}_{Ph}(Z,t;Q,\tau)=\left\{\sum_{\substack{n=1\\m=1\\k=1}}^{\infty}D^{T}_{n,m,k}\,e^{-\lambda_{n,m,k}\frac{\lambda_c}{c_p}[J(t-\tau)]}\frac{8}{abc}W(x,\xi;y,\eta;z,\zeta)\right\} \tag{2.3.12}$$

for the reciprocal constant image of the points [7–8], [11], [2–6]

$$\underset{\in \bar{\Omega}_{(a,b,c)t}}{Z(x,y,z,t)} \xleftrightarrow[\text{constant image}]{\text{Reciprocal}} \underset{\in \Omega_{Rt}}{Q(\xi,\eta,\zeta,\tau)} \tag{2.3.13}$$

3. The phenomenal solution of the rotational field

3.1. The field vector velocity (variable point $Q'(\xi',\eta',\beta',\tau')$).

As an introduction of our considerations let us recall the partial differential equations of the field vector velocity in the forms

$$\frac{DC}{Dt} = \mp C \operatorname{div} \mathbf{v} \tag{3.1.1}$$

$$C\frac{DT}{Dt} = \mp CT \operatorname{div} \mathbf{v} \tag{3.1.2}$$

$$C\frac{D\mathbf{v}}{Dt} = \eta_w \nabla^2 \mathbf{v} \mp Cv \operatorname{div} \mathbf{v} + \frac{\eta_w}{3} \operatorname{grad} \operatorname{div} \mathbf{v} \tag{3.1.3}$$

where:

$$\frac{D}{Dt} = \frac{\partial}{\partial t} \pm \mathbf{v} \operatorname{grad} \quad [10], [1-6].$$

Applying eq. (3.1.1) multiplied both handsides by the field vector velocity "v" into the eq. (3.1.3) this equation can be presented as:

$$C\frac{D\mathbf{v}}{Dt} - \mathbf{v}\frac{DC}{Dt} = \eta_w \nabla^2 \mathbf{v} + \frac{\eta_w}{3} \operatorname{grad} \operatorname{div} \mathbf{v}. \tag{3.1.4}$$

The assumption of the reciprocal constant image for the volume element $\bar{\Omega}_{(a,b,c)t}$ being in the statical state, the eq. (3.1.4) obtains its system form [7–8], [10]

$$\eta_w \nabla^2 \mathbf{v} + \frac{\eta_w}{3} \operatorname{grad} \operatorname{div} \mathbf{v} = 0 \tag{3.1.5}$$

and

$$C\frac{D\mathbf{v}}{Dt} - \mathbf{v}\frac{DC}{Dt} = 0. \tag{3.1.6}$$

As a consequence of the application of the rule [9]

$$\operatorname{rot} \operatorname{rot} \boldsymbol{\nu} = \operatorname{grad} \operatorname{div} \mathbf{v} - \nabla^2 \mathbf{v} \tag{3.1.7}$$

to the eq. (3.1.5) this equation gets its new form

$$\left(\eta_w + \frac{\eta_w}{3}\right) \operatorname{grad} \operatorname{div} \mathbf{v} - \eta_w \operatorname{rot} \operatorname{rot} \boldsymbol{\nu} = 0. \tag{3.1.8}$$

Subsequently the eq. (3.18) is undergoing to the transfromation with the result

$$\overline{\nabla}^2 v = 0 = \alpha \operatorname{grad} \operatorname{div} v - \beta \operatorname{rot} \operatorname{rot} v \tag{3.1.9}$$

where:

$$\alpha = \eta_w + \frac{\eta_w}{3} \quad \text{and} \quad \beta = \eta_w .$$

In the statical state, for $t=0$ every constant coordinates point $Z(x,y,z,t)$ possesses the field vector velocity $v_0(x,y,z,0)=v_0$ and the statical distribution of the field vector velocity is given by the formula

$$\mathrm{v}(\mathrm{Z},0)=\mathrm{v}(\mathrm{x},\mathrm{y},\mathrm{z},0)=\iiint_{\Omega_R} \Gamma_G^W(\mathrm{Z},\mathrm{Q}')\,\mathrm{d}\xi'\,\mathrm{d}\eta'\,\mathrm{d}\zeta' . \tag{3.1.10}$$

The formula (3.1.10) for the following cases of Green tensor for $\bar{\Omega}_{(a,b,c)}$ remains valid [12], [2–6]:

– the isotropic case, see Appendix 1

$$\Gamma_G^W(\mathrm{Z},\mathrm{Q}') = \Gamma_G(\mathrm{Z},\mathrm{Q}') \tag{3.1.11}$$

– the anisotropic case, see Appendix 2

$$\Gamma_G^W(\mathrm{Z},\mathrm{Q}') = \Gamma_{GA}(\mathrm{Z},\mathrm{Q}') . \tag{3.1.12}$$

Furthering we need to consider the formula governing the dynamical aspects of the field vector velocity, and for the realization of this task let us recall the following formula

$$C\frac{\mathrm{D}v}{\mathrm{Dt}} - \mathrm{v}\frac{\mathrm{D}C}{\mathrm{Dt}} = 0 \tag{3.1.13}$$

which is subsequently integrated to the form

$$\ln|\mathrm{N}| + \ln|\mathrm{v}| = \ln|\mathrm{C}| . \tag{3.1.14}$$

The eq. (3.1.14) can be modified to

$$\mathrm{v}(t) = \frac{C^z(t)}{N_1} \tag{3.1.15}$$

where: $C^z(t)$ – the source function of the concentration in the point $Z(x,y,z,t)$, for $x=a=\text{const}$, $y=b=\text{const}$, $z=c=\text{const}$.

$$\mathrm{C}^z(\mathrm{t}) = \mathrm{C}_0 \pm \int_0^t \mathrm{V}_B\,\mathrm{G}\,\mathrm{dt} \tag{3.1.16}$$

N_1 – the integration constant taken so that $\mathrm{N}_1 = \mathrm{v}(\mathrm{Z},0)$ from the eq. (3.1.10).

The resultant solution of the field vector velocity distribution by the source function of the field vector velocity

$$v^z(t)=v_0\pm\int_0^t M(t)\,dt \tag{3.1.17}$$

is given by the formula

$$v(Z,t)=\int_0^t\iiint_{\Omega_R}\left\{\frac{1}{N_1}\left[C_0\pm\int_0^{t\to(-\tau)}V_B G\,d\tau\right]\Gamma_G^W(Z,Q')\right\}\left[v_0\pm\pm\int_0^{t\to\tau}M(t)\,dt\right]d\xi'\,d\eta'\,d\zeta'\,d\tau \tag{3.1.18}$$

For the solution (3.1.18) the phenomenal Green function of the field vector velocity is

$$G_{Ph}^v(Z,t;Q;\tau)=\left\{\frac{1}{N_1}\left[C_0\pm\int_0^{t\to(t-\tau)}V_B G\,d\tau\right]\Gamma_G^W(Z,Q')\right\} \tag{3.1.19}$$

as the reciprocal interpretation of the constant and variable coordinates points [7–8], [11], [2–6]

$$\underset{\in\bar{\Omega}_{(a,b,c)t}}{Z(x,y,z,t)}\quad\underset{\text{constant image}}{\Longleftrightarrow\!\!\Longrightarrow}\quad\underset{\in\Omega_{Rt}}{Q'(\xi;\eta;\beta;\tau)} \tag{3.1.20}$$

4. The complete solution of the constitutive distributed parameter model by the existence of the initial conditions

The complete solution of the constitutive distributed parameter model bases on the state vectors of mass/charge, energy and momentum coordinates of the source and phenomenal solutions of the potential and rotational fields by the initial conditions. This chapters idea according to the chapters (A.1, A.2 and A.3), has its precisely made interpretation as follows:

Complete form of solution		Source functions in point Z(x,y,z,t)	
Concentration	=	Concentration changes (formula (A.1.1))	+
Temperature		Temperature changes (formula (A.1.2))	
Field vector velocity		Field vector velocity changes (formula (A.1.3))	

Diffusion Q→Z

Heat transfer Q→Z

Field vector velocity Q'→Z

$$+\left[\begin{array}{l}\text{Concentration changes}\\ \text{Temperature changes (formula (A.2.2.7))}\\ *\end{array}\right]+\left[\begin{array}{l}*\\ \text{Temperature changes (formula (A.2.3.11))}\\ *\end{array}\right]+$$

$$+\left[\begin{array}{l}\text{Concentration continuity (formula (A.3.1.1))}\\ \text{Temperature continuity (formula (A.3.1.2))}\\ \text{Field vector velocity changes (formula (A.3.1.18))}\end{array}\right]$$

* — the places without relations between physical phenomena.

(4.1)

B. EXISTENCE OF THE INITIAL AND BOUNDARY CONDITIONS

1. The surface boundary source problems

Let us assume the existence of the boundary functions for: — mass/charge $S(t)$, — thermic energy $U(t)$-scalar, — momentum $W(t)$-wector, on the outside oriented by the normal outside surface orientation vector $n_R^{(Z)}$ surface F_{Rt} of the working space volume-time Ω_{Rt}. While the system of the partial differential constitutive state equations (I) on the surface F_{Rt} is valid so that for the point $R(x,y,z,t)\in F_{Rt}$, for $x=\text{const}$, $y=\text{const}$, $z=\text{const}$ is:

$$(\mathrm{I}^{\mathrm{b}})\qquad \left.\begin{array}{ll}\dfrac{dC}{dt}=\mp S(t) & (1)\\ C\dfrac{dT}{dt}=\mp\dfrac{1}{c_p}U(t) & (2)\\ C\dfrac{dv}{dt}=\mp W(t) & (3)\end{array}\right\}\ \begin{matrix}t\\ \\ \\ 0\end{matrix}$$

Boundary source functions;

$$C^b(t),\ T^b(t),\ \mathbf{v}^b(t).$$

ASSUMPTION 1. The surface F_{Rt} is defined as having cuboidal form of constant space dimensions.

Boundary initial conditions: C_{bo}, T_{bo}, v_{bo}.

For the considerations of this paper one can assume that

$$(\mathrm{I}^{\mathrm{bi}})\qquad \begin{array}{ll}C_o=C_{bo} & (4)\\ T_o=T_{bo} & (5)\\ \mathbf{v}_o=\mathbf{v}_{bo} & (6)\end{array}$$

and the initial conditions of the inside of Ω_{Rt} are identical to those boundary conditions on the surface F_{Rt}.

Remark 2. For the system (I^b) validity of the REMARK 1 does not exist.

In the continuation of our considerations a necessary problem is to introduce:

Assumption 2. The boundary source fonctions on F_{Rt} generate the identical physical phenomena as existing inside working space volume-time Ω_{Rt}.

This assumption can be written as

$$[L^{bs}] \xrightarrow[\substack{\text{Physical}\\ \text{phenomena}\\ \text{generation,}}]{} [L^{cp}] \tag{AS2.1}$$

where: $[L^{bs}]$ – boundary source state vector F_{Rt},
$[L^{cp}]$ – state vector inside working space volume–time Ω_{Rt}.

2. The boundary source functions

The integration of the boundary source system of time ordinary differential equations (I^b) makes possible to obtain:

– for the concentration boundary source function from the eq. (I^b1) one gets

$$C^b(t) = C_{bo} \mp \int_0^t S(t)dt, \tag{2.1}$$

– for the temperature boundary source function from the eq. (I^b2) we have

$$T^b(t) = T_{bo} \mp \frac{1}{c_p} \int_0^t \frac{U(t)}{C^b(t)} dt \tag{2.2}$$

and consequently,

– for the field vector velocity boundary source function according to the eq. (I^b3) takes the form

$$v^b(t) = v_{bo} \mp \int_0^t \frac{W(t)}{C^b(t)} dt. \tag{2.3}$$

The boundary source functions (2.1), (2.2) and (2.3) are the sources for the generation of the physical phenomena from the source on the surface F_{Rt} in direction to the inside of Ω_{Rt}. Obtained above boundary source functions [$C^b(t)$, $T^b(t)$, $v^b(t)$] represent the source state vector for adequate physical phenomena of the potential and rotational fields.

The above presented ideas lead to:

Assumption 3. The given above ASSUMPTION 2 possesses its detailed based on the state-vectors form [2–6]:

$$\begin{bmatrix} C^{pot}(Q,\tau) \\ T^{pot}(Q,\tau) \\ \bigstar \end{bmatrix} \xleftarrow{\text{Potential fields}} \begin{bmatrix} C^{b}(t) \\ T^{b}(t) \\ v^{b}(t) \end{bmatrix} \xrightarrow{\text{Rotational field}} \begin{bmatrix} C^{rot}(Q;\tau) \\ T^{rot}(Q;\tau) \\ v^{rot}(Q;\tau) \end{bmatrix} \quad \text{(AS3.1)}$$

point $Q(\xi,\eta,\zeta,\tau) \in \Omega_{Rt}$ — point $R(x,y,z,t)$, $x=\text{const}, y=\text{const}, z=\text{const} \in F_{Rt}$ — point $Q'(\xi';\eta';\beta';\tau') \in \Omega_{Rt}$

where:
$\bigstar$ – the place without relation between physical phenomena.

3. The phenomenal solutions by the existence of the initial and boundary conditions

3.1. The phenomenal solutions of the potential fields

3.1.1. The diffusion concentration transport (variable point $Q(\xi,\eta,\zeta,\tau)$

From the chapter A.2.1. the phenomenal Green function of the initial conditions has for Ω_{Rt} the below form

$$G^{D}_{Ph}(Z,t;Q,\tau)=\sum_{\substack{n=1\\m=1\\k=1}}^{\infty} D^{C}_{n,m,k}\, e^{-\lambda_{n,m,k} D_s (t-\tau)} \frac{8}{abc} W(x,\xi;y,\eta;z,\zeta) \qquad (3.1.1.1)$$

with the integration constant given by the eq. (A.2.1.16).

On the basis of the formula (3.1.1.1), assuming that F_{Rt} has cuboidal shape one can determine the surface phenomenal Green functions:
– for "xy" surface

$$G^{Dxy}_{Ph}(R^{xy},\, t;Q^{\xi\eta},\tau)=\sum_{\substack{n=1\\m=1}}^{\infty} D^{Cxy}_{n,m}\, e^{-\lambda_{n,m} D_s (t-\tau)} \frac{4}{ab} W(x,\xi;y,\eta) \qquad (3.1.1.2)$$

where:

$$D^{Cxy}_{n,m}=\iint_{F_{Rxy}} C_{bo} \sqrt{\frac{4}{ab}}\, W(x,\xi;y,\eta)\, d\xi\, d\eta\,, \qquad (3.1.1.3)$$

– for "yz" surface

$$G^{Dyz}_{Ph}(R^{yz},\, t;Q^{\eta\zeta},\tau)=\sum_{\substack{m=1\\k=1}}^{\infty} D^{Cyz}_{m,k}\, e^{-\lambda_{m,k} D_s (t-\tau)} \frac{4}{bc} W(y,\eta;z,\zeta) \qquad (3.1.1.4)$$

where:

$$D_{m,k}^{Cyz} = \iint_{F_{R,yz}} C_{bo} \sqrt{\frac{4}{bc}}\, W(y,\eta;z,\zeta)\, d\eta\, d\zeta\,, \tag{3.1.1.5}$$

– for "xz" surface

$$G_{Ph}^{Dxz}(R^{xz},\, t;\, Q^{\xi\zeta},\tau) = \left\{ \sum_{\substack{n=1\\k=1}}^{\infty} D_{n,k}^{Cxz}\, e^{-\lambda_{n,k} D_e (t-\tau)}\, \frac{4}{ac}\, W(x,\xi;z,\zeta) \right\}, \tag{3.1.1.6}$$

where:

$$D_{n,k}^{Cxz} = \iint_{F_{R,xz}} C_{bo} \sqrt{\frac{4}{ac}}\, W(x,\xi;z,\zeta)\, d\xi\, d\zeta\,. \tag{3.1.1.7}$$

As a consequence of the application of the above considerations the complete form of the surface phenomenal Green function has been written:

$$G_{Ph}^{DS}(R,t;Q,\tau) = 2[G_{Ph}^{Dxy}(R^{xy},t;Q^{\xi\eta},\tau) + G_{Ph}^{Dyz}(R^{yz},t;Q^{\eta\zeta},\tau) + G_{Ph}^{Dxz}(R^{xz},t;Q^{\xi\zeta},\tau)]. \tag{3.1.1.8}$$

With the complete phenomenal Green functions at hand, of time and volume problems eq. (3.1.1.1) and surface problems eq. (3.1.1.8) with adequate source functions:
– for the volume-time problems see eq. (A.1.1), and
– for the surface-time problems given by the eq. (B.2.1), the resultant solution of the topic diffusion concentration transport

$$C(Z,R;t) = \int_0^t \iiint_{\Omega_R} G_{Ph}^{D}(Z,t;Q,\tau)\, (C_0 \pm \int_0^{t\to\tau} V_B G dt)\, d\xi d\eta d\zeta d\tau + \int_0^t \oiint_{F_R} G_{Ph}^{DS}(R,t;Q,\tau) \left[C_{bo} \mp \int_0^{t\to\tau} S(t) dt \right] d\xi d\eta d\zeta d\tau. \tag{3.1.1.9}$$

3.1.2. The diffusion enthalpy transport (variable point $Q(\xi,\eta,\beta,\tau)$)

The content of the chapter B.3.1.1. allows on the formulation of the solution of the problem of the diffusion enthalpy transport:

$$T(Z,R;t)=T_0(Z,t)+\frac{H_C}{c_p}\left[\ln\left|\int_0^t\iiint_{\Omega_R} G_{Ph}^{D}(Z,t;Q,\tau)\left(C_0\pm\int_0^{t\to\tau}V_B G\,dt\right)d\xi\,d\eta\,d\zeta\,d\tau\right|-\right.$$

$$-\ln\left|C_0\right|\Big]+T_{bo}(R,t)+\frac{H_C}{c_p}\left[\ln\left|\int_0^t\oint_{F_R} G_{Ph}^{DS}(R,t;Q,\tau)\Big[C_{bo}\mp\right.\right.$$

$$\left.\left.-\int_0^{t\to\tau}S(t)dt\Big]d\xi\,d\eta\,d\zeta\,d\tau\right|-\ln\left|C_{bo}\right|\right]. \qquad (3.1.2.1)$$

In the above formula the phenomenal Green functions are determined in the chapter B.3.1.1. as the eq. (3.1.1.1) – the working space volume-time problems and the eq. (3.1.1.8) the working surface-time problems.

3.1.3. The heat transfer temperature transport (variable point $Q(\xi,\eta,\zeta,\tau)$)

The phenomenal Green function of the heat transfer phenomenon for the initial and working space volume-time problems consequently to the chapter A.2.3. is given by the formula:

$$G_{Ph}^{T}(Z,t;Q,\tau)=\sum_{\substack{n=1\\m=1\\k=1}}^{\infty} D_{n,m,k}^{T}\,e^{-\lambda_{n,m,k}\frac{\lambda_C}{c_p}[J(t-\tau)}\,W(x,\xi;y,\eta;z,\zeta) \qquad (3.1.3.1)$$

where $D_{n,m,k}^{T}$ is shown by the eq. (A.2.3.9) and $[J(t-\tau)]$ is taken from the eq. (A.2.3.8). From the assumed cuboidal shape of the surface F_{Rt} there is

$$[J^s(t)]=\int\frac{dt}{C_{bo}\mp\int_0^t S(t)dt} \qquad (3.1.3.2)$$

– for "xy" surface

$$G_{Ph}^{Txy}(R^{xy},t;Q^{\xi\eta},\tau)=\left\{\sum_{\substack{n=1\\m=1}}^{\infty} D_{n,m}^{Txy}\,e^{-\lambda_{n,m}\frac{\lambda_C}{c_p}[J^s(t-\tau)]}\frac{4}{ab}\,W(x,\xi;y,\eta)\right\} \qquad (3.1.3.3.)$$

and

$$D_{n,m}^{Txy}=\iint_{F_{Rxy}} T_{bo}\sqrt{\frac{4}{ab}}\,W(x,\xi;y,\eta)\,d\xi\,d\eta, \qquad (3.1.3.4)$$

– for "yz" surface

$$G_{Ph}^{Tyz}(R^{yz},t;Q^{\eta\zeta},\tau)=\left\{\sum_{\substack{m=1\\k=1}}^{\infty} D_{m,k}^{Tyz}\,e^{-\lambda_{m,k}\frac{\lambda_C}{c_p}[J^s(t-\tau)}\frac{4}{bc}\,W(y,\eta;z,\zeta)\right\} \qquad (3.1.3.5.)$$

and

$$D_{m,k}^{Tyz} = \iint_{F_{Ryz}} T_{bo} \sqrt{\frac{4}{bc}}\, W(y,\eta;z,\zeta)\, d\eta\, d\zeta, \tag{3.1.3.6}$$

– for "xz" surface

$$G_{Ph}^{Txz}(R^{xz},t;Q^{\xi\zeta},\tau) = \left\{ \sum_{\substack{n=1\\k=1}}^{\infty} D_{n,k}^{Txz}\, e^{-\lambda_{n,k}\frac{\lambda_c}{c_p}[J^n(t-\tau)}\, \frac{4}{ac}\, W(x,\xi;z,\zeta) \right\} \tag{3.1.3.7.}$$

and

$$D_{n,k}^{Txz} = \iint_{F_{Rxz}} T_{bo} \sqrt{\frac{4}{ac}}\, W(x,\xi;z,\zeta)\, d\xi\, d\zeta, \tag{3.1.3.8}$$

As the result of above formulae the complete form of the surface phenomenal Green function for the working surface F_{Rt} can be written

$$D_{Ph}^{TS}(R,t;Q,\tau) = 2\left[G_{Ph}^{Txy}(R^{xy},t;Q^{\xi\eta},\tau) + G_{Ph}^{Tyz}(R^{yz},t;Q^{\eta\zeta},\tau) + G_{Ph}^{Txz}(R^{xz},t;Q^{\xi\zeta},\tau)\right]. \tag{3.1.3.9}$$

The phenomenal Green functions (3.1.3.1) and (3.1.3.9) with their source functions:
– the volume-time source function given by the eq. (A.1.2)
– the surface-time source function the eq. (B.2.2)
allow to formulate the initial and boundary conditions heat transfer solution:

$$T(Z,R;t) = \int_0^t \iiint_{\Omega_R} G_{Ph}^{T}(Z,t;Q,\tau)\left[T_0 \pm \frac{H_B}{c_p} \ln\left|\frac{C^z(\tau)}{C_0}\right|\right] d\xi\, d\eta\, d\zeta\, d\tau +$$

$$+ \int_0^t \oiint_{F_R} G_{Ph}^{TS}(R,t;Q,\tau)\left[T_{bo} \mp \frac{1}{c_p} \int_0^{t\to\tau} \frac{U(t)}{C^b(t)}\, dt\right] d\xi\, d\eta\, d\beta\, d\tau. \tag{3.1.3.10}$$

3.2. The phenomenal solution of the rotational field

3.2.1. The field vector velocity (variable point $Q'(\xi',\eta',\zeta',\tau)$).

The outset of our considerations is the working space volume-time field vector velocity general Green function from eq. (A.3.1.19)

$$G_{Ph}^{V}(Z,t;Q;\tau) = \frac{1}{N_1}\left[C_0 \pm \int_0^{t\to(t-\tau)} V_B G\, dt\right] \Gamma_G^W(Z,Q') \tag{3.2.1.1}$$

and contains two different cases:

– the isotropic case, see Appendix 1

$$\Gamma_G^W(Z,Q') = \Gamma_G(Z,Q') \tag{3.2.1.2}$$

– the anisotropic case, see Appendix 2

$$\Gamma_G^W(Z,Q') = \Gamma_{GA}(Z,Q'). \tag{3.2.1.3}$$

The working space volume-time phenomenal Green function (3.1.1.1) possesses its following surface-time coordinates:

– for "xy" surface

$$G_{Ph}^{vxy}(R^{xy},t;Q'^{\xi',\eta'}\tau\frac{1}{N_1^{xy}}\left[C_{b0}\mp\int_0^{t\to(t-\tau)}S(t)dt\right]\Gamma_G^{Wxy}(R^{xz},Q'^{\xi'\eta'}) \tag{3.2.1.4}$$

where:

$$N_1^{xy}=\iint_{F_{R,xy}} v_0^{xy}\,\Gamma_G^{Wxy}(R^{xy},Q'^{\xi'\eta'})\,d\xi'\,d\eta' \tag{3.2.1.5}$$

– for "yz" surface

$$G_{Ph}^{vyz}(R^{yz},t;Q'^{\eta',\xi'}\tau=\frac{1}{N_1^{yz}}\left[C_{b0}\mp\int_0^{t\to(t-\tau)}S(t)dt\right]\Gamma_G^{Wyz}(R^{yz},Q'^{\eta'\zeta'}) \tag{3.2.1.6}$$

where:

$$N_1^{yz}=\iint_{F_{R,yz}} v_0^{yz}\,\Gamma_G^{Wyz}(R^{yz},Q'^{\eta'\zeta'})\,d\eta'\,d\zeta' \tag{3.2.1.7}$$

– for "xz" surface

$$G_{Ph}^{vxz}(R^{xz},t;Q'^{\zeta'\zeta'},\tau)=\frac{1}{N_1^{xz}}\left[C_{b0}\mp\int_0^{t\to(t-\tau)}S(t)dt\right]\Gamma_G^{Wxz}(R^{xz},Q'^{\xi'\zeta'}) \tag{3.2.1.8}$$

where:

$$N_1^{xz}=\iint_{F_{R,xz}} v_0^{xz}\,\Gamma_G^{Wxz}(R^{xz},Q'^{\xi'\zeta'})\,d\xi'\,d\zeta'. \tag{3.2.1.9}$$

With the above surface time considerations at hand the complete form of the working surface-time phenomenal Green function can be stated:

$$D_{Ph}^{vS}(R,t;Q';\tau)=2\left[G_{Ph}^{vxy}(R^{xy},t;Q'^{\xi'\eta'},\tau)+G_{Ph}^{vyz}(R^{yz},Q'^{\eta'\zeta'},\tau)+\right.$$
$$\left.+G_{Ph}^{vxz}(R^{xz},t;Q',^{\xi',\zeta'},\tau)\right] \tag{3.2.1.10}$$

The field vector velocity phenomenal Green functions of working space wolume-time the eq. (3.2.1.1) and working surface-time the eq. (3.2.1.10) pertinent to their source problems:

– the volume-time source function, the eq. (A.1.3),
– the surface-time source function, the eq. (B.2.3), make possible the formulation of the complete form of the field vector velocity initial and boundary conditions distribution:

$$v(Z,R,t)=\int_0^t\iiint_{\Omega_R} G_{Ph}^{v}(Z,t;Q',\tau)\left[v_0\pm\int_0^{t\to\tau}\mathbf{M}(t)\,dt\right]d\xi'\,d\eta'\,d\zeta'\,d\tau'+$$

$$\int_0^t\oiint_{F_R} G_{Ph}^{vS}(R,t;Q',\tau)\left[v_{bo}\mp\int_0^{t\to\tau}\frac{\mathbf{W}(t)}{C^b(t)}dt\right]d\xi'\,d\eta'\,d\zeta'\,d\tau \qquad (3.2.1.11)$$

with additionally

$$\frac{DC}{Dt}=\mp C\,\mathrm{div}\,v$$

$$C\frac{DT}{Dt}=\mp CT\,\mathrm{div}\,v. \qquad (3.2.1.13)$$

4. The complete solution of the constitutive distributed parameter model by the existence of the initial and boundary conditions

The content of the chapter B.3 has been taken to build the complete solution of the whole constitutive distributed parameter model based on the state vectors of the concentration, thermic energy and momentum coordinates of the volume and boundary source functions and the solutions of the phenomenal partial differential equations of the potential and rotational fields. This concept has its formal expression:

Complete form of solution | Volume source functions in point Z(x,y,z,t)

$$\begin{bmatrix}\text{Concentration}\\ \text{Temperature}\\ \text{Field vector velocity}\end{bmatrix}=\begin{bmatrix}\text{Concentration changes (formula (A.1.1))}\\ \text{Temperature changes (formula (A.1.2))}\\ \text{Field vector velocity changes (formula (A.1.3))}\end{bmatrix}+$$

Boundary source functions in point R(x,y,z,t)

$$+\begin{bmatrix}\text{Concentration changes (formula (B.2.1))}\\ \text{Temperature changes (formula (B.2.2))}\\ \text{Field vector velocity changes (formula (B.2.3))}\end{bmatrix}+$$

Difusion
$Z \leftarrow Q \rightarrow R$

Heat transfer
$Z \leftarrow Q \rightarrow R$

$$+\left[\begin{array}{c}\text{Concentration changes}\\ \text{(formula B.3.1.1.8)}\\ \text{Temperature changes}\\ \text{(formula B.3.1.2.1)}\\ *\end{array}\right]+\left[\begin{array}{c}*\\ \text{Temperature changes}\\ \text{(formula B.3.1.3.10)}\\ *\end{array}\right]+$$

Field vector velocity
$Z \leftarrow Q' \rightarrow R$

$$+\left[\begin{array}{l}\text{Concentration continuity}\\ \text{(formula B.3.2.1.12)}\\ \text{Temperature continuity}\\ \text{(formula B.3.2.1.13)}\\ \text{Field vector velocity changes}\\ \text{(formula B.3.2.1.11)}\end{array}\right]$$

* — the places without relations between physical phenomena.

6. THE ANALYSIS OF THE ANALYTICAL SOLUTIONS OF THE CONSTITUTIVE DISTRIBUTED PARAMETER MODEL

Due to the chapters A and B we have the solution of the initial conditions — performed by the formula (A.4.1) and the solution of the initial and boundary conditions having its performance by the eq. (B.4.1). Both of them fulfil the rule of constitutive invariance based on the existence of the solution state vectors in:

$$[L^{cp}] \in \left[\begin{array}{l}\text{Every point}\\ Z(x,y,z,t)\\ \in \overline{\Omega}_{(a,b,c)t}\end{array}\right] \longleftrightarrow \left[\begin{array}{l}\text{Every point}\\ R(x,y,z)t\\ \in F_{Rt}\end{array}\right] \ni [L^{bs}]$$

with $F_{(a,b,c)t}$, $\mathbf{n}^{(z)}$ — with Ω_{Rt}, $\mathbf{n}_R^{(z)}$

$$\left[\begin{array}{l}\text{Common physical}\\ \text{phenomena of potential}\\ \text{and rotational field in}\\ \Omega_{Rt}\end{array}\right] \in [L^{Ph}]\,. \tag{6.1}$$

The state vectors of the formula (5.1) have all identical mass/charge, thermic energy and momentum coordinates if the physical phenomena of adequate influences exist — see diffusion and heat transfer influences. The problems of state vectors can be shown as follows:

— the source state vector $[L^{cp}]$ acting in every point $Z(x,y,z) \in \overline{\Omega}_{(a,b,c)t}$ is related to the physical phenomena of the potential and rotational fields state vector $[L^{Ph}]$ by the

circulation of the normal outside surface orientation vector $\mathbf{n}^{(z)}$ being decisive for the summations of the effects of the physical phenomena of the potential and rotational fields in adequate mass/charge, energy and momentum balances,

– the boundary source state vector $[L^{bs}]$ existing for every point $R(x,y,z,t) \in \mathbf{F}_{Rt}$ has connection to the physical phenomena of the potential and rotational fields $[L^{Ph}]$ according to the circulation of the normal outside surface orientation vector $n_{Rt}^{(z)}$ having underlying significance for the summations of the effects of the physical phenomena of the potential and rotational fields pertinent to the mass/charge, energy and momentum balances,

– the physical phenomena state vector $[L^{Ph}]$ pertinent to the summations of the effects of the physical phenomena of the potential and rotational fields in adequate mass/charge, energy and momentum balances is a consequence of the activity of the source state vector $[L^{sp}]$ or can be extorted by the boundary source state vector $[L^{bs}]$.

The phenomenal distributed parameter control problems for the constitutive distributed parameter model of the real process contains two different aspects:

I. – the working space volume-time influences by the use of the initial conditions, called $[L^{sp}]$,

II. – the surface-time of Ω_{Rt} influences by the application of the boundary conditions $[L^{bs}]$ or their initial conditions $[L_0^{bs}]$ existing on the brim $\mathbf{F}_{Rt}$ for the time $t=0$.

With the assumption that the initial conditions on the surface-time $\mathbf{F}_{Rt}$ are identical to those existing inside working space volume-time Ω_{Rt} – see mentioned system formula $[I^{bi}]$ the above explained influences I°° and II°° remain valid.

There exists a possibility to generate by the circulation of $\mathbf{n}_R^{(z)}$, so signed potential and rotational fields that:

– for the characteristics defined by the initial conditions

$$[DI] = [A_s] + [A_{pot}] + [A_{rot}] , \tag{6.2}$$

where:

$[A_s]$ – the state vector of the source functions,

$[A_{pot}]$ – the state vector of the phenomenal solutions of the potential fields,

$[A_{rot}]$ – the state vector of the phenomenal solutions of the rotational field,

and consequently,

– for the characteristics determined by the initial and boundary conditions,

$$[DIB] = [DI] + [B_s] + [B_{pot}] + [B_{rot}] \tag{6.3}$$

where:

$[B_s]$ – the state vector of the boundary source functions in $R(x,y,z,t)$

$[B_{pot}]$ – the state vector of the boundary phenomenal solutions of the potential fields only for the point $R(x,y,z,t)$,

$[B_{rot}]$ – the state vector of the boundary phenomenal solutions of the rotational field only for the point R(x,y,z,t).

The index of controlability for Ω_{Rt} physical problems is constructed by the boundary conditions approach which is given by the rule:

$$[DS]=[DIB]-[DI] \tag{6.4}$$

The formula (5.3) has its interpretation pertinent to the cases defined as follows:

$$A \pm B \quad \text{or} \quad A \underset{②}{\overset{①}{\rightleftarrows}} B \tag{6.5}$$

with "+" or "–" signs and: ① – generation of the physical phenomena, ② – the initial conditions.

The controlability problem defined by the use of the boundary conditions is formulated as follows [13–17], [2–6]:

$$[DS]=[B_s]+[B_{pot}]+[B_{rot}] \; . \tag{6.6}$$

The most important feature for the formula (5.5) is that the control problems are realized by the use of the dimensions "a,b,c" pertinent to the dimensions of the locally selected volume-time element $\overline{\Omega}_{(a,b,c)t}$ because the amplitude of the phenomenal solutions are related to the all of them. This approach is a result of the Assumption 1 that the norm of the eigenfunctions is: $N_e=1$.

On the basis of the formulae (5.3), (5.4) and (5.5) we can propose some control ideas by the use of the boundary conditions:

– the classic control theory – with the space and time elements,
– the hierarchy control theory – for example: optimal control, adaptive control different criterions – with space and time parts.

The constitutive interpretation of the formulae (6.4), (6.5) and (6.6) assures:

– validity of the boundary control problems for the point $Z(x,y,z,t) \in \overline{\Omega}_{(a,b,c)t} \in \Omega_{Rt}$ by the source generation of the physical phenomena in the point $R(x,y,z,t) \in F_{Rt} \in \Omega_{Rt}$,
– invariance of the constitutive distributed parameter model to its physical coefficients inside Ω_{Rt} and on the brim F_{Rt}, pertinent to the systems of mass/charge, energy and momentum partial differential equations of the continuity inside Ω_{Rt} and on its brim F_{Rt}.

7. CONCLUSIONS

The article represents an attempt to the formalization of the constitutive theory of the solution of the partial differential constitutive state equation although the considerations may be extended for the system of partial differential constitutive state equations. All characteristic constitutive steps of the solution method of partial differential constitutive state equations such as:

– source decomposition,
– homogeneous phenomenal decomposition,
– the splice of the phenomenal Green function with the source function pertinent to the physical phenomenon,
– state vector of mass/charge, energy and momentum coordinates of the summation of the effects of the influences of the potential and rotational fields, have been explained in detailed form.

In the article influences of the phenomenal links on the mass/charge, energy and momentum complete solution coordinates have been discussed. As an example for the presentation of the general solution method concentration, thermic energy and momentum coordinates state vector solution for the constitutive distributed parameter model deduced in Part I of this article [1] has been considered. This solution presents working space-time distributed parameter dynamics of the state variables from the constitutive distributed parameter dynamics and possesses the following properties:
– the high, related to the physical coefficients precision of the information,
– the interpretation of the information by the use of the physical phenomena,
– the complete information about the state variables inside Ω_{Rt} and on its brim F_{Rt},
– possibility of the realization of the yield and quality different kinds control of the products of the considered processes by the use of their single physical phenomena generated from the boundary conditions.

The important feature of the considered phenomenal solutions of the potential fields is that all of them are related to the dimensions "a,b,c" of $\overline{\Omega}_{(a,b,c)t}$. These dimensions are adjustable to the kind of the kinetics of the analyzed reaction or processing of the interested media. Although in this article we consider the continuous media having source space and time memories like: – electronic lamps and other electronic apparatuses, – electrochemical processes and others, all based on electrons and ions transport WITHOUT ELECTRICAL PARAMETERS, as an example of the constitutive way of the constitutive distributed parameter modelling, this approach can be applied to many other important real processes. The solutions of the constitutive distributed parameters obtained in the article have significance for:
A. **existence of the initial conditions** – by the identification of the considered processes and optimization of the working space volume-time Ω_{Rt},
B. **existence of the initial and boundary conditions** – for the phenomenally distributed parameter control by the use of the single physical phenomena generated from the existence and activity of the boundary conditions on F_{Rt}.

In the article an original state vector controllability $[DS]$ according to the formulae (5.3) and (5.5) has been introduced. The state vector of the controllability contains boundary phenomenal distributed parameter control influences on the yield and quality aspects of the products of the considered processes. As a consequence of the presented features of the constitutive distributed parameter solution of the real processes the area of the application of this approach should be very wide. This

opinion comes from the analysis of the discussions appearing during the International Federation of Automatic Control Meetings where Authors present in their publications [14] idea called "A Bridge Between Control Science and Technology" as a necessary tool to make the real processes distributed parameter control more and more effective. The analysis of the considerations of this article enables us to formulate the opinion that the considerations of this article fulfil all requirements in this area.

NOTATION

$\overline{\Omega}_{(a,b,c)t}$ – the locally selected volume-time element (cuboidal form) around the processing point Z(x,y,z,t), and x=a, y=b, z=c – constant coordinates m^3-s

$F_{(x,y,z)t}$ – the outside oriented surface for $\overline{\Omega}_{(a,b,c)t}$ m^3-s

$\mathbf{n}^{(z)}$ – the normal outside surface orientation vector for $F_{(x,y,z)t}$ which circulation has underlying significance for the summation of effects of potential and rotational fields in adequate balances,

Ω_{Rt} – working space voloume-time for the real processes m^3-s

F_{Rt} – the outside oriented surface for Ω_{Rt} to which belongs the boundary conditions point $R(x,y,z,t)$ m^2-s

$\mathbf{n}_R^{(z)}$ – the normal outside surface orientation vector for F_{Rt} which circulation is in relation to $\mathbf{n}^{(z)}$ and the boundary control task by the phenomenal boundary conditions,

C – the concentration of the singlecomponent processing medium $\frac{kg}{m^3}$

T – the temperature of the singlecomponent processing medium °K

$\mathbf{v}$ – the field vector velocity of the singlecomponent processing medium $\frac{m}{s}$

D_C – the diffusion coefficient for the singlecomponent processing medium $\frac{m^2}{s}$

H_C – the own enthalpy for the concentration "C" $\frac{J}{kg}$

λ_C – the heat transfer coefficient for the singlecomponent processing medium $\frac{W}{m°K}$

c_p – the specific heat of the single component processing medium $\frac{J}{kg°K}$

η_w – the coefficient of the dynamic viscosity $\frac{N_s}{m^2}$

V_B – the intensity of the generation of the mass/charge $\frac{mol}{s}$

G – the own molar mass for the mass/charge generation $\frac{kg}{mol\,m^3}$

H_B – the specific enthalpy for the intensity of mass/charge generation $\frac{J}{kg}$

$M(t)$ – the force generation function of mass/charge $\frac{N}{m^3}$

REFERENCES

1. W. Niemiec: *Part I of this article*
2. W. Niemiec: *On the constitutive theory of modelling and information for distributed parameter control of the continuous mass crystallization process.* Part II: *Complete information for phenomenal distributed parameter control of continuous mass crystallization process.* Third Int'l Conference on Liquid Metal Engineering and Technology in Energy Production, Oxford England, 9–13 April, 1984, Paper No. 192
3. W. Niemiec: *On modelling and information for control of distributed parameter chemical processes in fluid phase.* 3 IFAC Symposium "Control of Distributed Parameter Systems", Toulouse, France, 29.VI–2.VII, 1982, Session 23
4. W. Niemiec: *A mathematical model of distributed parameters of the continuous mass crystallization process for the adaptive control.* PhD Thesis. Silesian Technical University 1979
5. W. Niemiec: *The constitutive theory of modelling and information for phenomenal distributed parameter control of multicomponent chemical processes in gas, fluid and solid phase.* Part II. *The complete information for phenomenal distributed parameter control of multicomponent chemical processes in gas, fluid and solid phase.* 7th Miami Int'l Conference on Alternative Energy Sources, Miami Beach Florida, USA, 9–11 December, 1985, Session "Hydrocarbons/Energy transfer", pp. 589–598
6. W. Niemiec: *A mathematical model of the distributed parameters of the continuous mass crystallization process for the adaptive control.* Part II: *The complete information for the phenomenal distributed parameter control and adaptive features of the continuous mass crystallization process.* Poznańskie Towarzystwo Przyjaciół Nauk, Wydział Nauk Technicznych, Prace Komisji Automatyki i Informatyki, Tom XV – 1989, pp. 119–162
7. W. Nowacki: *Dynamiczne zagadnienia termosprężystości.* PWN, Warszawa 1966
8. W. Nowacki, Z. Olesiak: *Termodyfuzja w ciałach stałych.* PWN, Warszawa 1991
9. T. Trajdos: *Matematyka dla inżynierów.* PWN, Warszawa 1974
10. B. Średniawa: *Hydrodynamika i teoria sprężystości.* PWN, Warszawa 1974
11. A. Tichonow, A. Samarski: *Równania fizyki matematycznej.* PWN, Warszawa 1966
12. W. Kupradze: *Wybrane zagadnienia teorii sprężystości i termosprężystości.* Wyd. PAN, Wrocław–Warszawa–Kraków 1970
13. S. Węgrzyn: *Podstawy automatyki.* PWN, Warszawa 1974
14. *IFAC Newsletters* 1980–1992
15. W. Niemiec: *Zagadnienia sterowania adaptacyjnego procesów przemysłowych.* Zeszyty Naukowe Politechniki Śląskiej, "Automatyka",1981, z. 50
16. W. Niemiec: *Struktura sterowania adaptacyjnego procesów przemysłowych.* Zeszyty Naukowe Politechniki Śląskiej "Automatyka", 1981, z. 51
17. W. Niemiec: *Stabilność sterowania adaptacyjnego układów z modelem odniesienia.* Sem. "Metody badania systemów technicznych". Cracow Technical University, 1981

APPENDICES

APPENDIX 1. THE STATICAL ISOTROPIC GREEN TENSOR FOR THE VOLUME ELEMENT $\overline{\Omega}_{(a,b,c)}$ [12], [2–6]

For the tensor of the statical deformations of the volume element $\overline{\Omega}_{(a,b,c)}$

$$\varepsilon_{iklm}=\alpha\delta_{ik}\delta_{lm}+\beta(\delta_{il}\delta_{km}+\delta_{im}\delta_{kl}) \tag{1}$$

the statical isotropic Green tensor has the general form

$$\Gamma_G(Z,Q')=\frac{1}{8\Pi}\left[\frac{1}{\alpha}\Gamma_\alpha(Z,Q')+\frac{1}{\beta}\Gamma_\beta(Z,Q')\right]. \tag{2}$$

The component $\Gamma_\alpha(Z,Q')$ is connected to the operation [grad divv] and is written by the formula

$$\Gamma_\alpha(Z,Q')=\frac{1}{|\mathbf{r}|}I-\frac{\mathbf{r}\times\mathbf{r}}{|\mathbf{r}|^3} \tag{3}$$

where: I – unitary matrix, $\mathbf{r}=\mathbf{r}_Z-\mathbf{r}_{Q'}$.

Consequently the component $\Gamma_\beta(Z,Q')$ belongs to the operation [–rot rotv]

$$\Gamma_\beta(Z,Q')=\frac{1}{|\mathbf{r}|}I+\frac{\mathbf{r}\times\mathbf{r}}{|\mathbf{r}|^3}. \tag{4}$$

The coefficients α and β are defined by the formulae

$$\alpha=a+b \text{ and } \beta=a \text{ with } a=\eta_W \text{ and } b=\frac{1}{3}\eta_W \tag{5}$$

and the above presented information makes possible to state the isotropic Green tensor for the volume element $\overline{\Omega}_{(a,b,c)}$

$$\Gamma_G(Z,Q')=\frac{1}{32\Pi\eta_W}\left[7\frac{1}{|\mathbf{r}|}I+\frac{\mathbf{r}\times\mathbf{r}}{|\mathbf{r}|^3}\right] \tag{6}$$

where: I – unitary matrix, η_w – coefficient of the dynamic viscosity, $\mathbf{r}=\mathbf{r}_Z-\mathbf{r}_{Q'}$ – radius of the activity of the viscosity forces.

APPENDIX 2. THE STATICAL ANISOTROPIC GREEN TENSOR FOR THE VOLUME ELEMENT $\overline{\Omega}_{(a,b,c)}$ [12], [2–6]

We define the material coefficients of the medium as:

$$\alpha=a+b \text{ and } \beta=a. \tag{1}$$

The elasticity tensor of the deformations of the volume element $\overline{\Omega}_{(a,b,c)}$ is written in the form

$$\varepsilon_{iklm}\nabla_k\nabla_m V_i(Z)=-N_{Vi}(Z) \tag{2}$$

and ε_{iklm} – the elasticity tensor of the volume element $\overline{\Omega}_{(a,b,c)}$, $V_l(Z)$ – the field of the deformations of the volume element $\overline{\Omega}_{(a,b,c)}$, $N_{Vi}(Z)$ – the field of the forces of the deformations of the volume element $\overline{\Omega}_{(a,b,c)}$. Now let us introduce the following operation

$$\varepsilon_{iklm}\nabla_k\nabla_m = \hat{\varepsilon}_{il} \tag{3}$$

which modifies the eq. (2) to its another form

$$\hat{\varepsilon}_{il}V_l(Z) = -N_{Vi}(Z)\,. \tag{4}$$

The field of the deformations $V_l(Z)$ of the locally selected volume element $\overline{\Omega}_{(a,b,c)}$ is:

$$V_l(Z) = \iiint_{\Omega_R} \Gamma_{ln}(Z-Q')N_{Vn}(Q')d\Omega_R(Q') \tag{5}$$

where $\Gamma_{ln}(Z-Q')$ – the statical Green tensor of the volume element $\overline{\Omega}_{(a,b,c)}$. We can write the eq. (4) in the form

$$\hat{\varepsilon}_{il}\Gamma_{ln} = -\delta_{in}\delta_{\overline{\Omega}}(Z) \tag{6}$$

and the eq. (5) can be rewritten in the splice form

$$V_l = \Gamma_{ln} * N_{Vn}\,. \tag{7}$$

Now we introduce the tensor series form of

$$\varepsilon_{iklm} = \sum_{r=0}^{\infty} \varphi^r \underset{r}{\varepsilon}_{iklm} \tag{8}$$

and

$$\underset{0}{\varepsilon}_{iklm} = \alpha\delta_{ik}\delta_{lm} + \beta(\delta_{il}\delta_{km} + \delta_{im}\delta_{kl}) \tag{9}$$

with consequently the series form of the statical Green tensor

$$\Gamma_{ln} = \sum_{r=0}^{\infty} \varphi^r \underset{r}{\Gamma}_{ln}\,. \tag{10}$$

Making use of the eq. (8) and the eq. (3) with the eq. (10), the eq. (6) obtains its final form

$$\left[\sum_{r=0}^{\infty} \varphi^r \underset{r}{\hat{\varepsilon}}_{il}\right]\left[\sum_{r=0}^{\infty} \varphi^r \underset{r}{\Gamma}_{ln}\right] = -\delta_{in}\delta_{\overline{\Omega}}(Z) \tag{11}$$

After development of the series (8) and (10) in the eq. (11), comparing the parts being in the identical power of small parameter φ, the anisotropic Green tensor can be written as:

$$\Gamma_{GA}(Z,Q') = \Gamma_{ln}(Z,Q') = \underset{0}{\Gamma}_{lp}(Z,Q') * \underset{1}{\hat{\varepsilon}}_{pq}(Z,Q')\,\underset{k-1}{\Gamma}_{qn}(Z,Q') \tag{12}$$

for $k = 1(1)w$ and $\underset{0}{\Gamma}_{lp}(Z,Q') = \Gamma_G(Z,Q')$ where w – number of power of the small parameter.

W. NIEMIEC

O KONSTYTUTYWNYM ROZŁOŻONYM PARAMETRYCZNIE MODELOWANIU JEDNOSKŁADNIKOWYCH RZECZYWISTYCH PROCESÓW CZĘŚĆ II. ROZWIĄZANIE RÓWNAŃ RÓŻNICZKOWYCH CZĄSTKOWYCH KONSTYTUTYWNYCH STANU DLA JEDNOSKŁADNIKOWYCH RZECZYWISTYCH PROCESÓW

Streszczenie

Artykuł jest poświęcony prezentacji konstytutywnego podejścia do rozwiązania ogólnej postaci równania różniczkowego cząstkowego konstytutywnego stanu wyprowadzonego w CZĘŚĆ I tego artykułu. Ten ogólny sposób konstytutywnego rozwiązania został następnie użyty do rozwiązania wyprowadzonego w CZĘŚĆ I układu równań różniczkowych cząstkowych konstytutywnych stanu dla masy/ładunku, energii termicznej i pędu jako przykładu. Rozważono dwa przypadki tego przykładowego rozwiązania przy:

A. – istnieniu warunków początkowych,

B. – istnieniu warunków początkowych i brzegowych,

dla wszystkich jego zjawisk fizykalnych. Zdefiniowano, wprowadzono do rozważań i przedyskutowano użycie brzegowego wskaźnika sterowalności na bazie postaci wektora stanu o współrzędnych masa/ładunek, energia termiczna i pęd.

KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJI, 1993, 39, z. 1, ss. 187-198

621.3.087.92

Pojemność wzajemna dwóch przetworników międzypalczastych w obudowie metalowej

EUGENIUSZ DANICKI, DARIUSZ BOGUCKI

Instytut Podstawowych Problemów Techniki PAN, Warszawa

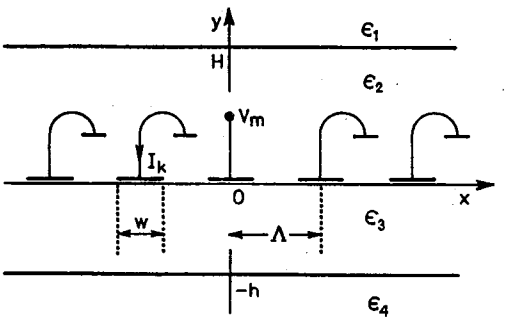
Otrzymano 1992.07.01

Autoryzowano do druku 1992.08.28

Rozważany jest periodyczny układ idealnie przewodzących elektrod umieszczonych w warstwowym ośrodku dielektrycznym. Analizowany jest problem przenoszenia sygnału drogą pojemnościową pomiędzy przetwornikami międzypalczastymi w układach z akustyczną falą powierzchniową. Wyniki mogą też być stosowane w analizie pojemnościowego czujnika odległości bazującego na konstrukcji kondensatora międzypalczastego.

1. WPROWADZENIE

Przechodzenie sygnału drogą elektromagnetyczną (pojemnościową) między dwoma przetwornikami międzypalczastymi w urządzeniach z akustyczną falą powierzchniową (AFP) jest źródłem sygnału fałszywego, który może być znaczny w filtrach AFP o małym opóźnieniu [1]. W artykule analizowane jest modelowe zagadnienie nieskończonego, okresowego układu idealnie przewodzących elektrod umieszczonych w warstwowym ośrodku dielektrycznym jak na rysunku 1. Okres elektrod



Rys. 1. Rozważany układ elektrod w warstwowej strukturze dielektrycznej

jest Λ, szerokość każdej z nich równa jest $\sqrt{\omega}$. Paski umieszczone są w płaszczyźnie $y=0$ (pomiędzy warstwami o współczynnikach przenikalności dielektrycznej ε_2 i ε_3), równolegle do osi z.

Niech elektroda m będzie podłączona do źródła o potencjale V_m a pozostałe niech będą zwarte do masy (Rys. 1). Prąd dopływający do elektrody k jest

$$i_k = y_{km} V_m \tag{1}$$

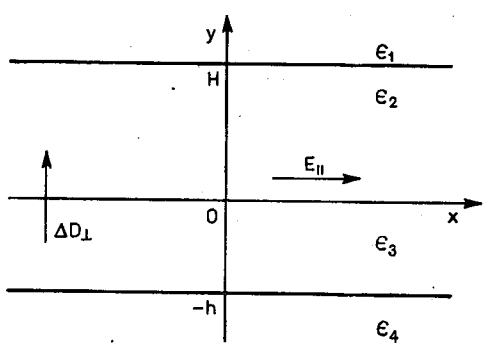
gdzie y_{km} jest transadmitancją elektrod k i m. W przypadku ośrodków półnieskończonych, (niewarstwowych) y_{km} wyznaczono w [2]. W artykule została wykorzystana ta sama metoda ale uogólniono ją na przypadek dielektryka warstwowego, ograniczonego przez półpłaszczyzny metalowe, modelujące ścianki obudowy metalowej filtru z AFP.

Przedstawione w niniejszym artykule rezultaty mogą być użyteczne w projektowaniu filtrów z AFP w aspekcie minimalizacji pojemnościowego sygnału fałszywego (transmisja sygnału „elektromagnetycznego" między przetwornikiem wejściowym a wyjściowym). Prezentowane wyniki mogą być też wykorzystane przy konstrukcji pojemnościowego czujnika odległości zbudowanego na bazie kondensatora międzypalczastego o analogicznej konstrukcji do przetwornika międzypalczastego ale wykonanego na podłożu niepiezoelektrycznym [4]. W czujniku takim wykorzystuje się zależność współczynników y_{km} od odległości i przenikalności dielektrycznej zbliżanego do elektrod płaskiego przedmiotu.

W następnym punkcie rozważana jest tzw. efektywna przenikalność dielektryczna dla podłoża warstwowego, charakteryzująca to podłoże pod względem zależności między potencjałem a ladunkiem na powierzchni podłoża. W kolejnym rozdziale krótko przedstawiona jest metoda analizy prezentowanego zagadnienia (może ona być też stosowana przy analizie kondensatora międzypalczastego).

2. EFEKTYWNA PRZENIKALNOŚĆ DIELEKTRYCZNA

Rozważmy dielektryczną strukturę warstwową przedstawioną na Rys. 2. Analogicznie jak uczyniono to w [3], [2], [5] można elektryczne własności badanej struktury odnieść do płaszczyzny $y=0$ korzystając z pojęcia tzw. efektywnej przeni-



Rys. 2. Warstwowa struktura dielektryczna

kalności dielektrycznej zdefiniowanej dla amplitud zespolonych $\exp(-jkx+j\omega t)$ gdzie k – liczba falowa, ω – częstość kołowa, na powierzchni $y=0$ następująco

$$\varepsilon_f(k)=jS_k\frac{\Delta D_\perp}{E_\parallel}\bigg|_{y=0} \tag{2}$$

gdzie $E_\parallel$ jest składową pola elektrycznego styczną do płaszczyzny $y=0$, zaś $\Delta D_\perp=D_y(y=0^+)-D_y(y=0^-)$ jest skokiem indukcji elektrycznej równej gęstości ładunku na płaszczyźnie $y=0$. Funkcja S_k wprowadzona dla wygody, jest zdefiniowana następująco

$$S^k = \begin{cases} +1 \text{ dla } k\geqslant 0 \\ -1 \text{ dla } k<0 \end{cases},$$

Zakładając powyższą harmoniczną postać rozwiązania na pole elektryczne dla $y=0$, potencjał w każdej z warstw może być zapisany w postaci

$$\phi = (\Phi' e^{+|k|y}+\Phi'' e^{-|k|y})\, e^{-jkx}\, e^{j\omega t}. \tag{3}$$

Przy czym stała $\Phi'=0$ w warstwie ε_1 (dla $y>H$), zaś $\Phi''=0$ dla $y<-h$ tj. w warstwie ε_4 (wynika to z konieczności znikania potencjału odpowiednio w $+\infty$ i $-\infty$).

Ponieważ $E_x=-\partial\phi/\partial x$ i $D_y=-\varepsilon_0\varepsilon_i\partial\phi/\partial y$, w i-tej warstwie o przenikalności dielektrycznej ε_i to warunki brzegowe na granicach między warstwami mają postać

$$\begin{aligned}
&- \text{ dla } y \quad H \qquad D_y(y=H^+)-D_y(y=H^-) = 0\\
&\qquad\qquad\quad E_x(y=H^+)=E_x(y=H^-)\\
&- \text{ dla } y=0 \qquad D_y(y=0^+)-D_y(y=0^-)=\Delta D_\perp\\
&\qquad\qquad\quad E_x(y=H^+)=E_x(y=H^-)=E_\parallel\\
&- \text{ dla } y=-h \quad D_y(y=-h^+)-D_y(y=-h^-)=0\\
&\qquad\qquad\quad E_x(y=-h^+)=E_x(y=-h^-)\,.
\end{aligned}$$

Ostatecznie rozwiązując wynikające z tych równań warunki na Φ' i Φ'' dostajemy poszukiwaną przenikalność dielektryczną efektywną w postaci

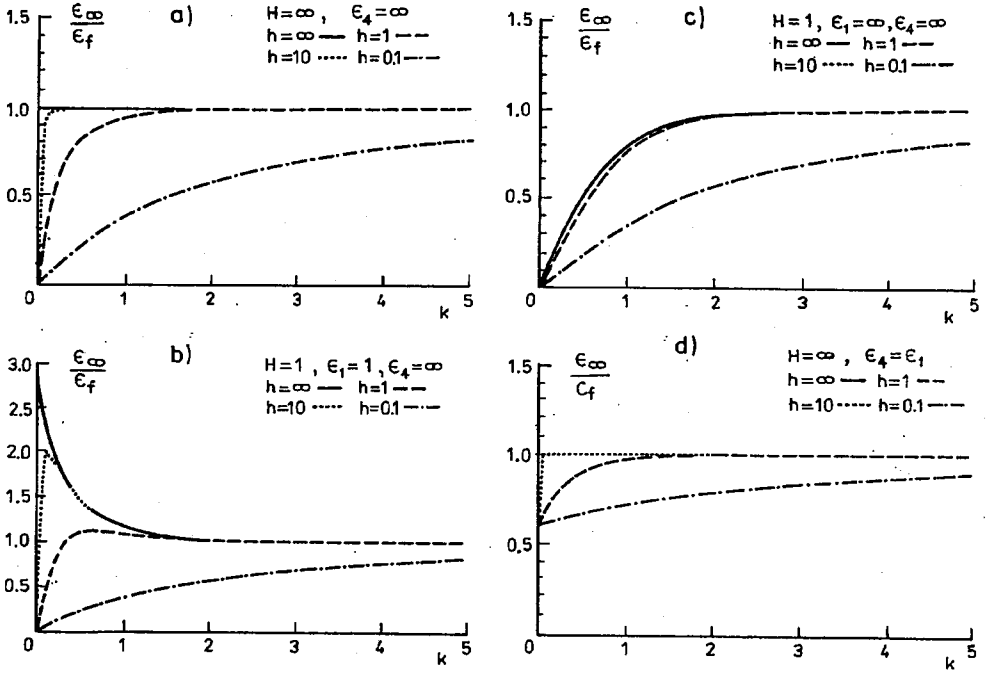
$$\varepsilon_f(k)=\varepsilon_0\cdot\left\{\varepsilon_2\frac{\varepsilon_1+\varepsilon_2\tanh|k|H}{\varepsilon_2+\varepsilon_1\tanh|k|H}+\varepsilon_3\frac{\varepsilon_4+\varepsilon_3\tanh|k|h}{\varepsilon_3+\varepsilon_4\tanh|k|h}\right\}. \tag{4}$$

Istotny dla dalszej analizy jest fakt, że dla dużych wartości k zachodzi

$$\varepsilon_f(k) \xrightarrow{k\to\infty} \varepsilon_\infty = \varepsilon_0(\varepsilon_2+\varepsilon_3). \tag{5}$$

Rysunek 3 przedstawia przebieg $\varepsilon_\infty/\varepsilon_f(k)$ dla różnych wartości H, h i przenikalności dielektrycznych poszczególnych warstw (funkcja ta będzie wykorzystywana do konstrukcji rozwiązania w dalszej części artykułu). Zakładając $\varepsilon_2=4.5$ (szkło kwarcowe) i $\varepsilon_3=1$ rozważano przypadki

a) $H=\infty$ (półprzestrzeń), $\varepsilon_4=\infty$ (metal idealny)
b) $H=1$, $\varepsilon_4=\infty$, $\varepsilon_1=1$
c) $H=1$, $\varepsilon_4=\infty$, $\varepsilon_1=\infty$
d) $H=\infty$, $\varepsilon_4=\infty$, $\varepsilon_4=\varepsilon_1$.



Rys. 3. Przebieg funkcji $\varepsilon_\infty/\varepsilon_f(k)$ dla różnych wariantów struktury dielektrycznej

3. ANALIZA PERIODYCZNEGO UKŁADU ELEKTROD

W analizowanej strukturze (Rys. 1) na płaszczyźnie $y=0$ mamy

$$\begin{aligned} E_{\parallel} &= 0 \quad \text{na paskach metalowych} \\ \Delta D_{\perp} &= 0 \quad \text{pomiędzy paskami.} \end{aligned} \tag{6}$$

Są to mieszane warunki brzegowe na poszukiwane pole elektryczne między elektrodami i gęstość ładunku ($\Delta D_\perp$) na elektrodach. Zgodnie z twierdzeniem Floqueta rozwiązania poszukuje się w postaci [3], [2]

$$E_{\parallel} = \sum_{n=-\infty}^{\infty} E_n e^{-j(r+nK)} e^{j\omega t} \tag{7}$$

$$\Delta D_{\perp} = \sum_{n=-\infty}^{\infty} D_n e^{-j(r+nK)} e^{j\omega t} \tag{8}$$

gdzie $K=2\pi/\Lambda$, $r\in(0;K)$ (czynnik $\exp(j\omega t)$ będzie pomijany w dalszej części artykułu). Współczynniki D_n powiązane są z E_n poprzez efektywną przenikalność dielektryczną, wyprowadzoną w poprzednim punkcie:

$$D_n = -jS_k \varepsilon_n E_n \tag{9}$$

przy czym w miejsce k należy podstawić $r+n\mathrm{K}$ oraz

$$\varepsilon_n = \varepsilon_f(r+n\mathrm{K}). \tag{10}$$

Rozwiązania na E_n i D_n można zapisać w następującej postaci wykorzystującej pewne własności funkcji Legendre'a [3]

$$\mathrm{E}_n = \sum_m \alpha_m \mathrm{S}_{n-m} \mathrm{P}_{n-m}(\cos\Delta) \tag{11}$$

$$\mathrm{D}_n = \sum_m \beta_m \mathrm{P}_{n-m}(\cos\Delta), \tag{12}$$

gdzie $\Delta = \pi w/\Lambda$, α_m i β_m, są dowolne. Sumowanie po m przebiega w pewnych dużych lecz skończonych granicach zależnie od zakładanej dokładności.

Ponieważ dla dużych k mamy $\varepsilon_f(k) \to \varepsilon_\infty$, możemy znaleźć takie liczby całkowite N_1 i N_2, że tylko dla $n \in [-\mathrm{N}_1, \mathrm{N}_2]$ jest

$$\varepsilon_n = \varepsilon_f(r+n\mathrm{K}) \neq \varepsilon_\infty \tag{13}$$

podczas gdy dla $n \leqslant -\mathrm{N}_1 - 1$ lub $n \geqslant \mathrm{N}_2 + 1$ można położyć $\varepsilon_n \approx \varepsilon_\infty$. W rezultacie z porównania (11) i (12) dla dużych n otrzymujemy

$$\beta_m = \varepsilon_\infty \alpha_m .$$

Dalsza analiza [3] pokazuje, że sumowanie po m należy przeprowadzać w granicach $m \in [\mathrm{N}_1, \mathrm{N}_2 + 1]$.

Relacje (11), (12) i (13) dają następujący układ równań na niewiadome α_m

$$\alpha_m \left(\mathrm{S}_{n-m} - \frac{\varepsilon_\infty}{\varepsilon_n} \mathrm{S}_n \right) \mathrm{P}_{n-m}(\cos\Delta) = 0 \tag{14}$$

z którego można wyznaczyć wszystkie stałe α_m w zależności od jednej z nich np. α_0. Ta z kolei określona jest przez dodatkowe zależności dotyczące potencjałów elektrod i całkowitych prądów elektrod [2].

Wyznaczmy admitancję paska jako stosunek prądu płynącego w elektrodzie do jej potencjału dla danej liczby falowej r.

$$\mathrm{Y}(r) = \frac{\hat{\mathrm{I}}(r)}{\hat{\mathrm{V}}(r)} = j\omega \mathrm{C}(r) \tag{15}$$

przy czym potencjał $\hat{\mathrm{V}}(r)$ i całkowity prąd paska $\hat{\mathrm{I}}(r)$ położonego w $x=0$ wyznaczane są następująco

$$\hat{\mathrm{V}}(r) = \sum_{n=-\infty}^{\infty} \frac{\mathrm{E}_n}{j(r+n\mathrm{K})} \tag{16}$$

$$\hat{\mathrm{I}}(r) = j\omega \int_{w/2}^{w/2} \Delta \mathrm{D}_\perp \mathrm{d}x, \tag{17}$$

postępując analogicznie jak w [3] dostajemy ($v=r/K$)

$$\hat{V}(r) = \frac{-j}{K}\frac{\pi}{\sin \pi v}\alpha_0 \sum_m (-1)^m \frac{\alpha_m}{\alpha_0} P_{m+v-1}(\cos\Delta) \tag{18}$$

$$\hat{I}(r) = 2\pi \frac{\omega}{K}\alpha_0 \varepsilon_\infty \sum_m \frac{\alpha_m}{\alpha_0} P_{m+v-1}(\cos\Delta). \tag{19}$$

Ostatecznie dostajemy $Y(r)=j\omega C(r)$ gdzie

$$C(r)=2\varepsilon_\infty \sin(\pi v)\frac{\sum_m \frac{\alpha_m}{\alpha_0} P_{m+v-1}\cos\Delta}{\sum_m (-1)^m \frac{\alpha_m}{\alpha_0} P_{m+v-1}\cos\Delta}. \tag{20}$$

Potencjał i prąd elektrod (w odniesieniu do jednostki ich długości) są wyznaczane z odwrotnej transformaty Fouriera poprzez całki [2], [5]

$$i_k = \frac{1}{K}\int_0^K \hat{I}(r)e^{jk\Lambda r}\,dr \tag{21}$$

$$v_l = \frac{1}{K}\int_0^K \hat{V}(r)e^{jl\Lambda r}\,dr. \tag{22}$$

Kładąc $v_l=\delta_{lm}V_m$ (zasilana jest tylko elektroda o numerze m, wszystkie pozostałe są uziemione) z równania (22) wyznaczamy nieznane $\alpha_0(r)$, które podstawione do (21) daje szukaną transadmitancję (1)

$$y_{km} = j\omega C_{|k-m|} \tag{23}$$

$$C_{|k-m|} = \frac{1}{K}\int_0^K C(r)e^{j(k-m)\Lambda}\,dr \tag{24}$$

Obliczenie całki (24) jest w ogólnym przypadku możliwe tylko numerycznie. Jednak w przypadku gdy $\varepsilon_f(k)=\text{const}$ tzn. gdy elektrody umieszczone na granicy dwóch półprzestrzeni dielektrycznych, C_n można wyznaczyć ściśle [2], [5]

$$C_n = \frac{-4\varepsilon_\infty}{\pi(4n^2-1)} \tag{25}$$

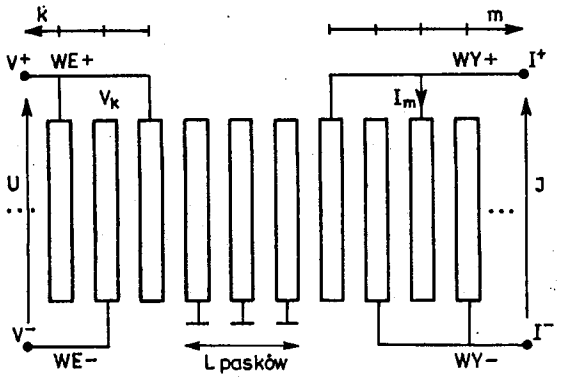
C_n dla $n\neq 0$ jest ujemne, gdyż kierunek prądu płynącego w uziemionych elektrodach jest przeciwny do kierunku prądu elektrody zasilanej.

W prezentowanych przykładach numerycznych współczynniki C_n wyznaczano za pomocą algorytmu FFT.

4. WYNIKI NUMERYCZNE

4.1. POJEMNOŚĆ WZAJEMNA DWÓCH PRZETWORNIKÓW

We wszystkich obliczeniach przyjęto następujące założenia
- $\Delta = 0$ tj. $w = \Lambda/2$
- $\varepsilon_2 = 4.5$ tj. elektrody umieszczone są na podłożu ze szkła kwarcowego
- $\varepsilon_3 = 1$ (próżnia)
- h i H normowano względem Λ: $\bar{h} = h/\Lambda$ i $\bar{H} = H/\Lambda$



Rys. 4. Układ dwóch przetworników międzypalczastych (półnieskończonych)

Rozważmy konfigurację przedstawioną na rysunku 4. Mamy tu dwa półnieskończone przetworniki międzypalczaste rozdzielone L elektrodami uziemionymi. Pojemność wzajemną układu tych przetworników definiujemy jako stosunek $C_X = J/j\omega U$ gdzie
- V jest napięciem przyłożonym do przetwornika nadawczego WE,
- J jest prądem przetwornika odbiorczego WY.

Pojemność wzajemna C_X odpowiada za przechodzenie sygnału fałszywego z wejścia na wyjście. Rozważmy C_X w funkcji L i konfiguracji połączeń obu przetworników.

Przetwornik nadawczy (źródło sygnału fałszywego) może być zasilany w dwojaki sposób:
- w konfiguracji antysymetrycznej: $V^+ = V$ and $V^- = 0$ (napięcie jest przyłożone tylko do wejścia WE+, wejście WE− jest uziemione),
- w konfiguracji symetrycznej: $V^+ = V/2$ and $V^- = -V/2$ (napięcie jest podawane na wejścia w przeciwfazie np. przez transformator symetryzujący).

Anbalogiczne konfiguracje są rozważane w przypadku przetwornika odbiorczego:
- w konfiguracji antysymetrycznej: $J = I^+$ (prąd wyjściowy jest zbierany wyłącznie z wyjścia WY+),

— w konfiguracji symetrycznej: $J=(I^+-I^-)/2$ przy spełnieniu relacji $I^+=-I^-$.

Prądy I^+ i I^- wyznaczane są nastepująco

$$I^+ = \sum_m A_m I_m \tag{26}$$

$$I^- = \sum_m B_m I_m \tag{27}$$

zaś

$$I_m = j\omega \sum_m V_k \cdot C_{m+k+L-1} \tag{28}$$

gdzie

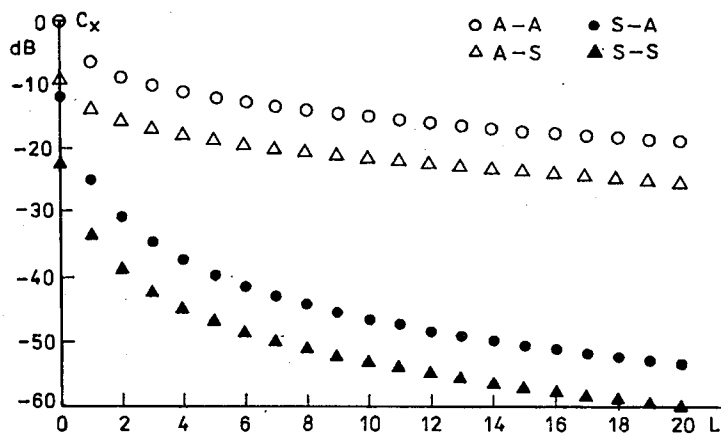
$$A_m = \begin{cases} 1 \text{ dla elektrod połączonych z WY+}, \\ 0 \text{ w przeciwnym przypadku}, \end{cases} \tag{29}$$

$$B_m = \begin{cases} -1 \text{ dla elektrod połączonych z WY−}, \\ 0 \text{ w przeciwnym przypadku}, \end{cases} \tag{30}$$

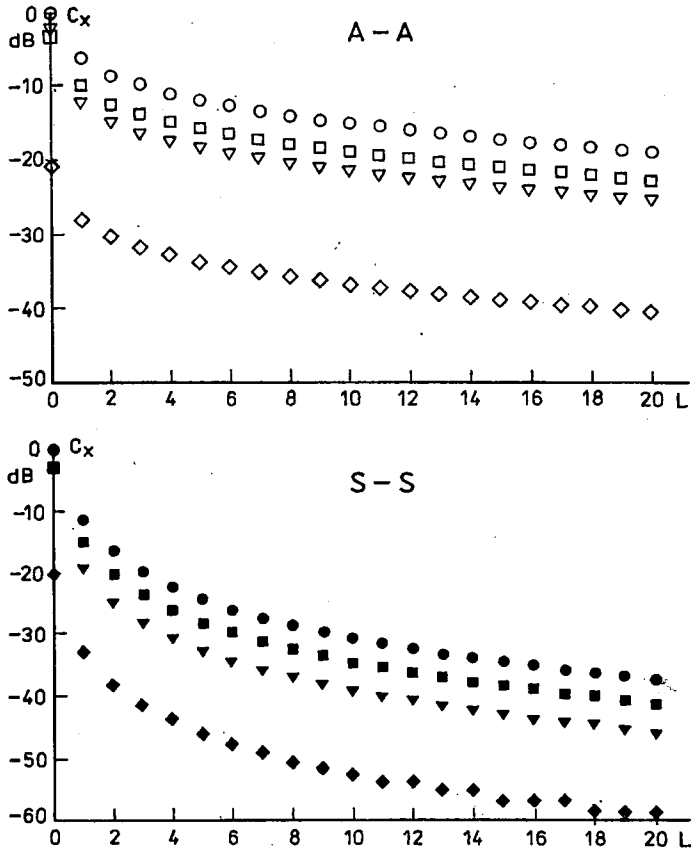
$$V_k = \begin{cases} V^+ \text{ dla elektrod połączonych z WE+}, \\ V^- \text{ dla elektrod połączonych z WE−}. \end{cases} \tag{31}$$

Na Rys. 5. przedstawiono wyniki odniesione do wartości maksymalnej na wykresie w skali decybelowej, dla różnych konfiguracji przetworników (pierwsza litera opisuje konfigurację przetwornika nadawczego, druga — odbiorczego).

Rezultaty nie zależą od układu dielektryków, w którym umieszczone są elektrody: największą pojemność wzajemną mamy w konfiguracji A—A (oba przetworniki są połączone antysymetrycznie), najmniejszą w S—S (oba przetworniki połączone symetrycznie).



Rys. 5. Pojemność wzajemna dwóch przetworników dla różnych konfiguracji ich połączeń (skala dB)



Rys. 6. Porównanie pojemności wzajemnej dla różnych ośrodków a) w konfiguracji A–A, b) w konfiguracji S–S (skala dB)

Rys. 6 daje porównanie pojemności wzajemnej dla konfiguracji A–A i S–S w czterech przypadkach:

– elektrody umieszczono na półprzestrzeni dielektrycznej i $\bar{h}=\infty$ (wyniki przedstawiono kółkami),

– podobnie ale $\bar{h}=1$ i $\varepsilon_4=\infty$ (kwadraty),

– elektrody na płycie dielektrycznej o $\bar{H}=1$ i $\varepsilon_1=1$ (trójkąty),

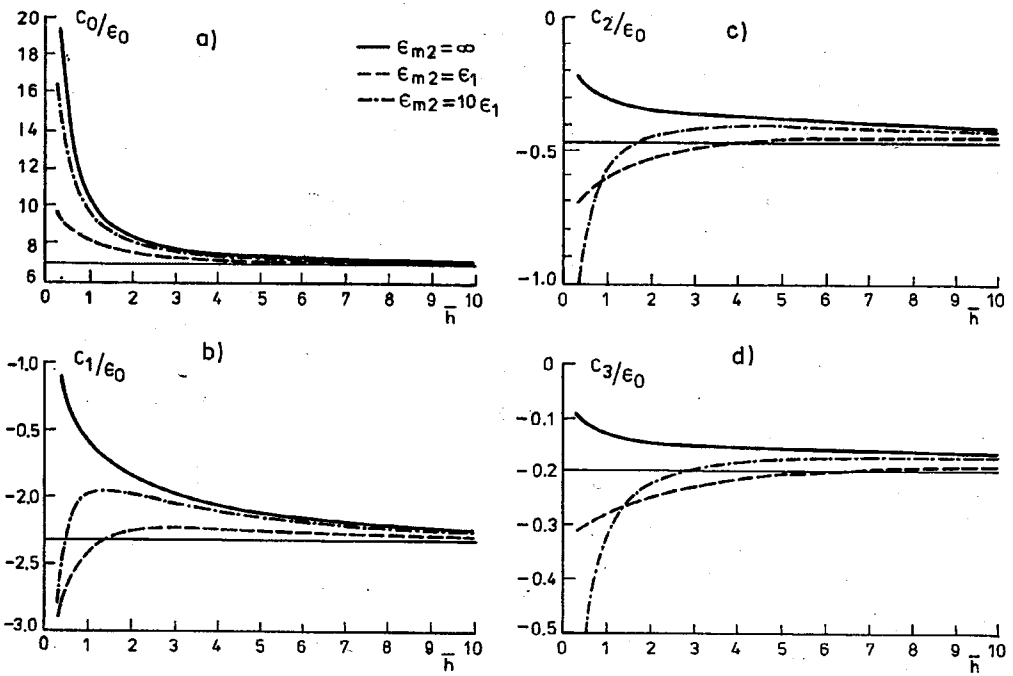
– podobnie ale $\varepsilon_1=\infty$ („*kara*").

Wszystkie wyniki są w dB (normowano do wartości maksymalnej na wykresie). W obu przypadkach (A–A i S–S) pojemność wzajemna jest najmniejsza dla płyty jednostronnie metalizowanej, największa zaś dla elektrod na półprzestrzeni dielektrycznej.

4.2. POJEMNOŚCIOWY CZUJNIK ODLEGŁOŚCI

Rozważmy C_n w funkcji $\bar{H}$, $\bar{h}$ oraz ε_i poszczególnych warstw. Ponieważ C_n szybko maleje z n rozważymy jedynie C_n dla bliskich elektrod tj. dla $n=0,...,3$.

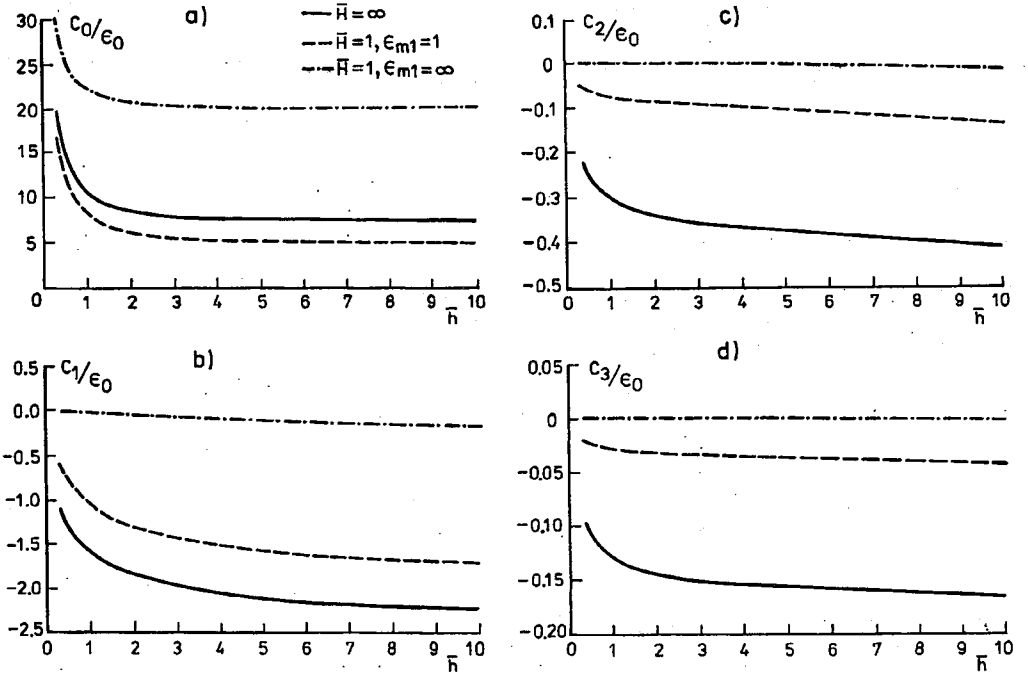
Na rysunku 7 przedstawiono wyniki numeryczne dla $\bar{H}=\infty$ i $0<\bar{h}<10$ (elektrody umieszczono na półprzestrzeni dielektrycznej ze szkła kwarcowego — sytuacja odpowiada Rys. 3a. Przedstawiono wyniki obliczeń dla kolejno $\varepsilon_4=\infty$, $10\varepsilon_2$, ε_2, linia ciągła reprezentuje przypadek $\varepsilon_4=1$. C_n na Rys. 7 unormowano względem ε_0.



Rys. 7. Przebieg C_n w funkcji $\bar{h}$ dla elektrod umieszczonych na półprzestrzeni dielektrycznej

Dla dużych wartości $\bar{h}$ wyniki dla $\varepsilon_4=\infty$ jak i $\varepsilon_4<\infty$ są podobne: prąd elektrody zasilanej (C_0) narasta wraz ze zmniejszaniem $\bar{h}$, prądy elektrod uziemionych maleją z powodu zamykania linii pola elektrycznego w ośrodku na granicy $y=-\bar{h}$. Jednak dla małych $\bar{h}$ sytuacja jest odmienna: prądy elektrod uziemionych w przypadku $\varepsilon_4=\infty$ maleją nadal co do modułu wraz ze zmniejszaniem $\bar{h}$, natomiast dla $\varepsilon_4<\infty$ z powodu wnikania pola elektrycznego w głąb dielektryka ε_4, ten prąd narasta co do modułu. W skrajnym przypadku tj. dla $\bar{h}=0$ mamy $C_0=\infty$ i $C_n=0$ dla $\varepsilon_4=\infty$. Ponieważ $\varepsilon_\infty(\bar{h}=0)>\varepsilon_\infty(\bar{h}=\infty)$ krzywe na Rys. 7 dla $n\neq 0$ w przypadku $\varepsilon_4<\infty$ wykazują charakterystyczne wygięcie.

Na rysunku 8 przedstawiono dla porównania rezultaty dla $\varepsilon_4=\infty$ dla elektrod umieszczonych na różnych podłożach: półprzestrzeni dielektrycznej, płyty dielektrycznej o $\bar{H}=1$ z górną powierzchnią swobodną tj. $\varepsilon_1=1$ i na płycie o $\bar{H}=1$ ale



Rys. 8. Porównanie C_n w funkcji $\bar{h}$ dla elektrod umieszczonych w różnych ośrodkach

z górną powierzchnią metalizowaną tj. $\varepsilon_1 = \infty$. Przypadki te odpowiadają odpowiednim wykresom efektywnej przenikalności dielektrycznej z rysunku 3.

Widać, że dla elektrod położonych na cienkiej ($\bar{H}=1$) płytce dielektrycznej metalizowanej po jednej stronie współczynniki C_n są dla $n \neq 0$ znacznie mniejsze niż dla elektrod umieszczonych na półprzestrzeni dielektrycznej czy płyty niemetalizowanej. Ma to istotne znaczenie w minimalizacji pojemności wzajemnej przetworników.

W przypadku czujnika zbudowanego na bazie kondensatora międzypalczastego dla otrzymania pełnych zależności jego pojemności od odległości przedmiotu i jego przenikalności dielektrycznej ε_4, należy przeprowadzić sumowanie po współczynnikach C_n analogicznie jak to czyniono przy analizie pojemności wzajemnej w poprzednim rozdziale.

PODSUMOWANIE

– Pojemność wzajemna dwóch przetworników jest najmniejsza gdy oba przetworniki pracują w konfiguracji symetrycznej.

– Inny sposób na zminimalizowanie tej pojemności to zamknięcie układu między dwiema płytami metalowymi.

– Parametry czujnika zależą od jego konfiguracji: możliwe jest wykorzystanie wszystkich lub niektórych C_n w tworzeniu sygnału wejściowego. Ogólnie najlepiej jest umieścić elektrody na szkle o małej przenikalności dielektrycznej.

BIBLIOGRAFIA

1. D.P. Morgan: *Surface-Wave Devices for Signal Processing.* Amsterdam: Elsevier, 1985
2. E. Danicki: *Unified theory of interdigital transducer and saw reflectors.* J. Techn. Phys., 1980, vol. 21, no. 3, pp. 387–403
3. K. Bløtekjaer, K. Ingebrigtsen, H. Skeie: *A method for analysing waves in structures consisting of metal strips on dispersive media.* IEEE Trans. Electron. Devices, 1973, vol. 20, no. 12, pp. 1133–1138
4. B.A. Auld: *informacja prywatna,* 1989
5. E. Danicki, D. Gafka: *Propagation, generation and detection of SAW in multiperiodic system of metal strips on a piezoelectric substrate.* JASA, 1991, vol. 89, no. 1, pp. 27–38

E. DANICKI, D. BOGUCKI

MUTUAL CAPACITANCE OF TWO INTERDIGITAL TRANSDUCERS IN THE METAL CASE

Summary

A periodic system of ideally conducting electrodes deposited in layered dielectric media is analysed. A capacitive electromagnetic crosstalk between interdigital transducers in surface acoustic wave devices is investigated. The results are also applicable in designing of capacitive distance sensor based on the interdigital capacitor.

KWARTALNIK ELEKTRONIKI I TELEKOMUNIKACJI, 1993, 39, z. 1, ss. 199-218

519.281
621.391.83

Błąd podstawowy w metodzie szczególnego próbkowania

JERZY SAWICKI

Katedra Miernictwa Elektrycznego, Politechnika Gdańska

Otrzymano 1992.07.08

Autoryzowano do druku 1992.09.02

Metoda szczególnego próbkowania służy do pomiaru wektora harmonicznej podstawowej przebiegu odkształconego. Podstawy teoretyczne tej metody zostały opublikowane w pracy [5]. Podano tam jedynie wstępne dane na temat osiągalnej dokładności, bowiem problem ten jest mocno złożony. Błąd pomiaru składa się z dwu zasadniczych części. Jedna z nich wynika z faktu, że niektóre wyższe harmoniczne, występujące w badanym przebiegu, pozostają niewyeliminowane. Ta przyczyna powoduje występowanie tak zwanego *„błędu podstawowego"*. Nieidealne właściwości przetwarzania analogowo-cyfrowego prowadzą do pojawienia się innej niedokładności, wywołującej tak zwany *„błąd digitalizacji"*. W niniejszym opracowaniu wyprowadzono zależności, które umożliwiają znalezienie granicznych wartości błędu podstawowego, jakie mogą wystąpić w najbardziej niesprzyjających warunkach. Obliczono także odnośne wartości błędu dla kilku regularnych przebiegów odkształconych, odpowiadające różnym realizacjom metody.

1. WSTĘP

W pracy [5] wykazano, że watość błędu podstawowego zależy od wielu czynników. Istotną rolę grają tu nie tylko kształt badanego przebiegu oraz właściwości zastosowanej realizacji metody — lecz także moment początkowy x, w którym rozpoczyna się proces pobierania serii próbek. Ogólnie biorąc, argument odpowiadający wspomnianemu momentowi początkowemu nie posiada określonej wartości, bowiem wynika on z przypadkowego przesunięcia czasowego pomiędzy przebiegiem badanym i działaniem urządzenia sterującego procesem pomiarowym. W dalszej części niniejszych rozważań wszystkie argumenty i kąty wyrażane są w [deg].

Amplituda A_1 i faza α_1 szukanego wektora harmonicznej podstawowej wyznaczane są w rozważanej metodzie z wartości dwu sum: Y_a oraz Y_r wyprowadzonych w publikacji [5]. Przy obliczaniu wyniku pomiaru przyjmuje się, że powyższe

sumy są wolne od wpływu wszelkich innych składowych oprócz podstawowej. W rzeczywistości jednak założenie to może nie być spełnione i wówczas pomiar daje przybliżone rezultaty A'_1, α'_1. W ten sposób powstają:

a) względny błąd amplitudy δ, odpowiadający zależności

$$A'_1 = (1+\delta)^* A_1 \tag{1}$$

b) bezwzględny błąd fazy ε, który jest określony związkiem

$$\alpha'_1 = \alpha_1 + \varepsilon \tag{2}$$

2. ZALEŻNOŚCI PODSTAWOWE

Na podstawie rozważań, przedstawionych w [5], można napisać:

$$\begin{aligned} C(1)\,A'_1 \cos(\alpha'_1 + x + \beta) &= +Y_\alpha \\ C(1)\,A'_1 \sin(\alpha'_1 + x + \beta) &= -Y_r, \end{aligned} \tag{3}$$

gdzie $C(1)$ oraz β oznaczają stałe, zależne od wybranej realizacji, natomiast x jest momentem początkowym serii próbkowań. Należy tu zaznaczyć, iż rozważania niniejsze odnoszą się do *„wariantu drugiego"* metody [5], który wykazuje najlepsze właściwości. Wyżej wspomniane sumy są w rzeczywistości określone zależnościami:

$$\begin{aligned} +Y_\alpha &= C(1)\,a_1 \cos(\alpha_1 + x + \beta) + \\ &+ \sum_{k=2} C(2k-1)\,A_{2k-1} \cos[\alpha_{2k-1} + (2k-1)(x+\beta)] \end{aligned} \tag{4a}$$

$$\begin{aligned} -Y_r &= C(1)\,A_1 \sin(\alpha_1 + x + \beta) + \\ &+ \sum_{k=2} -1^{k+1}\,C(2k-1)\,A_{2k-1} \sin[\alpha_{2k-1} + (2k-1)(x+\beta)]. \end{aligned} \tag{4b}$$

Powyższe wyrażenia są wolne od składowej stałej oraz wszystkich harmonicznych parzystych i to bez względu na zastosowaną realizację metody. Przez porównanie wzorów (3) i (4) oraz przy uwzględnieniu zależności (1), (2) otrzymuje się wyrażenia umożliwiające obliczenie błędu amplitudy i fazy. Celem uproszczenia zapisu dalszych rozważań wprowadza się teraz oznaczenia pomocnicze:

$$\frac{C(2k-1)}{C(1)} * \frac{A_{2k-1}}{A_1} = g_{2k-1} \tag{5}$$

$$\alpha_{2k-1} + (2k-1)(x+\beta) = \varphi_{2k-1}. \tag{6}$$

Należy tutaj zauważyć, iż stała $C(1)$ oraz wszystkie amplitudy A_1, A_{2k-1} są wartościami dodatnimi. Inaczej rzecz się przedstawia w odniesieniu do stałych $C(2k-1)$, które dla pewnych harmonicznych są w danej realizacji ujemne. Przy $k=1$, czyli dla harmonicznej podstawowej, jest $g_1 = \pm 1$. Wobec tego otrzymuje się następujące wyrażenia: