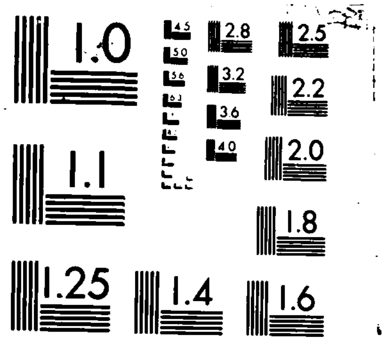
AD-A177 348 JOINT SERVICES ELECTRONICS PROGRAM: BASIC RESEARCH IN ELECRRONICS (JSEP) (U) TEXAS UNIV AT AUSTIN ELECTRONICS RESEARCH CENTER E J POWERS 31 DEC 86 1/2

UNCLASSIFIED AFOSR-TR-87-0091-APP F49620-86-C-0045 F/G 9/5 NL


1.0
4.5
5.0
5.6
6.3
2.8
2.5
3.2
2.2
3.6
4.0
2.0
1.1
1.8
1.25
1.4
1.6

1

ANNUAL REPORT ON ELECTRONICS RESEARCH AT THE UNIVERSITY OF TEXAS AT AUSTIN

APPENDIX - JSEP SPONSORED PUBLICATIONS

NO. 34 - Appendix

For the period April 1, 1986 through December 31, 1986

JOINT SERVICES ELECTRONICS PROGRAM

Research Contract ~~AFOSR~~ F49620-86-C-0045

Approved for public release;
distribution unlimited.

December 31, 1986

ELECTRONICS RESEARCH CENTER

Bureau of Engineering Research
The University of Texas at Austin
Austin, Texas 78712-1084

The Electronics Research Center at The University of Texas at Austin consists of interdisciplinary laboratories in which graduate faculty members, Master and PhD candidates from numerous academic disciplines conduct research. The disciplines represented in this report include information electronics, solid state electronics, quantum electronics, and electromagnetics.

The research summarized in this report was supported by the Department of Defense's JOINT SERVICES ELECTRONICS PROGRAM (U.S. Army, U.S. Navy and the U.S. Air Force) through the Research Contract AFOSR F49620-86-C-0045. This program is monitored by the Department of Defense's JSEP Technical Coordinating Committee consisting of representatives from the U.S. Army Research Office, Office of Naval Research and the U.S. Air Force Office of Scientific Research.

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION
UNCLASSIFIED

1b. RESTRICTIVE MARKINGS
AD-A177 348

2a. SECURITY CLASSIFICATION AUTHORITY

2b. DECLASSIFICATION/DOWNGRADING SCHEDULE

3. DISTRIBUTION/AVAILABILITY OF REPORT
Approved for public release;
distribution unlimited.

4. PERFORMING ORGANIZATION REPORT NUMBER(S)
N/A

5. MONITORING ORGANIZATION REPORT NUMBER(S)
AFOSR-TR- 87-0091

6a. NAME OF PERFORMING ORGANIZATION
University of Texas

6b. OFFICE SYMBOL (If applicable)

7a. NAME OF MONITORING ORGANIZATION
AFOSR

6c. ADDRESS (City, State and ZIP Code)
Bureau of Engineering Research
Austin Texas 78712

7b. ADDRESS (City, State and ZIP Code)
Bldg 410
Bolling AFB DC 20332

8a. NAME OF FUNDING/SPONSORING ORGANIZATION
7a

8b. OFFICE SYMBOL (If applicable)
NE

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
~~same as 5~~ F49620-86-C-0045

8c. ADDRESS (City, State and ZIP Code)
same as 7b

10. SOURCE OF FUNDING NOS.

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT NO.
61102F	2305	A9	(JSEP)

appendix

11. TITLE (Include Security Classification)
Basic Research in Electronics (JSEP)

12. PERSONAL AUTHOR(S)
Prof Powers

13a. TYPE OF REPORT
Annual

13b. TIME COVERED
FROM 4-1-86 TO 12-31-86

14. DATE OF REPORT (Yr., Mo., Day)
Dec 31 86

15. PAGE COUNT
81

16. SUPPLEMENTARY NOTATION

17. COSATI CODES

FIELD	GROUP	SUB. GR.

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

NARR. --- PROG. - FROM U)
4-1-86 to 12-31-86

This paper develops a systematic approach to select a time-dependent state transformation which can map a linear time-variant (LTV) digital filter to an equivalent filter having diagonal state-feedback matrices. Due to the struct- simplicity of the diagonal systems, this time-dependent state transformation is a convenient tool for analyzing recursive LTV filters expressible in the state-variable form. In this paper, we discuss both the theoretical basis and the application of this diagonalization procedure. The properties of two types of recursive LTV filters are examined by using this state transformation technique. Based upon the separable properties of the impulse responses with major class of recursive LTV filters. This technique, through suboptimal can substantially reduce the computation required in the synthesis procedure.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT
UNCLASSIFIED/UNLIMITED □ SAME AS RPT. □ DTIC USERS □

21. ABSTRACT SECURITY CLASSIFICATION
UUUU

22a. NAME OF RESPONSIBLE INDIVIDUAL
Dr. G.L. Witt

22b. TELEPHONE NUMBER (Include Area Code)
767-4931

22c. OFFICE SYMBOL
NE

INT. J. CONTROL, 1986, VOL. 44, NO. 4, 1103–1124

Approximate and local linearizability of non-linear discrete-time systems

HONG-GI LEE† and STEVEN I. MARCUS†

We consider a single-input non-linear discrete-time system of the form

$$\Sigma: \quad x(t+1) = f(x(t), u(t))$$

where $x \in \mathsf{R}^N$, $u \in \mathsf{R}$, and $f(x,u): \mathsf{R}^{N+1} \to \mathsf{R}^N$ is a C^∞ R^N-valued function. Necessary and sufficient conditions for approximate linearizability are given for Σ. We also give necessary and sufficient conditions for local linearizability. Finally, we present analogous results for multi-input non-linear discrete-time systems.

1. Introduction

We consider a single-input non-linear discrete-time system of the form

$$\Sigma: \quad x(t+1) = f(x(t), u(t)) \tag{1}$$

where $x \in \mathsf{R}^N$, $u \in \mathsf{R}$, and $f(x, u): \mathsf{R}^{N+1} \to \mathsf{R}^N$ is a C^∞ R^N-valued function.

Many authors have studied (local or global) linearization (Cheng *et al.* 1985, Hunt and Su 1981, Jakubczyk and Respondek 1980, Krener 1973, Su 1982) and approximate linearization (Krener 1984) by state feedback and coordinate change for non-linear continuous-time systems. In this paper we discuss necessary conditions and sufficient conditions for local linearization and approximate linearization by state feedback and coordinate change for non-linear discrete-time systems. A necessary and sufficient condition for local linearization has recently been found by Grizzle (1985 c); a result equivalent to this is proved in our Theorem 5. These conditions are very similar to those available for continuous-time systems, but they are more difficult to calculate than our sufficient condition in Theorem 4. Other related work on non-linear discrete-time systems can be found in Grizzle (1985 a, b), Grizzle and Nijmeijer (1985), Monaco and Normand-Cyrot (1983 a, b, 1984).

Definition 1

A point (x_e, u_e) such that $f(x_e, u_e) = x_e$ is called an *equilibrium point*.

Now consider the following linear discrete-time system Σ_0:

$$\Sigma_0: \quad y(t+1) = Ay(t) + bv(t) = g(y(t), v(t))$$

Received 23 January 1986.

† Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, Texas 78712, U.S.A.

where

$$A = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \ddots & \vdots \\ 0 & 0 & 0 & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & 0 \\ \vdots & \vdots & \vdots & & 1 \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix} \quad (N \times N \text{ matrix})$$

$$b = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \quad (N \times 1 \text{ matrix})$$

Similarly to the continuous-time case (Krener 1984, Su 1982), we can define local linearizability and approximate linearizability for a discrete-time system. Let (x_e, u_e) be an equilibrium point of Σ.

Definition 2

Σ is said to be *locally linearizable at* (x_e, u_e) if there exist an open neighbourhood U ($\subset \mathbb{R}^{N+1}$) of (x_e, u_e) and a diffeomorphism $T: U \to T(U)$ such that

(i) $\bar{T} = (T_1, T_2, \dots, T_N)$ are functions of $x_1, x_2, \dots, x_N$ only,

(ii) $T(x_e, u_e) = 0_{(N+1)\times 1}$,

(iii) $\bar{T} \circ f = g \circ T$.

If we let $(y(t)^{\mathrm{T}} v(t))^{\mathrm{T}} = T(x(t), u(t))$ then $y(t)$ and $v(t)$ satisfy Σ_0. Definition 2 indicates that we want to find a diffeomorphism T such that the following diagram commutes:

$$\begin{tikzcd} \mathbb{R}^N \arrow[d, "\bar{T}"'] & \mathbb{R}^{N+1} \arrow[l, "f"'] \arrow[d, "T"] \\ \mathbb{R}^N & \mathbb{R}^{N+1} \arrow[l, "g"] \end{tikzcd}$$

Once we find such a diffeomorphism, we can apply linear system theory instead of non-linear system theory.

Definition 3

Σ is said to be *approximately linearizable with order* ρ if there exist an open neighbourhood U ($\subset \mathbb{R}^{N+1}$) of (x_e, u_e) and a diffeomorphism $T: U \to T(U)$ such that

(i) $\bar{T} = (T_1, T_2, \dots, T_N)$ are functions of $x_1, x_2, \dots, x_N$ only,

(ii) $T(x_e, u_e) = 0_{(N+1)\times 1}$, and

(iii) $\bar{T} \circ f = g \circ T + O(x - x_e, u - u_e)^{\rho+1}$.

Thus in Definition 3 we consider the following nearly linear discrete-time system:

$$\Sigma_0': \quad y(t+1) = Ay(t) + bv(t) + O(x - x_e, u - u_e)^{\rho+1}$$

where the $N \times N$ matrix A and $N \times 1$ matrix b are the same as Σ_0. Clearly, local linearizability at (x_e, u_e) implies approximate linearizability with arbitrary order.

In § 2 some background material is reviewed and notation is defined. In § 3 necessary and sufficient conditions for approximate linearizability will be given for the system (1). Also, we shall give necessary and sufficient conditions for local linearizability. We can define local linearizability and approximate linearizability for multi-input discrete-time systems similarly to Definitions 2 and 3. Then the multi-input case will be discussed in § 4.

2. Preliminaries

In this section notations and definitions to be used later will be mentioned. The Kronecker product is very useful in the field of matrix calculus (Graham 1981). First, define the Kronecker product $\otimes$ by

$$\underset{p \times q}{A} \otimes \underset{m \times n}{B} = \begin{bmatrix} a_{11}B & a_{12}B & \dots & a_{1q}B \\ a_{21}B & a_{22}B & \dots & a_{2q}B \\ \vdots & \vdots & & \vdots \\ a_{p1}B & a_{p2}B & \dots & a_{pq}B \end{bmatrix}_{(pm) \times (qn)}$$

where a_{ij} is the (i, j)-component of the $p \times q$ matrix A.

Define the derivative of a matrix with respect to a matrix by

$$D_A B = \begin{bmatrix} \frac{\partial}{\partial a_{11}}B & \frac{\partial}{\partial a_{12}}B & \dots & \frac{\partial}{\partial a_{1q}}B \\ \frac{\partial}{\partial a_{21}}B & \frac{\partial}{\partial a_{22}}B & \dots & \frac{\partial}{\partial a_{2q}}B \\ \vdots & \vdots & & \vdots \\ \frac{\partial}{\partial a_{p1}}B & \frac{\partial}{\partial a_{p2}}B & \dots & \frac{\partial}{\partial a_{pq}}B \end{bmatrix}_{(mp) \times (nq)}$$

We also define

$$D_A^0 B = B$$

$$D_A^1 B = D_A B$$

$$D_A^{i+1} B = D_A(D_A^i B) \quad \text{for } i \geqslant 1$$

Let $h(x)$ be a scalar real-valued function of $x \in \mathbb{R}^N$. Then $(D_x^k h)(x)$ and $(D_{x^T}^k h)(x)$ are $N^k \times 1$ and $1 \times N^k$ vectors respectively.

Fact (Vetter 1970, 1971)

Using the definition of Kronecker product and derivative operations on matrices, Taylor's formula can be expressed by

$$h(x) = h(0) + \sum_{k=1}^{l} \frac{1}{k!}(D_{x^T}^k h(x))_{x=0}(x \otimes x \otimes \dots \otimes x) + R_{l+1}(x^*)$$

where $R_{l+1}(x^*)$ is a remainder term.

Now define the $N^k \times N^k$ permutation matrix $U_{i_1,i_2,\ldots,i_k}$ as follows: the $((a_{i_1}-1)N^{k-1}+(a_{i_2}-1)N^{k-2}+\ldots+(a_{i_{k-1}}-1)N+a_{i_k})$th column of $U_{i_1,i_2,\ldots,i_k}$ is the $((a_1-1)N^{k-1}+(a_2-1)N^{k-2}+\ldots+(a_{k-1}-1)N+a_k)$th column of the $N^k \times N^k$ identity matrix $(I_{N^k \times N^k})$ for $1 \leqslant a_1, a_2, \ldots, a_k \leqslant N$ (the $\{a_i\}$ are related to the 'base N' representation of the column). Here $\{i_1, i_2, \ldots, i_k\}$ is a permutation of $\{1, 2, \ldots, k\}$. For example, when $N=2$ and $k=3$

$$U_{123} = I_{8 \times 8}$$

and

$$U_{321} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Let A be a $p \times N^k$ matrix. Define the operator $\underset{k!}{\oplus}$ by

$$\underset{k!}{\oplus} A = A\left(\sum_{\substack{\text{all permutations} \\ \{i_1,i_2,\ldots,i_k\} \text{ of } \{1,2,\ldots,k\}}} U_{i_1,i_2,\ldots,i_k}\right)$$

For example, when A is a $p \times N^3$ matrix

$$\underset{3!}{\oplus} A = A(U_{123} + U_{132} + U_{213} + U_{231} + U_{312} + U_{321})$$

Let

$$\left\{\left(\frac{\partial f}{\partial u}\right)_{(0,0)}, \left(\frac{\partial f}{\partial x}\right)_{(0,0)}\left(\frac{\partial f}{\partial u}\right)_{(0,0)}, \ldots, \left(\frac{\partial f}{\partial x}\right)^{N-1}_{(0,0)}\left(\frac{\partial f}{\partial u}\right)_{(0,0)}\right\}$$

be linearly independent; that is, they form a basis for $\mathbb{R}^N$. Define $\zeta: \mathbb{R}^N \to \mathbb{R}$ by $\zeta(v) = \alpha_N$, where v is a $1 \times N$ row vector and

$$v^{\mathrm{T}} = \sum_{i=1}^{N} \alpha_i \left(\frac{\partial f}{\partial x}\right)^{i-1}_{(0,0)}\left(\frac{\partial f}{\partial u}\right)_{(0,0)}$$

That is, $\zeta(v)$ is the last coefficient of v^{T} with respect to the basis $\{w_1, w_2, \ldots, w_N\}$, where $w_1 = (\partial f/\partial x)^{i-1}_{(0,0)}(\partial f/\partial u)_{(0,0)}$, $1 \leqslant i \leqslant N$. Also define $\bar{\zeta}: \mathbb{R}^{p \times N} \to \mathbb{R}^p$ by

$$\bar{\zeta}(V) = \begin{bmatrix} \zeta(v_1) \\ \zeta(v_2) \\ \vdots \\ \zeta(v_p) \end{bmatrix}$$

where v_i is the ith row of V.

3. Single-input case

In this section our main results will be given. If $f(x, u)$ has an equilibrium point, without loss of generality, we can assume that $f(0, 0) = 0$; for, if not, let $\tilde{x} = x - x_e$ and $\tilde{u} = u - u_e$. Then $\tilde{x}(t + 1) = \tilde{f}(\tilde{x}(t), \tilde{u}(t)) \triangleq f(\tilde{x} + x_e, \tilde{u} + u_e) - x_e$ with $\tilde{f}(0, 0) = 0$.

Let

$$\hat{f}^1(x, u) = f(x, u)$$

$$\hat{f}^{i+1}(x, u) = f(\hat{f}^i(x, u), 0) \quad \text{for } 1 \leqslant i \leqslant N - 1$$

$\hat{f}^i(x, u)$ represents the effect of an input u at $t = 0$ on the state at $t = i$. $\hat{f}^i(x, u)$ is essential for solving many problems arising in discrete-time non-linear systems.

Lemma 1

Σ is locally linearizable at $(0, 0)$ if and only if there exists a C^{∞} function $h : W(\subset \mathbb{R}^N) \to \mathbb{R}$ such that

(i) W is an open neighbourhood of $0 \in \mathbb{R}^N$

(ii) $D_u(h \circ \hat{f}^i) \equiv 0$ on some neighbourhood of $0 \in \mathbb{R}^{N+1}$ for $1 \leqslant i \leqslant N - 1$

(iii) $$\det \begin{bmatrix} \left(\dfrac{\partial h}{\partial x}\right)_{x=0} \\ \left(\dfrac{\partial (h \circ \hat{f})}{\partial x}\right)_{(0,0)} \\ \vdots \\ \left(\dfrac{\partial (h \circ \hat{f}^{N-1})}{\partial x}\right)_{(0,0)} \end{bmatrix} \neq 0$$

(iv) $(D_u(h \circ \hat{f}^N))_{(0,0)} \neq 0$

(v) $h(0) = 0$

Proof

Necessity. Suppose that Σ is locally linearizable. Then we have a diffeomorphism T. Let $h(x) = T_1(x)$. (Since $T_1(x, u)$ depends only on x, we can write $T_1(x)$ instead of $T_1(x, u)$.) Note that $T_2 = T_1 \circ f$. Since $D_u(T_2) \equiv 0$ on some neighbourhood of the origin, $D_u(T_1 \circ f) \equiv 0$ on some neighbourhood of the origin. From now on, for convenience, we shall omit 'on some neighbourhood of the origin'. Note that $T_3 = T_2 \circ f = T_1 \circ \hat{f}^2$. (Actually, we can write $T_3 = T_1 \circ f^2$, because $T_1 \circ f$ depends only on x. But $\hat{f}^2$ is used, for consistency of notation.) Since $D_u(T_3) \equiv 0$, $D_u(T_1 \circ \hat{f}^2) \equiv 0$. Proceeding in this manner, since $T_N = T_{N-1} \circ f = \ldots = T_1 \circ \hat{f}^{N-1}$ and $D_u(T_N) \equiv 0$, $D_u(T_1 \circ \hat{f}^{N-1}) \equiv 0$. Thus we have shown that $D_u(T_1 \circ \hat{f}^i) \equiv 0$, for $1 \leqslant i \leqslant N - 1$. Since T is a diffeomorphism,

$T_i = T_1 \circ \hat{f}^{i-1}$ for $2 \leqslant i \leqslant N+1$, and $T_1, T_2, \ldots, T_N$ depend only on x,

$$\det \begin{bmatrix} \left(\dfrac{\partial h}{\partial x}\right)_{x=0} \\ \left(\dfrac{\partial (h \circ \hat{f})}{\partial x}\right)_{(0,0)} \\ \vdots \\ \left(\dfrac{\partial (h \circ \hat{f}^{N-1})}{\partial x}\right)_{(0,0)} \end{bmatrix} \neq 0$$

and $D_u(h \circ \hat{f}^N) \neq 0$. Since $T_1(0,0) = 0$, $h(0) = 0$.

Sufficiency. Suppose that there exists $h : \mathbb{R}^N \to \mathbb{R}$ satisfying the given conditions. Let $T_i(x) = h \circ \hat{f}^{i-1}$ for $1 \leqslant i \leqslant N+1$. Then it can be easily checked that $\bar{T} \circ f = g \circ T$ and $T(0,0) = 0$. Since $\det((\partial T/\partial(x,u))_{(0,0)}) \neq 0$ there exists an open neighbourhood U of $(0,0)$ such that $T: U \to T(U)$ is a diffeomorphism by the inverse function theorem. □

Let $\xi = \begin{pmatrix} x \\ u \end{pmatrix}$.

Lemma 2

Σ is *approximately linearizable* with order ρ if and only if there exists a C^∞ function $h : W(\subset \mathbb{R}^N) \to \mathbb{R}$ such that

(i) W is an open neighbourhood of $0 \in \mathbb{R}^N$

(ii) $(D^j_\xi D_u(h \circ \hat{f}^i))_{(0,0)} = 0_{(N+1)^j \times 1}$ for $1 \leqslant i \leqslant N-1$ and $0 \leqslant j \leqslant \rho - 1$

(iii) $$\det \begin{bmatrix} \left(\dfrac{\partial h}{\partial x}\right)_{x=0} \\ \left(\dfrac{\partial (h \circ \hat{f})}{\partial x}\right)_{(0,0)} \\ \vdots \\ \left(\dfrac{\partial (h \circ \hat{f}^{N-1})}{\partial x}\right)_{(0,0)} \end{bmatrix} \neq 0$$

(iv) $(D_u(h \circ \hat{f}^N))_{(0,0)} \neq 0$

(v) $h(0) = 0$

Proof

Necessity. Suppose Σ is approximately linearizable with order ρ. Let $h(x) = T_1(x)$. By definition, $T_1 \circ f(x,u) = T_2(x) + O(x,u)^{\rho+1}$. So $(D^j_\xi(D_u(h \circ f))_{(0,0)} = 0$ for $0 \leqslant j \leqslant \rho - 1$. Note that $T_1 \circ \hat{f}^2 = T_2 \circ f + O(x,u)^{\rho+1}$. Since $T_2 \circ f(x,u) = T_3(x) + O(x,u)^{\rho+1}$ by definition, $T_1 \circ \hat{f}^2 = T_3(x) + O(x,u)^{\rho+1}$. Thus $(D^j_\xi D_u(h \circ \hat{f}^2))_{(0,0)} = 0$ for $0 \leqslant j \leqslant \rho - 1$. Proceeding in this manner, we can show that $(D^j_\xi D_u(h \circ \hat{f}^i))_{(0,0)} = 0$ for $1 \leqslant i \leqslant N - 1$

and $0 \leqslant j \leqslant \rho - 1$. Note that $(\partial T_i/\partial x)_{(0,0)} = (\partial(h \circ \hat{f}^{i-1})/\partial x)_{(0,0)}$ for $1 \leqslant i \leqslant N$. Since

$$\det\begin{bmatrix} \left(\dfrac{\partial T_1}{\partial x}\right)_{(0,0)} \\ \left(\dfrac{\partial T_2}{\partial x}\right)_{(0,0)} \\ \vdots \\ \left(\dfrac{\partial T_N}{\partial x}\right)_{(0,0)} \end{bmatrix} \neq 0, \quad \det\begin{bmatrix} \left(\dfrac{\partial h}{\partial x}\right)_{x=0} \\ \left(\dfrac{\partial(h \circ \hat{f})}{\partial x}\right)_{(0,0)} \\ \vdots \\ \left(\dfrac{\partial(h \circ \hat{f}^{N-1})}{\partial x}\right)_{(0,0)} \end{bmatrix} \neq 0$$

It can easily be shown that $T_1 \circ \hat{f}^N(x, u) = T_{N+1}(x, u) + 0(x, u)^{\rho+1}$. Thus $(D_u(h \circ \hat{f}^N))_{(0,0)} = (D_u T_{N+1}(x, u))_{(0,0)} \neq 0$. Finally, $h(0) = T_1(0) = 0$.

Sufficiency (by construction). Let

$$T_1(x) = h(x)$$

$$T_2(x) = \sum_{k=1}^{\rho} \frac{1}{k!}(D^k_{x^\mathrm{T}}(h \circ \hat{f}))_{(0,0)}\overbrace{(x \otimes x \otimes \ldots \otimes x)}^{k\,\text{times}}$$

$$T_3(x) = \sum_{k=1}^{\rho} \frac{1}{k!}(D^k_{x^\mathrm{T}}(h \circ \hat{f}^2))_{(0,0)}(x \otimes x \otimes \ldots \otimes x)$$

$$\vdots$$

$$T_N(x) = \sum_{k=1}^{\rho} \frac{1}{k!}(D^k_{x^\mathrm{T}}(h \circ \hat{f}^{N-1}))_{(0,0)}(x \otimes x \otimes \ldots \otimes x)$$

$$T_{N+1}(x, u) = T_1 \circ \hat{f}^N(x, u)$$

Then it can easily be checked that T as defined above satisfies the conditions of Definition 3.

Now note that

$$(D^m_\xi D_u(h \circ \hat{f}^i))_{(0,0)} = \sum_{l=0}^{m} (B^i_{m,l})_{(0,0)}(D^{l+1}_x h)_{x=0} \tag{2}$$

where

$$B^i_{m,0} = D^m_\xi(D_u \hat{f}^i)^\mathrm{T}$$

and

$$B^i_{m,l} = \sum_{k_1=1}^{m-l+1} \sum_{k_2=1}^{k_1} \cdots \sum_{k_l=1}^{k_{l-1}} D^{m-l+1-k_1}_\xi(D_\xi \hat{f}^{i\mathrm{T}} \otimes D^{k_1-k_2}_\xi(D_\xi \hat{f}^{i\mathrm{T}} \otimes D^{k_2-k_3}_\xi)$$
$$\times (D_\xi \hat{f}^{i\mathrm{T}} \otimes \ldots \otimes D^{k_{l-1}-k_l}_\xi(D_\xi \hat{f}^{i\mathrm{T}} \otimes D^{k_l-1}_\xi D_u \hat{f}^{i\mathrm{T}}) \ldots)) \quad \text{for } 1 \leqslant l \leqslant m$$

(For a proof of (2) see the Appendix.) Let

$$A_k = \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1k} \\ A_{21} & A_{22} & \dots & A_{2k} \\ \vdots & \vdots & & \vdots \\ A_{k1} & A_{k2} & \dots & A_{kk} \end{bmatrix}$$

where A_{ij} is an $(N-1)(N+1)^i \times N^{j+1}$ submatrix defined by

$$A_{ij} = 0_{(N-1)(N+1)^i \times N^{j+1}} \qquad \text{if } i < j$$

$$A_{ij} = \bigoplus_{(j+1)!} \begin{bmatrix} (B^1_{ij})_{(0,0)} \\ (B^2_{ij})_{(0,0)} \\ \vdots \\ (B^{N-1}_{ij})_{(0,0)} \end{bmatrix} \qquad \text{if } i \geqslant j$$

Let the $(N-1)(N+1)^i \times 1$ vector β^i be defined by

$$\beta^i = \tilde{\zeta}\left(\begin{bmatrix} (B^1_{i,0})_{(0,0)} \\ (B^2_{i,0})_{(0,0)} \\ \vdots \\ (B^{N-1}_{i,0})_{(0,0)} \end{bmatrix}\right)$$

(See § 2 for the definitions of $\bigoplus_{j!}$ and $\tilde{\zeta}$.) Also, let $\beta_k = (\beta^{1\mathrm{T}} \quad \beta^{2\mathrm{T}} \quad \dots \quad \beta^{k\mathrm{T}})^{\mathrm{T}}$.

With these preliminaries, we can state our main theorems.

Theorem 3

Σ is approximately linearizable with order ρ ($\geqslant 2$) if and only if

(i) $\left\{\left(\frac{\partial f}{\partial u}\right)_{(0,0)}, \left(\frac{\partial f}{\partial x}\right)_{(0,0)}\left(\frac{\partial f}{\partial u}\right)_{(0,0)}, \dots, \left(\frac{\partial f}{\partial x}\right)^{N-1}_{(0,0)}\left(\frac{\partial f}{\partial u}\right)_{(0,0)}\right\}$ are linearly independent

(ii) $\beta_{\rho-1} \in \mathrm{Image}\,(A_{\rho-1})$

Proof

Necessity. Suppose that Σ is approximately linearizable with order ρ. Then there exists a function $h(x)$ satisfying (i)–(v) of Lemma 2; in particular,

$$(D_u(h \circ \bar{f}^i))_{(0,0)} = \left(\frac{\partial h}{\partial x}\right)_{x=0}\left(\frac{\partial f}{\partial x}\right)^{i-1}_{(0,0)}\left(\frac{\partial f}{\partial u}\right)_{(0,0)} = 0 \quad \text{for } 1 \leqslant i \leqslant N-1$$

and

$$\left(\frac{\partial h}{\partial x}\right)_{x=0}\left(\frac{\partial f}{\partial x}\right)^{N-1}_{(0,0)}\left(\frac{\partial f}{\partial u}\right)_{(0,0)} \neq 0$$

Assume that

$$\left\{\left(\frac{\partial f}{\partial u}\right)_{(0,0)}, \left(\frac{\partial f}{\partial x}\right)_{(0,0)}\left(\frac{\partial f}{\partial u}\right)_{(0,0)}, \dots, \left(\frac{\partial f}{\partial x}\right)^{N-1}_{(0,0)}\left(\frac{\partial f}{\partial u}\right)_{(0,0)}\right\}$$

are not linearly independent. Then there exists k such that $1 \leqslant k \leqslant N-1$ and

$$\left(\frac{\partial f}{\partial x}\right)^{k}_{(0,0)}\left(\frac{\partial f}{\partial u}\right)_{(0,0)} = \sum_{j=0}^{k-1} \alpha_j \left(\frac{\partial f}{\partial x}\right)^{j}_{(0,0)}\left(\frac{\partial f}{\partial u}\right)_{(0,0)} \quad \text{for some constants } \{\alpha_j\}_{j=0}^{k-1}$$

Thus

$$\left(\frac{\partial f}{\partial x}\right)^{N-1}\left(\frac{\partial f}{\partial u}\right)_{(0,0)} = \sum_{j=0}^{k-1} \alpha_j \left(\frac{\partial f}{\partial x}\right)^{N-1-k+j}_{(0,0)}\left(\frac{\partial f}{\partial u}\right)_{(0,0)}$$

and

$$\left(\frac{\partial h}{\partial x}\right)_{x=0}\left(\frac{\partial f}{\partial x}\right)^{N-1}_{(0,0)}\left(\frac{\partial f}{\partial u}\right)_{(0,0)} = 0$$

This is a contradiction, which implies that

$$\left\{\left(\frac{\partial f}{\partial u}\right)_{(0,0)}, \left(\frac{\partial f}{\partial x}\right)_{(0,0)}\left(\frac{\partial f}{\partial u}\right)_{(0,0)}, \dots, \left(\frac{\partial f}{\partial x}\right)^{N-1}_{(0,0)}\left(\frac{\partial f}{\partial u}\right)_{(0,0)}\right\}$$

are linearly independent.

Recall that $(D_u(h \circ \hat{f}^i))_{(0,0)} = (D_\xi^m D_u(h \circ \hat{f}^i))_{(0,0)} = 0$ for $m = 0$ and $1 \leqslant i \leqslant N-1$. Since

$$\left\{\left(\frac{\partial f}{\partial u}\right)_{(0,0)}, \left(\frac{\partial f}{\partial x}\right)_{(0,0)}\left(\frac{\partial f}{\partial u}\right)_{(0,0)}, \dots, \left(\frac{\partial f}{\partial x}\right)^{N-1}_{(0,0)}\left(\frac{\partial f}{\partial u}\right)_{(0,0)}\right\}$$

are linearly independent, $(D_x h)_{x=0}$ is uniquely determined up to a constant multiple (i.e. $(D_x h)_{x=0} = \alpha c$, where the scalar $\alpha\ (\neq 0)$ is arbitrary and the $N \times 1$ column vector c satisfies $c^{\mathrm{T}}(\partial f/\partial x)^i_{(0,0)}(\partial f/\partial u)_{(0,0)} = 0$ for $0 \leqslant i \leqslant N-2$, and $c^{\mathrm{T}}(\partial f/\partial x)^{N-1}_{(0,0)}(\partial f/\partial u)_{(0,0)} = 1$).

Now by (ii) and (iv) of Lemma 2, $(D_\xi^m D_u(h \circ \hat{f}^i))_{(0,0)} = 0$ for $1 \leqslant m \leqslant \rho - 1$ and $1 \leqslant i \leqslant N-1$. From (2) we obtain

$$\begin{bmatrix} B^1_{1,1} & 0 & 0 & \dots & 0 \\ B^2_{1,1} & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ B^{N-1}_{1,1} & 0 & 0 & \dots & 0 \\ B^1_{2,1} & B^1_{2,2} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ B^{N-1}_{2,1} & B^{N-1}_{2,2} & 0 & \dots & 0 \\ \vdots & \vdots & & & \vdots \\ & & & & 0 \\ B^1_{\rho-1,1} & B^1_{\rho-1,2} & & \dots & B^1_{\rho-1,\rho-1} \\ \vdots & \vdots & & & \vdots \\ B^{N-1}_{\rho-1,1} & B^{N-1}_{\rho-1,2} & & \dots & B^{N-1}_{\rho-1,\rho-1} \end{bmatrix}_{(0,0)} \begin{bmatrix} (D_x^2 h)_{x=0} \\ (D_x^3 h)_{x=0} \\ \vdots \\ (D_x^\rho h)_{x=0} \end{bmatrix} = - \begin{bmatrix} B^1_{1,0} \\ B^2_{1,0} \\ \vdots \\ B^{N-1}_{1,0} \\ B^1_{2,0} \\ \vdots \\ B^{N-1}_{2,0} \\ \vdots \\ B^1_{\rho-1,0} \\ \vdots \\ B^{N-1}_{\rho-1,0} \end{bmatrix}_{(0,0)} (D_x^1 h)_{x=0} \tag{3}$$

Since $(B^1_{1,0})_{(0,0)}(D^1_x h)_{x=0} = (D^1_\zeta D_u \hat{f}^{\mathrm{T}})_{(0,0)}(D^1_x h)_{x=0} = \alpha\zeta((D^1_\zeta D_u \hat{f}^{\mathrm{T}})_{(0,0)})$

$$\begin{bmatrix} B^1_{1,0} \\ B^2_{1,0} \\ \vdots \\ B^{N-1}_{1,0} \end{bmatrix}_{(0,0)} (D^1_x h)_{x=0} = \alpha\beta^1$$

Thus the right-hand side of (3) is $-\alpha[(\beta^1)^{\mathrm{T}} \; \dots \; (\beta^{\rho-1})^{\mathrm{T}}]^{\mathrm{T}} = -\alpha\beta^{\mathrm{T}}_{\rho-1}$. It follows that $\beta_{\rho-1}$ is in the image of the matrix on the left-hand side of (3). However, the $\{D^k_x h\}$ are constrained because, for example, $\partial^2 h/\partial x_i^2 \partial x_j = \partial^2 h/\partial x_j \partial x_i$. Hence the stronger condition $\beta_{\rho-1} \in \mathrm{Image}\,(A_{\rho-1})$ holds, as is proved in the Appendix in Lemma A.2.

Sufficiency. Suppose that (i) and (ii) above are true. By (i) there exists an $N \times 1$ vector C_1 such that $C^{\mathrm{T}}_1(\partial f/\partial x)^i_{(0,0)}(\partial f/\partial u)_{(0,0)} = 0$ for $0 \leqslant i \leqslant N-2$ and $C^{\mathrm{T}}_1(\partial f/\partial x)^{N-1}_{(0,0)}(\partial f/\partial u)_{(0,0)} = 1$. By (ii) there exist $C_2, C_3, \dots, C_\rho$ such that

$$A_{\rho-1}\begin{bmatrix} \frac{1}{2!}C_2 \\ \frac{1}{3!}C_3 \\ \vdots \\ \frac{1}{\rho!}C_\rho \end{bmatrix} = -\beta_{\rho-1}$$

where C_i is an $N^i \times 1$ vector. Let

$$h(x) = \sum_{i=1}^{\rho} \frac{1}{i!} C^{\mathrm{T}}_i \underbrace{(x \otimes x \otimes \dots \otimes x)}_{i\,\text{times}}$$

Then it can be easily checked that $(D^j_\zeta D_u(h \circ \hat{f}^i))_{(0,0)} = 0$ for $1 \leqslant i \leqslant N-1$ and $0 \leqslant j \leqslant \rho - 1$. Clearly $(D_u(h \circ \hat{f}^N))_{(0,0)} = (D^1_x h)^{\mathrm{T}}_{x=0}(\partial f/\partial x)^{N-1}_{(0,0)}(\partial f/\partial u)_{(0,0)} = 1 \neq 0$.

Now assume that

$$\det\begin{bmatrix} \left(\dfrac{\partial h}{\partial x}\right)_{x=0} \\ \left(\dfrac{\partial(h \circ \hat{f})}{\partial x}\right)_{(0,0)} \\ \vdots \\ \left(\dfrac{\partial(h \circ \hat{f}^{N-1})}{\partial x}\right)_{(0,0)} \end{bmatrix} = 0$$

Then there exists k such that $1 \leqslant k \leqslant N-1$ and

$$(D^1_x h)^{\mathrm{T}}_{x=0}\left(\frac{\partial f}{\partial x}\right)^k_{(0,0)} = \sum_{i=0}^{k-1} \alpha_i (D^1_x h)^{\mathrm{T}}_{x=0}\left(\frac{\partial f}{\partial x}\right)^i_{(0,0)} \quad \text{for some } \{\alpha_i\}^{k-1}_{i=0}$$

Thus

$$(D_x^1 h)^{\mathrm{T}}_{x=0}\left(\frac{\partial f}{\partial x}\right)^{N-1}_{(0,0)}\left(\frac{\partial f}{\partial u}\right)_{(0,0)} = \sum_{i=0}^{k-1} \alpha_i (D_x^1 h)^{\mathrm{T}}_{x=0}\left(\frac{\partial f}{\partial x}\right)^{N-1-k+i}_{(0,0)}\left(\frac{\partial f}{\partial u}\right)_{(0,0)} = 0$$

This is a contraction, which implies (iii) of Lemma 2. Hence, by Lemma 2, Σ is approximately linearizable with order ρ. □

Remark

Σ is approximately linearizable with order 1 if and only if (i) of Theorem 3 holds, just as in the continuous case (Krener 1984).

Now a sufficient condition for local linearizability is given in the following theorem.

Theorem 4

Suppose that $f(x, u)$ of Σ is an analytic $\mathbb{R}^N$-valued function. Σ is locally linearizable at (0, 0) if

(i) $\left\{\left(\frac{\partial f}{\partial u}\right)_{(0,0)}, \left(\frac{\partial f}{\partial x}\right)_{(0,0)}\left(\frac{\partial f}{\partial u}\right)_{(0,0)}, \ldots, \left(\frac{\partial f}{\partial x}\right)^{N-1}_{(0,0)}\left(\frac{\partial f}{\partial u}\right)_{(0,0)}\right\}$ are linearly independent,

(ii) there exists k ($<\infty$) such that $\beta_l \in \operatorname{span}(C_l^k)$ for all $l \geqslant 1$, where C_l^k is composed of the first k columns of A_l.

Proof

By (i) there exists an $N \times 1$ vector c_1 such that $c_1^{\mathrm{T}}(\partial f/\partial x)^i_{(0,0)}(\partial f/\partial u)_{(0,0)} = 0$ for $0 \leqslant i \leqslant N-2$ and $c_1^{\mathrm{T}}(\partial f/\partial x)^{N-1}_{(0,0)}(\partial f/\partial u)_{(0,0)} = 1$. By (ii) there exist $c_2, c_3, \ldots, c_j$ such that $j < \infty$ and

$$A_l \begin{bmatrix} \frac{1}{2!}c_2 \\ \frac{1}{3!}c_3 \\ \vdots \\ \frac{1}{j!}c_j \\ 0 \\ \vdots \\ 0 \end{bmatrix} = -\beta_l \quad \text{for } l \geqslant j$$

where c_i is an $N^i \times 1$ vector. Let

$$h(x) = \sum_{i=1}^{j} \frac{1}{i!} c_i^{\mathrm{T}} \overbrace{(x \otimes x \otimes \ldots \otimes x)}^{i\,\text{terms}}$$

Then it can easily be checked that $(D_\xi^s D_u(h \circ \hat{f}^i))_{(0,0)} = 0$ for $1 \leqslant i \leqslant N-1$ and $s \geqslant 0$. Since both $h(x)$ and $\hat{f}^i$ are analytic, $h \circ \hat{f}^i(x,u)$ are analytic, for $1 \leqslant i \leqslant N-1$. Thus $D_u(h \circ \hat{f}^i) \equiv 0$ for $1 \leqslant i \leqslant N-1$. As in the sufficiency proof in Theorem 3, it can be shown that h satisfies the other conditions of Lemma 1. □

It is easy to see that (ii) of Theorem 4 implies (ii) of Theorem 3.

Remark

Conditions (i) and (ii) are also necessary for local linearizability at (0, 0) whenever $f(x,u)$ is polynomial and a polynomial $T(x,u)$ is sought.

In the following theorem we give necessary and sufficient conditions for local linearizability.

Theorem 5

Σ is locally linearizable at (0, 0) if and only if

(i) $$\left\{\left(\frac{\partial f}{\partial u}\right)_{(0,0)}, \left(\frac{\partial f}{\partial x}\right)_{(0,0)}\left(\frac{\partial f}{\partial u}\right)_{(0,0)}, \ldots, \left(\frac{\partial f}{\partial x}\right)^{N-1}_{(0,0)}\left(\frac{\partial f}{\partial u}\right)_{(0,0)}\right\}$$

are linearly independent,

(ii) there exists an open neighbourhood U of $0 \in \mathbb{R}^N$ such that $\Delta_i \equiv f_*(\Delta_0) + f_*^2(\Delta_0) + \ldots + f_*^i(\Delta_0)$ are well-defined i-dimensional involutive distributions on U for $1 \leqslant i \leqslant N-1$, where $\Delta_0 \equiv \text{span}\{\partial/\partial u\}$.

Proof

Necessity. For condition (i) see Theorem 3. By Lemma 1 there exists $h(x)$ such that conditions (i)–(v) of Lemma 1 are satisfied. Let $y_1 = h(x)$, $y_2 = h \circ f(x,u), \ldots$, and $y_N = h \circ f^{N-1}(x,u)$. Then by (ii) and (iii) of Lemma 1 we can choose $(y_1(x), y_2(x), \ldots, y_N(x))$ as new coordinates on an open neighbourhood U of $0 \in \mathbb{R}^N$. By (ii) and (iv) of Lemma 1, $f_*(\Delta_0) = \text{span}\{\partial/\partial y_N\}$. Similarly, $f_*(\Delta_0) + f_*^2(\Delta_0) = \text{span}\{\partial/\partial y_{N-1}, \partial/\partial y_N\}$, and by induction, $\Delta_{N-1} = \text{span}\{\partial/\partial y_2, \partial/\partial y_3, \ldots, \partial/\partial y_N\}$. Hence Δ_{N-1} is an involutive distribution.

Sufficiency. By Frobenius' Theorem there exists a C^∞ function $h: \mathbb{R}^N \to \mathbb{R}$ such that $\Delta_{N-1}(h) \equiv 0$ and $(\partial h/\partial x)_{x=0} \neq 0$. Therefore $f_*^i(\Delta_0)(h) \equiv 0$ for $1 \leqslant i \leqslant N-1$. By (i), since $(\partial h/\partial x)_{x=0} \neq 0$, $(\partial/\partial u)(h \circ f^N(x,u))_{(0,0)} \neq 0$. It is easy to see that $h(x)$ satisfies the conditions of Lemma 1. □

Remarks

(*a*) Condition (ii) of Theorem 5 can be replaced by (ii)′ $\ker f_* + (\pi_*^{-1} f_*)^i(\Delta_0)$ is involutive for $0 \leqslant i \leqslant N-2$, where $\pi(x,u) \equiv x$, $(\pi_*^{-1} f_*)^0(\Delta_0) \equiv \Delta_0$, and $(\pi_*^{-1} f_*)^j(\Delta_0) \equiv (\pi_*^{-1} f_*)((\pi_*^{-1} f_*)^{j-1}(\Delta_0))$ for $j \geqslant 1$ (for this see Grizzle 1985 c, Lemma 2.1).

(*b*) A more geometric necessary and sufficient condition for approximate linearizability of order ρ can also be obtained (Lee 1986); however, the conditions of Theorem 3 are much easier to check.

Example 1

Consider the following discrete-time non linear system

$$\Sigma: \begin{bmatrix} x_1(t+1) \\ x_2(t+1) \end{bmatrix} = \begin{bmatrix} x_2(t) + 2x_1(t)u(t) + u^2(t) \\ x_1(t) + u(t) \end{bmatrix}$$

Since $(\partial f/\partial u)_{(0,0)} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ and [illegible]

Note that

$$\beta^1 = \tfrac{2}{1!}(D_1 D_2 [illegible]) \quad \Big([illegible] \Big)$$

and $\beta^i = 0_{[illegible]\times 1}$ for $i \geq 2$

$$(B^1_{11})_{(0,0)} = (D_2 f^1)_{(0,0)} \otimes (D_2 [illegible]) [illegible]$$

$$A_1 = \bigoplus (B^1_{11})_{(0,0)} = \begin{bmatrix} 0 & 0 & 0 & 2 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}$$

It can be easily checked that all elements of the 4th column of A_l are 0 for $l \geq 2$. Thus $\beta_l \in \operatorname{span}(C_l^4)$ for $l \geq 1$, where C_l^4 is composed of the first four columns of A_l. Therefore (ii) of Theorem 4 is also satisfied. Hence Σ is locally linearizable at (0, 0). Actually we can construct a diffeomorphism $T = (T_1, T_2, T_3)$ in the way that is given in the proof of Theorem 4. Since

$$A_l \begin{bmatrix} \frac{1}{2!}\begin{bmatrix} 0 \\ 0 \\ 0 \\ -2 \end{bmatrix} \\ \frac{1}{3!}0_{8\times 1} \\ \vdots \\ \frac{1}{(l+1)!}0_{2^{l+1}\times 1} \end{bmatrix} = -\beta_l \quad \text{for } l \geq 1$$

$c_2^{\mathrm{T}} = (0 \quad 0 \quad 0 \quad -2)$. Clearly $c_1^{\mathrm{T}} = (1 \quad 0)$. Thus

$$T_1(x) = x_1 - \frac{2}{2!}x_2^2 = x_1 - x_2^2$$

$$T_2(x) = T_1 \circ f(x, u) = x_2 - x_1^2$$

$$T_3(x, u) = T_1 \circ \hat{f}^2(x, u) = x_1 + u - (x_2 + 2x_1u + u^2)^2$$

Example 2

Consider

$$\Sigma\colon \begin{bmatrix} x_1(t+1) \\ x_2(t+1) \end{bmatrix} = \begin{bmatrix} x_2(t) + 2x_1(t)u(t) + x_1(t)^2u(t) + u(t)^2 \\ x_1(t) + u(t) \end{bmatrix} = f(x(t), u(t))$$

Clearly (i) of Theorem 3 is satisfied, because $(\partial f/\partial u)_{(0,0)}$ and $(\partial f/\partial x)_{(0,0)}(\partial f/\partial u)_{(0,0)}$ are the same as in Example 1. Since

$$(D^1_\xi D_u f^{\mathrm{T}})_{(0,0)} = \begin{bmatrix} 2 & 0 \\ 0 & 0 \\ 2 & 0 \end{bmatrix}, \quad \beta^1 = \tilde{\zeta}((D_\xi D_u f^{\mathrm{T}})_{(0,0)}) = \begin{bmatrix} 2 \\ 0 \\ 2 \end{bmatrix}$$

Since

$$(D^2_\xi D_u f^{\mathrm{T}})_{(0,0)} = \begin{bmatrix} 2 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad \beta^2 = (2\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0)^{\mathrm{T}}$$

Also we have

$$A_1 = \begin{bmatrix} 0 & 0 & 0 & 2 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}$$

$$\begin{aligned}(B^1_{22})_{(0,0)} &= (D_\xi f^{\mathrm{T}})_{(0,0)} \otimes (D_\xi f^{\mathrm{T}})_{(0,0)} \otimes (D_u f^{\mathrm{T}})_{(0,0)} \\ &= \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \otimes \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \otimes (0 \quad 1) \\ &= \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}\end{aligned}$$

therefore $A_{22} = \bigoplus_{3!} (B^1_{22})_{(0,0)}$

$$= \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 6 \\ 0 & 0 & 0 & 2 & 0 & 2 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 6 \\ 0 & 0 & 0 & 2 & 0 & 2 & 2 & 0 \\ 0 & 2 & 2 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 2 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 6 \\ 0 & 0 & 0 & 2 & 0 & 2 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 6 \end{bmatrix}$$

$$(B^1_{21})_{(0,0)} = (D_\xi(D_\xi f^{\mathrm{T}} \otimes D_u f^{\mathrm{T}}))_{(0,0)} + (D_\xi f^{\mathrm{T}} \otimes D_\xi D_u f^{\mathrm{T}})_{(0,0)}$$

$$= (D^2_\xi f^{\mathrm{T}} \otimes D_u f^{\mathrm{T}})_{(0,0)} + \begin{bmatrix} D_\xi f^{\mathrm{T}} \otimes D_{x_1} D_u f^{\mathrm{T}} \\ D_\xi f^{\mathrm{T}} \otimes D_{x_2} D_u f^{\mathrm{T}} \\ D_\xi f^{\mathrm{T}} \otimes D_u D_u f^{\mathrm{T}} \end{bmatrix}_{(0,0)} + (D_\xi f^{\mathrm{T}} \otimes D_\xi D_u f^{\mathrm{T}})_{(0,0)}$$

$$= \begin{bmatrix} 0 & 0 & 4 & 0 \\ 2 & 0 & 0 & 0 \\ 0 & 2 & 4 & 0 \\ 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 \\ 0 & 2 & 4 & 0 \\ 2 & 0 & 0 & 0 \\ 0 & 2 & 4 & 0 \end{bmatrix}$$

$$A_{21} = \bigoplus_{2!} (B^1_{21})_{(0,0)} = \begin{bmatrix} 0 & 4 & 4 & 0 \\ 4 & 0 & 0 & 0 \\ 0 & 6 & 6 & 0 \\ 4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 4 & 0 & 0 & 0 \\ 0 & 6 & 6 & 0 \\ 4 & 0 & 0 & 0 \\ 0 & 6 & 6 & 0 \end{bmatrix}$$

Since $\beta_1 \in \mathrm{Image}(A_1)$, Σ is approximately linearizable with $\rho = 2$. However, since $\beta_2 \notin \mathrm{Image}(A_2)$, it is not approximately linearizable with $\rho = 3$. Thus it is also not locally linearizable. Let

$$T_1 = x_1 - x_2^2$$

$$T_2 = x_2 - x_1^2$$

$$T_3 = x_1 + u - (x_2 + 2x_1 u + x_1^2 u + u^2)^2$$

Then

$$\begin{aligned} y_1(t+1) &\triangleq T_1(x(t+1)) = x_2(t) + 2x_1(t)u(t) + x_1(t)^2 u(t) + u(t)^2 - (x_1(t) + u(t))^2 \\ &= x_2(t) - x_1(t)^2 + x_1(t)^2 u(t) = T_2(x(t)) + 0(x, u)^3 \\ &= y_2(t) + 0(x, u)^3 \\ y_2(t+1) &\triangleq T_2(x(t+1)) = T_3(x(t), u(t)) \triangleq v(t) \end{aligned}$$

4. Multi-input case

The results in § 3 can be easily generalized to the multi-input case. Thus in this section we give (without proof) a sufficient condition for local linearizability and a necessary and sufficient condition for approximate linearizability by state feedback and coordinate change for a multi-input non-linear discrete-time system (for proof see Lee 1986).

Consider a multi-input non-linear discrete-time system of the form

$$\Sigma: \quad x(t+1) = f(x(t), u(t)) \tag{4}$$

where $x(t) \in \mathbb{R}^N$, $u(t) \in \mathbb{R}^m$, and $f(x, u): \mathbb{R}^{N+m} \to \mathbb{R}^N$ is a C^∞ $\mathbb{R}^N$-valued function. Also, consider the following multi-input linear discrete-time system Σ_0:

$$\Sigma_0: \quad y(t+1) = Ay(t) + Bv(t) = g(y(t), v(t))$$

where $y(t) \in \mathbb{R}^N$, $v(t) \in \mathbb{R}^m$, $A = \text{block diag}\{A_{11}, A_{22}, \ldots, A_{mm}\}$

$$A_{ii} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix} \quad (K_i \times K_i \text{ matrix})$$

$$\sum_{i=1}^{m} K_i = N,$$

$B = \text{block diag}\{b_1, b_2, \ldots, b_m\}$

$$b_i = (0 \quad \ldots \quad 0 \quad 1)^T \quad (K_i \times 1 \text{ matrix})$$

Definition 4

Σ is said to be *locally linearizable at* (x_e, u_e) if there exist indices $\{K_i\}_{i=1}^m$, an open neighbourhood $U\ (\subset \mathbb{R}^{N+m})$ of an equilibrium point (x_e, u_e) and a diffeomorphism $T: U \to T(U)$ such that

(i) $\bar{T} = (T_1, T_2, \ldots, T_N)$ are functions of $x_1, x_2, \ldots, x_N$ only,

(ii) $T(x_e, u_e) = 0_{(N+m)\times 1}$

(iii) $\bar{T} \circ f = g \circ T$

If we let

$$\begin{pmatrix} y(t) \\ v(t) \end{pmatrix} = T(x(t), u(t))$$

then $y(t)$ and $v(t)$ satisfy the relation Σ_0.

Definition 5

Σ is said to be *approximately linearizable with order* ρ if there exist indices $\{K_i\}_{i=1}^{m}$, an open neighbourhood $U\ (\subset \mathbb{R}^{N+m})$ of an equilibrium point (x_e, u_e) and a diffeomorphism $T: U \to T(U)$ such that

(i) $\bar{T} = (T_1, T_2, \ldots, T_N)$ are functions of $x_1, x_2, \ldots, x_N$ only,

(ii) $T(x_e, u_e) = 0_{(N+m)\times 1}$

(iii) $\bar{T} \circ f = g \circ T + O(x - x_e, u - u_e)^{\rho+1}$

Thus in Definition 5 we consider the following nearly linear multi-input discrete-time system:

$$\Sigma_0' : \quad y(t+1) = Ay(t) + Bv(t) + O(x - x_e, u - u_e)^{\rho+1}$$

where the $N \times N$ matrix A and $N \times m$ matrix B are as in Σ_0.

Now we state the generalized version of Lemmas 1 and 2 and Theorems 3 and 4. Just as in the single-input case, we can assume $f(0, 0) = 0$ without loss of generality, if f has an equilibrium point. Also, we define $\hat{f}^i(x, u)$ in the same way as in the previous section.

Lemma 5

Σ is locally linearizable at (0, 0) if and only if there exist $\{K_i\}_{i=1}^{m}$ and C^∞ functions $h_1(x), h_2(x), \ldots, h_m(x): W(\subset \mathbb{R}^N) \to \mathbb{R}$ such that

(i) W is an open neighbourhood of $0 \in \mathbb{R}^N$,

(ii) $D_u(h_j \circ \hat{f}^i) \equiv 0 \quad$ for $1 \leqslant j \leqslant m$ and $1 \leqslant i \leqslant k_j - 1$

(iii)

$$\det \begin{bmatrix} \left(\dfrac{\partial h_1}{\partial x}\right)_{x=0} \\ \left(\dfrac{\partial h_1 \circ f}{\partial x}\right)_{(0,0)} \\ \left(\dfrac{\partial h_1 \circ f^{K_1-1}}{\partial x}\right)_{(0,0)} \\ \vdots \\ \left(\dfrac{\partial h_m}{\partial x}\right)_{x=0} \\ \vdots \\ \left(\dfrac{\partial (h_m \circ f^{K_m-1})}{\partial x}\right)_{(0,0)} \end{bmatrix} \neq 0$$

(iv)

$$\det \begin{bmatrix} \left(\dfrac{\partial(h_1 \circ f^{K_1})}{\partial u}\right)_{(0,0)} \\ \vdots \\ \left(\dfrac{\partial(h_m \circ f^{K_m})}{\partial u}\right)_{(0,0)} \end{bmatrix} \neq 0$$

(v) $h_j(0) = 0 \quad \text{for } 1 \leqslant j \leqslant m$

Let $\xi = (x^{\mathrm{T}} \quad u^{\mathrm{T}})^{\mathrm{T}}$. Thus ξ is a $(N+m) \times 1$ vector.

Lemma 6

Σ is approximately linearizable with order ρ if and only if there exist $\{K_i\}_{i=1}^{m}$ and C^{∞} functions $h_1(x), h_2(x), \ldots, h_m(x): W(\subset \mathbb{R}^N) \to \mathbb{R}$ such that

(i) W is an open neighbourhood of $0 \in \mathbb{R}^N$,

(ii) $(D_{\xi}^{k} D_u (h_j \circ f^i))_{(0,0)} = 0 \quad \text{for } 1 \leqslant j \leqslant m,\ 1 \leqslant i \leqslant k_j - 1,\ 0 \leqslant k \leqslant \rho - 1$

(iii), (iv) and (v) of Lemma 5 are satisfied.

Let

$$E = \left\{ \left(\frac{\partial f}{\partial u_1}\right)_{(0,0)}, \left(\frac{\partial f}{\partial x}\right)_{(0,0)} \left(\frac{\partial f}{\partial u_1}\right)_{(0,0)}, \ldots, \left(\frac{\partial f}{\partial x}\right)_{(0,0)}^{K_1 - 1} \left(\frac{\partial f}{\partial u_1}\right)_{(0,0)}, \left(\frac{\partial f}{\partial u_2}\right)_{(0,0)}, \ldots, \right.$$
$$\left. \left(\frac{\partial f}{\partial x}\right)_{(0,0)}^{K_2 - 1} \left(\frac{\partial f}{\partial u_2}\right)_{(0,0)}, \ldots, \left(\frac{\partial f}{\partial u_m}\right)_{(0,0)}, \ldots, \left(\frac{\partial f}{\partial x}\right)_{(0,0)}^{K_m - 1} \left(\frac{\partial f}{\partial u_m}\right)_{(0,0)} \right\}$$

$$E_i = \left\{ \left(\frac{\partial f}{\partial u_1}\right)_{(0,0)}, \left(\frac{\partial f}{\partial x}\right)_{(0,0)} \left(\frac{\partial f}{\partial u_1}\right)_{(0,0)}, \ldots, \left(\frac{\partial f}{\partial x}\right)_{(0,0)}^{K_i - 2} \left(\frac{\partial f}{\partial u_1}\right)_{(0,0)}, \left(\frac{\partial f}{\partial u_2}\right)_{(0,0)}, \ldots, \right.$$
$$\left. \left(\frac{\partial f}{\partial x}\right)_{(0,0)}^{K_i - 2} \left(\frac{\partial f}{\partial u_2}\right)_{(0,0)}, \ldots, \left(\frac{\partial f}{\partial u_m}\right)_{(0,0)}, \ldots, \left(\frac{\partial f}{\partial x}\right)_{(0,0)}^{K_i - 2} \left(\frac{\partial f}{\partial u_m}\right)_{(0,0)} \right\}, \quad i = 1, \ldots, m$$

Suppose that the elements of E are linearly independent; that is, they form a basis for $\mathbb{R}^N$. Let $\sigma_i = \sum_{j=1}^{i} K_j$ for $1 \leqslant i \leqslant m$. Define $\zeta^i(v): \mathbb{R}^N \to \mathbb{R}$ by $\zeta^i(v) = \alpha_{\sigma_i}$, where v is a $1 \times N$ row vector and

$$v^{\mathrm{T}} = \alpha_1 \left(\frac{\partial f}{\partial u_1}\right)_{(0,0)} + \ldots + \alpha_{\sigma_1} \left(\frac{\partial f}{\partial x}\right)_{(0,0)}^{K_1 - 1} \left(\frac{\partial f}{\partial u_1}\right)_{(0,0)} + \alpha_{\sigma_1 + 1} \left(\frac{\partial f}{\partial u_2}\right)_{(0,0)} + \ldots$$
$$+ \alpha_{\sigma_2} \left(\frac{\partial f}{\partial x}\right)_{(0,0)}^{K_2 - 1} \left(\frac{\partial f}{\partial u_2}\right)_{(0,0)} + \ldots + \alpha_{\sigma_{m-1} + 1} \left(\frac{\partial f}{\partial u_m}\right)_{(0,0)} + \ldots$$
$$+ \alpha_{\sigma_m} \left(\frac{\partial f}{\partial x}\right)_{(0,0)}^{K_m - 1} \left(\frac{\partial f}{\partial u_m}\right)_{(0,0)}$$

Also, define $\tilde{\zeta}^i: \mathbb{R}^{p \times N} \to \mathbb{R}^p$ for $i = 1, 2, \ldots, m$ by

$$\tilde{\zeta}^i(V) = (\zeta^i(v_1) \quad \zeta^i(v_2) \quad \ldots \quad \zeta^i(v_p))^{\mathrm{T}}$$

where v_j is the jth row of V. Let

$$\beta^i_j = \zeta^i\left(\begin{bmatrix} (D^j_\zeta D_u f^{\mathrm{T}})_{(0,0)} \\ (D^j_\zeta D_u \hat{f}^{2\mathrm{T}})_{(0,0)} \\ \vdots \\ (D^j_\zeta D_u \hat{f}^{K_i-1\,\mathrm{T}})_{(0,0)} \end{bmatrix}\right)$$

Also, let $\gamma^i_k = ((\beta^i_1)^{\mathrm{T}} \quad (\beta^i_2)^{\mathrm{T}} \quad \dots \quad (\beta^i_k)^{\mathrm{T}})^{\mathrm{T}}$. Let

$$D^l_k = \begin{bmatrix} D^1_{11} & D^1_{12} & \dots & D^1_{1k} \\ D^1_{21} & D^1_{22} & \dots & D^1_{2k} \\ \vdots & \vdots & & \vdots \\ D^1_{k1} & D^1_{k2} & \dots & D^1_{kk} \end{bmatrix}$$

where

$$D^l_{ij} = 0_{m(k_l-1)(N+m)^i \times N^{j+1}} \qquad \text{if } i < j$$

$$D^l_{ij} = \bigoplus_{(j+1)!} \begin{bmatrix} (B^1_{ij})_{(0,0)} \\ (B^2_{ij})_{(0,0)} \\ \vdots \\ (B^{k_l-1}_{ij})_{(0,0)} \end{bmatrix} \qquad \text{if } i \geqslant j$$

Theorem 7

Σ is approximately linearizable with order ρ ($\geqslant 2$) if and only if there exist $\{K_i\}^m_{i=1}$ such that

(i) the elements of E are linearly independent,

(ii) span E_i = span $(E_i \cap E)$ for $1 \leqslant i \leqslant m$

(iii) $\gamma^l_{\rho-1} \in \operatorname{Image}(D^l_{\rho-1})$ for $1 \leqslant l \leqslant m$

Remark

Σ is approximately linearizable with order 1 if and only if (i) and (ii) of Theorem 7 hold, just as in the continuous-time case (Krener 1984). If $m = 1$ (single-input case) then $K_1 = N$. Thus (i) of Theorem 7 is the same as (i) of Theorem 3. Since $E_1 = E_1 \cap E$, (ii) of Theorem 7 is trivially satisfied. Since the operator ζ^1 is the same as the operator ζ in the previous section, $\gamma^1_{\rho-1} = \beta_{\rho-1}$. Since $D^1_{\rho-1} = A_{\rho-1}$, (iii) of Theorem 7 is the same as (ii) of Theorem 3. Therefore Theorem 7 is a generalized version of Theorem 3.

Now a sufficient condition for local linearizability is given in the following theorem.

Theorem 8

Suppose that $f(x, u)$ of Σ is an analytic $\mathbb{R}^N$-valued function. Σ is locally linearizable at (0, 0) if

(i) the elements of E are linearly independent,

(ii) span $E_i = \text{span}\,(E_i \cap E)$ for $1 \leqslant i \leqslant m$

(iii) there exists $k\,(<\infty)$ such that $\gamma_i^l \in \text{span}\,((F_i^l)^k)$ for $1 \leqslant l \leqslant m$ and $i \geqslant 1$, where $(F_i^l)^k$ is composed of the first k columns of D_i^l.

Given the system (4), we choose the Kronecker indices $\{K_i\}_{i=1}^m$ in a similar way to the continuous-time case (Hunt and Su 1981). First we form the matrix

$$\begin{bmatrix} \left(\dfrac{\partial f}{\partial u_1}\right)_{(0,0)} & \left(\dfrac{\partial f}{\partial u_2}\right)_{(0,0)} & \cdots & \left(\dfrac{\partial f}{\partial u_m}\right)_{(0,0)} \\ \left(\dfrac{\partial f}{\partial x}\right)_{(0,0)}\left(\dfrac{\partial f}{\partial u_1}\right)_{(0,0)} & \left(\dfrac{\partial f}{\partial x}\right)_{(0,0)}\left(\dfrac{\partial f}{\partial u_2}\right)_{(0,0)} & \cdots & \left(\dfrac{\partial f}{\partial x}\right)_{(0,0)}\left(\dfrac{\partial f}{\partial u_m}\right)_{(0,0)} \\ \vdots & \vdots & & \vdots \\ \left(\dfrac{\partial f}{\partial x}\right)_{(0,0)}^{N-1}\left(\dfrac{\partial f}{\partial u_1}\right)_{(0,0)} & \left(\dfrac{\partial f}{\partial x}\right)_{(0,0)}^{N-1}\left(\dfrac{\partial f}{\partial u_2}\right)_{(0,0)} & \cdots & \left(\dfrac{\partial f}{\partial x}\right)_{(0,0)}^{N-1}\left(\dfrac{\partial f}{\partial u_m}\right)_{(0,0)} \end{bmatrix}$$

Let α_i be the number of linearly independent vectors in the first $i+1$ rows for $0 \leqslant i \leqslant N-1$. Take $\gamma_0 = \alpha_0$ and $\gamma_i = \alpha_i - \alpha_{i-1}$ for $1 \leqslant i \leqslant N-1$, and define K_i to be the number of γ_j with $\gamma_j \geqslant i$.

Acknowledgment

The authors would like to thank A. Arapostathis and J. W. Grizzle for many helpful discussions involving the material presented in this paper.

This research was supported in part by the Air Force Office of Scientific Research under grant AFOSR-86-0029, in part by the National Science Foundation under grant ECS-8412100, and in part by the Joint Services Electronics Program under contract F49602-82-C-0033.

Appendix

Lemma A.1

Equation (2) holds.

Proof

Clearly,

$$D_\xi^m D_u(h \circ \hat{f}^i) = \sum_{l=0}^{m} (B_{m,l}^i)_{(x,u)} (D_x^{l+1} h)_{\hat{f}^i(x,u)}$$

$$D_\xi^{m+1} D_u(h \circ \hat{f}^i) = \sum_{l=0}^{m+1} (B_{m+1,l}^i)_{(x,u)} (D_x^{l+1} h)_{\hat{f}^i(x,u)}$$

where $\{\beta_{m,l}^i\}$ are to be determined. Then

$$B_{m+1,0}^i = D_\xi(B_{m,0}^i) \tag{A 1}$$

$$B_{m+1,l}^i = D_\xi(B_{m,l}^i) + D_\xi \hat{f}^{iT} \otimes B_{m,l-1}^i \quad \text{for } 1 \leqslant l \leqslant m \tag{A 2}$$

$$B_{m+1,m+1}^i = D_\xi \hat{f}^{iT} \otimes B_{m,m}^i \tag{A 3}$$

By (A 1), since $B_{0,0}^i = D_u \hat{f}^{iT}$, $B_{m,0}^i = D_\xi^m D_u \hat{f}^{iT}$ for $m \geqslant 1$. Note that (2) is true when

$m = 1$. Now suppose that (2) is true for $m \leqslant p$. Let $1 \leqslant l \leqslant p$. Then, by (A 2),

$$\begin{aligned}\mathbf{B}^i_{p+1,l} &= D_\xi(B^i_{p,l}) + D_\xi \hat f^{iT} \otimes B^i_{p,l-1}\\ &= \sum_{k_1=1}^{p-l+1}\sum_{k_2=1}^{k_1}\cdots\sum_{k_l=1}^{k_{l-1}} D_\xi^{p+1-l+1-k_1}(D_\xi \hat f^{iT}\otimes D_\xi^{k_1-k_2}(D_\xi \hat f^{iT}\otimes\cdots\otimes D_\xi^{k_{l-1}-k_l}\\ &\quad (D_\xi \hat f^{iT}\otimes D_\xi^{k_l-1}D_u \hat f^{iT})\ldots)) + \sum_{k_1=1}^{p+1-l+1}\sum_{k_2=1}^{k_1}\cdots\sum_{k_{l-1}=1}^{k_{l-2}} D_\xi \hat f^{iT}\otimes D^{p+1-l+1-k_1}\\ &\quad (D_\xi \hat f^{iT}\otimes D_\xi^{k_1-k_2}(D_\xi \hat f^{iT}\otimes\cdots\otimes D_\xi^{k_{l-2}-k_{l-1}}(D_\xi \hat f^{iT}\otimes D_\xi^{k_{l-1}-1}D_u \hat f^{iT})\ldots))\end{aligned}$$

Changing the dummy variables $k_1, k_2, \ldots, k_{l-1}$ of the second term into $k_2, k_3, \ldots, k_l$, respectively, the second term becomes

$$\begin{aligned}&\sum_{k_2=1}^{p+1-l+1}\sum_{k_3=1}^{k_2}\cdots\sum_{k_l=1}^{k_{l-1}} D_\xi f^{iT}\otimes D_\xi^{p+1-l+1-k_2}(D_\xi f^{iT}\otimes D_\xi^{k_2-k_3}(D_\xi f^{iT}\otimes\cdots\otimes(D_\xi^{k_{l-1}-k_l}\\ &\quad (D_\xi \hat f^{iT}\otimes D_\xi^{k_l-1}\mathrm{D}_u \hat f^{iT})\ldots)) = \sum_{k_1=p+1-l+1}^{p+1-l+1}\sum_{k_2=1}^{k_1}\sum_{k_3=1}^{k_2}\cdots\sum_{k_l=1}^{k_{l-1}} D^{p+1-l+1-k_1}\\ &\quad (D_\xi \hat f^{iT}\otimes D_\xi^{k_1-k_2}(D_\xi \hat f^{iT}\otimes D_\xi^{k_2-k_3}(D_\xi \hat f^{iT}\otimes\cdots\otimes D_\xi^{k_{l-1}-k_l}(D_\xi \hat f^{iT}\otimes D_\xi^{k_l-1}D_u \hat f^{iT})\ldots))\end{aligned}$$

Thus (2) is true for $m = p + 1$ and $1 \leqslant l \leqslant p$. By (A 3), it is easy to see that (2) is true for $m = l = p + 1$. Hence (2) is true for $m = p + 1$. By induction, (2) is true for $m \geqslant 1$. □

Let $h(x)\colon \mathbb{R}^N \to \mathbb{R}$ be a C^∞ function.

Lemma A.2

If

$$(S_1 \quad S_2 \quad \ldots \quad S_k)\begin{bmatrix}(D_x^2 h)_{x=0}\\ (D_x^3 h)_{x=0}\\ \vdots\\ (D_x^{k+1} h)_{x=0}\end{bmatrix} = d_{p\times 1} \qquad \text{(A 4)}$$

where S_i is a $p \times N^{i+1}$ matrix for $1 \leqslant i \leqslant k$, then $d \in \operatorname{Image}(B)$, where

$$B = \left(\left(\bigoplus_{2'} S_1\right), \left(\bigoplus_{3'} S_2\right), \ldots, \left(\bigoplus_{(k+1)'} S_k\right)\right)$$

Proof

Equation (A 4) is equivalent to

$$S_1(D_x^2 h)_{x=0} + S_2(D_x^3 h)_{x=0} + \ldots + S_k(D_x^{k+1} h)_{x=0} = d$$

Consider

$$S_1(D_x^2 h)_{x=0} = d'$$

Note that $(\partial^2 h/\partial x_i \partial x_j)_{x=0} \triangleq h_{ij} = h_{ji} \triangleq (\partial^2 h/\partial x_j \partial x_i)_{x=0}$ for $1 \leqslant i \leqslant N$ and $1 \leqslant j \leqslant N$. Let $(s_1)_i$ be the ith column of S_1. Then, since $h_{ij} = h_{ji}$,

$$\begin{aligned}S_1(D_x^2 h)_{x=0} &= \sum_{a_1=1}^{N} (s_1)_{(a_1-1)N+a_1} h_{a_1 a_1} + \sum_{a_1=1}^{N}\sum_{a_2=a_1+1}^{N}((s_1)_{(a_1-1)N+a_2}\\ &\quad + (s_1)_{(a_2-1)N+a_1}) h_{a_1 a_2} = d'\end{aligned}$$

Therefore $d' \in \operatorname{span}(Q)$, where

$$Q = \left(\bigcup_{a_1 = 1}^{N} \bigcup_{a_2 = a_1 + 1}^{N} \{ s_{(a_1 - 1)N + a_2} + s_{(a_2 - 1)N + a_1} \} \right) \cup \left(\bigcup_{a_1 = 1}^{N} \{ s_{(a_1 - 1)N + a_1} \} \right)$$

Now consider the matrix S_1' defined by

$$S_1' = \bigoplus_{2'} S_1 = S_1(U_{12} + U_{21})$$

It is easy to see that the $((a_1 - 1)N + a_2)$th column of S_1'

$$(s_1')_{(a_1 - 1)N + a_2} = \begin{cases} 2(s_1)_{(a_1 - 1)N + a_1} & \text{if } a_1 = a_2 \\ (s_1)_{(a_1 - 1)N + a_2} + (s_1)_{(a_2 - 1)N + a_1} & \text{if } a_1 \neq a_2 \end{cases}$$

where $1 \leqslant a_1 \leqslant N$ and $1 \leqslant a_2 \leqslant N$. Clearly $\operatorname{Image}(S_1') = \operatorname{span}(Q)$. Similar arguments can be applied for $(D_x^3 h)_{x=0}, \ldots, (D_x^{k+1} h)_{x=0}$. Therefore $d \in \operatorname{Image}(B)$.

REFERENCES

CHENG, D., TARN, T., and ISIDORI, A., 1985, *I.E.E.E. Trans. autom. Control*, **30**, 808.

GRAHAM, A., 1981, *Kronecker Products and Matrix Calculus with Applications* (New York: Ellis Horwood).

GRIZZLE, J. W., 1985 a, *I.E.E.L. Trans. autom. Control*, **30**, 868; 1985 b, Local input–output decoupling of discrete time nonlinear systems. Preprint University of Illinois; 1985 c, Feedback linearization of discrete-time systems. Presented at the *7th Int. Conf. on Analysis and Optimization of Systems*, Antibes, France.

GRIZZLE, W., and NIJMEIJER, H., 1985, Zeros at infinity for nonlinear discrete time systems. Preprint Memorandum No. 500, Department of Applied Mathematics, Twente University of Technology.

HUNT, L., and SU, R., 1981, Local transformations for multi-input nonlinear systems. *Proc. Joint Automatic Control Conf.*, Charlottesville, Virginia.

JAKUBCZYK, B., and RESPONDEK, W., 1980, *Bull. Acad. Polon. Sci. Ser. Math. Astron. Physics*, **28**, 517.

KRENER, A. J., 1973, *SIAM J. Control*, **11**, 670; 1984, *Syst. Control Lett.*, **5**, 181.

LEE, H.-G., 1986, On discrete time nonlinear control systems. Ph.D. dissertation, Department of Electrical and Computer Engineering, University of Texas at Austin.

MONACO, S., and NORMAND-CYROT, D., 1983 a, *Int. J. Control*, **38**, 245; 1983 b, Formal power series and input–output linearization of nonlinear discrete-time systems. *Proc. 22nd I.E.E.E. Conf. on Decision and Control*, San Antonio, Texas, pp. 655–660; 1984, *Syst. Control Lett.*, **5**, 191.

SU, R., 1982, *Syst. Control Lett.*, **2**, 48.

VETTER, W. J., 1970, *I.E.E.E. Trans. autom. Control*, **15**, 241; 1971, *Ibid.*, **16**, 113.

A Structure-Independent Approach to the Analysis and Synthesis of Recursive Linear Time-Variant Digital Filters

TZONG-YEU LEOU, STUDENT MEMBER, IEEE, AND J. K. AGGARWAL, FELLOW, IEEE

Abstract —This paper develops a systematic approach to select a time-dependent state transformation which can map a linear time-variant (LTV) digital filter to an equivalent filter having diagonal state-feedback matrices. Due to the structural simplicity of the diagonal systems, this time-dependent state transformation is a convenient tool for analyzing recursive LTV filters expressible in the state-variable form. In this paper, we discuss both the theoretical basis and the application of this diagonalization procedure. The properties of two types of recursive LTV filters are examined by using this state transformation technique. Based upon the separable properties of the impulse responses, we have explored a new algorithm for synthesizing desired impulse responses with a major class of recursive LTV filters. This technique, though suboptimal, can substantially reduce the computation required in the synthesis procedure.

I. Introduction

IN RECENT YEARS, there has been considerable interest in the analysis and synthesis of linear time-variant (LTV) digital filters for processing signals whose characteristics change significantly with time. Generally speaking, a straightforward extension of a synthesis technique for linear time-invariant (LTI) filters is sufficient for implementing LTV filters which have finite-duration impulse responses. However, the use of a recursive structure has the advantages of saving computation time and storage space, if the duration of the desired impulse response of an LTV filter is relatively long. But the synthesis of a recursive LTV filter is difficult because the characteristics of an LTV filter are related to the filter coefficients in a complicated fashion, except for certain filters implemented with some special structures [1]. Some researchers have suggested a very simple but somewhat heuristic synthesis method, which is based on the implementation of the frozen-time transfer function of an LTV filter [2], [3]. However, as illustrated in [4], noticeable differences between the desired and the realized filter characteristics of an LTV filter exist, unless the filter coefficients change very slowly with time.

The basic properties of a continuous-time LTV system realizable as a differential equation have been investigated in the literature [5] [7]. Recently, the properties of the discrete-time counterpart have also been reported [1], [8]. But the previous studies of discrete-time recursive LTV filters are limited to those filters realized with time-variant difference equations. Since there are many other conceivable structures which can be utilized to implement an LTV filter, a structure-independent approach to the analysis and synthesis of recursive LTV filters is desirable. Unlike the implementation of a recursive LTI filter, the structure selected for implementing a recursive LTV filter is an important factor in determining the characteristics of the realized filter.

The main objective of this paper is to develop a structure-independent approach to analyze and synthesize recursive LTV filters. Therefore, we express the basic model of recursive LTV filters in the state-variable form, which is capable of representing most recursive LTV filters. A time-dependent state transformation is devised to reduce the complexity of the state-variable model of LTV filters. We then examine the basic properties of recursive LTV filters by utilizing the time-dependent state transformation. Further, we explore the solutions to the synthesis problem of realizing an LTV impulse response with a finite-order recursive digital filter.

In Section II, we first introduce several descriptions of LTV digital filters, and discuss the properties of those representations relevant to the synthesis of recursive LTV filters. In Section III, attention is devoted to the analysis of LTV filters represented in the state-variable form. A time-dependent state transformation that diagonalizes the state-feedback matrices has been developed so that an LTV filter expressed in the state-variable form can be transformed into a filter consisting of K parallel first-order filters. This leads to a general expression for the impulse response realizable via a recursive LTV filter. In Section IV, we illustrate how this diagonalization procedure can be applied to analyze the properties of LTV filters realized with different filter structures. In Section V, we formulate the time-domain synthesis problem of a recursive LTV filter by minimizing the squared difference between the desired impulse response and the impulse response realizable as a major class of recursive LTV filters. The numerical difficulty of obtaining the optimal solution is examined. An efficient suboptimal algorithm based on the minimization of the localized squared difference between

Manuscript received November 26, 1984; revised November 14, 1985. This work was supported by the Department of Defense Joint Services Electronics Program under the Air Force Office of Scientific Research Contract F49620-82-C-0033.

The authors are with the Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712.

IEEE Log Number 8608528.

0098-4094/86/0700-0687$01.00 © 1986 IEEE

the desired impulse response and the realized impulse response is also developed. A numerical example has been selected as an illustration of this synthesis algorithm.

II. Characterizations of LTV Filters

Generally speaking, most methods for describing LTV filters are evolved from those used for LTI filters. A common time-domain description for LTV filters is the time-variant impulse response, which is defined as the output measured at the instant n in response to a unit-sample input applied at instant m. Then, the input $x(n)$ and the output $y(n)$ are related to the impulse response $h(n,m)$ by the summation

$$y(n) = \sum_{m=-\infty}^{\infty} h(n,m)x(m) \tag{1}$$

where the filter is said to be causal if the impulse response $h(n,m)$ satisfies

$$h(n,m) = 0, \qquad \text{for } n < m. \tag{2}$$

If one considers the computation and storage requirements of implementing an LTV filter having a long-duration impulse response, it is desirable to synthesize the filter using some recursive filter structure. A widely used structure for implementing LTV filters is a time-variant difference equation, which relates the output sequence $y(n)$ to the input sequence $x(n)$ by

$$\sum_{i=0}^{K} a_i(n)y(n-i) = \sum_{i=0}^{L} b_i(n)x(n-i) \tag{3}$$

where $a_0(n) \neq 0$ for all n and the order of the difference equation is equal to K if $a_K(n) \neq 0$ for some n. By adopting the direct form II structure [9] used in the synthesis of LTI filters, one can define another LTV filter structure in terms of a two-stage difference equation

$$w(n) = -\sum_{i=1}^{K} a_i(n)w(n-i) + x(n)$$

$$y(n) = \sum_{i=0}^{K-1} c_i(n)w(n-i). \tag{4}$$

The block diagram of the direct form II realization of LTV filters is shown in Fig. 1. Many other filter structures can be developed in the same way. However, unlike the case of representing recursive LTI filters, it is difficult to establish the explicit relationships among the recursive LTV filters realized with different structures.

For the purpose of analyzing recursive LTV filters synthesized with a variety of structures, we focus our attention on the state-variable representation in the present study. Assuming that $x(n)$ and $y(n)$ denote, respectively, the input and the output of the filter, the input-output relationship of a recursive LTV filter can be expressed in terms of the state equations

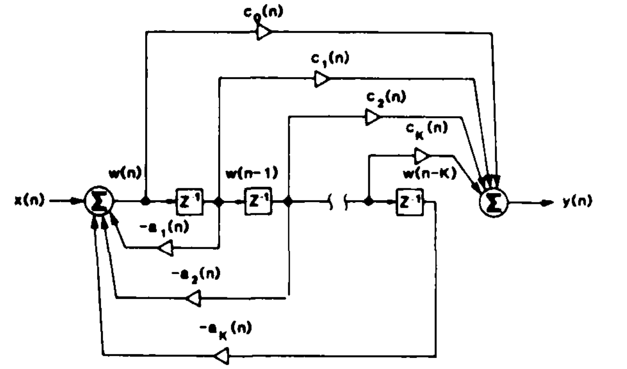


Fig. 1. Block diagram of the direct form II realization of a recursive LTV filter.

$$W(n) = A(n)W(n-1) + B(n)x(n)$$

$$y(n) = C(n)W(n) \tag{5}$$

where $W(n)$ is the state vector and $A(n)$, $B(n)$ and $C(n)$ are matrices of appropriate dimensions. It is clear that a recursive LTV filter realized via a difference equation can be easily expressed in the state-variable form by choosing an appropriate set of variables as the state vector. Further, the state-variable form is very suitable for representing a recursive LTV filter with multiple inputs and multiple outputs.

The basic properties of an LTV filter realizable as a time-variant difference equation has been explored by Huang and Aggarwal [1]. The main result of Huang and Aggarwal's work relevant to our present study is stated below. The time-variant impulse response of a recursive LTV difference equation given in (3) is a Kth-order causal separable sequence of the form

$$h(n,m) = \begin{cases} \sum_{i=1}^{K} u_i(n)v_i(m), & n \geq m \\ 0, & \text{elsewhere} \end{cases} \tag{6}$$

where $u_i(n)$, $i = 1, 2, \cdots, K$ are K independent solutions of

$$\sum_{i=0}^{K} a_i(n)y(n-i) = 0 \tag{7}$$

and $v_i(m)$ is given by

$$v_i(m) = \sum_{k=0}^{L} b_k(m+k)\frac{D_i(m+k)}{D(m+k)} \tag{8}$$

where $D(m)$ denotes the determinant of

$$\begin{bmatrix} u_1(m) & u_2(m) & \cdots & u_{K-1}(m) & u_K(m) \\ u_1(m-1) & u_2(m-1) & \cdots & u_{K-1}(m-1) & u_K(m-1) \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ u_1(m-K+1) & & \cdots & u_{K-1}(m-K+1) & u_K(m-K+1) \end{bmatrix}. \tag{9}$$

and $D_i(m)$ is the cofactor of the element $u_i(m)$ of the determinant $D(m)$. An implicit formulation of this result can also be found in [8]. However, after careful examination, we find that the expression in (6) holds true for all n and m only when $L \leqslant K-1$ and $a_K(n) \neq 0$ for all n. Therefore, we suspect that there exists a more generalized expression for the impulse response of an LTV difference equation. In addition to the time-variant difference equation, we are also interested in analyzing LTV filters realized with other filter structures. Consequently, the development in this paper is based upon the state-variable model for LTV filters. The detailed derivations associated with the state-variable model are discussed in the following section.

III. Diagonalizing Transformation for State-Variable Model

The technique of state transformation is a very convenient tool for analyzing linear models represented in the state-variable form. In most cases, we want to reduce the state-feedback matrix to a diagonal matrix or a diagonal-shape matrix consisting of Jordan blocks. Methods for selecting state transformation matrices for a continuous-time system have been discussed in [10]. But these procedures are not applicable to discrete-time filters having singular state-transition matrices. Therefore, we devote our effort to develop a new procedure for selecting transformation matrices for a general discrete-time LTV filter.

Let us first restrict our consideration to a single-input single-output LTV filter. For the Kth-order filter given by (5), the matrices $\boldsymbol{A}(n)$, $\boldsymbol{B}(n)$ and $\boldsymbol{C}(n)$ are of dimension $K \times K$, $K \times 1$, and $1 \times K$, respectively. After the input $x(n)$ in (5) is substituted with a shifted unit impulse $\delta(n-m)$, the impulse response of the filter can be derived as

$$h(n,m) = \begin{cases} \boldsymbol{C}(n)\left[\prod_{j=0}^{n-m-1} \boldsymbol{A}(n-j)\right]\boldsymbol{B}(m), & n \geqslant m+1 \\ \boldsymbol{C}(n)\boldsymbol{B}(n), & n = m \\ 0, & n < m. \end{cases} \tag{10}$$

Because the impulse response given in (10) is a complicated expression involving multiple matrix multiplications, it is difficult to examine the properties of the impulse response without further simplification.

One way to circumvent this difficulty is to select a time-dependent state transformation that can transform all the state-feedback matrices $\boldsymbol{A}(n)$ into diagonal matrices. Assume the new state vector $\boldsymbol{W}^*(n)$ is related to the original state vector $\boldsymbol{W}(n)$ by

$$\boldsymbol{W}(n) = \boldsymbol{P}(n)\boldsymbol{W}^*(n) \tag{11}$$

where $\boldsymbol{P}(n)$ is a nonsingular transformation matrix. After substituting (11) into the original state equation (5) and rearranging the result, we can define an equivalent filter characterized by the state equations

$$\begin{aligned} \boldsymbol{W}^*(n) &= \boldsymbol{A}^*(n)\boldsymbol{W}^*(n-1) + \boldsymbol{B}^*(n)x(n) \\ y(n) &= \boldsymbol{C}^*(n)\boldsymbol{W}^*(n) \end{aligned} \tag{12}$$

where

$$\begin{aligned} \boldsymbol{A}^*(n) &= \boldsymbol{P}^{-1}(n)\boldsymbol{A}(n)\boldsymbol{P}(n-1) \\ \boldsymbol{B}^*(n) &= \boldsymbol{P}^{-1}(n)\boldsymbol{B}(n) \\ \boldsymbol{C}^*(n) &= \boldsymbol{C}(n)\boldsymbol{P}(n). \end{aligned} \tag{13}$$

Consequently, the expression for the impulse response given in (10) can also be rewritten for the equivalent filter as

$$h(n,m) = \begin{cases} \boldsymbol{C}^*(n)\left[\prod_{j=0}^{n-m-1} \boldsymbol{A}^*(n-j)\right]\boldsymbol{B}^*(m), & n \geqslant m+1 \\ \boldsymbol{C}^*(n)\boldsymbol{B}^*(n), & n = m \\ 0, & n < m. \end{cases} \tag{14}$$

If all the state-feedback matrices $\boldsymbol{A}^*(n)$ are of diagonal form, then the original filter has been decomposed into K parallel first-order filters. In this way, we can represent each recursive LTV filter in terms of an equivalent diagonalized filter. Hence the diagonalizing transformation is an useful tool for the analysis of recursive LTV filters of *different structures*. Next, we will show that such a diagonalizing transformation exists for every recursive LTV filter represented in the state-variable form.

Assuming that the system matrices of an LTV filter are defined for all n such that $M_0 \leqslant n \leqslant N_0$, we should be able to choose the transformation matrices $\boldsymbol{P}(n)$ for $M_0 - 1 \leqslant n \leqslant N_0$ in order to completely define the equivalent filter. Let us define the forward state-transition matrix $\Phi(n,m)$ of the original filter as

$$\Phi(n,m) = \begin{cases} \prod_{j=0}^{n-m-1} \boldsymbol{A}(n-j), & n > m \\ \boldsymbol{I}, & n = m \end{cases} \tag{15}$$

and let $S_j(n)$ denote the linear vector space consisting of all the column vector q such that

$$\Phi(n+j,n)q = 0. \tag{16}$$

We also choose $\boldsymbol{A}(M_0-1) = \boldsymbol{A}(N_0+1) = \boldsymbol{0}$ to facilitate the illustration of this algorithm. Then the transformation matrices $\boldsymbol{P}(n)$ may be chosen in sequence by carrying out the following steps from $n = M_0 - 1$ to $n = N_0$:

1) Let $P_i(n)$ denote the ith column of the transformation matrix $\boldsymbol{P}(n)$. Then choose $P_i(n) = \boldsymbol{A}(n)P_i(n-1)$ if $\boldsymbol{A}(n)P_i(n-1) \neq 0$, where $i = 1, 2, \cdots, K$. Let $I(n)$ denote the set of index i such that $\boldsymbol{A}(n)P_i(n-1) \neq 0$ and let $\Omega(n)$ denote the linear expansion of the column vectors $P_i(n)$ for $i \in I(n)$. The rest of the column vectors in $\boldsymbol{P}(n)$ are to be determined in the next step.

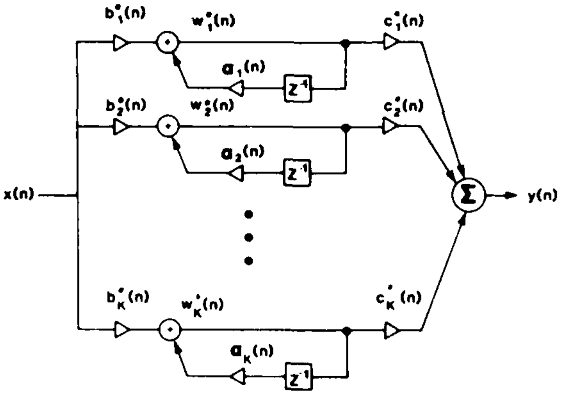


Fig. 2 Block diagram of the equivalent filter having a diagonal state-feedback matrix.

2) For $j = 1, 2, \cdots, N_0 - n + 1$, define

$$S_j^T(n) = \text{the linear space generated by the vectors in } S_j(n) \cup \Omega(n) \quad (17)$$

and

$$\begin{aligned} T_j(n) &= S_j(n) \perp S_{j-1}^T(n) \\ &= \{ q \mid q \in S_j(n) \text{ and } q^T p = 0 \\ &\qquad \text{for every } p \in S_{j-1}^T(n) \}. \end{aligned} \quad (18)$$

Let $\{ g_{k,j}(n) \mid 1 \leqslant k \leqslant K_j(n) \}$ be a basis of the subspace $T_j(n)$ and $\{ \eta_{k,j}(n),\ k = 1, 2, \cdots, K_j(n) \}$ be an arbitrary vector set in $S_{j-1}(n)$. For $j = 1, 2, \cdots, N_0 - n + 1$, we choose $g_{k,j}(n) + \eta_{k,j}(n)$, $1 \leqslant k \leqslant K_j(n)$ as column vectors of the transformation matrix $\boldsymbol{P}(n)$ if $T_j(n) \neq \{0\}$. The column vectors selected in this step will exactly fill the empty columns of $\boldsymbol{P}(n)$ which have not been determined in step 1.

The basic properties of the state transformation defined by the procedure stated above can be summarized in the following two theorems. The proofs are given in the Appendix.

Theorem 1: Assuming that a discrete LTV filter represented in the state-variable form of (5) is defined in the interval $M_0 \leqslant n \leqslant N_0$, then the state transformation matrix $\boldsymbol{P}(n)$ selected with the iterative procedure stated above has the following properties:

1) The matrix $\boldsymbol{P}(n)$ is of full rank.

2) Each column of $\boldsymbol{P}(n)$ belongs to one of the disjointed sets

$$\begin{aligned} \Delta S_j(n) &= S_j(n) - S_{j-1}(n) \\ &= \{ q \mid q \in S_j(n) \text{ and } q \notin S_{j-1}(n) \} \end{aligned} \quad (19)$$

where $j = 1, 2, \cdots, N_0 - n + 1$. The number of the column vectors of $\boldsymbol{P}(n)$ belonging to ΔS_j is equal to the difference between the dimensions of $S_j(n)$ and $S_{j-1}(n)$.

Theorem 2: Under the same assumptions of Theorem 1, there exists an equivalent filter given in (12) such that the new state-feedback matrix is of the form

$$\boldsymbol{A}^*(n) = \operatorname{diag}[\alpha_1(n) \quad \alpha_2(n) \quad \cdots \quad \alpha_K(n)] = \begin{bmatrix} \alpha_1(n) & & & \boldsymbol{0} \\ & \alpha_2(n) & & \\ & & \ddots & \\ \boldsymbol{0} & & & \alpha_K(n) \end{bmatrix} \quad (20)$$

where $\alpha_i(n) = 0$ or 1 for $i = 1, 2, \cdots, K$.

The essence of these two theorems is that once an LTV filter of the state variable form is completely specified in an interval, there exists a time-dependent state transformation which can define an equivalent filter having diagonal state-feedback matrices in the same interval. However, the state transformation thus determined is not unique because in step 2 of the selection procedure, we have some freedom in choosing the column vectors and assigning them to respective column positions. Even though selecting different sets of transformation matrices may generate different diagonalized filters, the input–output characteristics of these filters remain the same.

A general expression for the impulse response can be derived from the result in Theorem 2. Since the state-feedback matrix of the equivalent filter has the form of (20), the corresponding forward state-transition matrix of the equivalent filter is given as

$$\Phi^*(n,m) = \operatorname{diag}[\beta_1(n,m) \quad \beta_2(n,m) \quad \cdots \quad \beta_K(n,m)] \quad (21)$$

where

$$\beta_i(n,m) = \begin{cases} \prod_{j=0}^{n-m-1} \alpha_i(n-j), & n \geqslant m+1 \\ 1, & n = m \\ 0, & n < m \end{cases} \quad (22)$$

for $i = 1, 2, \cdots, K$. After substituting (21) into (14), we can derive the impulse response of the LTV filter as

$$\begin{aligned} h(n,m) &= \boldsymbol{C}^*(n)\Phi^*(n,m)\boldsymbol{B}^*(m) \\ &= \sum_{i=1}^{K} c_i^*(n)\beta_i(n,m)b_i^*(m) \end{aligned} \quad (23)$$

where $\boldsymbol{C}^*(n) = \boldsymbol{C}(n)\boldsymbol{P}(n)$, $\boldsymbol{B}^*(m) = \boldsymbol{P}^{-1}(m)\boldsymbol{B}(m)$ and $c_i^*(n)$ and $b_i^*(m)$ denote ith elements of $\boldsymbol{C}^*(n)$ and $\boldsymbol{B}^*(m)$, respectively. This result can also be easily verified by examining the block diagram of the equivalent LTV filter shown in Fig. 2.

The expression for the time-variant impulse response shown in (23) somewhat resembles the result given in (6), which is obtained specifically for time-variant difference equations. However, an additional weighting factor $\beta_i(n,m)$, which is a function of m and n, has been included in the summation of (23). Further, for some LTV filters, certain terms in (23) might cancel. This leads to the situation that fewer terms in (23) really contribute to the impulse response. This property will be illustrated later by applying the diagonalization procedure to an LTV filter realized as a time-variant difference equation.

IV. Analysis of LTV Filters via Diagonalizing Transformation

Now, the usefulness of this diagonalizing transformation is illustrated by applying it to several realization schemes for recursive LTV filters. The analysis of LTV filters using the diagonalizing transformation is usually accomplished by the following steps. First, we represent an LTV filter in the state-variable form. Then we can obtain a set of transformation matrices by using the diagonalizing procedure discussed in Section III. The properties of the original filter can be extracted from the system matrices of the corresponding diagonalized filter.

The first example is an LTV filter whose state-feedback matrix $A(n)$ is nonsingular for all n such that $M_0 \leqslant n \leqslant N_0$. Following the selection procedure discussed in Section III, we can choose the transformation matrix $P(M_0 - 1)$ as any nonsingular $K \times K$ matrix P_0. Then, just by carrying out step 1 of the selection procedure, we can select the rest of the transformation matrices as

$$P(n) = \left\{ \prod_{j=0}^{n-M_0} A(n-j) \right\} P_0, \qquad \text{for } M_0 \leqslant n \leqslant N_0. \tag{24}$$

After substituting (24) into (13), we can derive the system matrices of the equivalent filter as

$$\begin{aligned} A^*(n) &= P^{-1}(n)A(n)P(n-1) = I \\ B^*(n) &= P^{-1}(n)B(n) \\ C^*(n) &= C(n)P(n) \end{aligned} \tag{25}$$

where $M_0 \leqslant n < N_0$. The impulse response of such a filter can be obtained from (23) as

$$h(n,m) = \begin{cases} \sum_{i=1}^{K} c_i^*(n) b_i^*(m), & n \geqslant m \\ 0, & n < m. \end{cases} \tag{26}$$

A typical example of such an LTV filter is the direct form II realization of the recursive LTV filter given in (4).

The second example is the time-variant difference equation shown in (3), where $a_0(n)$ is assumed to be 1. In this case, the state vector may be chosen as

$$\begin{aligned} W(n) = [&y(n-K+1) \ \cdots \ y(n-1) \ \ y(n) \mid \\ &x(n-L+1) \ \cdots \ x(n-1) \ \ x(n)]^T \end{aligned} \tag{27}$$

Note that the dimension of the state vector is $K + L$. Then, the corresponding system matrices of the filter are found to be

$$\begin{aligned} B(n) &= [0 \ \cdots \ 0 \ \ b_0(n) \mid 0 \ \ 0 \ \cdots \ 0 \ \ 1]^T \\ C(n) &= [0 \ \ 0 \ \cdots \ 1 \mid 0 \ \ 0 \ \cdots \ 0] \end{aligned}$$

and

$$A(n) = \left[\begin{array}{c|c} A_{11}(n) & A_{12}(n) \\ \hline A_{21}(n) & A_{22}(n) \end{array} \right] \tag{28}$$

where

$$A_{11}(n) = \begin{bmatrix} 0 & 1 & 0 & & \mathbf{0} \\ & & 1 & & \\ & & & \ddots & \\ & \mathbf{0} & & & 1 \\ -a_K(n) & & \cdots & -a_2(n) & -a_1(n) \end{bmatrix}$$

$$A_{12}(n) = \begin{bmatrix} & & \mathbf{0} & \\ b_L(n) & \cdots & b_2(n) & b_1(n) \end{bmatrix}$$

$$A_{21}(n) = [0]$$

$$A_{22}(n) = \begin{bmatrix} 0 & 1 & & \mathbf{0} \\ & & 1 & \\ & & & \ddots \\ & \mathbf{0} & & 1 \\ & & & 0 \end{bmatrix}.$$

By using the diagonalization procedure discussed earlier, the transformation matrix $P(n)$ may be chosen as

$$P(n) = \left[\begin{array}{c|c} P_{11}(n) & P_{12}(n) \\ \hline \mathbf{0} & I_S(n-M_0) \end{array} \right] \tag{29}$$

where

$$P_{11}(n) = \begin{cases} \prod_{j=0}^{n-M_0} A(n-j), & n \geqslant M_0 \\ I, & n = M_0 - 1 \end{cases}$$

$$I_S(j) = [I_S(0)]^{j+1}$$

$$I_S(0) = \begin{bmatrix} 0 & 1 & & \mathbf{0} \\ & & \ddots & \\ 0 & \mathbf{0} & & 1 \\ 1 & 0 & & 0 \end{bmatrix}$$

$$\begin{aligned} P_{12}(n) = &- P_{11}(n) \sum_{i=1}^{\text{MIN}} \left[\sum_{j=1}^{i} P_{11}^{-1}(n+j) b_{L-((i-j))}(n+j) \right] \\ &\cdot \Delta[L, ((n+i-M_0))+1] \\ &+ Q(n) \left[I - \sum_{i=1}^{\text{MIN}} I_0(n+i-M_0) \right] \end{aligned}$$

$$Q(n) = \begin{cases} P_{11}(n) \sum_{i=M_0}^{n} P_{11}^{-1}(i) A_{12}(i) I_S(i-M_0-1), & \\ & N_0 \geqslant n \geqslant M_0 \\ \mathbf{0}, & n = M_0 - 1 \end{cases}$$

$$\text{MIN} = \text{minimum}\{L, N_0 - n\}$$

$$I_0(j) = I_S^{-1}(j) \Delta[K, K] I_S(j)$$

where $\Delta[i, j]$ is an $L \times L$ matrix having a single nonzero element 1 at the position (i, j) and $((i))$ denotes the L-modulus of i.

After substituting (29) into (13), we obtain the corresponding system matrices of the equivalent diagonalized filter as

$$\begin{aligned} \boldsymbol{A}^*(n) &= \boldsymbol{P}^{-1}(n)\boldsymbol{A}(n)\boldsymbol{P}(n-1) \\ &= \left[\begin{array}{c|c} \boldsymbol{I} & \boldsymbol{0} \\ \hline \boldsymbol{0} & \boldsymbol{I}-\boldsymbol{I}_0(n-M_0) \end{array}\right] \\ \boldsymbol{B}^*(n) &= \boldsymbol{P}^{-1}(n)\boldsymbol{B}(n) \\ \boldsymbol{C}^*(n) &= \boldsymbol{C}(n)\boldsymbol{P}(n). \end{aligned} \tag{30}$$

Then the impulse response of this filter can be obtained by substituting (30) into (23). After some simplification of the resulting expression, the impulse response is of the form

$$h(n,m) = \begin{cases} \displaystyle\sum_{j=0}^{\min\{L,\,n-m\}} [0\cdots0 \quad 1]\boldsymbol{P}_{11}(n)\boldsymbol{P}_{11}^{-1}(m+j) \\ \quad b_j(m+j)[0\cdots0 \quad 1]^T, & n \geqslant m \\ 0, & n < m \end{cases} \tag{31}$$

which is somewhat different from the expression given in (6). But, with further simplification, it can be shown that (31) is equivalent to (6) for $L \leqslant K-1$. From the definition of $\boldsymbol{P}_{11}(n)$ given in (29), it is clear that the upper $K-1$ rows of $\boldsymbol{P}_{11}(n)$ are obtained by shifting each of the corresponding rows of $\boldsymbol{P}_{11}(n-1)$ upward by one row. This leads to the relation

$$[0 \quad 0\cdots1]\boldsymbol{P}_{11}(n)\boldsymbol{P}_{11}^{-1}(m+j)[0 \quad 0\cdots1]^T = 0 \tag{32}$$

for $n-m+1 \leqslant j \leqslant n-m+K-1$. After substituting (32) for $n-m+1 \leqslant j \leqslant L$ into (31), we have

$$h(n,m) = \begin{cases} \displaystyle\sum_{j=0}^{L} b_j(m+j)[0\cdots1]\boldsymbol{P}_{11}(n)\boldsymbol{P}_{11}^{-1}(m+j) \\ \quad [0\cdots1]^T, & n \geqslant m \\ 0, & n < m. \end{cases} \tag{33}$$

This expression is equivalent to the result given in (6).

From the above discussion, we have shown that the diagonalizing transformation is a valuable tool for the analysis of digital LTV filters. Therefore, a systematic approach based upon the diagonalizing transformation can be developed to reduce the complexity of the LTV filters synthesized with a variety of structures. Since the original filter is state-to-state equivalent to the corresponding diagonalized filter, the use of the diagonalizing transformation is promising in such areas as the stability analysis and the roundoff noise analysis of LTV filters.

V. Synthesis of Recursive LTV Filters

In this section, we examine the time-domain solutions to the deterministic synthesis problem of recursive LTV filters. Our main objective is to develop techniques for synthesizing a desired time-variant impulse response with a recursive LTV filter such that the difference between the desired and the synthesized impulse responses is minimized according to some error criterion. In a practical problem, the desired impulse response may be determined either by solving a statistical filtering problem or by using some empirical rules. Since the computation for obtaining the true optimal solution grows rapidly as the order of the filter or the duration of the impulse response increases, it is necessary to develop efficient solutions for the synthesis of recursive LTV filters. In this section, we introduce an efficient suboptimal technique which is based upon the minimization of the squared difference between the desired impulse response and the synthesized impulse response in localized regions. A numerical example is selected to illustrate this new synthesis technique.

It is quite difficult to synthesize a desired impulse response in terms of impulse responses given in (23) that have arbitrary $\beta_i(n,m)$'s. In order to obtain a manageable formulation of the synthesis problem, we restrict our consideration to a subset of recursive LTV filters whose impulse responses are causal Kth-order separable functions, i.e.,

$$h(n,m) = \begin{cases} \displaystyle\sum_{i=1}^{K} u_i(n)v_i(m), & n \geqslant m \\ 0, & \text{elsewhere} \end{cases} \tag{34}$$

where $\{u_i(n);\ i=1,\cdots,K\}$ and $\{v_i(m);\ i=1,\cdots,K\}$ are two sets of independent functions. Even though (34) is a special case of the expression given in (23), (34) is still capable of representing a major class of the recursive LTV filters. In particular, the impulse responses of most recursive LTV filters discussed in Section IV satisfy (34). Assuming that the desired causal impulse response is $h_D(n,m)$, the synthesis of the recursive LTV filter can then be formulated as the minimization of the squared error function

$$D(\Theta) = \sum_{n=0}^{N}\sum_{m=0}^{n}\left[h_D(n,m) - \sum_{i=1}^{K} u_i(n)v_i(m)\right]^2 \tag{35}$$

where $\Theta = \{u_i(n), v_i(n) | 1 \leqslant i \leqslant K,\ 0 \leqslant n \leqslant N\}$ is a set of unknown variables and K is the order of the recursive LTV filter. Once we determine the optimal Θ which minimizes the cost function, the coefficients of the recursive LTV filter synthesized with a particular filter structure can be derived from the optimal Θ. An unrestricted nonlinear optimization algorithm may be applied to find the solution of this optimization problem. The coefficients $u_i(n)$'s and $v_i(m)$'s are considered as the unknown variables in the optimization process. However, since the number of unknown parameters is proportional to the order of the filter times the duration of the impulse response, it is impractical to find the true optimal solution of (35) for an LTV filter with a large K or a large N. Therefore, it is useful to develop some efficient suboptimal techniques for solving this nonlinear minimization problem.

In Huang and Aggarwal's work [11], the causal condition in (35) is removed so that a straightforward algorithm can be applied to solve this filter synthesis problem. The new error function is defined as

$$D'(\Theta) = \sum_{n=0}^{N} \sum_{m=0}^{N} \left[h_D(n,m) - \sum_{i=1}^{K} u_i(n) v_i(m) \right]^2. \quad (36)$$

The minimization of (36) can easily be solved by using a procedure that was originally developed for finding the spectral decomposition of a matrix [12]. The required computation is approximately equivalent to that of obtaining K most dominant eigenvalues and eigenvectors of an $(N+1)\times(N+1)$ matrix.

The main drawback of using the noncausal error function is that the performance of this synthesis procedure depends largely on how we choose the desired impulse response in the noncausal region. In order to make the result obtained by the spectral decomposition technique close to the optimal one, the function $h_D(n,m)$ in the region $\{(n,m); 0 \leqslant n \leqslant m \leqslant N\}$ must be selected such that

$$\sum_{n=0}^{N} \sum_{m=n+1}^{N} [h_D(n,m) - h_{\text{OPT}}(n,m)]^2 \simeq 0 \quad (37)$$

where $h_{\text{OPT}}(n,m)$ is the noncausal separable impulse response defined by the optimal Θ that minimizes (35). Since there are no explicit rules for making a good guess of the noncausal part of the desired impulse response, the spectral decomposition technique often achieves less than satisfactory result. It usually takes a high-order recursive LTV filter to make a good approximation of the desired impulse response.

To circumvent the problems in the nonlinear optimization technique and the spectral decomposition technique, a new suboptimal technique for solving the synthesis problem of recursive LTV filters is formulated by minimizing (35) in a localized sense. The solution of $u_i(n)$ for $i = 1,\cdots,K$ is obtained by minimizing the localized error function

$$D_l(n) = \sum_{m=0}^{n} \left[h_D(n,m) - \sum_{i=1}^{K} u_i(n) v_i(m) \right]^2 \quad (38)$$

under the conditions that

$$h_D(n,m) = \sum_{i=1}^{K} u_i(n) v_i(m),$$
$$\text{for } m = n, n-1, \cdots, n-K+1. \quad (39)$$

With an index change in (39), we can express the linear equations for obtaining $v_i(m)$, $i = 1,\cdots,K$ as

$$h_D(m+j,m) = \sum_{i=1}^{K} u_i(m+j) v_i(m),$$
$$\text{for } j = 0,1,\cdots,K-1. \quad (40)$$

After substituting (39) into (38), we have that

$$D_l(n) = \sum_{m=0}^{n-K} \left[h_D(n,m) - \sum_{i=1}^{K} u_i(n) v_i(m) \right]^2. \quad (41)$$

Differentiating (41) with respect to $u_i(n)$ and setting the result equal to zero, we can find a set of linear equations for $u_i(n)$'s

$$\sum_{k=1}^{K} u_k(n) \left[\sum_{m=0}^{n-K} v_i(m) v_k(m) \right]$$
$$= \sum_{m=0}^{n-K} h_D(n,m) v_i(m), \qquad i = 1,\cdots,K. \quad (42)$$

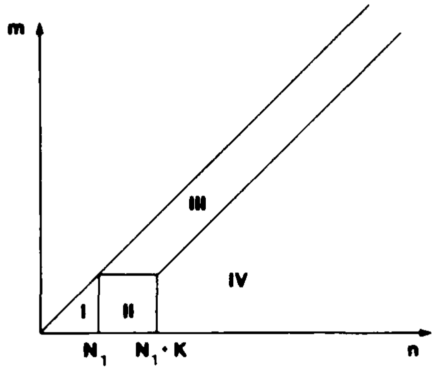


Fig. 3. Regions of the impulse response where the error function is evaluated in different steps of the localized technique.

Note that the coefficients $u_k(n)$'s in (42) can be easily calculated once we determine the values of $v_i(m)$ for $i = 1,\cdots,K$ and $m = 0,1,\cdots,n-K$. Therefore, by using (40) and (42) iteratively, we can determine a suboptimal set of $u_i(n)$'s and $v_i(n)$'s.

The constraints given in (39) ensure that the synthesized impulse response is the same as the desired one in the region $0 \leqslant n - m \leqslant K-1$. Hence, the localized synthesis technique will favor the synthesis of a impulse response having dominant components along the diagonal line $n = m$. Now let us summarize the complete algorithm of this suboptimal solution as follows:

1) Apply the nonlinear optimization technique to find the optimal solution of (35) in a small interval $[0, N_1]$. For example, the Fletcher-Powell technique [13] works well for a small N_1.

2) Use (42) to find $u_i(n)$ for $n = N_1+1,\cdots,N_1+K$ and $i = 1,\cdots,K$.

3) For $j = 1,2,\cdots,N-N_1$, use (40) and (42) to obtain the solutions for $v_i(N_1+j)$ and $u_i(N_1+K+j)$, $i = 1,\cdots,K$.

In fact, this localized minimization technique has divided the domain of the impulse response into four distinct regions as shown in Fig. 3. Regions I and II denote, respectively, the areas where steps 1 and 2 of the algorithm are applied. And regions III and IV represent, respectively, the areas where (40) and (42) are applied at step 3 of the algorithm. The computation of this localized minimization algorithm at each sampling point is equivalent to that of solving two Kth-order linear equations.

We have tested both the spectral decomposition technique and the localized minimization technique with a variety of time-variant impulse responses. The results show

TABLE I
PERFORMANCE INDICES OF THE SYNTHESIS ALGORITHMS FOR LTV FILTERS

	K	$\bar{D}(K)$	
γ		SPECTRAL DECOMPOSITION	LOCALIZED MINIMIZATION
0.1	1	0.74114	0.17202
	2	0.58230	0.01803
	3	0.46670	0.00552
	4	0.36559	0.00273
0.05	1	0.88489	0.19549
	2	0.77337	0.03869
	3	0.68461	0.00403
	4	0.60011	0.00074

that the localized minimization technique consistently performs better than the spectral decomposition technique. A numerical example is selected to illustrate this situation. The impulse response of the desired filter is chosen as

$$h_D(n,m) = h_c(n/16, m/16) \tag{43}$$

where

$$h_c(t,\tau) = \begin{cases} \exp(-[0.02(t-\tau)^2 + 0.1t] \\ \quad \operatorname{sinc}[2(t-\tau)(1-\gamma t)], & t \geqslant \tau \\ 0, & \text{elsewhere} \end{cases} \tag{44}$$

and $\operatorname{sinc}(x) = \sin(\pi x)/(\pi x)$. The domain of the desired impulse response under consideration is limited to the region where $0 \leqslant n \leqslant 128$ and $0 \leqslant m \leqslant 128$. Basically, the desired impulse response is a truncated sinc function along the axis $n + m = 0$, but with time-variant frequency contents and an accelerated decay. A normalized squared error function

$$\hat{D}(K) = \left[\sum_{n=0}^{128} \sum_{m=0}^{n} (h(n,m) - h_D(n,m))^2\right] \Bigg/ \left[\sum_{n=0}^{128} \sum_{m=0}^{n} h_D^2(n,m)\right] \tag{45}$$

is selected as the measure of the performance of the synthesis algorithm. A lower value of the normalized error function means a better approximation of the desired impulse response. Two sets of the normalized errors have been obtained: the first set is for the spectral decomposition technique in which the noncausal half of the impulse response is assumed to be symmetrical to the causal half; and the second set is for the localized minimization technique. Table I lists the normalized error indices obtained for the desired impulse response functions with γ values 0.1 and 0.05. The desired impulse response with $\gamma = 0.05$, the impulse response of a fourth-order filter synthesized with the localized minimization technique, and the difference between the two are shown in Fig. 4.

It is clearly seen from Table I that the localized synthesis technique achieves much better results than the spectral

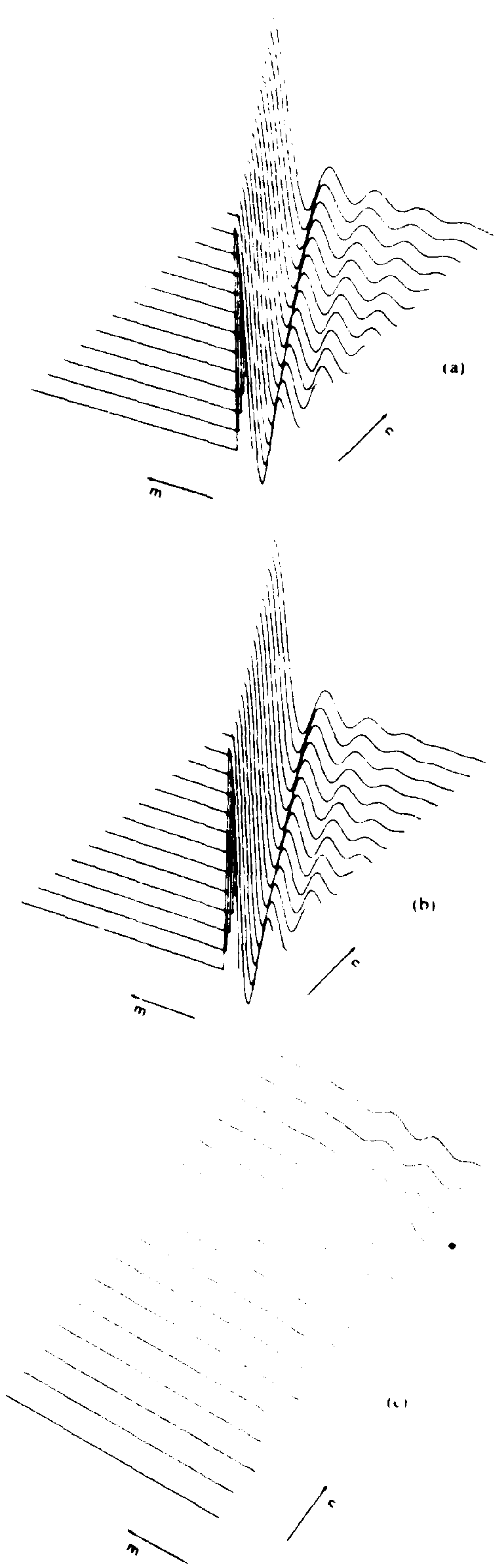


Fig. 4(a) Desired impulse response for $\gamma = 0.05$ in the numerical example. (b) Impulse response synthesized with a fourth-order LTV filter using the localized minimization technique. (c) Difference between the desired impulse response and the synthesized impulse response.

decomposition technique does. We also attempt to find the optimal solution of minimizing (35) with a general nonlinear minimization algorithm so that we can evaluate the results obtained by those suboptimal techniques. After experimenting with a number of initial conditions, we find that the performance index settles at a much larger value than that obtained by the localized technique. This seems to confirm the conjecture that it is quite difficult to use a general nonlinear minimization technique for finding the optimal solution of a synthesis problem which has moderate values of filter order K and impulse response duration N.

VI. Conclusions

It has been demonstrated that the diagonalizing transformation is a helpful tool for the analysis and synthesis of recursive LTV filters. Because there is a state-to-state equivalence between the original filter and the corresponding diagonalized filter, the diagonalizing transformation can be useful in other research areas related to the recursive LTV systems. In particular, the observability and the controllability properties of a recursive LTV filter are easily obtainable from the system matrices of the equivalent diagonalized filter. Since the diagonalizing transformation procedure introduced in this paper works only for LTV filters defined in a finite interval, further investigation is warranted on whether we can generalize the diagonalizing procedure such that it is applicable to filters defined in an infinite interval.

Appendix

Proof for Theorem 1

It is easily seen that $S_j(n)$, $S_j^I(n)$ and $T_j(n)$ are linear subspaces of R^K for $j = 0, 1, \cdots, N_0 - n + 1$, and that $\{0\} = S_0(n) \subseteq S_1(n) \subseteq \cdots \subseteq S_{N_0-n+1}(n) = R^K$. Then, $\{\Delta S_j(n), j = 0, 1, \cdots, N_0 - n\}$ becomes a collection of disjointed sets. Further, if $\{g_{j1}(n), g_{j2}(n), \cdots, g_{jk}(n)\}$ is a basis of $T_j(n)$ and any set $\{\eta_{j1}(n), \eta_{j2}(n), \cdots, \eta_{jk}(n)\}$ is contained in $S_{j-1}(n)$, then $\{g_{j1}(n) + \eta_{j1}(n), g_{j2}(n) + \eta_{j2}(n), \cdots, g_{jk}(n) + \eta_{jk}(n)\}$ is a collection of independent column vectors in the set $\Delta S_j(n)$. And the number of vectors in this collection is equal to the dimension of the subspace $T_j(n)$, which is equivalent to the difference between the dimension of $S_j(n)$ and the dimension of the intersection of $S_{j-1}^I(n)$ and $S_j(n)$. Furthermore, it is clear that the vector set $\{g_{j1}(n) + \eta_{j1}(n), g_{j2}(n) + \eta_{j2}(n), \cdots, g_{jk}(n) + \eta_{jk}(n)\}$ together with any set of independent vectors in $S_{j-1}^I(n)$ constitute a new independent vector set.

Then, Theorem 1 can be shown by induction. At $n = M_0 - 1$, we have that $A(M_0 - 1) = 0$ which leads to that $\Omega(M_0 - 1) = \{0\}$ and that $S_j^I(M_0 - 1) = S_j(M_0 - 1)$. Since $A(M_0 - 1)q = 0$ for all q, all the column vectors of $P(M_0 - 1)$ are selected in step 2 of the diagonal transformation procedure. From the discussion in the previous paragraph we know that each column of $P(M_0 - 1)$ belongs to one of the disjointed sets $\Delta S_j(n)$, $j = 1, \cdots, N_0 - n + 1$, and these columns are independent. Because the dimension of the set $T_j(M_0 - 1)$ is the difference between the dimensions of $S_j(M_0 - 1)$ and $S_{j-1}(M_0 - 1)$, it is clear that the number of column vectors of $P(M_0 - 1)$ belonging to $\Delta S_j(M_0 - 1)$ is equal to the difference between the dimensions of the subspaces $S_j(M_0 - 1)$ and $S_{j-1}(M_0 - 1)$.

Assume the statements in Theorem 1 are true for $n = N$, we need to show that they are also true for $n = N + 1$. In step 1 of the selection procedure, we choose $P_j(N + 1) = A(N + 1)P_j(N)$ if $A(N + 1)P_j(N) \neq 0$. Assume that the dimension of the subspace $S_1(N)$ is equal to K_1. Since $S_1(N) - \{0\}$ is equivalent to the set $\Delta S_1(N)$, there are K_1 column vectors of $P(N)$ belonging to $S_1(N)$. Therefore, there are $K - K_1$ column vectors of $P(N + 1)$ selected in the step 1. And these $K - K_1$ vectors must be independent because the matrix $A(N + 1)$ has rank $K - K_1$ and matrix $P(N)$ has rank K. Further, each of these $K - K_1$ vectors selected in step 1 satisfy one of the conditions

$$\Phi(N + j + 1, N + 1)q = 0 \quad \text{and} \quad \Phi(N + j, N + 1)q \neq 0 \tag{46}$$

for $j = 1, 2, \cdots, N_0 - N$. Therefore, each of the vectors selected in step 1 belongs to one of the sets $\Delta S_j(N + 1)$, $j = 1, 2, \cdots, N_0 - N$ and the number of vectors belonging to each set is equal to the difference between the dimension of the intersection of the sets $S_{j-1}^I(N + 1)$ and $S_j(N + 1)$ and the dimension of the set $S_{j-1}(N + 1)$. In the step 2 of the selection procedure, each column vector selected also belongs to one of the sets $\Delta S_j(N + 1)$ for $j = 1, 2, \cdots, N_0 - N$. From the discussion at the beginning of this proof, the number of column vectors belonging to $\Delta S_j(N + 1)$ is the difference between and the dimension of $S_j(N + 1)$ and the dimension of the intersection of $S_{j-1}^I(N + 1)$ and $S_j(N + 1)$.

After reviewing the properties of the column vectors in both steps of the selection process, we have that the number of column vectors belonging to $\Delta S_j(N + 1)$ is equal to the difference between the dimensions of $S_j(N + 1)$ and $S_{j-1}(N + 1)$ for $j = 1, 2, \cdots, N_0 - N$. Further, the new vectors selected at each stage of step 2, together with the column vectors of $P(N + 1)$ previously chosen, still constitute a set of independent vectors. Thus the matrix $P(N + 1)$ is a nonsingular matrix. Now we have shown that the statements in Theorem 1 holds true for the matrix $P(N + 1)$. By induction, the statement in Theorem 1 is true for all n such that $M_0 - 1 \leq n \leq N_0$.

Proof for Theorem 2

By using the result obtained in Theorem 1, we can find the transformation matrices $P(n)$ for $n = M_0 - 1, M_0, \cdots, N_0$, and these matrices have the properties stated in Theorem 1. Assume that matrix $A(n)$ has rank K_n. For $M_0 < n < N_0$, we have that there are K_n independent column vectors $P_j(n - 1)$ such that

$$P_j(n) = A(n)P_j(n - 1) \neq 0 \tag{47}$$

and the rest of the columns satisfy that

$$A(n)P_j(n - 1) = 0 \tag{48}$$

From these two equations, we have that $P_i^{-1}(n)A(n)P_i(n-1) = \Delta[i]$ if $A(n)P_i(n-1) \neq 0$ and $P_i^{-1}(n)A(n)P_i(n-1) = 0$ if $A(n)P_i(n-1) = 0$, where $\Delta[i]$ denotes a column vector having the only nonzero element 1 at its ith position. Therefore, the equivalent feedback matrix corresponding to the time-variant state transformation obtained in Theorem 1 is a diagonal matrix with the diagonal elements of the value 0 or 1.

References

[1] N. C. Huang and J. K. Aggarwal, "On linear shift-variant digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-27, pp. 672-679, Aug. 1980.

[2] Z. J. Nikolic, "A recursive time-varying band-pass filter," *Geophysics*, vol. 40, pp. 520-526, June 1975.

[3] R. A. Stein, "Design of recursive digital filters with continuously variable passbands," *Proc. IEEE Int. Symp. on Circuits and Systems* (Chicago, IL), pp. 424-427, Apr. 1981.

[4] T. Y. Leou and J. K. Aggarwal, "Recursive implementation of LTV filters—Frozen-time transfer function versus generalized transfer function," *Proc. IEEE*, vol. 72, pp. 980-981, July 1984.

[5] J. B. Cruz, "On the realizability of linear differential systems," *IRE Trans. Circuit Theory*, vol. CT-7, pp. 347-348, Sept. 1960.

[6] L. A. Zadeh, "Time-varying networks, I," *Proc. IRE*, vol. 51, pp. 1488-1503, Oct. 1961.

[7] H. D'Angelo, *Linear Time-Varying Systems: Analysis and Synthesis.* Boston, MA: Allyn and Bacon, 1970.

[8] Y. Grenier, "Time-dependent ARMA modeling of nonstationary signals," *IEEE Trans. Acoust., Speech and Signal Processing*, vol. ASSP-31, pp. 899-911, Aug. 1983.

[9] L. R. Rabiner and B. Gold, *Theory and Application of Digital Signal Processing.* Englewood Cliffs, NJ: Prentice-Hall, 1978.

[10] M. Y. Wu, "Transformation of a linear time-varying system into a linear time-invariant system," *Int. J. Contr.*, vol. 27, no. 4, pp. 589-602, 1978.

[11] N. C. Huang and J. K. Aggarwal, "Synthesis of recursive linear shift-variant digital filters," *IEEE Trans. Circuits Systems*, vol. CAS-30, pp. 29-36, Jan. 1983.

[12] S. R. Searle, *Matrix Algebra Useful for Statistics.* New York: Wiley, 1982.

[13] R. Fletcher and M. J. D. Powell, "A rapid convergent descent method for minimization," *Comput. J.*, vol. 6, no. 2, pp. 163-168, 1963.

✠

Tzong-Yeu Leou (S'81) was born in Taiwan on January 9, 1956. He received the B.S. and M.S. degrees in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1977 and 1980, respectively.

From 1979 to 1980 he worked as a Teaching Assistant at National Taiwan University, Taipei, Taiwan. From 1980 to 1982 he was a Teaching Assistant at the Department of Electrical and Computer Engineering at the University of Texas at Austin, Austin, Texas. Since 1981 he has been a Research Assistant at the Laboratory for Image and Signal Analysis at the University of Texas at Austin, Austin, Texas. His main research interests are in digital signal processing, filter design, and system theory.

✠

J. K. Aggarwal (S'62-M'65-SM'74-F'76) received the B.S. in mathematics and physics from the University of Bombay (1956), the B. Eng. from the University of Liverpool, England (1960) and the M.S. and Ph.D. at the University of Illinois, Urbana in 1961 and 1964, respectively.

He joined The University of Texas in 1964 as an Assistant Professor and has since held positions as Associate Professor (1968) and Professor (1972). Currently he is the John J. McKetta Energy Professor of Electrical and Computer Engineering and Computer Science at The University of Texas at Austin. Further, he was a visiting assistant professor at Brown University, Providence, RI (1968), and a visiting associate professor at the University of California, Berkeley, California, during 1969-1970. He has published numerous technical papers and several books, *Notes on Nonlinear Systems* (1972), *Nonlinear Systems: Stability Analysis* (1977), *Computer Methods in Image Analysis* (1977), *Digital Signal Processing* (1979), and *Deconvolution of Seismic Data* (1982). His current research interests are image processing and computer vision. He was co-editor of the special issues on Digital Filtering and Image Processing (IEEE CAS Transactions, March 1975), and on Motion and Time Varying Imagery (IEEE PAMI Transactions, November 1980) and editor of the two volume special issue of Computer Vision, Graphics and Image Processing on Motion (CVGIP, January and February 1983). Currently he is an associate editor of the journals *Pattern Recognition*, *Image and Vision Computing*, and *Computer Vision, Graphics and Image Processing*. Further, he is a member of the IEEE Transnational Relations Committee and of the Editorial Board of IEEE Press. He was the General Chairman for the IEEE Computer Society Conference on Pattern Recognition and Image Processing, Dallas 1981 and was the program Chairman for the First Conference on Artificial Intelligence Applications sponsored by IEEE Computer Society and AAAI held in Denver 1984.

Dr. Aggarwal is a member of ACM, AAAI, The International Society for Optical Engineering, Pattern Recognition Society, and Eta Kappa Nu.

Reprinted from JOURNAL OF THE ELECTROCHEMICAL SOCIETY
Vol 133, No 2, February 1986
Printed in U S A
Copyright 1986

Temperature Response of GaAs in a Rapid Thermal Annealing System

T. R. Block, C. W. Farley,* and B. G. Streetman*

Microelectronics Research Center, Department of Electrical and Computer Engineering, The University of Texas, Austin, Texas 78712

Rapid thermal annealing (RTA) systems are used to heat semiconductors for short periods of time, usually seconds. This technique has a variety of applications in the processing of Si and III-V compound semiconductors (1). In an RTA system which uses radiant energy to heat the semiconductors, one might expect to find variations in the temperature response of materials due to their different absorption properties. An investigation of GaAs and Si shows that this is indeed the case, and that the temperature variation is quite pronounced.

Experimental

The RTA system used in these experiments is similar to that described by Davies and Kennedy (2), and consists of a heating chamber containing two elliptically shaped cavities which focus light produced by two 2 kW tungsten filament quartz lamps onto a sample suspended at the center of the chamber. A microprocessor controls power to the lamps with a thermocouple for temperature feedback. Samples are suspended on very thin (~250 μm thick) silica slides inside a silica tube which allows annealing in a selected gas ambient.

Semi-insulating (100) GaAs doped with Cr, p-type (100) GaAs doped with 7×10^{17} cm^{-3} Zn, and n-type (100) Si doped with 4×10^{15} P were cut into samples ~8 mm × 8 mm. Holes were etched into the GaAs using a 20% bromine-methanol etch and a mask of CVD deposited SiO_2, and into the Si using a solution of pyrocatechol, ethylene diamine, and water (3). 200 nm of CVD SiO_2 containing 7% P was then deposited on each sample. K-type thermocouples (0.002 in. diam wire) were glued into the holes with a mixture of Aremco no. 516 cement (ZrO) and either ground GaAs or Si using a procedure similar to that suggested by Cohen *et al.* (4). Such small thermocouples were used to insure that the temperature response of the samples would be little affected by the thermocouples.

Throughout the experiments, the GaAs samples were either thermally isolated from each other or placed on a larger piece of silicon to thermally connect them, as shown in Fig. 1. Temperature feedback to the controller was provided by the Si sample. These two configurations were then submitted to a heat cycle consisting of a preheat at 300°C for 30s to stabilize initial conditions, then an instantaneous change of setpoint to 750°C, a hold at this temperature for 30s (counted after the sample is within 30°C of 750°C), and then a change of the setpoint back to zero. The heat cycling was done in a stagnant N_2 atmosphere.

Results and Discussion

Temperature response curves for the thermally isolated samples are shown in Fig. 2. Curve (a) is the programmed setpoint and curve (b) is the thermocouple output of the Si sample. The Zn doped and semi-insulating GaAs samples are shown as curve (c) and (d), respectively.

Clearly, GaAs couples to the radiant flux differently than Si, exhibiting a drastically different temperature response. Comparison of doped *vs.* semi-insulating GaAs in Fig. 2 shows the tremendous effect of free carrier concentration on the temperature response, an effect also found in silicon by Seidel *et al.* (5). However this is not enough to explain the data completely. At 750°C the free carrier concentration of Si is greater than that of the doped GaAs, yet GaAs is at a higher temperature. This effect must be due to basic differences in the bulk material properties, for example direct *vs.* indirect bandgap, size of

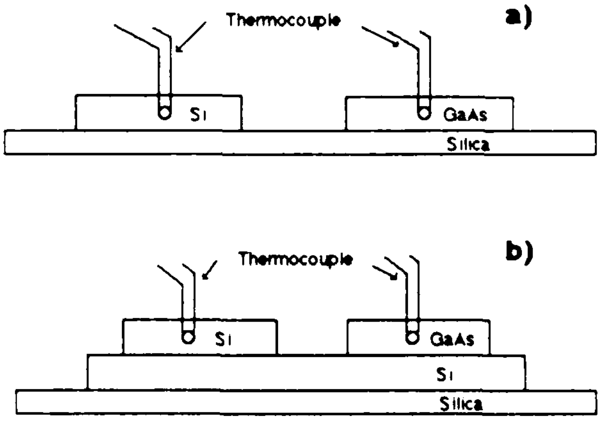


Fig. 1. Schematic diagram of sample configuration for (a) thermally isolated and (b) thermally connected cases.

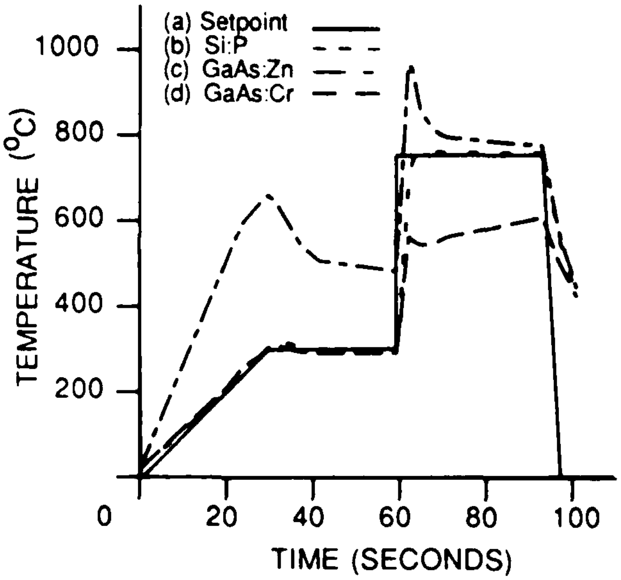


Fig. 2. Temperature response curve for thermally isolated samples: (a) temperature setpoint, (b) P doped Si, (c) Zn doped GaAs, and (d) Cr doped GaAs.

*Electrochemical Society Active Member.

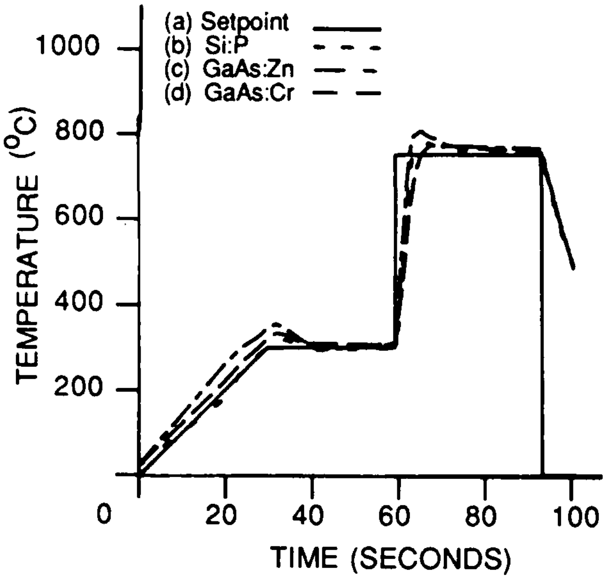


Fig. 3. Temperature response curve for thermally connected samples: (a) temperature setpoint, (b) P doped Si, (c) Zn doped GaAs, and (d) Cr doped GaAs.

bandgap, optical absorption parameters, and thermal conductivity. Theoretical work by Borisenko *et al.* (6) predicts different heating rates for materials with different emissivities. Another factor may be a shift in the frequency distribution of the tungsten lamps' output at different power levels (due to blackbody radiation at different temperatures). As the lamp power is cut back to hold the sample at temperature, the light shifts to longer wavelengths, which may affect the relative amounts of band-to-band *vs.* free carrier excitation.

Results for the thermally connected samples are shown in Fig. 3. Curves (a) and (b) are again the programmed setpoint and the Si sample response, (c) and (d) the Zn doped and semi-insulating GaAs samples.

When the materials are put in thermal contact, their temperature behavior becomes similar. The significant temperature overshoot of setpoint is due to thermal lags in the control loop from the configuration in Fig. 1b. In actual anneals, the large piece of Si pictured in Fig. 1b would contain the thermocouple. In such a setup, we have observed negligible temperature overshoot and the temperature difference between GaAs and Si is usually less than 15°C at 750°C. It should be noted that this annealing arrangement requires good thermal contact of the sample to the large Si susceptor. We also found the temperature response to be sensitive to gas flow in the system. A high flow rate resulted in a GaAs temperature 50°C higher than the temperature of the Si susceptor at 750°C. The source of this pheno nenon is not clear; the anneals were therefore performed in a stagnant atmosphere.

These results point to an important fact: GaAs behaves quite differently from Si in a rapid thermal annealing system. Consequently, the temperature of the GaAs must be used to control the system. This can be done either directly, by using a pyrometer looking at the GaAs or indirectly, by having the GaAs in isothermal contact to a susceptor to which a thermocouple is attached. The latter method is the less complicated of the two and avoids problems pyrometers themselves face in a radiant RTA system (7).

Acknowledgments

The authors acknowledge support from the Joint Services Electronics Program under contract No. AFOSR F49620-82-C-0033 and the National Science Foundation under grant no. ECS-8420002. We also wish to thank D. Neikirk and S. Lester for useful discussion during the course of the experiments.

Manuscript submitted Oct. 4, 1985.

The University of Texas assisted in meeting the publication costs of this article.

REFERENCES

1. J. F. Gibbons, D. M. Dobkin, M. E. Greiner, J. L. Hoyt, and W. G. Opyd, *Mater. Res. Soc. Symp. Proc.*, **23**, 37 (1984).
2. D. E. Davies and E. F. Kennedy, *Electronics Lett.*, **18**, no. 7, 282 (1982).
3. K. E. Bean, *IEEE Trans. Electron. Devices*, **ed-25**, 1185 (1978).
4. S. A. Cohen, T. O. Sedgewick, and J. L. Speidell, *Mater. Res. Soc. Symp. Proc.*, **23**, 321 (1984).
5. T. E. Seidel, D. E. Lischner, C. S. Pai, and S. S. Lau, *J. Appl. Phys.*, **57**, (4), 1317 (1985).
6. V. E. Borisenko, V. V. Gribkovskii, V. A. Labunov, and S. G. Yudin, *Phys. Status Solidi A*, **86**, 573 (1984).
7. C. Russo, *Nucl. Instrum. Methods*, **B6**, 298 (1985).

Reprinted from JOURNAL OF THE ELECTROCHEMICAL SOCIETY
Vol. 133, No. 10, October 1986
Printed in U.S.A.
Copyright 1986

Factors Influencing the Photoluminescence Intensity of InP

S. D. Lester,* T. S. Kim, and B. G. Streetman**

Department of Electrical and Computer Engineering, Microelectronics Research Center, The University of Texas, Austin, Texas 78712

Room temperature photoluminescence (PL) has recently been used as a tool to assess the quality of InP substrates during and after various processing steps and the quality of InP/insulator interfaces. It has generally been suggested that high PL intensities reflect high quality bulk material or high quality interfaces. For example, band edge PL intensity has been used to assess the effectiveness of annealing treatments for activating ion-implanted dopants[1] and has been used in a number of studies to monitor processing steps used during the fabrication of MIS devices.[2] In the former studies it was suggested that high PL intensities reflect high quality annealing (a bulk property) and in the latter studies it was suggested that the PL intensity of n-type InP yields a reliable estimate of the interface state density in the upper part of the band gap. In this communication we discuss factors which influence band-edge PL intensity and point out that great care must be taken in interpreting such data.

Although many factors are involved in determining PL intensity, three important influencing factors are bulk parameters (i.e., mobility, lifetimes, doping level, etc.), the surface recombination velocity (S), and the presence of a space charge region (electric field) at the surface. The first of these reflects bulk crystal quality and the second two are determined by surface properties. The relative influence of these factors in determining the PL intensity of a given sample is, in general, very difficult to determine; thus, processing-induced changes in PL intensity can easily be misinterpreted. We note that PL intensity changes have been used as a quantitative measure of GaAs surface properties.[3,4]

Although PL has long been recognized as a near-surface probe of bulk material, surface effects can very strongly influence the intensity of band-edge luminescence. This surface sensitivity is dramatic in the case of n-type InP where both liquid and gas ambients have been shown to have pronounced effects on PL intensities. In the case of n-InP immersed in chemical solutions, it has been demonstrated,[5] and confirmed in our laboratory, that PL intensities can be changed by three orders of magnitude. For example, in-situ measurements of n-InP alternately flushed with DI water and dilute HF show that the band-edge intensity can be reversibly varied by ~1000x in a very short time. Apparently, different chemical treatments also leave the InP surface with varying degrees of "stability" to subsequent changes in PL intensity caused by exposure to air or low temperature annealing.[5,6]

Another example of surface effects on PL intensity is the case of InP exposed to various gas ambients. Fig. 1 shows the band-edge PL intensity of an undoped ($n=5x10^{15}$ cm^{-3}) sample repeatedly exposed to oxygen and nitrogen. As the figure shows, the PL intensity is reduced in oxygen, increased in nitrogen, and can be cycled repeatedly. The size of these ambient-induced PL changes are influenced by many factors, including pressure (flowrate), substrate doping concentration and type, the laser intensity, humidity, and the history of the sample. Some of these effects have been reported previously,[7,8] and a more detailed description will be presented elsewhere.[9] We also note that GaAs is sensitive to ambient effects, but to a lesser degree than InP. Like the effects of chemical solutions, these ambient effects are very substantial (in certain cases the intensity can be change by >5x) and illustrate the high surface sensitivity of band edge PL intensity. The fact that PL intensity is so strongly influenced by these surface effects (including interface effects at InP/dielectric interfaces) indicates that great care must be taken in extracting bulk information from PL intensity data. We also note that PL intensities measured at low temperatures (samples immersed in liquid He) are sensitive to sample surface properties prior to cooling. To safely compare the PL intensities of a number of samples it is therefore important to insure that the samples have nearly identical surfaces.

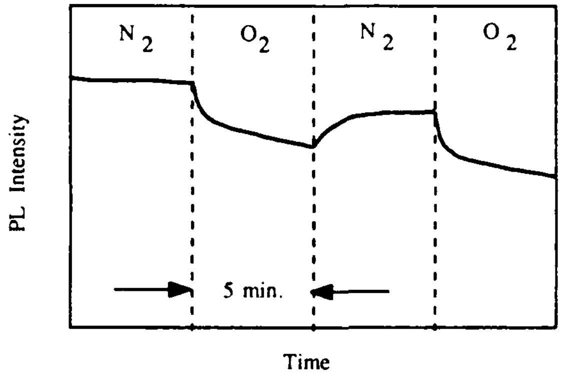


Fig. 1. Band-edge PL response of $n=5x10^{15}$ cm^{-3} InP to ambient changes.

Another notable feature of Fig. 1 is the gradual reduction of PL intensity. This slow trend is at least partially reversible and has previously been attributed to oxidation.[7] We

*Electrochemical Society Student Member
**Electrochemical Society Active Member

have found that this slow trend is a direct result of the illumination process and can substantially reduce the PL intensity under high laser power levels. Therefore, it should be recognized that the illumination process itself can markedly alter the surface properties and consequently the PL intensity of InP, so care should be taken to account for it when taking room temperature PL data (i.e., samples should receive comparable amounts of laser excitation prior to the start of the measurement).

The previous examples have shown that surface properties have an important influence on PL intensity and need to be considered if bulk information is to be extracted from PL intensity data. On the other hand, these surface effects can be used as a tool for studying InP surfaces and interfaces. If surface information is to be obtained it is important to distinguish between the effects of band bending changes and surface recombination velocity changes on measured PL intensities. As suggested by Aspnes,[10] to do this, it is extremely helpful to have a second measurement technique which can give an independent measure of the surface Fermi level (E_{Fs}). Possible techniques for this include Raman or photoemission spectroscopy or the use of PL with two excitation wavelengths; however, the simplest example of such a technique is to measure the resistance of a thin film resistor which will have a resistivity that is a function of the depth of the space charge region at the surface.

Figure 2 shows the resistivity of an n-type resistor structure in oxygen and nitrogen. The structure was made by implanting Si (10^{12} cm^{-2} @ 150 keV) into a ~3 x 5 mm InP:Fe sample and alloying In/Sn ohmic contacts. Figure 2a shows the resistance under illumination and Fig. 2b shows the resistance in complete darkness. It is clear from these figures that the resistance of this structure, like the PL intensity, is reversibly changed with ambient. Since the resistance under illumination depends on the surface recombination velocity, Fig. 2a is not sufficient to indicate whether a change in S or a change in band bending is responsible for the PL intensity and resistance changes. However, Fig. 2b clearly indicates that ambient changes result in changes in the surface Fermi level (depletion depth). This change in E_{Fs} and the depletion depth then results in the PL intensity response shown in Fig. 1. Such PL intensity variations have previously been attributed solely to changes in S.[7] Our data, of course, does not exclude a change in S, and in fact S should be a function of Fermi level position. However, this does demonstrate that the surface Fermi level is affected by ambient changes and that changes in E_{Fs} must be considered when interpreting PL intensity variations.

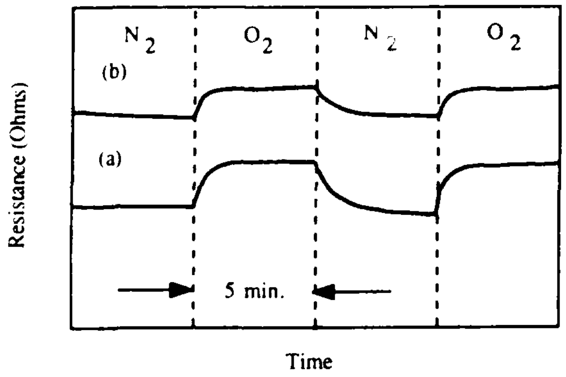


Fig. 2 Resistivity response of an n-type resistor (a) under illumination and, (b) in complete darkness.

In summary, room temperature PL can be an extremely useful technique for investigating bulk and surface properties of InP. However, since the intensity of the band-edge transition is strongly affected by a number of factors, great care must be taken in interpreting intensity data. If bulk information is to be obtained, samples must be prepared with identical surfaces and if meaningful surface information is sought, a second measurement should be used to separate the effects of band bending from the effects of changes in the surface recombination velocity.

Acknowledgement:

The authors acknowledge support from the Joint Services Electronics Program under contract No. AFOSR F49620-82-C-0033 and the Texas Advanced Technology Program.

References:

1. D. Kirillov, J. L. Merz, R. Kalish, and S. Shatas, J. Appl. Phys. **57**, 531 (1985).
2. S. Krawczyk, B. Bailly, B. Sautreuil, R. Blanchet, and P. Viktorovitch, Electron. Lett. **20**, 255 (1984).
3. J. M. Woodall, G. D. Pettit, T. Chappell, and H. J. Hovel, J. Vac. Sci. Technol. **16**, 1389 (1979).
4. S. D. Offsey, J. M. Woodall, A. C. Warren, P. D. Kirchner, T. I. Chappell, and G. D. Pettit, Appl. Phys. Lett. **48**, 475 (1986).
5. S. K. Krawczyk and G. Hollinger, Appl. Phys. Lett. **45**, 870 (1984).
6. H. Nagai and Y. Noguchi, J. Appl. Phys. **50**, 1544 (1979).
7. H. Nagai and Y. Noguchi, Appl. Phys. Lett. **33**, 312 (1978).
8. H. Nagai, S. Tohno, and Y. Mizushima, J. Appl. Phys. **50**, 5446 (1979).
9. S. D. Lester, T. S. Kim, and B. G. Streetman, unpublished.
10. D. E. Aspnes, Surf. Sci. **132**, 406 (1983).

Manuscript submitted May 14, 1986; revised manuscript received July 23, 1986.

The University of Texas assisted in meeting the publication costs of this article.

PHYSICAL REVIEW B VOLUME 33, NUMBER 2 15 JANUARY 1986

Experimental observation of adsorbate orbital splitting at single-crystal metal surfaces

Marshall Onellion and J. L. Erskine
Department of Physics, University of Texas, Austin, Texas 78712

(Received 30 May 1985)

Splitting of the $5P_{3/2}$ component of the photoexcited Xe ion doublet is observed on the (110) planes of several metal surfaces. This effect is shown to originate from a true "crystal-field" effect, not as a consequence of adatom-adatom interactions. The splitting therefore provides a probe of local fields at the screened ion.

One of the more important parameters associated with surface phenomena is the local potential. Recent efforts to more thoroughly understand local surface potentials include calculations of the orbital splitting of atoms approaching a metal surface,[1] studies of dipole moments and polarizabilities of adsorbed atoms[2] (both ground-state effects), and analysis of charge transfer and screening effects at metal surfaces which accompany various photon and electron excitation phenomena.[3-5] One particular focus of work on this problem has involved excitation properties of rare-gas atoms[6-11] on metal surfaces, i.e., systems which have well-defined ground states.

Waclawski and Herbst[6] conducted one of the first photoemission investigations of a rare-gas monolayer on a metal surface. They reported significant broadening of the $5P_{3/2}$ component of the spin-orbit split $5p$ level of Xe physisorbed on W(100), and attributed the broadening to unresolved splitting resulting from the surface crystal field. This interpretation stimulated several model calculations[7-9] and additional experiments[10] aimed at testing the hypothesis of surface crystal fields in more detail. These calculations, and subsequent experiments,[10,12] which were conducted under more carefully controlled conditions, have shown that the broadening of the $5P_{3/2}$ line observed by Waclawski and Herbst at full monolayer coverages was not due to crystal-field or image charge effects.[7,8] In this case, the line broadening resulted from Xe-Xe interactions as shown by angle-resolved photoemission determination of the Xe-band dispersion throughout the surface Brillouin zone and comparison with results of simple tight-binding calculations.

More recent experiments by Opila and Gomer[13] have again raised the issue of surface image dipole or crystal-field effects in relation to the photoemission spectra of physisorbed rare-gas atoms on metal surfaces. In these carefully conducted experiments a very convincing case is presented in support of the existence of a mechanism, unrelated to adatom-adatom interactions, which splits the $5P_{3/2}$ line of rare-gas atoms on W(110). This result is in contrast to the null result obtained by Erskine[10] under similar experimental conditions for Xe on W(100).

We have recently conducted extensive photoemission studies of physisorbed rare-gas atoms on single crystal NiAl alloy surfaces[14] to investigate surface stoichiometry and local work functions as probed by the photoemission of adsorbed xenon (PAX) technique.[15] During this study we observed splitting of the $5P_{3/2}$ component of photoexcited Xe atoms on the NiAl(110) surface, but not on the (100) surface of the same ordered alloy. The splitting was observed using experimental conditions under which Xe-Xe interactions are negligible (i.e., low coverages), leading to the conclusion that the splitting is similar in nature to that reported by Opila and Gomer. We have conducted additional experiments on Ni(110) and Ni(100) surfaces which yield the same result, i.e., that the lower-symmetry surface produces splitting of the $5P_{3/2}$ line. The purpose of the present Brief Report is to present these results which suggest that splitting of the $J=\frac{3}{2}$ component of photoexcited Xe atoms physisorbed at low coverage on (110) surfaces results from the lower coordination symmetry of the adsorption site.

Experiments reported here were conducted using an Auger-photoelectron spectrometer[10] equipped with low-energy electron diffraction (LEED) optics and a cold stage manipulator capable of sample temperatures ranging from 1200 to below 30 K. The $\frac{3}{8}$-in.-diam × $\frac{1}{16}$-in.-thick NiAl(110) samples were aligned to ±1° using x-ray Laue techniques, and spark cut and mechanically polished using alumina powder to 0.05-μm grit. *In situ* cleaning using Ne ion sputtering (500 eV, 10 μA/cm^2) and annealing to 800 °C yielded clean, well-ordered surfaces. Auger analysis of the clean surfaces, our work[14] in which work-function changes were studied, and chemisorption experiments[16] involving CO indicate that the well-annealed NiAl(110) surfaces exhibit a stoichiometry (ratio of Ni to Al) equal to the bulk value (i.e., 1 to 1). Recent LEED studies of this surface suggest a reconstruction involving atomic rippling,[17] which consists of small displacement of the surface atoms (~ 0.08 Å) perpendicular to the surface. Extensive angle-resolved photoemission studies[18] of the NiAl(110) surface using synchrotron radiation have yielded bulk band structure in good agreement with calculations. These experiments constitute additional characterization of the NiAl(110) surface.

Our interpretation of the splitting of the $5P_{3/2}$ peak in terms of local-substrate-related fields rather than adsorbate-adsorbate coupling relies on accurate knowledge of the surface conditions, including substrate order and composition, which was just discussed, as well as adatom coverage and spatial distributions. Adatom concentration was accurately calibrated at integral monolayer coverages by analysis of multipeak spectra resulting from layer-dependent *NOO* Auger energy shifts,[10] and checked using $5P$ valence level shifts observed in photoemission spectra. Uniform Xe layers of $n=1$, 2, and 3 monolayers could be obtained by adsorption followed by carefully monitored annealing. Submonolayer coverages were determined from work-function changes, which are roughly linear in the 0–0.5-ML range, and from the intensity of the angle-integrated photoemission peaks relative to the NiAl d bands measured in the same configuration used in calibration experiments. Sample temperatures during and after adsorption were maintained below 40 K to ensure low surface mobility of the adsorbed rare gases. No evidence of island or cluster formation of Xe atoms was observed in photoemission (i.e., $k_{\parallel}$ dispersion of peaks) or by LEED analysis at these temperatures. Weak halos having sixfold symmetry were observed in LEED studies of low coverage Xe films only after annealing to ~ 100 K, indicating that temperatures in this range are required to induce island formation.

Figure 1 displays angle-resolved electron energy distribution curves (EDC's) for various coverages of physisorbed Xe on NiAl(110) at $T \sim 30$ K. Various features[18] of the d bands of NiAl along the Λ direction of the three-dimensional Brillouin zone are apparent in the energy range within 5 eV of the Fermi energy, E_F. Submonolayer Xe spectra exhibit two primary peaks corresponding to the $5P_{1/2}$, $5P_{3/2}$ states of the ion. Spectra for coverages greater than one monolayer exhibit two additional peaks which increase in strength with coverage. These are the second layer peaks, which are shifted to higher binding energy due to less effective screening of the ion (relaxation shifts), and which were used in thickness calibration.

Close inspection of the $5P_{3/2}$-derived peaks corresponding to low coverage reveals that it is not symmetric as is the $5P_{1/2}$ peak at equal coverage. Figure 2 illustrates, on an expanded scale, the two peaks for equal coverages of Xe on NiAl(110) and on NiAl(100). One does not expect to be able to resolve the actual crystal-field splitting of the $P_{3/2}$ level because the *intrinsic* broadening of the lines due to relaxation mechanisms related to the presence of the metal surface is approximately equal to the splitting. The inset of Fig. 2 compares results of curve fitting the two EDC's using three Gaussian functions, assuming that the $5P_{3/2}$ component in each case is composed of two Gaussians having the same width as the $5P_{1/2}$ component. This analysis shows that splitting of the $5P_{3/2}$ line is approximately 0.35 eV.

There are no detectable shifts in peak positions as a function of Xe coverage at low coverages. This coverage independence suggests that splitting of the $5P_{3/2}$ line results from coupling between the screened ion and the substrate rather than with neighboring Xe atoms. The annealing experiments which established the temperature at which islands did form also eliminated the possibility that the independence of binding energy with coverage resulted from islands or clusters of constant density at all coverage.

We have carried out corresponding experiments involving low coverage Xe layers on Ni(100) and Ni(110). These experiments yield a similar splitting of the $5P_{3/2}$ peak for Xe on the Ni(110) surface but not on the Ni(100) surface. Close inspection of experimental results of Jacobi and Rotermund,[19] which were obtained under similar experimental conditions, also reveals a split $5P_{3/2}$ level for low coverages of Xe on Ni(110), in agreement with our results. Our previous search for crystal-field splitting on the W(100) surface revealed no splitting of the $5P_{3/2}$ level,[10] but photoemission studies of Xe of Opila and Gomer on the W(110) surface exhibit splitting of this level.[19] Based on these experimental data,[7] it appears that the splitting of the $5P_{3/2}$ state could be related to the reduced symmetry of the adsorption site on the (110) surfaces.

To validate this possibility, one must be convinced that the (110) and the (100) faces of W, Ni, and β-NiAl can and probably do yield qualitatively different local environments

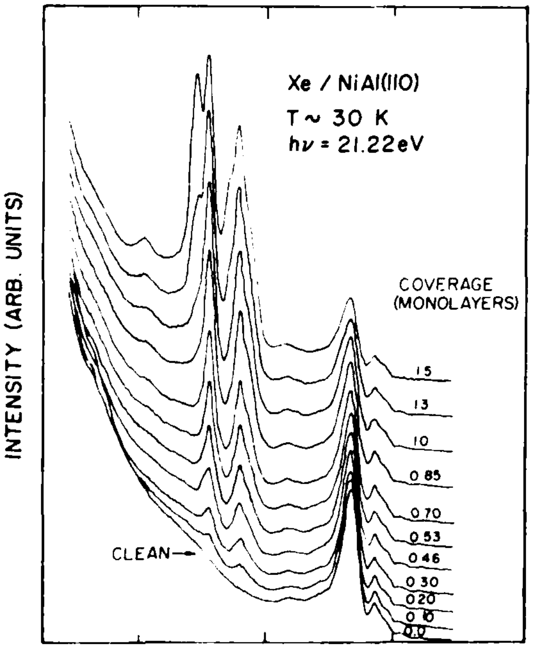


FIG. 1. Angle-resolved photoemission spectra for Xe adsorbed on NiAl(110) as a function of Xe coverage.

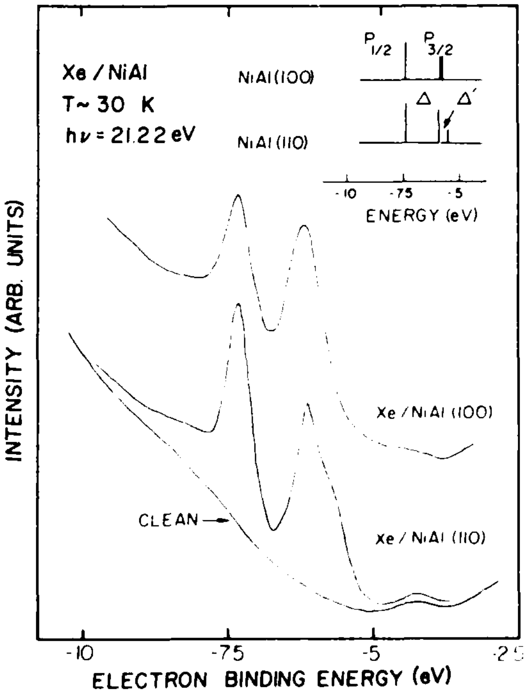


FIG. 2. Expanded scale of angle-resolved photoemission spectra for Xe adsorbed on NiAl(110) and NiAl(100) surfaces. Inset, line intensities as determined by Gaussian fitting of the data. Values of the splitting are Xe on NiAl(110), $\Delta = 1.22 \pm 0.04$ eV, $\Delta' = 0.37 \pm 0 4$ eV; Xe on NiAl(100), $\Delta = 1.20 \pm 0.04$ eV, $\Delta' = 0.06 \pm 0.06$ eV

TABLE I. Polar-angle dependence of $5P_{1/2}$ and $5P_{3/2}$ linewidths of photoemission spectra for Xe adsorbed on NiAl(110). Photon energy $h\nu = 21.22$ eV, sample temperature $T \sim 30$ K. Angle is measured from the sample normal along the $[1\bar{1}0]$ direction; center and width energies are in eV. Angular resolution ±4°.

	$5P_{1/2}$		$5P_{3/2}$	
Polar angle	Center (below ϵ_F)	Width	Center (below ϵ_F)	Width
0°	7.05	0.54	5.83	0.75
10°	7.05	0.53	5.83	0.70
20°	7.05	0.52	5.83	0.67
30°	7.05	0.50	5.83	0.63
40°	7.05	0.50	5.83	0.63

for adsorbed Xe. For physisorbed atoms, at *low coverage*, the preferred site will most likely be the deepest hollow sites in a surface unit cell. For bcc W(100), Xe should physisorb at the C_{4v} fourfold hollow site; for W(110) the corresponding site has C_{2v} symmetry. In the case of fcc Ni, again, the preferred surface site on the (100) face will have C_{4v} symmetry, and C_{2v} symmetry on the (110) face. The crystal structure of β-NiAl is the CaF_2 (cubic) structure. Here again, the (100) surface offers only C_{4v} sites, whereas the (110) surface offers C_{2v} sites. Qualitatively, the argument of a symmetry based origin of the splitting appears valid.

Herbst[9] has investigated theoretically the angular dependence of photoelectrons from atoms adsorbed on metal surfaces, taking into account the effects of the substrate atoms. Although none of the specific results obtained by Herbst apply directly to Xe adsorbed in the C_{2v} site on NiAl, one of the general characteristics of the model should apply. This characteristic is that the polar-angle variation of photoelectron emission associated with component lines of a crystal-field split level will be different.

Table I illustrates the experimentally determined polar angle dependence of the $5P_{1/2}$, $5P_{3/2}$ linewidths for Xe on NiAl(110). The intrinsic broadening is too large to clearly resolve the $5P_{3/2}$ state splitting (as shown in Fig. 2). However, two features are clear from our polar-angle-variation data. First, the binding energies of the $5P_{1/2}$ and $5P_{3/2}$ states are independent of polar angle, confirming that lateral interactions (which would produce band dispersion) are not present. Second, the $5P_{1/2}$ linewidth is nearly constant, whereas the $5P_{3/2}$ linewidth changes significantly, as would be expected if the intensity ratio of the component lines changed. This constitutes additional evidence of a local crystal-field origin of the splitting.

This work was supported by the National Science Foundation, Grant No. DMR-83-04368, and by the Joint Services Electronics Program.

[1] P. V. S. Rao and J. T. Waber, Surf. Sci. **28**, 299 (1971).

[2] T. C. Chiang, G. Kaindl, and D. E. Eastman, Solid State Commun. **41**, 661 (1982).

[3] N. D. Lang, Phys. Rev. Lett. **46**, 842 (1981).

[4] N. C. Giles and C. M. Varma, Phys. Rev. B **23**, 5600 (1981).

[5] N. D. Lang and A. R. Williams, Phys. Rev. B **16**, 2408 (1977).

[6] B. J. Waclawski and J. F. Herbst, Phys. Rev. Lett. **35**, 1594 (1975).

[7] P. R. Antoniewicz, Phys. Rev. Lett. **38**, 374 (1977).

[8] J. A. D. Matthew and M. G. Devey, J. Phys. C **9**, L413 (1976).

[9] J. F. Herbst, Phys. Rev. B **15**, 3720 (1977).

[10] J. L. Erskine, Phys. Rev. B **24**, 2236 (1981).

[11] K. Horn, M. Scheffler, and A. M. Bradshaw, Phys. Rev. Lett. **41**, 822 (1978).

[12] M. Scheffler, K. Horn, A. M. Bradshaw, and K. Kambe, Surf. Sci. **80**, 69 (1979).

[13] R. Opila and R. Gomer, Surf. Sci. **127**, 569 (1983).

[14] M. Onellion and J. L. Erskine (unpublished).

[15] K. Wandelt, J. Vac. Sci. Technol. A **2**, 802 (1984).

[16] W. K. Ford (private communication).

[17] H. L. Davis and J. R. Noonan, Phys. Rev. Lett. **54**, 566 (1985).

[18] W. K. Ford, E. W. Plummer, J. L. Erskine, and D. Pease, Bull. Am. Phys. Soc. **29**, 523 (1984).

[19] K. Jacobi and H. H. Rotermund, Surf. Sci. **116**, 435 (1982).

PHYSICAL REVIEW B VOLUME 33, NUMBER 10 15 MAY 1986

Electronic structure and properties of epitaxial Fe on Cu(100): Theory and experiment

M. F. Onellion
Department of Physics, University of Texas, Austin, Texas 78712

C. L. Fu
Department of Physics and Astronomy, Northwestern University, Evanston, Illinois 60201

M. A. Thompson and J. L. Erskine
Department of Physics, University of Texas, Austin, Texas 78712

A. J. Freeman
Department of Physics and Astronomy, Northwestern University, Evanston, Illinois 60201

(Received 27 February 1986)

Results of a combined experimental-theoretical study of the electronic structures and properties of epitaxial Fe on Cu(100) are reported. Angle-resolved photoelectron spectroscopy is used to determine the electronic structure of one, two, and four layers of epitaxial Fe on Cu(100). The experimentally determined two-dimensional energy bands of $p(1\times1)$Fe monolayers and bilayers verify predictions of local spin-density full-potential linearized augmented plane-wave calculations. Changes in electronic properties with coverage and the evolution of the bulk electronic structure of the substrate-stabilized fcc iron are described.

Advances in (i) novel sample preparation techniques, in particular epitaxial growth of layered structures, and (ii) in local spin-density electronic-structure theory of surfaces and interfaces are stimulating new interest and excitement in the field of thin-film magnetism. Taken together, they offer unique opportunities for developing new magnetic materials[1] as well as advancing our understanding of fundamental magnetic interactions in solids.[2] The opportunities for advancing our fundamental understanding of magnetic phenomena are particularly attractive in the subfield of solid-state physics in which carefully characterized materials of known structure are experimentally studied and the results coordinated with first-principles calculations.

In a recent publication[3] we reported experimental results for epitaxial $p(1\times1)$Ni films on Cu(100). This study presented some of the first detailed experimental evidence showing that high-quality epitaxial magnetic films could be grown on metallic single-crystal surfaces,[4] and that the two-dimensional electronic structure and magnetic exchange splitting of these films could be determined with sufficient accuracy to provide meaningful tests of the predictions of first-principles calculations.[5]

The present Rapid Communication continues to explore the prospects of advancing our understanding of two-dimensional magnetic structures based on the interplay between *ab initio* first-principles calculations and photoemission studies of novel thin-film structures fabricated by molecular-beam epitaxy. Our experimental results confirm the predictions of the computational studies,[5] and indicate that the $p(1\times1)$Fe on Cu(100) system is a *second* suitable candidate for detailed experiments in which electronic and magnetic properties can be probed by the rapidly increasing number of surface and spin sensitive spectroscopic techniques.[6]

Our experiments were performed at the Synchrotron Radiation Facility in Stoughton, Wisconsin. The 1-m stainless-steel Seya-Namioka monochrometer was used to dispense radiation from the Tantalus storage ring, and an angle-resolving photoelectron spectrometer, described previously,[3,7] was used to prepare the epitaxial crystals and to obtain the photoelectron spectra. Our sample preparation techniques and surface characterization methods were also identical to those described previously.[3]

The theoretical electronic structures were determined from local spin-density functional theory by means of the highly precise all-electron full-potential linearized augmented plane-wave (FLAPW) method.[8] The surfaces are modeled by a single-slab geometry with Fe layer(s) atop a five-layer Cu(001) substrate: The stacking has atoms in the fourfold hollow site of adjacent atomic planes. The Fe-Cu interlayer spacing was determined from total energy calculations. We find that Fe forms an ordered overlayer on Cu(001) with an Fe-Cu interlayer spacing which is very close (within 0.05 Å) to that of the substrate. This result is confirmed by low-energy electron-diffraction (LEED) studies described later. For the case of two monolayers $p(1\times1)$Fe/Cu(001), the problem is complicated by the magnetic coupling between the Fe layers. The magnetic ground state was therefore determined from a comparison of spin-polarization energy between various magnetic states. The ferromagnetic coupled bilayers are found to have the lowest total energy, whereas the antiferromagnetic coupling between the Fe layers exists as a metastable state with a total energy 0.2 eV above the ferromagnetic state.

Despite the presence of the nonmagnetic Cu substrate, strongly enhanced magnetic moments localized at the Fe site are found from these calculations: (1) $2.85\mu_B$ for 1 Fe/Cu(001); and (2) $2.83\mu_B$ and $2.58\mu_B$ for the surface and interface Fe layers, respectively, for the two monolayer coverage. The Fe-derived localized interface states and the narrowing of the *d* band appears to be the mechanism driving the enhancement of the magnetic moments over the value ($2.12\mu_B$) in bulk Fe.

Previous experimental work[9] has established that the excellent bulk lattice constant match permits pseudomorphic growth of fcc Fe on Cu(001). Evidence for pseudomorphic growth is based on transmission electron microscopy studies of 1000 Å Fe films grown on Cu(100) films in high vacuum (10^{-7} torr). Unlike the Ni on Cu(100) system, which we previously studied,[3] apparently, no structure studies of 1–5-monolayer films grown on well-characterized substrate surfaces in 10^{-10}-torr vacuum are currently available for the Fe on Cu(100) system. Therefore, we have conducted an extensive investigation of the growth properties and structure of thin Fe layers on Cu(100) surfaces. Our LEED and Auger analysis of this system are reported elsewhere.[10] The important results of this work relevant to the present paper are (1) $p(1\times1)$Fe grows on Cu(100) as an extension of the substrate with identical (±0.1 Å) lattice constant, (2) thicker films (up to 4 layers) appear to form excellent epitaxial layers of fcc iron stabilized by the Cu(100) substrate and having a lattice constant equal to the substrate, (3) interdiffusion on Fe and Cu at the interface is not apparent for substrate temperatures below 250 °C, and (4) the epitaxial growth appears to be dominated by a layer-by-layer mechanism at substrate temperatures of 150 °C.

Figure 1 displays representative angle-resolved photoemission spectra for one and two monolayers of $p(1\times1)$Fe on Cu(100) along the $\bar{\Gamma}$-$\bar{X}$ direction of the two-dimensional Brillouin zone. All of our energy distribution curves (EDC's) for epitaxial layers were taken using s-polarized light with the A vector along a symmetry axis of the crystal and with the emitted electrons detected either in that mirror plane (even symmetry) or perpendicular to it (odd symmetry). All spectra were taken at room temperature (300 K) at an energy resolution (monochromator plus analyzer) of approximately 100 meV. Approximately 200 spectra for one- and two-monolayer Fe films on Cu(100) were taken in even and odd geometry for $k_{\parallel}$ along $\bar{\Gamma}$-$\bar{X}$ and $\bar{\Gamma}$-$\bar{M}$ directions of the two-dimensional Brillouin zone. Several photon energies were used. Except for some minor variations of binding energies with film thickness (discussed below), spectra corresponding to a given symmetry and $k_{\parallel}$ value were consistent.

Figure 2 presents the two-dimensional electronic structure of one- and two-layer films obtained from our photoemission data. Solid lines and dashed lines in Fig. 2 represent calculated[5] majority spin and minority spin bands for a $p(1\times1)$Fe film on a five-layer Cu(100) slab which have over 50% of their wave function derived from Fe basis functions. These are the specific two-dimensional energy bands to which our experiments should be most sensitive.

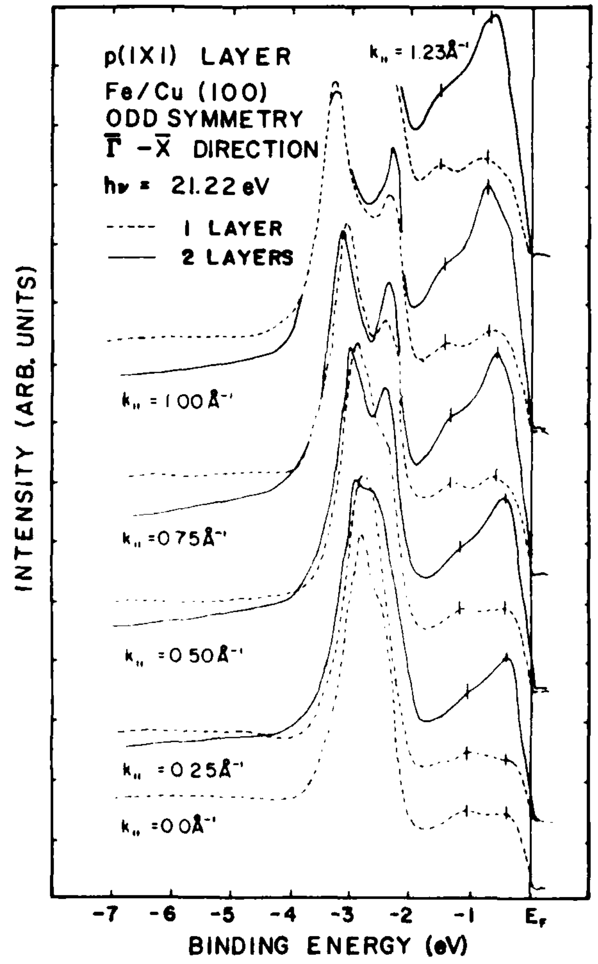


FIG. 1. Angle-resolved photoemission spectra for one- and two-layer $p(1\times1)$Fe films on Cu(100). Values of $k_{\parallel}$ correspond to the $\bar{\Gamma}$-$\bar{X}$ direction of the two-dimensional Brillouin zone.

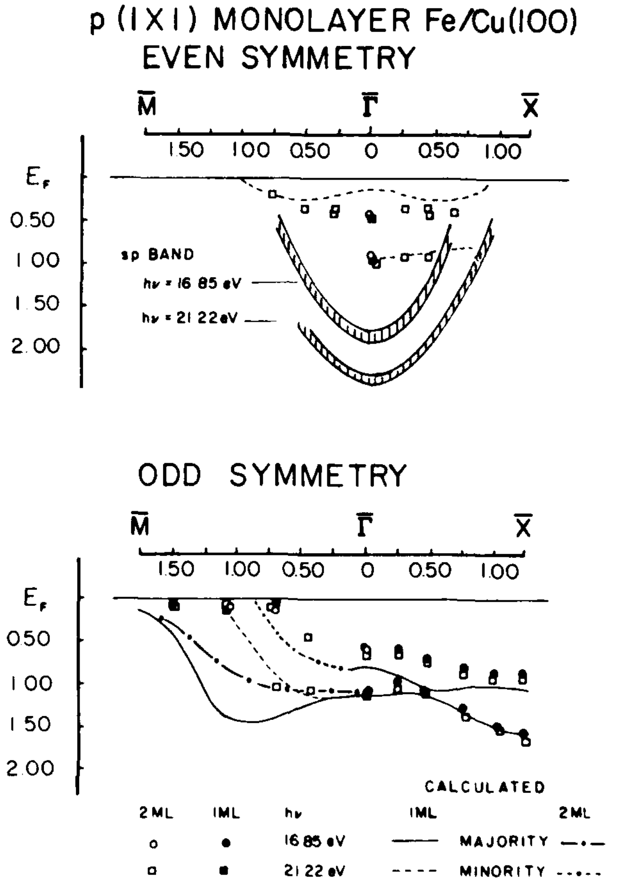


FIG. 2. Two-dimensional electronic structure of $p(1\times1)$Fe on Cu(100). The two broad curves indicate the regions of binding energy and $k_{\parallel}$ where a prominent structure resulting from the Cu sp band is observed. Light solid and dashed curves represent calculated (Ref. 5) surface Fe bands having over 50% surface character. Data are represented by empty (two-monolayer films) and solid (one-monolayer films) circles ($h\nu = 16.85$ eV) and rectangles ($h\nu = 21.11$ eV).

The broad shaded areas in the even symmetry bands represent regions of $k_\parallel$ where prominent features associated with the Cu *sp* band sweep through the angle-resolved energy-distribution curves (EDC's) near E_F. These structures were identified by $k_\parallel$ scans on clean Cu(100) surfaces; their presence precludes the study of even symmetry Fe bands in several regions of the two-dimensional Brillouin zone at specific photon energies. Interference of the *sp* band in studies of $p(1\times1)$Ni on Cu(100) (Ref. 3) and on Ag(100) (Ref. 11) is less of a problem because the overlayer Ni *d*-band features are much narrower than corresponding Fe features, and are therefore more easily distinguished from the *sp*-band features.

Some results of the theoretical studies are shown in Fig. 2. It is clear from the comparison that the calculations are in reasonably good agreement with the results shown in Fig. 2 in the regions of the two-dimensional Brillouin zone effectively probed by our experiments. Based on the calculations, the most prominent features observed in even symmetry geometry in the vicinity of E_F should be minority spin bands. Clearly, spin-polarized photoemission studies of this system will be informative. Our experimental results around $\bar{\Gamma}$, where Fe-derived features are not masked by the Cu *sp* band, support the number of bands and their binding energies predicted by the calculations. Odd symmetry bands observed along $\bar{\Gamma}$-$\bar{X}$ also appear to agree rather well with the calculated results, although there are significant differences (0.25 eV) between measured and calculated binding energies. In general, we find that the measured bands lie closer to the Fermi energy than the calculations predict for the majority and minority spin bands. This result could imply that the magnetic exchange splitting is actually smaller than predicted theoretically, or could be a manifestation of the importance of many-body effects (i.e., core-hole relaxation and correlation). Our previous experimental studies of the bulk[7] and surface[12] electronic properties of ferromagnetic Fe have shown, however, that correlation effects appear to be small; we therefore assume that these effects are also of minor importance in the Fe overlayer system.

One source of the discrepancy between the experiments and the theoretical results could be due to the fact that our measurements were conducted at 300 K, and our calculations correspond to $T=0$ K. We have not yet attempted any low-temperature experiments or experiments above $T=300$ K to detect possible temperature dependences in the results. A second possible source of disagreement between our experimental results and the calculations are some subtle layer dependencies we have observed in electronic properties of the Fe layers. Our LEED studies revealed two domain (2×1) LEED patterns in the coverage range between 0.5 and 1.0 monolayers.[10] These structures were not observed above full monolayer coverage, but their presence at low coverage suggests a competition between the substrate stabilized fcc structure and the possible tendency for Fe to form some other structure, in particular the normal bcc structure of bulk Fe. Displacement of iron atoms too small to detect by our LEED measurements is not inconceivable, and such displacements would cause significant changes in the electronic properties, particularly in the $\bar{\Gamma}$-$\bar{M}$ portion of the two-dimensional Brillouin zone. As noted in detail elsewhere,[10] while we have eliminated film growth models that assume the second layer begins before the first layer is more than 85% complete, we cannot discriminate between film growth in which islands of the second layer begin to nucleate when there is less than 15% of the first layer remaining to be completed.

The calculated bands for two layers of Fe and Cu(001) are modified from those of the monolayer coverage. The overall Fe surface layer derived features and dispersions remain similar to those shown plotted in Fig. 1. However, the localization of the surface state wave functions on the surface Fe layer becomes less pronounced due to the hybridization between surface and subsurface Fe layers. This change of bonding also manifests itself in a reduced upward dispersion of odd symmetry bands (dominated by d_{xz} and d_{yz} character) along the $\bar{\Gamma}$-$\bar{M}$ direction (cf. Fig. 2). Furthermore, the calculation predicts the existence of additional majority spin bands having a strong admixture of wave functions localized on the subsurface Fe layer, in particular, odd symmetry bands with binding energy ≈ 2 eV along $\bar{\Gamma}$-$\bar{X}$, and an upward dispersion band along $\bar{\Gamma}$-$\bar{M}$ (not shown in Fig. 2). Interference from the Cu *sp* band precludes observing the new even symmetry band for two monolayers, and the broad peak widths associated with the Fe overlayers renders it very difficult to differentiate small differences in spectra for one- or two-monolayers. Since the size of the

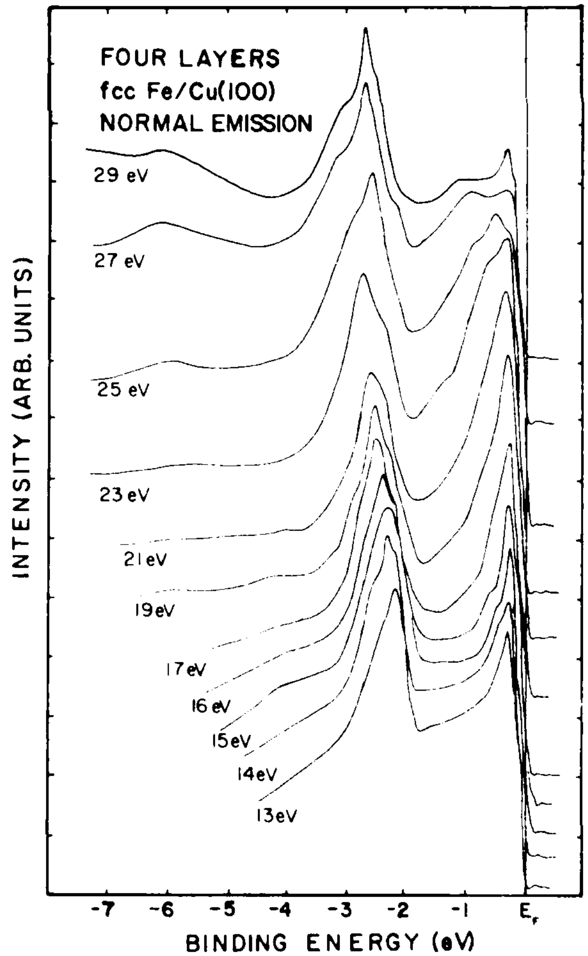


FIG. 3. Normal emission photoemission spectra for four-layer fcc Fe films stabilized on the Cu(100) surface. Peaks near E_F result from the Fe film and exhibit clear evidence of direct bulk interband transitions.

magnetic moments for the surface Fe layers remains almost identical for one- or two-monolayer coverage, the exchange splitting (ΔE_{ex}) is expected to be the same for both cases ($\Delta E_{ex} \approx 2.65 \pm 0.05$ eV).

Figure 3 displays normal emission spectra for four layers of Fe on Cu(100). Our LEED analysis[10] has shown that high-quality fcc films of epitaxial Fe form at this coverage. Features in the spectra near E_F are definitely due to emission from the Fe overlayer. This assignment was checked by obtaining corresponding spectra for clean Cu(100). Peaks near E_F exhibit clear dispersion with photon energy ($k_\perp$), indicating strong influence of direct bulk transitions. Dipole selection rules limit the symmetry of initial states probed in normal emission to Δ_1 and Δ_2 symmetry. The dispersion of the peaks near E_F with $k_\perp$ is consistent with the calculated bulk bands[13] of fcc Fe along the Γ-X direction of the three-dimensional Brillouin zone. It is therefore possible to carry out detailed energy band measurements of the bulk band structure of the fcc phase of ferromagnetic iron stabilized on Cu(100).[14]

In summary, our results have identified a second thin-film magnetic system in which considerable success has been achieved in the synthesis of the film, in its characterization, and in obtaining accurate electronic structure information. The $p(1 \times 1)$ Fe on Cu(100) appears to represent an additional excellent model system in which to explore the relationship between magnetism and electronic structure from the point of view of thin-film magnetism (two-dimensional magnetism), and the magnetism of a new artificially stabilized bulk phase (fcc Fe).

We wish to thank the staff of the Synchrotron Radiation Center, Stoughton, Wisconsin, particularly Roger Otte, for their support. This work was supported at the University of Texas by the National Science Foundation under Grant No. DMR-83-04368, and the University of Texas Joint Services Electronics Program AFOSR-F49620-82-C-0033, and at Northwestern University by the Office of Naval Research under Grant No. N 00014-81-K-0438, and by a computing grant from CRAY Research, Inc. The Tantalus storage ring is also supported by the National Science Foundation. We wish to thank Dr. Alex Ignatiev for the loan of a spot photometer. Finally, we wish to thank J. C. Hermanson for making his calculations available to us prior to publication.

[1] R. M. White, Science **229**, 11 (1985).

[2] U. Gradmann, Appl. Phys. Lett. **3**, 161 (1974), and references therein.

[3] M. Thompson and J. L. Erskine, Phys. Rev. B **31**, 6832 (1985).

[4] Epitaxy of metals is not new; see, for example, the review article by E. Bauer, Appl. Surf. Sci. **11/12**, 479 (1982). Our experiments are among the first to explore the electronic properties and magnetic exchange splitting of epitaxial magnetic films using photoemission. One previous study of a similar nature has been reported by R. Miranda, F. Yndurain, D. Chandesris, D. Lecante, and Y. Petroff, Phys. Rev. B **25**, 527 (1982).

[5] C. L. Fu, A. J. Freeman, and T. Oguchi, Phys. Rev. Lett. **54**, 2700 (1985).

[6] J. M. Kirschner, *Springer Tracts in Modern Physics* (Springer, New York, 1985), Vol 108.

[7] A. M. Turner, A. W. Donoho, and J. L. Erskine, Phys. Rev. B **29**, 2986 (1984), and references therein.

[8] E. Wimmer, H. Krakauer, M. Weinert, and A. J. Freeman, Phys. Rev. B **24**, 864 (1981), and references therein.

[9] K. W. A. Jesser and J. W. Matthews, Philos. Mag. **15**, 1097 (1967).

[10] M. F. Onellion, M. A. Thompson, J. L. Erskine, C. B. Duke, and A. Paton (unpublished).

[11] M. A. Thompson, M. F. Onellion, and J. L. Erskine (unpublished).

[12] A. M. Turner and J. L. Erskine, Phys. Rev. B **30**, 6675 (1984); **28**, 5628 (1983).

[13] D. Bagayoko and J. Calaway, Phys. Rev. B **28**, 5419 (1983).

[14] M. A. Thompson and J. L. Erskine (unpublished).

Reprinted from Applied Optics, Vol. 25, Page 3864, November 1, 1986

Laser-induced damage and ion emission of GaAs at 1.06 μm

Austin L. Huang, Michael F. Becker, and Rodger M. Walser

This study focused on the multipulse laser damage and the subdamage threshold ion emission of GaAs. The initial goals were to determine the pulse-dependent damage threshold and to correlate ion emission with surface damage. A Q-switched Nd:YAG laser was used to irradiate the ⟨100⟩ GaAs samples. Using values of N from 1 to 100, we obtained accumulation curves based on 50% damage probability. Corresponding damage threshold fluences were 0.4–0.8 J/cm^2 for $N > 1$ and 1.5 J/cm^2 for $N = 1$. We observed large site-to-site fluctuations in ion emission and found the onset of emission at 0.2 J/cm^2 for all cases. Once surface damage occurred, ion emission increased greatly. The observed behavior supports a surface cleaning model for the ion emission which precedes surface damage. Measurements of linear and nonlinear free carrier absorption were made, but no anomalous absorption was observed.

I. Introduction

The interaction between laser radiation and solids has been a perplexing problem for many years. Often the lifetime of an optical device is determined by its susceptibility to optical damage. For example, it is still a significant problem that the performance of GaAs injection lasers degrades at different rates with respect to power level thus implying an accumulation effect.[1] There has been much controversy and unexplained phenomena associated with the energy transfer mechanism of normal laser damage as well as the physical nature of surface damage.

The objectives of the experiments reported here were to characterize the statistical nature of surface damage for GaAs, particularly for multiple pulses on one site (N-on-1), to observe the relationship of charged particle emission to surface damage, and to observe the damage morphology of GaAs. In silicon it has been reported that charged particle emission is coincident with surface damage.[2] A relation between charged particle emission and surface damage has yet to be reported for GaAs. To explore the statistical nature of surface damage of GaAs, we performed various single-shot and N-on-1 subthreshold laser tests. We measured the positive charged particles emitted during each laser pulse to correlate the charge emission events with surface damage. Theoretical calculations for a thermal model, as well as a plasma production model, have been carried out to determine the mechanism responsible for the surface damage.

II. Samples and Apparatus

The GaAs samples used in the experiment were supplied by the microwave integrated circuit production group at Texas Instruments in Dallas, TX. The samples had a ⟨100⟩ orientation and were very lightly doped with chromium ($\approx 1 \times 10^{16}$ cm^{-3}) but otherwise undoped and unannealed. These samples were characterized by a resistivity of $>1 \times 10^{7}$ Ω-cm and an etch pit density of 40–60×10^{3} cm^{-2}. Only the front face of the wafer was polished to optical quality.

To prepare the sample for the vacuum tank we sequentially cleaned the GaAs wafers in boiling solvents of trichlorethylene, acetone, and methanol to assure that all contaminants had been removed. While in each solvent, the wafer was ultrasonically cleaned before proceeding. After the final cleaning in methanol, the wafer was rinsed in deionized water (>5 min at room temperature) and placed into the vacuum tank.

To measure the linear and nonlinear absorption in the transmission tests, the GaAs wafers were given an optical quality finish on both sides. The backside of the wafer was mechanically polished in a two-step process. Rough polishing with a grinding pad and 6-μm diamond ferrous paste was used to initially buff the wafer. The finishing pad used a liquid 0.05-μm alumina nonferrous suspension as a grinding media. After polishing, the wafer was chemically cleaned as described above.

The experimental system is shown schematically in Fig. 1. It utilized a Q-switched Nd:YAG laser with a full width at half-maximum (FWHM) pulse length of 45 ns, TEM$_{00}$ transverse mode, and wavelength of 1.064 μm. Although the pulse envelope was Gaussain, the laser was not single longitudinal mode. A knife-edge scanning technique was used to measure the focused spot diameter on the sample surface, which was found to be 580 μm. The laser energy fluctuated from

The authors are with University of Texas at Austin, Austin, Texas 78712.

Received 18 April 1986.
0003-6935/86/213864-07$02.00/0.

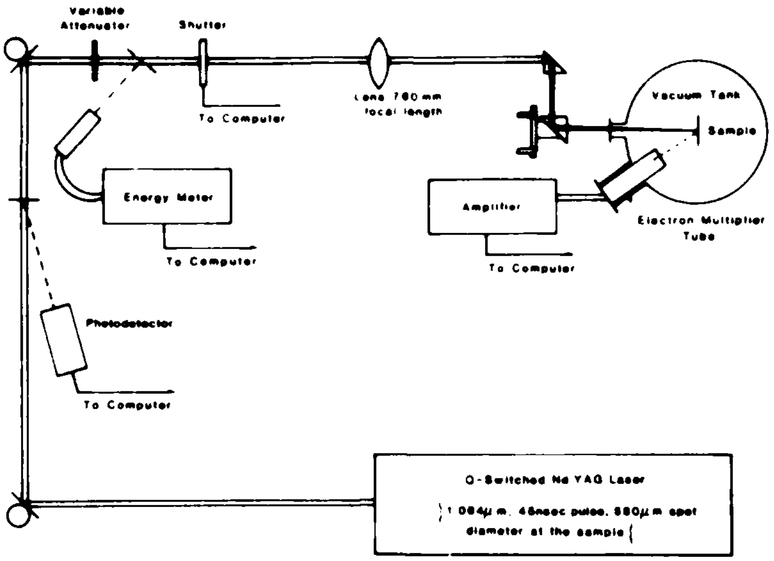


Fig. 1. Experimental setup.

pulse to pulse resulting in a 2–5% standard deviation in energy for each data run. The laser was operated at a 15-Hz repetition rate, and the incident energy was attenuated by rotating a halfwave plate in front of a polarizer.

A computer was utilized to control and count the laser pulses striking the GaAs sample and to record the charge collected. The computer had a Z80 microprocessor and supported several data acquisition ports via sample and hold amplifiers and a multichannel A-D converter. The delay for the sample and hold amplifiers was set to 130 μs in hardware. This assured that the integrated values read for the charge emission were near the peak value but were collected well after the laser noise burst. The electronics were triggered by a vacuum photodiode. A shutter, controlled by one of the computer's digital output ports, was used to select pulses for sample irradiation. To allow time for data computation and storage betwee pulses, the shutter selected every third pulse from the laser train.

A 760-mm focal length lens focused the beam onto the sample. The long focal length gave a region of constant spot size ~3 mm deep at the focal plane and permitted greater error in the placement of the sample. A system of 90° prisms was placed after the lens to scan the beam across the sample in the vacuum tank.

To detect charged particle emission the sample and detection device were placed into the vacuum environment. The system was dry pumped to achieve experimental pressures of $<1 \times 10^{-6}$ Torr. A Hamamatsu electron multiplier tube model R596 served as the ion detector and was placed inside the vacuum system at a 36° angle to the laser beam path. The sample surface was normal to the optical path. The output of the electron multiplier was connected to a capacitive voltage divider and amplifier. The dynamic range of the data acquisition system was increased by making one channel 100× less sensitive than the other. The electron multiplier current gain was estimated to be $\sim 0.75 \times 10^6$ at a voltage of -2 kV at the first dynode. A data value of 1000 in the more sensitive A-D converter channel corresponded to a charge of $\sim 1.5 \times 10^{-17}$ C. This -2-kV bias potential not only determined the gain of the electron multiplier, but it created an attractive potential sufficient to collect all emitted positive ions.

III. Experimental Data

Three experiments were conducted on the cleaned samples: (1) an emission scan of the GaAs sample at a fixed laser fluence; (2) an N-on-1 emission/damage threshold test; and (3) a transmission test. All the tests were performed in vacuum with the exception of the transmission test. The calibration and beam profiles were checked at the start and end of each experimental session to assure accurate beam fluence measurements.

The emission scan of the GaAs sample checked the uniformity of charge emission under constant laser fluence. The scan spots were separated by 0.8 mm, and forty-two samples were taken. The sample was irradiated with a constant fluence of 0.63 J/cm^2 (± 7%). The results in Fig. 2 show a bilevel emission contour. A different pattern of emission variations was observed on each sample. Site-to-site fluctuations of the defect density in GaAs have been previously noted[3] and could disrupt the statistics of charged particle emission and the emission threshold. To minimize the effects of these site-to-site variations, we conducted the emission/damage tests over a small area of the wafer.

Other experiments were concerned with the multipulse and single-pulse irradiance of GaAs. In the multipulse experiments, fluences below the single-pulse damage threshold were used to search for accumulation effects associated with either emission or damage. The objectives of these experiments were (1) to identify accumulation effects associated with surface dam-

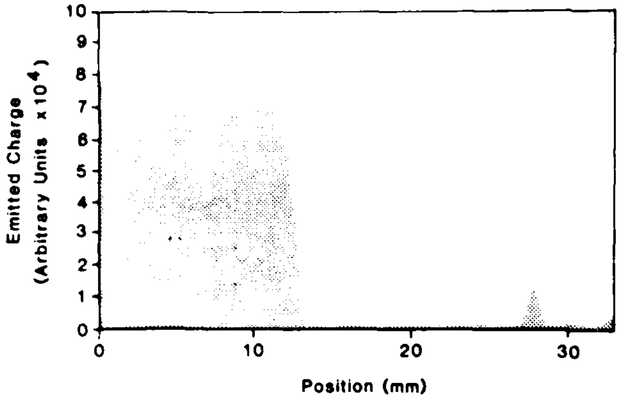


Fig. 2. Charged particle emission vs position on wafer for a constant laser fluence of 0.63 J/cm^2.

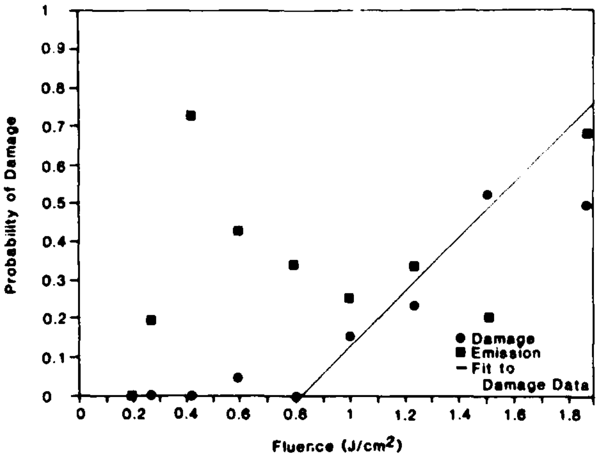


Fig. 3. Damage and emission probabilities vs fluence for $N = 1$.

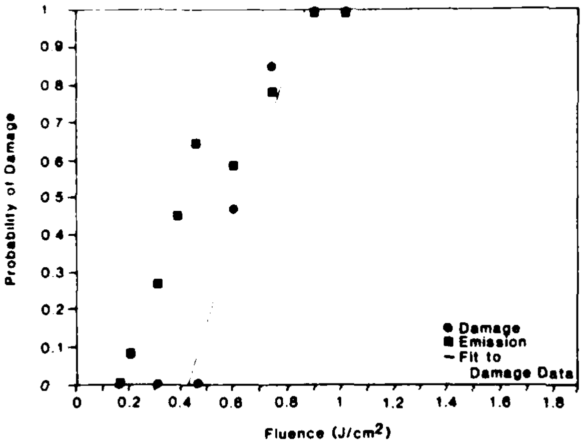


Fig. 4. Damage and emission probabilities vs fluence for $N = 10$.

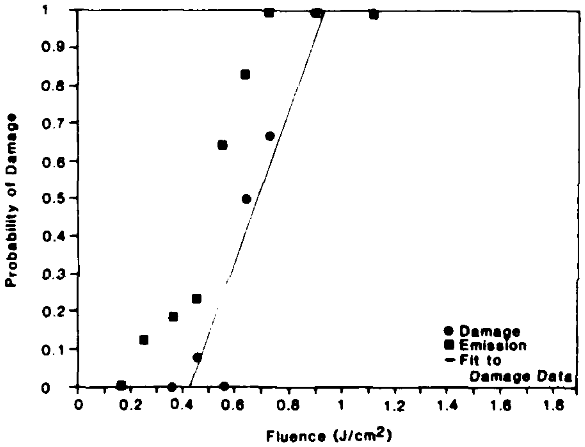


Fig. 5. Damage and emission probabilities vs fluence for $N = 100$.

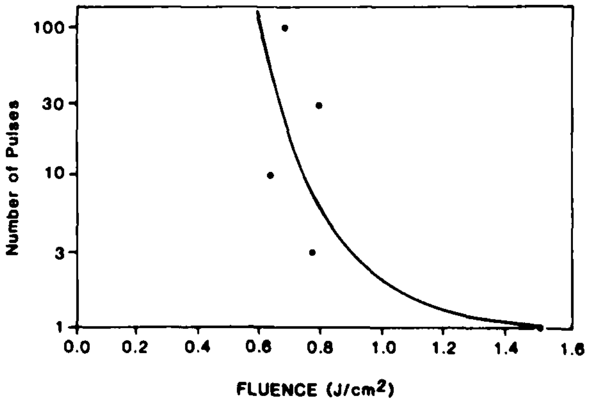


Fig. 6. Number of pulses required to reach 50% damage probability vs fluence. The line was hand drawn to aid the viewer.

age, (2) to measure a damage threshold based on 50% probability statistics, (3) to search for correlations between surface damage and charged particle emission, and (4) to determine the damage morphology as a function of fluence and number of pulses.

For each value of the number of pulses N used (N = 1, 3, 10, 30, and 100) we irradiated 150 to 200 different sites to obtain the statistics of surface damage. Surface damage was identified by searching for surface changes with a Normarski optical microscope at 200×. The damage data for N = 1, 10, and 100 are plotted in Figs. 3–5, respectively. In each plot we used a linear curve fitting program to obtain the damage probability. From the linear fit we obtained the 50% probability fluence for each value of N and used this value in the accumulation plot in Fig. 6.

For all values of N, we observed the onset of charged particle emission at an average fluence of 0.2 J/cm^2 (±0.07 J/cm^2). The values for the onset of emission and the 50% damage threshold are given in Table I. From these damage fluence values and their plot in Fig. 6, it is apparent that an accumulation effect was present in that the threshold decreased after the first

Table I. Onset of Charge Emission and 50% Damage Threshold Fluence for *N*-on-1 Experiments

N pulses	Charged particle emission onset fluence (J/cm²)	50% Damage fluence (J/cm²)
1	0.18	1.52
3	0.15	0.78
10	0.17	0.64
30	0.32	0.79
100	0.17	0.68

pulse. As the number of laser pulses incident on the wafer increased, the 50% damage fluence decreased and then leveled off. It is not clear from this scattered data whether the decrease continues slowly for $N > 1$ or the curve is level. Scatter in the threshold data and the related observation of low slopes in the damage probability vs fluence curves have been associated with the presence of the local defects.[4] This possibility will be correlated with damage mcrphology in a later section. Similar curves to the one shown in Fig. 6, but showing a monotonic decrease in threshold, have been reported for metals,[5] polymethylmethacrylate, and modified polymethylmethacrylate.[6] The single-shot 50% damage fluence for GaAs is in agreement with values previously measured.[7,8]

The positive charged particle emission data were collected on a pulse-by-pulse basis. A typical profile for $N = 30$ is shown in Fig. 7. The site from which this emission profile was obtained was damaged. From these data we note that the first few pulses (from two up to ten depending on the case) induced a charge emission that decreased as the pulse number increased. After this, the emission increased greatly. Since the microscopy and determination of damage were done after termination of the experiment, we were unable to fix accurately the pulse number at which damage was initiated. Visual observation of the initial flash was a somewhat insensitive measure of damage initiation. We interpret the decrease in the first few pulses as a surface cleaning effect. These first pulses cleaned any residue either left from the chemical cleaning process or ejected from previously laser irradiated sites. This emission could also be due in part to the depletion of the more volatile specie, As, from the surface atomic layers. Figure 7 also shows that, once the site experiences surface damage, charge emission is greatly increased. The emission at non-damaged sites always decreased with increasing number of pulses.

Figures 8 and 9 are SEM micrographs showing the development of laser-induced surface damage morphologies. The initial surface was featureless and without contrast. In the Normarski microscope, the initial surface change appeared as a depression with an area equal to that of the laser spot. Within this area, there were several very small pitted regions (similar to those shown in Fig. 8 but not as well defined). As the fluence increased, or the number of pulses increased, the initial surface damage evolved into the melt pits shown in Fig. 8. Similar damage pits have been seen in

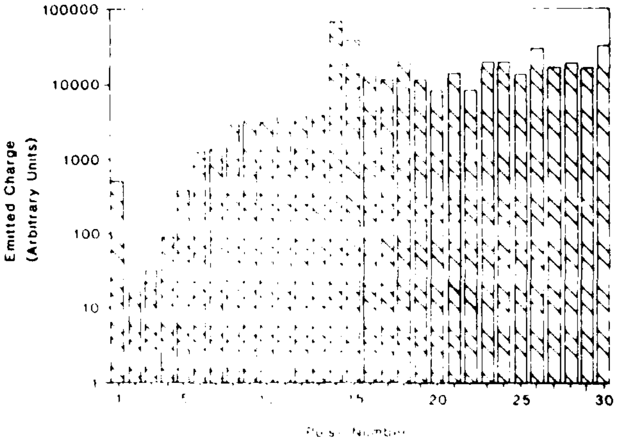


Fig. 7. Pulse by pulse charged particle emission for $N = 30$ at an average fluence of [illegible] J/cm².

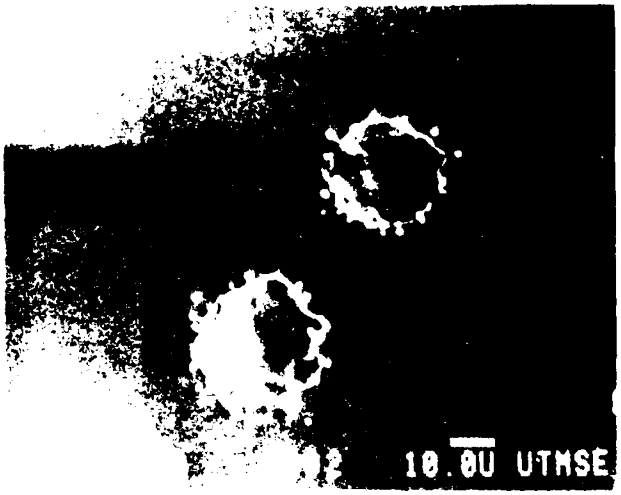


Fig. 8. SEM micrograph of damage due to three pulses at an average fluence of 0.51 J/cm².

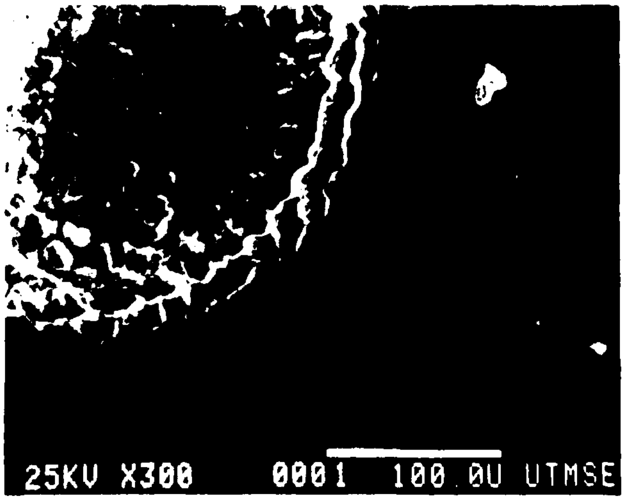


Fig. 9. SEM micrograph of damage due to thirty pulses at an average fluence of 0.38 J/cm².

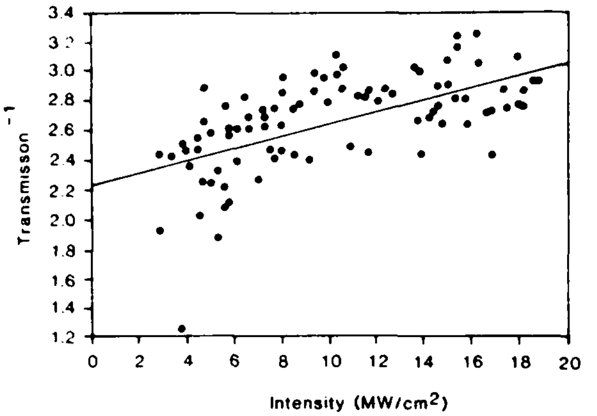


Fig. 10. Inverse transmission vs intensity for a GaAs wafer polished on both sides. The solid line represents a least-squares linear fit to the data.

GaAs at 1.064 μm.[1,7,8] The damage evolution ends with severe cratering as shown in Fig. 9. The theory of Fauchet and Siegman[9] would seem to apply in this case. They postulate that laser initiated ripple patterns on the surface of silicon and GaAs are formed from the interaction of the incident laser wave front with scattered optical surface waves. At 1.064 μm, however, with a pulse length of 45 ns and spot diameter of 580 μm, no ripple patterns were seen in our experiments. Accordingly, the longer pulse lengths may inhibit ripple formation.

IV. Discussion and Models

Thermal models have been used in the past to explain both laser annealing processes and laser damage phenomena. These models assumed uniform heating, and the possibility of inhomogeneities was not included. The necessity of inhomogeneous processes is made apparent by examining the laser absorption in the sample. From the slope and intercept of the linear fit of the inverse transmission vs intensity shown in Fig. 10, the linear and nonlinear absorption behavior may be determined. Although small values of linear absorption coefficient α_0 are not measured accurately by this method, a value for α_0 is a by-product of the curve fitting procedure. The linear absorption coefficient was found using the equation

$$T_0 = (1 - R)^2 \exp(-\alpha_0 L) , \tag{1}$$

where T_0 is the zero fluence transmission intercept, R is the reflectivity at the air–gallium arsenide wafer interface, and L is the thickness of the GaAs sample. In our case T_0 was 0.45 (from Fig. 10), R was calculated to be 0.306 based on tabulated refractive-index data, and L = 0.0635 cm. Substituting these values into Eq. (1) and solving yield α_0 = 1.2 cm^{-1} (±40%). The free carrier or nonlinear absorption can be determined from a linearized equation for inverse transmission vs intensity as shown in the following analysis. The coupled differential equations describing the carrier density and optical intensity are

$$dN/dt = \alpha_0 I/hf + \beta_0 I^2/2hf - N/t_c , \tag{2}$$

$$dI/dz = -\alpha_0 I - \beta_0 I^2 - \sigma N I , \tag{3}$$

where $N = N(x,y,z,t)$ is the time- and space-dependent carrier density, $I = I(x,y,z,t)$ is the time- and space-dependent optical intensity, hf is the photon energy, t_c is the free carrier lifetime, β_0 is the two-photon absorption coefficient, and σ is the free carrier optical cross section. In Eq. (2), carrier relaxation due to Auger recombination has been neglected due to the low expected carrier densities. At 10 MW/cm^2, for example, two-photon absorption is negligible with respect to one photon and free carrier absorption; $\beta_0 I$ = 0.23 cm^{-1} and β_0 = 0.023 cm/MW.[10] For this reason, two-photon absorption will be neglected in both Eqs. (2) and (3). Also in Eq. (2) we have assumed that the one-photon absorption is impurity dominated and that only one charge carrier is generated per absorbed photon.

The solution to these equations will proceed much as described by Boggess *et al.*[11] except that in our case the optical pulse length is much longer than the free carrier lifetime, and the solution to Eq. (2) may be taken to be quasi-steady state. The solution to Eq. (2) for instantaneous carrier density is given by

$$N = t_c \alpha_0 I/hf . \tag{4}$$

This allows Eqs. (3) and (4) to be combined and solved. That is, Eq. (3) must be integrated over z as well as x,y, and t to obtain a solution in terms of toal transmitted energy, which is an experimentally observable parameter. These integrations may be performed as shown in Ref. 11 for Gaussian temporal and spatial profiles if the free carrier absorption is not excessively large. This approximation is possible if the intensity is less than a critical intensity for free carrier absorption as given by

$$I < I_c = \frac{hf}{\sigma t_c} \frac{1}{(1 - R)\,[1 - \exp(-\alpha_0 L)]} . \tag{5}$$

For the paramaters in our experiments, I_c = 17.3 MW/cm^2. The solution to Eq. (3) for this case is conventionally written in terms of the inverse total energy transmission T, which linearizes the relationship

$$T^{-1} = [(1 - R)^2 \exp(-\alpha_0 L)]^{-1} + \sqrt{\frac{\ln 2}{2\pi}}\, \frac{(\sigma t_c)[1 - \exp(-\alpha_0 L)]}{hf(1 - R)[\exp(-\alpha_0 L)]} \left(\frac{E}{\pi w_0^2 t_p}\right) . \tag{6}$$

where E is the total transmitted laser energy, t_p is the optical pulse width (FWHM), and w_0 is the laser spot radius (1/e^2 intensity). The term on the right in brackets may be thought of as an effective intensity and is the quantity plotted on the abscissa in Fig. 10.

A linear fit to the data in Fig. 10 gives a slope of 40.2 and a corresponding value of (σt_c) = 2.0 × 10^{-25} s cm^2. This product of two experimentally measurable quantities may be checked against published data, but unfortunately there is uncertainty and dispersion in the published values. For σ, values ranging from 5 × 10^{-20} (Ref. 12) to 5 × 10^{-17} cm^2 (Ref. 6) extrapolated to 1.06 μm have been reported. For t_c, one direct measurement gave 32 ns,[13] but this value is known to be strongly dependent on impurities and defects which increase the recombination rate. If we use a value of $t_c = 3.2 \times 10^{-8}$ s (one of the largest values reported, presumably for relatively pure material), our experiments predict a free carrier cross section of 6.3×10^{-18} cm^2, which is well within the range of reported values given above even if the free carrier lifetime were a factor of 10 smaller to account for a possibly larger deep level impurity density. In addition, the functional dependence of T^{-1} on intensity agrees with the theoretically predicted form as shown in Fig. 10. From this, we conclude that we have included the major homogeneous absorption mechanisms in our calculations and that no gross anomalous absorption was involved.

To determine the mechanism responsible for the laser-induced damage, we examine the temperature change of the surface. Both linear and free carrier absorption must be included in the heating model. Assuming that all the absorbed energy is thermalized without thermal or carrier diffusion, the peak surface temperature rise can be written

$$\Delta T = (dI/dx)_{\text{surf}} t_p / C_v \rho \, . \qquad (7)$$

where t_p is the pulse length, C_v is the heat capacity, and ρ is the mass density. In this approximate calculation we assume a rectangular pulse shape. Substituting Eqs. (3) and (4) into Eq. (7) and taking the surface reflectivity into account give

$$\Delta T = [\alpha_0 + (1 - R)\sigma t_c \alpha_0 I_0 / hf](1 - R) I_0 t_p / C_v \rho \, , \qquad (8)$$

where I_0 is the incident intensity. For GaAs at room temperature, we have $C_v = 0.327$ J/g K and $\rho = 5.32$ g/cm^3. Using a typical damage intensity for the single-shot case, $I_0 = 19$ MW/cm^2 and the measured value for (σt_c), we obtain a maximum temperature rise of 6.1°C. This diffusionless model assumes homogeneous energy absorption and temperature-independent reflection and linear and nonlinear absorption coefficients.

From the results of the previous calculation we conclude that uniform heating is not sufficient to explain the melting effects observed on the GaAs surface. A similar conclusion has been reported for GaAs by Grasyuk and Zubarev[14] and for silicon by Merkle *et al.*[15] and Becker *et al.*[2] However, this does not exclude the possibility of an abnormal absorption or heterogeneous absorption in defects at or near the GaAs surface.

An alternative process that might lead to damage is lattice disruption caused by a high-density carrier plasma produced during the laser–solid interaction. The possible relation of this process to silicon annealing has been reported previously.[16,17] To investigate this possibility, we calculated the peak number of carriers produced by the impurity absorption.

The peak carrier density produced is calculated from Eq. (4) where ambipolar diffusion was neglected because of the short pulse length. To obtain an upper bound for the carrier density, we used the incident intensity of a typical 1-on-1 damage pulse, $I = 19$ MW/cm^2. Accounting for the reflectivity, we obtained a peak carrier density of $N = 1.7 \times 10^{16}$ cm^{-3}. For comparison, the compensated intrinsic carrier concentrations were calculated. Accounting for the chromium dopant concentration ($\approx 1 \times 10^{16}$ cm^{-3}) and the resistivity of the sample ($\geq 1 \times 10^7$ Ω cm), we calculated an electron concentration of 7.35×10^7 cm^{-3} and a hole concentration of 1.56×10^9 cm^{-3}. Mobilities of 5000 and 300 cm^2/V s were used for the electrons and holes, respectively. Although the photogenerated carrier density is much greater than the compensated intrinsic density, it is still quite small compared to the densities where plasma effects become important, 10^{19}–10^{20} cm^{-3}. Therefore, we can disregard the plasma model from further consideration as a primary damage mechanism in GaAs.

The charged particle detector system monitored the pulse-by-pulse charge emitted from the GaAs sample. The magnitude of the emitted charge has been plotted Fig. 7 on a pulse-by-pulse basis for a site which exhibited surface damage. By examining the emission profiles, the emitted charge was observed to decrease after the first few pulses and later increase at least 2 orders of magnitude when damage occurred. The relatively small decreasing emission profile of the initial pulses appears to be a surface cleaning effect. The later increase of charge emission after the fifth pulse in this case is attributable to surface damage. For GaAs, only two types of charge emission profile were observed: either a cleaning effect without damage or a cleaning effect with damage. Damage in the absence of a cleaning effect was never observed, although at very high intensities the two profiles merged and became indistinguishable. This differs greatly from the emission characteristics of silicon as reported in Ref. 2. For silicon, emission of charged particles coincided with surface damage initiation, and no cleaning emission was observed. Furthermore, the emission of charge prior to damage in GaAs did not affect the observed damage behavior in any detectable way.

It appears that neither the uniform heating model nor the carrier pair production model can adequately describe the damage mechanism of GaAs. In addition, it does not seem possible to relate the nature of the damage to charge emission due to surface cleaning, although the latter appears to be a necessary, and possibly essential, precursor to damage. From these results we are led to speculate on other possible influences that may cause, or enhance, surface damage. Lattice defects near the surface of the GaAs wafer are possible energy absorption sites for nucleating damage. These lattice defects include anomalous vacancies, interstitials, and dislocations introduced during the crystal growth or surface preparation processes. A single-point defect seems an improbable physical area for absorbing sufficient energy to cause a melt spot. It is more probable that a cluster of point defects will act as an efficient absorption site. This would result in a surface damage morphology of random melt points within the beam diameter of the laser.

Other possible sites for the nucleation of damage are physical surface defects caused by processing or physical handling of the wafer that would scratch the surface. Surface defects on an optical surface will increase absorption and scattering, possibly creating absorption nuclei which could lead to damage. The potential for nucleation at physical defects have been previously speculated by Smith.[8] In our experiments, microscopic inspection before damage has eliminated the possibility of gross surface defects (>1 μm) as a cause of damage. One would not have expected these types of defect in integrated circuit quality material.

V. Conclusion and Comments

We observed accumulation effects in GaAs with multiple-pulse laser irradiance below the 1-on-1 damage threshold fluence. These experiments also yielded various N-on-1 laser damage thresholds derived from the statistics of the surface damage. In the experiments we also measured the linear and nonlinear optical absorption constants for the GaAs sample. With this information, calculations were performed to determine the mechanism responsible for the surface damage. Results for the uniform heating and carrier production models clearly show that these mechanisms could not be responsible for damage. The possibility remains that nonuniform localized carrier generation and heating may be involved in the damage process. An inhomogeneous heating model is supported by the observed damage morphologies. It is also possible that the sample contained absorbing defects, or defect clusters, that could grow and contribute to the damage process or surface defects (e.g., pits and scratches) that might cause a melt spot to nucleate. Although in our experiments the former is believed to be more likely.

By investigating the pulse-by-pulse history of the charged particle emission, a cleaning-type profile was discovered. This cleaning emission proved to have no observable influence on the subsequent damage. Apparently, adsorbates remaining on the wafer surface after cleaning did not act as absorbing sites to cause surface damage. The charge emission proved to be very noisy from site to site. Although the emission did not strongly correlate with the event of surface damage, once surface damage commenced an increase in charge emission was observed. This behavior is consistent with the idea of a hot partially ionized vapor being released from melt pits as damage proceeds.

Further investigation is needed to understand the mechanism of energy transfer that causes damage and emission. The relationship of the accumulation effect to lattice defects could prove to be a fundamental key in understanding the damage process. Another parameter which might influence the damage threshold is the dopant level of the semiconductor material.

We gratefully acknowledge support given by the DoD Joint Services Electronics Program under research contract AFOSR F49620-82-C-0033. We also thank John Beall at Texas Instruments for providing the GaAs samples.

References

1. H. Kressel and H. Mierop, "Catastrophic Degradation in GaAs Injection Laser," J. Appl. Phys. **38,** 5419 (1967).
2. M. F. Becker, Y. K. Jhee, M. Bordelon, and R. M. Walser, "Charged Particle Exoemission from Silicon during Multi-Pulse Laser Induced Damage," in *Fourteenth ASTM Laser Damage Symposium*, NBS Spec. Publ. 669 (1983); Y. K. Jhee, M. F. Becker, and R. M. Walser, "Charge Emission and Precursor Accumulation in the Multiple-Pulse Damage Regime of Silicon," J. Opt. Soc. Am. B **2,** 1626 (1985).
3. S. Dannefaer, B. Hogg, and D. Kerr, "Defect Characterization in Gallium Arsenide By Positron Annihilation," in *Thirteenth International Conference on Defects in Semiconductors*, L. C. Kimerling and J. M. Parsey, Jr., Eds. (American Institute of Mining, Metallurgical, and Petroleum Engineers, New York, 1985), pp. 1029–1033.
4. S. R. Foltyn, "Spotsize Effects in Laser Damage Testing," in *Fourteenth ASTM Laser Damage Symposium*, NBS Spec. Publ. 669 (1983).
5. C. S. Lee, N. Koumvakalis, and M. Bass, "A Theoretical Model For Multiple-Pulse Laser-Induced damage to Metal Mirrors," J. Appl. Phys. **54,** 5727 (1983).
6. A. A. Manenkov, G. A. Matyushin, V. S. Nechitailo, A. M. Prokhorov, and A. S. Tsaprilov, "On The Nature of Accumulation Effect in the Laser-Induced Damage to Optical Materials," at *Fourteenth ASTM Laser Damage Symposium*, NBS Spec. Publ. 669 (1983).
7. J. R. Meyer, M. R. Kruer, and F. J. Bartoli, "Optical Heating in Semiconductors: Laser Damage in Ge, Si, InSb, and GaAs," J. Appl. Phys. **51,** 5513 (1980).
8. J. L. Smith, "Surface Damage of GaAs from 0.694 and 1.06 mm Laser Radiation," J. Appl. Phys. **43,** 3399 (1972).
9. P. M. Fauchet and A. E. Siegman, "Surface Ripples on Silicon and Gallium Arsenide under Picosecond Laser Illumination," Appl. Phys. Lett. **40,** 824 (1982).
10. E. W. Van Stryland, M. A. Woodall, H. Vanherzeele, and M. J. Soileau, "Energy Band-Gap Dependence of Two-Photon Absorption," Opt. Lett. **10,** 490 (1985).
11. T. F. Boggess, K. M. Bohnert, K. Mansour, S. C. Moss, I. W. Boyd, and A. L. Smirl, "Simultaneous Measurement of the Two-photon Coefficient and Free-carrier Cross Section Above the Bandgap of Crystalline Silicon," IEEE J. Quantum Electron. **QE-22,** 360 (1986).
12. R. K. Willardson and A. C. Beer, *Semiconductors and Semimetals: Optical Properties of III-V Compounds, Vol. 3,* (Academic, New York, 1967), p. 409.
13. D. S. Chemla, D. A. B. Miller, P. W. Smith, A. C. Gossard, and W. Wiegmann, "Room Temperature Excitonic Nonlinear Absorption and Refraction in GaAs/AlGaAs Multiple Quantum Well Structures," IEEE J. Quantum. Electron. **QE-20,** 265 (1984).
14. A. Z. Grasyuk and I. G. Zubarev, "Interaction of Semiconductors with Intense Light Fluxes," Sov. Phys. Semicond. **3,** 576 (1969).
15. K. L. Merkle, R. H. Uebbing, H. Baumgart, and F. Phillipp, "Picosecond Laser Pulse Induced Damage in Crystalline Silicon," in *Laser and Electron-Beam Interactions with Solids*, B. R. Appleton and G. K. Celler, Eds. (Elsevier, New York, 1982), pp. 337–342.
16. J. A. van Vechten and A. D. Compaan, "Plasma Annealing State of Semiconductors: Plasmon Condensation to a Superconductivity Like State at 1000 K?," Solid State Commun. **39,** 867 (1981).
17. D. M. Kim, R. R. Shah, D. Von der Linde, and D. L. Crosthwait, "Picosecond Dynamics of Laser Annealing," in *Laser and Electron-Beam Interactions with Solids*, B. R. Appleton and G. K. Celler, Eds. (Elsevier, New York, 1982) pp. 85 90.

Modeling of Ion-Implanted GaAs MESFET's by the Finite-Element Method

N. SONG, DEAN P. NEIKIRK, MEMBER, IEEE, AND TATSUO ITOH, FELLOW, IEEE

Abstract—**We discuss the results of a new two-dimensional (2-D) finite-element model for GaAs MESFET's made by ion implantation. Several different devices are characterized by varying gate recess and doping profile. The simulation, in qualitative agreement with experimental findings, shows that a FET with a shallow gate recess exhibits a similar behavior to a FET with a deep implantation, i.e., an improvement in linearity, a higher pinch-off voltage, and a decrease in transconductance.**

I. Introduction

The PERFORMANCE of ion-implanted GaAs MESFET's is usually predicted [1], [2] using an analytical model such as the "two region model" [3]. To include diffusion current terms and transverse current terms more adequately, we have adopted a two-dimensional finite-element method (2-D FEM) to numerically simulate GaAs MESFET's with nonuniform doping profiles and various gate recesses. In particular, we have studied a model ion-implanted structure, varying both the depth of the implant relative to the source–drain surface and the depth of the gate recess. For these numerical simulations the FEM has several advantages over the finite-difference method, such as the flexibility of the mesh size used in the calculation and the inclusion of current conservation without the need of phantom nodes at insulatory boundaries where default Neumann boundary conditions are applied.

II. Simulation

Using a finite-element algorithm from a previous work [4], the coupled Poisson equation and current continuity equations in the device are solved to find the unknown potentials and electron concentrations under various bias conditions using a standard FEM formulation [5]. Once the electric field is calculated by taking an average for all Gaussian point values in an element, an effective mobility is obtained using the assumed velocity versus electric-field relationships [6]. Using the calculated electric field and velocity, a temperature is found from the energy transport equation [7] neglecting the time dependence of energy for a first-order approximation. With the effective mobility and temperature, the diffusion coefficient is calculated from Einstein's relation.

The time step used in the simulation is fixed at 0.01 ps, i.e., the dielectric relaxation time for material doped to a level similar to the peak of the doping profile. From these, I-V characteristics and equivalent circuit parameters have been calculated. Unlike the triangular elements preferred in many semiconductor simulations, the bilinear rectangular elements adopted here make it easier to manipulate the input and output data by block mesh generation. The boundary conditions are applied easily by the so-called penalty method [8]. It is found that satisfactory results are obtained with less than 500 nodes. For a typical node number of 307, it takes less than 2 s of execution time for each iteration on a CDC Dual Cyber 170/750 at the University of Texas at Austin.

III. Device Structure and Approximations

The basic device structure used in this simulation is shown in Fig. 1. We have assumed a finite active-layer depth, neglecting carrier transport far from the implant peak. The deep region is considered to be semi-insulating. This assumption is based on the reported observations that phenomena such as deep traps [2], [9], [10] contribute to mobility degradation in this deep active region. For all the simulations we have also assumed a Gaussian-like active-layer doping profile, with a constant peak carrier concentration of $2.5 \times 10^{17}\,\mathrm{cm}^{-3}$. First, the effects of gate recess changes were studied by holding the depth of the doping profile constant (as shown in Fig. 1(b)), with the peak located at 85 nm from the source–drain surface. Next, to study the effects of implant depth, the gate recess was set to zero and the surface location of the electrodes was varied instead, which results in a change in the gate-to-profile peak distance. Here the implanted doping profile is left unchanged.

The doping concentration below the source and drain contacts is fixed at $2 \times 10^{17}\,\mathrm{cm}^{-3}$, and below the Schottky gate contact is set to zero to approximate the built-in depletion region. The built-in gate voltage is assumed to be 0.8 V. It is assumed that the low field mobility is $3200\,\mathrm{cm}^2/\mathrm{V}\cdot\mathrm{s}$ (appropriate for electrons near the peak of the doping profile), the saturation velocity is 10^7 cm/s, and the threshold electric field is 4 kV/cm. In order to simplify the model we have taken these parameters to be independent of doping.

IV. Results and Discussion

The I-V characteristics obtained by varying the gate recess depth are shown in Fig. 2. The data for a gate depth of 55 nm are comparable to previous experimental results [11]. As the gate depth increases, the transconductance becomes larger, the magnitude of pinch-off voltage becomes smaller, and the linearity of transconductance variation becomes poorer. Similar phenomena are observed for the various implantation cases shown in Fig. 3. Here the I-V curves behave in a fashion quite

Manuscript received November 4, 1985; revised January 16, 1986. This work was supported by Joint Services Electronics Program under Contract F-49620-82-C-0033.

The authors are with the Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, TX 78712.

IEEE Log Number 8607960.

0741-3106/86/0400-0208$01.00 © 1986 IEEE

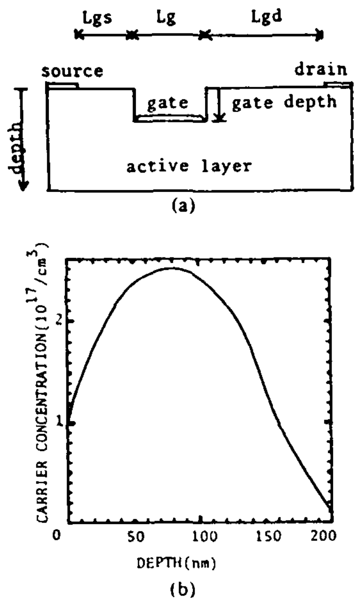


Fig. 1. Two-dimensional GaAs MESFET model. (a) Dimensions of device simulated. Drain and source length: 0.2 μm; gate length L_g: 0.5 μm; gate-to-source distance L_{gs}: 1 μm; gate-to-drain distance L_{gd}: 1.5 μm; gate width: 300 μm. (b) Carrier profile used in simulation. All the carrier profiles used have a peak concentration of 2.5×10^{17}/cm^3. For Fig. 2, the source and drain surface is located at 0 nm, and for Fig. 3 the source, drain, and gate surfaces are all located at 30, 55, and 70 nm for project ranges 55, 30, and 15 nm, respectively.

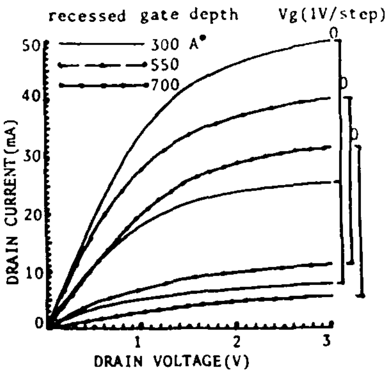


Fig. 2. *I-V* characteristics simulated by varying gate depth in recessed gate structure.

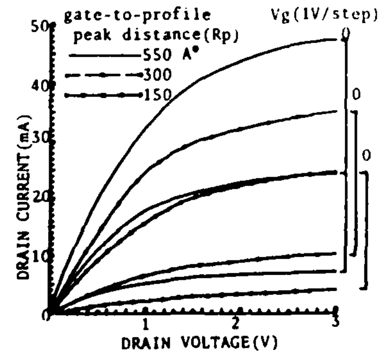


Fig. 3. *I-V* characteristics simulated by varying carrier profile in nonrecessed gate structure.

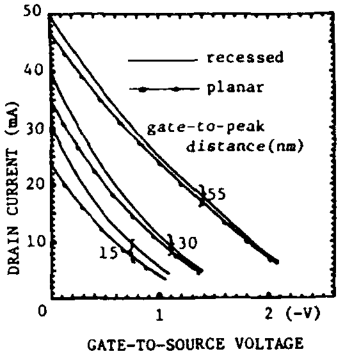


Fig. 4. Plot of drain-current variation versus gate-to-source voltage for Figs. 2 (recessed) and 3 (planar).

similar to the *I-V* curves for a gate recess with identical gate-to-profile peak distances. These phenomena are explicitly shown in Fig. 4 for one typical drain voltage. Here, the variation of the slope is related to the linearity. Such results as an improvement in linearity with a corresponding decrease in intermodulation distortion for deeper implants (i.e., larger gate-to-peak distance) have been reported previously [12]-[15]. This particular characteristic is important in large-signal applications, and therefore a trade-off between transconductance and linearity may be required in these applications.

The effects of changes in the tail of the active-layer doping profile have also been investigated. Current-voltage curves for devices with the same profile shown in Fig. 1(b) but with an additional tail 10-20 nm long (extending the active-layer doping concentration down to 3×10^{15} cm^{-3} from the original 1×10^{16}) have been calculated. It is found that the pinch-off voltage increases somewhat, while the transconductance decreases near pinch-off. These changes are thought to be related to the steepness of the doping profile near the substrate interface. The steeper profiles appear to give more linear behavior near pinch-off—an important consideration in the design of low noise devices.

Typical values of the small-signal transconductance and gate-to-source capacitance at zero gate bias are shown in Fig. 5. In the case of recessed gate device, it can be seen that the cutoff frequency $f_T(f_T = g_m/(2\pi C_{gs}))$ increases with gate recess depth. This effect can be explained as follows: as the distance between the gate and the active layer implant peak decreases, a smaller gate voltage swing is required to affect a given drain current change [2], thus increasing the transconductance g_m. At the same time gate-to-source capacitance C_{gs} decreases because the gate depletion layer extends beyond the implant peak as the recess becomes deeper. These effects combine to produce an increase in f_T. For very deep recesses (i.e., very small gate-to-implant peak separation) the transconductance begins to decrease, and the cutoff frequency f_T saturates at about 24 GHz. Although no graphical data are shown here, somewhat different results are obtained for gate bias voltages near pinch-off, where the depletion layer is always deeper than the implant peak. It is found that as the recess increases, C_{gs} increases rather than decreases as it does in the low gate bias case.

The results for planar devices obtained by varying gate-to-

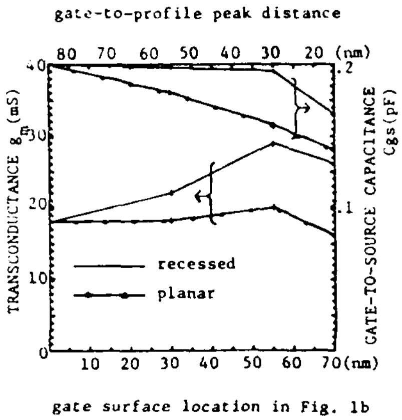


Fig. 5. Calculated results of transconductance and gate-to-source capacitance for Figs. 2 and 3.

peak distance show similar effects. The larger g_m at high drain current for the recessed gate structure, compared to the planar structure, is caused by the thicker active regions under the source and drain in the recessed gate structure.

V. Conclusions

A two-dimensional finite element method has been used to characterize ion-implanted GaAs MESFET's. This simulation allows accurate modeling of changes in device geometry, including the location of both the channel implant and the gate recess. Results of the model have shown that FET's with a shallow recess show similar behavior to those with a deep implant, i.e., an improvement in linearity, a higher pinch-off voltage, a decrease in transconductance, and a corresponding decrease in cutoff frequency. Depending on the device application, different choices for these parameters would be appropriate. Thus this simulation could be useful as a guide in the fabrication of optimized devices. Further calculations are in progress to determine bias-dependent parameter changes, which are especially important in large-signal applications.

References

[1] J. A. Higgens, R. L. Kuvas, F. H. Eisen, and D. R. Ch'en, "Low noise GaAs FET's prepared by ion implantation," *IEEE Trans. Electron Devices*, vol. ED-25, pp. 587-596, 1978.

[2] J. M. M. Golio and R. J. Trew, "Profile studies of ion-implanted MESFET's," *IEEE Trans. Microwave Theory Tech.*, vol. MTT-31, pp. 1066-1071, 1983.

[3] R. A. Pucel, H. A. Haus, and H. Statz, "Signal and noise properties of GaAs microwave FET's," in *Advances in Electronics and Electron* Physics, vol. 38. New York: Academic, 1975, pp. 195-265.

[4] N. Song and T. Itoh, "Accurate simulation of MESFET by finite element method including energy transport and substrate effects," in *Proc. 1985 European Microwave Conf.*, pp. 245-250.

[5] J. J. Barnes and R. J. Lomax, "Finite element methods in semiconductor device simulation," *IEEE Trans. Electron Devices*, vol. ED-24, pp. 1082-1089, 1977.

[6] K. Yamaguchi and H. Kodera, "Two-dimensional numerical analysis of stability criteria of GaAs FET's," *IEEE Trans. Electron Devices*, vol. ED-23, pp. 1283-1289, 1976.

[7] W. R. Curtice and Y. Yun, "A temperature model for the GaAs MESFET," *IEEE Trans. Electron Devices*, vol. ED-28, pp. 954-962, 1981.

[8] M. Utku and G. F. Carey, "Boundary penalty techniques," *Computer Meth. Appl. Mech. Eng.*, vol. 31, pp. 103-118, 1984.

[9] K. Lee, M. S. Shur, K. Lee, T. T. Vu, P. C. T. Roberts, and M. J. Helix, "Low field mobility in GaAs ion implanted FET's," *IEEE Trans. Electron Devices*, vol. ED-31, pp. 390-393, 1984.

[10] R. H. Wallis and P. K. Jay, "Deep levels and electron mobility near the active layer substrate interface in GaAs MESFET's," in *Semi-Insulating 3-5 Materials*, S. Makram-Ebeid and T. Brian, Eds. Nantwich, U.K.: Shiva, 1982, pp. 344-351.

[11] M. Feng, V. K. Eu, T. Zielinski, H. Kanber, and W. B. Henderson, "GaAs MESFET's made by ion implantation into MOCVD buffer layers," *IEEE Electron Device Lett.*, vol. EDL-5, pp. 18-20, 1984.

[12] R. E. Williams and D. W. Shaw, "GaAs FET's: Improved linearity and noise figure," *IEEE Trans. Electron Devices*, vol. ED-25, pp. 600-605, 1978.

[13] J. A. Higgens and R. L. Kuvas, "Analysis and improvement of intermodulation distortion in GaAs power FET's," *IEEE Trans. Microwave Theory Tech.*, vol. MTT-28, pp. 9-17, 1980.

[14] R. A. Pucel, "Profile design for distortion reduction in microwave FET's," *Electron Lett.*, vol. 14, no. 16, pp. 204-206, 1978.

[15] J. J. M. Dekkers, F. Ponse, and H. Beneking, "Buried channel GaAs MESFET's," *IEEE Trans. Electron Devices*, vol. ED-28, pp. 1065-1070, 1981.

Physica 20D (1986) 320–334
North-Holland, Amsterdam

ESTIMATION OF NONLINEAR TRANSFER FUNCTIONS FOR FULLY DEVELOPED TURBULENCE

Ch.P. RITZ and E.J. POWERS
University of Texas at Austin, Austin, TX 78712, USA

Received 13 March 1985
Revised manuscript received 3 September 1985

A statistical method for modeling the linear and quadratically nonlinear relationship between fluctuations monitored at two points in space or time in a turbulent medium is presented. This relationship is described with the aid of linear and quadratic transfer functions and the concept of coherency is extended to quantify the goodness of the quadratic model. A unique feature of the approach described in this paper is that it is valid for non-Gaussian "input" and "output" signals. The validity of the approach is demonstrated with simulation data. The method is applied to experimental data taken in the turbulent edge plasma of the TEXT tokamak. The results indicate a three wave process with energy transfer to large scale fluctuations. The estimation of transfer functions is a first step in quantitatively measuring coupling coefficients and the energy transfer.

1. Introduction

Many fluctuation phenomena in nature, as well as in various technical problems, can be reduced to relatively simple systems which are describable by a set of source ("input") signals and the response ("output") signals of the system. The system can then be considered as a "black box". By modeling this "black box" by an appropriate network, consisting of linear, quadratic and higher-order nonlinear elements, it is possible to gain considerable insight into the dynamics of the system under test.

The simplest such network consists of a single input, single output system which is purely linear, since the nonlinear contributions of the system are assumed to be negligible. The determination of the linear system model from the measured input and output signals is based upon cross-correlation techniques. Applications of the linear network led for example to the estimation of the dispersion relation in plasmas [1–4]. It is also often used to test complex electronic circuits. In this work we will discuss a system which can be described by a single input and single output which will be modeled in the spatial or temporal frequency domain by linear and quadratic elements of the form

$$Y_p = L_p X_p + \tfrac{1}{2} \sum_{\substack{p_1, p_2 \\ p = p_1 + p_2}} Q_p^{p_1, p_2} X_{p_1} X_{p_2} + \epsilon_p. \qquad (1)$$

Such a system is presented schematically in fig. 1. L_p and $Q_p^{p_1, p_2}$ are usually called linear and quadratic transfer functions and are generally complex quantities. The Fourier transforms of the measurable input signal $x(s)$ and of the output signal $y(s)$ are X_p and Y_p, respectively. The signals $x(s)$ and $y(s)$ are assumed to be zero mean stationary random processes. The error term ϵ_p is the Fourier transform of a process which is assumed to be statistically independent of the first two terms of eq. (1). It can be regarded as the error due to noise inherent in the measurement as well as systematic errors not described by linear and quadratic terms. The goal is to estimate the linear and quadratic transfer functions from the measured input and output signals.

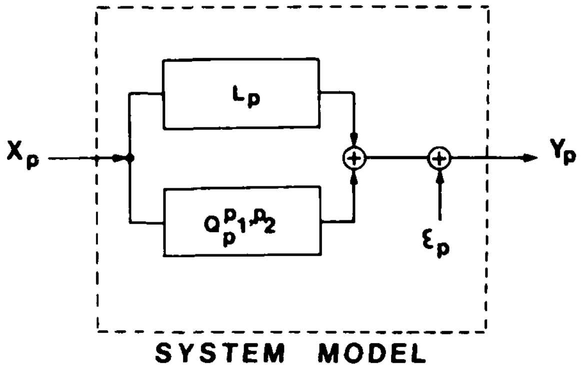


Fig. 1. Schematic model of the nonlinear system, given in eq. (1).

Indeed, we have motivation for modeling various nonlinear physical systems with such a quadratically nonlinear equation, e.g. eq. (1). Examples of applications include electronic networks, electromagnetics [5, 6], and parametric excitation in mooring dynamics [7–9], in structural vibrations [10] and in nonlinear optics [11]. Characteristic of these examples is that the spectral components X_p and Y_p stand for the temporal frequency spectra. On the other hand, many physical phenomena can be described by an equation similar to eq. (1) with Fourier components X_p and Y_p representing spatial frequencies, i.e. wave numbers. When we consider the Navier–Stokes equation for neutral fluids and Fourier transform in space, we find a wave-coupling equation which describes the temporal change of a spectral component of given wavenumber due to a linear mechanism (growth rate, dispersion) and due to three wave coupling [e.g., 12, 13]. The linear and quadratic nonlinear terms in this equation can be modeled by eq. (1). Similiar wave coupling equations describe the turbulent behavior in plasmas [14, 15] and in the solid state [16]. In several of these fields, competing theoretical models exist to describe the formation of the turbulent spectra. Therefore, it is useful to have an experimental measurement available with which the competing theories can be compared directly. In the past we reported preliminary experimental results which estimate the coupling coefficient and energy cascading in such a turbulent plasma [17]. The estimation of the nonlinear transfer function is a necessary first step to quantifying the coupling coefficients and the energy transfer. The approach has not yet been published.

Note that eq. (1) is the simplest variant of an equation which can be related to wave–wave coupling, as we assume that four-wave coupling and higher order processes are much weaker than three-wave coupling. In some cases when three-wave nonlinear coupling is forbidden by the dispersion properties of the system, as for example for surface gravity waves in water [18], higher-order terms must be included.

The determination of the transfer functions L_p and $Q_p^{p_1,p_2}$ is straight forward and well known for the case of a Gaussian input signal $x(s)$ [19–21]. Such a method can be used for analyzing a system that can be excited externally by a Gaussian noise source [20] and for systems with input signals which can be assumed to be Gaussian [21]. Many systems such as turbulent fluids and plasmas, however, do not allow such a restrictive assumption for the input signal. In order to obtain insight into the physical mechanisms of turbulence in these media, one can monitor the fluctuations at two points in time or space and study the change of the spectra between these points. In these cases and in general the input should not be considered to be a Gaussian process because of nonlinear history of the fluctuations.

The main objective of this work is to present a technique which enables one to estimate, in an efficient way, the linear and quadratic transfer functions from the measured fluctuation signals $x(s)$ and $y(s)$. These transfer functions serve as the fundamental quantities with which to estimate the growth rate, the wave–wave coupling coefficients and finally the energy transfer between different spectral components [17]. This paper focuses on the technique of estimating the transfer functions and qualitatively describing the physical results. A subsequent paper will cover the application of this concept to estimating quantitatively the coupling coefficient and energy cascading.

In section 2 we present a new technique to estimate transfer functions. For the case of non-Gaussian input signals the usual definition of the coherency must be generalized. The influence of noise and of systematic errors will also be briefly discussed. As the approach involves rather extensive computation, the convergence and accuracy of the model must be tested prior to applying it to real data. We present the results of such a test in section 3. In section 4, the method is applied to data from turbulence measured in the edge plasma of the TEXT tokamak. We estimate the linear and quadratic transfer functions between two points in space and interpret them in physical terms. Next we compare the linear transfer function with the one obtained by neglecting nonlinear contributions. Such a comparison is of interest since most present correlation methods are based on a linear assumption.

2. Method

The transfer functions for the quadratic nonlinear system given by eq. (1) can be derived in a straightforward way when a sufficient number of independent realizations of the spectra X_p and Y_p are available. The variable p represents for example the wavenumber or temporal frequency.

We first rewrite eq. (1) by using the symmetry relation, $Q_p^{p_1, p_2} = Q_p^{p_2, p_1}$ (Note that the spectral components X_{p_1} and X_{p_2} are interchangeable in eq. (1)). It is therefore sufficient to sum the quadratic terms over the frequency components with $p_1 \geq p_2$.

$$Y_p = L_p X_p + \sum_{\substack{p_1 \geq p_2 \\ p = p_1 + p_2}} Q_p^{p_1, p_2} X_{p_1} X_{p_2} + \varepsilon_p. \qquad (1a)$$

Multiplying eq. (1a) with the complex conjugate of the input signal X_p and computing the expected value by ensemble averaging $\langle\ \rangle$ over many statistically similar realizations, we obtain

$$L_p = \frac{\langle Y_p X_p^* \rangle - \sum_{p_1 \geq p_2} Q_p^{p_1, p_2} \langle X_p^* X_{p_1} X_{p_2} \rangle - \langle \varepsilon_p X_p^* \rangle}{\langle X_p X_p^* \rangle},$$

$$p = p_1 + p_2. \qquad (2)$$

Similarly, by multiplying Y_p by $X_{p_1'}^* X_{p_2'}^*$ and ensemble averaging, we have

$$\langle Y_p X_{p_1'}^* X_{p_2'}^* \rangle = L_p \langle X_p X_{p_1'}^* X_{p_2'}^* \rangle$$

$$+ \sum_{p_1 \geq p_2} Q_p^{p_1, p_2} \langle X_{p_1} X_{p_2} X_{p_1'}^* X_{p_2'}^* \rangle + \langle \varepsilon_p X_{p_1'}^* X_{p_2'}^* \rangle, \qquad (3)$$

with

$$p = p_1 + p_2 = p_1' + p_2'.$$

To simplify eq. (3) and, therefore, reduce the computation time we approximate the fourth-order moment $\langle X_{p_1} X_{p_2} X_{p_1'}^* X_{p_2'}^* \rangle$ with second-order moments $\langle |X_{p_1} X_{p_2}|^2 \rangle$ by neglecting components $\langle X_{p_1} X_{p_2} X_{p_1'}^* X_{p_2'}^* \rangle$ with $(p_1, p_2) \neq (p_1', p_2')$. However, we retain the third-order moment of the input signal $\langle X_{p_1 + p_2} X_{p_1}^* X_{p_2}^* \rangle$, as we assume that the input signal is non-Gaussian. Such an approach was proposed by Millionshchikov [22] and is used to close the system in weak turbulence theories [12–15]. This approach was also used in many strong turbulence theories [23, 24], in which the linear mode structure is strongly altered by the turbulence and for which $\Delta\omega/\omega$ is large. Comparison with simulations which do not use such a closure scheme confirmed the validity of this approximation also in strongly turbulent cases [24, 25]. Approximating the fourth-order moments requires that the fluctuations must be close to Gaussian distributed, a condition which is not generally valid for strong turbulence. In a future work it is planned to compute the complete set of fourth-order moments and solve for the transfer functions. This will allow us also to test the validity of the closure scheme on real data. Using the

Millionshchikov hypothesis we have

$$Q_p^{p_1,p_2} = \frac{\langle Y_p X_{p_1}^* X_{p_2}^* \rangle - L_p \langle X_p X_{p_1}^* X_{p_2}^* \rangle - \langle \varepsilon_p X_{p_1}^* X_{p_2}^* \rangle}{\langle |X_{p_1} X_{p_2}|^2 \rangle} ,$$

$$p = p_1 + p_2 . \quad (3a)$$

The second term in the numerator of equations (2) and (3a) vanish for a Gaussian input because then the skewness $(\langle x^3(s) \rangle / \langle x^2(s) \rangle^{3/2})$ is zero, therefore $\langle X_{p_1+p_2} X_{p_1}^* X_{p_2}^* \rangle = 0$. The second terms can thus be interpreted physically as corrections due to the nonlinear history of the fluctuation detected at the input. As long as the error term ε_p is zero mean and statistically independent with respect to the input signal, the error terms in eqs. (2) and (3a) vanish.

The set of dependent equations (2) and (3a) can be solved iteratively to find the transfer functions L_p and $Q_p^{p_1,p_2}$. As an initial guess we neglect the quadratic contribution in eq. (2) and find for the linear transfer function

$$L_p = \frac{\langle Y_p X_p^* \rangle}{\langle X_p X_p^* \rangle} . \quad (2a)$$

The number of iterations used for a given frequency p depends on the accuracy of the estimated moments. It is also sensitive to the stationarity of the signal to be analyzed and is thus larger for real data than for simulated data. The number of iterations is also dependent on the magnitude of the quadratic transfer function with respect to the linear one. The typical number of iterations needed for L_p to change by δ less than 1‰ is of order 5 for simulation data and of order 10 for measured data ($\delta < 1\%$). Note that the special case of a Gaussian input signal can be solved directly without an iterative procedure as the first term on the right-hand side of eq. (3) vanishes.

As indicated by equations (2) and (3a), L_p and $Q_p^{p_1,p_2}$ can be obtained by determination of the well-known statistical quantities in the pointed brackets, the auto-power spectrum $P_p = \langle X_p X_p^* \rangle$, the cross-power spectrum $C_p = \langle Y_p X_p^* \rangle$, the auto-bispectrum $B_a(p_1, p_2) = \langle X_{p_1+p_2} X_{p_1}^* X_{p_2}^* \rangle$ and the cross-bispectrum $B_c(p_1, p_2) = \langle Y_{p_1+p_2} X_{p_1}^* X_{p_2}^* \rangle$. The cross-power spectrum measures the statistical dependence of amplitude and phase of the same spectral components in the input and output. Therefore the cross-power spectrum must play an important role in detecting the dispersion relation and the growth rate. The auto- and cross-bispectra measure the statistical relationship of amplitude and phase between the spectral components p_1, p_2 and $p = p_1 + p_2$. If the waves at p_1, p_2 and p have statistically independent random phases, then the resulting (bi-)phase of the polar representation $(\theta_{p_1+p_2} - \theta_{p_1} - \theta_{p_2})$ will be random and the expected value of the bispectra converges to zero. If, however, a coherent phase relationship exists due to nonlinear coupling of these waves, the bispectra, averaged over many realizations, will reach a finite value. The auto- and cross-bispectra play, therefore, an important role in detecting three-wave coupling effects. The mathematical and statistical background of the estimation of these expected values with the aid of digital signal processing are thoroughly discussed in [26–28].

The amount of information gained from the transfer functions can be quite large. Therefore it is useful to introduce normalized quantities to provide an easier interpretation of the input–output relationship. A convenient normalization is the coherency, which gives the fraction of the power in the output signal which can be accounted for by the linear and the quadratic transfer function model. To define the coherencies in terms of the transfer functions, we multiply eq. (1a) by its complex conjugate, take a statistical average and divide the result by the output-power $\langle Y_p Y_p^* \rangle$. We find

$$1 = \gamma_L^2(p) + \gamma_Q^2(p) + \gamma_{LQ}^2(p) + \gamma_n^2(p) + \langle \text{error terms} \rangle , \quad (4a)$$

with

$$\gamma_L^2(p) = |L_p|^2 \frac{\langle X_p X_p^* \rangle}{\langle Y_p Y_p^* \rangle},$$

$$\gamma_Q^2(p) = \sum_{p_1 > p_2} |Q_p^{p_1, p_2}|^2 \frac{\langle |X_{p_1} X_{p_2}|^2 \rangle}{\langle Y_p Y_p^* \rangle},$$

$$\gamma_{LQ}^2(p) = \frac{2\,\mathrm{Re}\left(L_p \sum_{p_1 > p_2} [Q_p^{p_1, p_2}]^* \langle X_p X_{p_1}^* X_{p_2}^* \rangle \right)}{\langle Y_p Y_p^* \rangle},$$

$$\gamma_n^2(p) = \frac{\langle \epsilon_p \epsilon_p^* \rangle}{\langle Y_p Y_p^* \rangle}, \qquad (4b)$$

where

$$p = p_1 + p_2$$

The linear coherency $\gamma_L^2(p)$ and the quadratic coherency $\gamma_Q^2(p)$ denote the fraction of output power accounted for by the linear and quadratic transfer functions. The value under the summation in $\gamma_Q^2(p)$ is called the cross-bicoherency $b_{YXX}^2(p_1, p_2)$ and measures the portion of the power *of the output signal $Y_{p=p_1+p_2}$ which is phase locked* with X_{p_1} and X_{p_2} in the input signal. The coherency $\gamma_{LQ}^2(p)$ gives the portion of the output power, for which the response due to the linear transfer and due to the quadratic transfer are correlated. Note that this term arises because of the non-Gaussian input $x(s)$. The term $\gamma_n^2(p)$ represents the ratio of the output power due to noise, which can not be accounted for by the system. The error terms in eq. (4a) vanish as long as the error term ϵ_p of the output signal is zero mean and independent of the input signal of the system. The usual definitions of the coherencies [19] are bounded by zero and unity (as we shall show below for the case of a Gaussian input signal). However, the individual coherencies $\gamma_L^2(p)$, $\gamma_Q^2(p)$ and $\gamma_{LQ}^2(p)$ defined above are not necessarily bounded by unity. While $\gamma_L^2(p)$ and $\gamma_Q^2(p)$ must be larger than zero, the term $\gamma_{LQ}^2(p)$ is allowed to take on negative values as well. The noise term $\gamma_n^2(p)$ can take on any value between zero and unity. The "goodness of fit" of the model can then be characterized by the total coherency of the model,

$$\gamma^2(p) = \gamma_L^2(p) + \gamma_Q^2(p) + \gamma_{LQ}^2(p). \qquad (5)$$

Note that the above definition of the coherency converges to the commonly used definition [19, 20] when the input signal is Gaussian. In this case $\langle X_{p_1+p_2} X_{p_1}^* X_{p_2}^* \rangle = 0$ and the term $\gamma_{LQ}^2(p)$ vanishes. We find

$$\gamma_L^2(p) = \frac{|\langle Y_p X_p^* \rangle|^2}{\langle X_p Y_p^* \rangle \langle Y_p Y_p^* \rangle},$$

$$\gamma_Q^2(p) = \sum_{p_1 > p_2} \frac{|\langle Y_p X_{p_1}^* X_{p_2}^* \rangle|^2}{\langle |X_{p_1} X_{p_2}|^2 \rangle \langle Y_p Y_p^* \rangle}, \qquad (6)$$

$$\gamma_{LQ}^2(p) = 0,$$

where

$$p = p_1 + p_2.$$

The linear coherency $\gamma_L^2(p)$ is bounded by zero and unity as can be shown with the Schwartz inequality. Because of eq. (4a), $\gamma_Q^2(p)$ must also be bounded by unity. The application of eq. (6) to experimental data with a non-Gaussian input signal will thus lead to an erroneous result and can yield values for the total coherency which are greater than unity.

3. Simulation test

To test the validity of the approach described in the previous section, we have carried out a computer simulation. We start with analytically defined linear and quadratic transfer functions. For given non-Gaussian input signals we compute the corresponding output signals. Next, we apply the method of section 2 to estimate the linear and quadratic transfer functions from the input and output data, and then we compare the results with the expected values.

For this example, we define L_p and $Q_p^{p_1,p_2}$ as

$$L_p = 1.0 - 0.4\frac{p^2}{p_{\text{Nvq}}^2} + \text{i}0.8\frac{p}{p_{\text{Nyq}}},$$

$$Q_p(p_1, p_2) = \frac{\text{i}}{5p_{\text{Nvq}}^4}\,\frac{p_1 p_2(p_2^2 - p_1^2)}{1 + p^2/p_{\text{Nyq}}^2},$$

where

$$p = p_1 + p_2, \quad \text{i} = \sqrt{-1}\,.$$

The magnitude and shape of the quadratic transfer functions defined above are chosen arbitrarily, but are realistic in that the values are of the same order of magnitude as the ones predicted by the Hasegawa–Mima equation [15]. The linear transfer function is defined such that the input and output spectra are similar in shape as would be expected for a stationary state. The Nyquist frequency shall be abbreviated as p_{Nvq}*. The real and imaginary parts of L_p are illustrated in fig. 2a, 2b and the absolute value of $Q_p^{p_1,p_2}$ is shown in a contour plot in fig. 3a. To approximate the situation which actually occurs in a continuous medium, we consider five identical "black boxes" of the type shown in fig. 1, which are connected in series. A Gaussian signal is applied to the input of the first black box. The output, which is now non-Gaussian because of the nonlinear nature of the black box, becomes the input to the second black box and so on. For the simulation we utilize the input and output of the fifth black box. The input signal, X_p, for the estimation can thus be assumed to be approximately as "non-Gaussian" as the output signal Y_p.

In a second step we estimate the linear and quadratic transfer functions $\hat{L}_p$ and $\hat{Q}_p^{p_1,p_2}$ using the approach of section 2. As shown in figs. 2c, 2d and fig. 3b, the estimated transfer functions, are in good agreement with the true values. The symbol, "^", denotes an estimator. To save space, we compare here only the absolute value of the quadratic transfer function. The phase comparison would show a similar agreement. Note, however, that the phase information is also important for the interpretation of the nonlinear system. Because of the symmetry properties possessed by the quadratic transfer function, it is not necessary to plot $Q_p^{p_1,p_2}$ over the entire two-dimensional plane†. Fig. 3 gives the value of $|Q_p^{p_1,p_2}|$ at the frequency $p = p_1 + p_2$ due to wave-wave coupling with p_1 and p_2. The area, for which the transfer function is plotted, can be subdivided into three regions with essential differences in the physical content (I, II, III in fig. 3a). The triangular region (I) gives the quadratic transfer function at the highest frequency p involved in the interaction ($p > p_1, p_2$). Region (II)

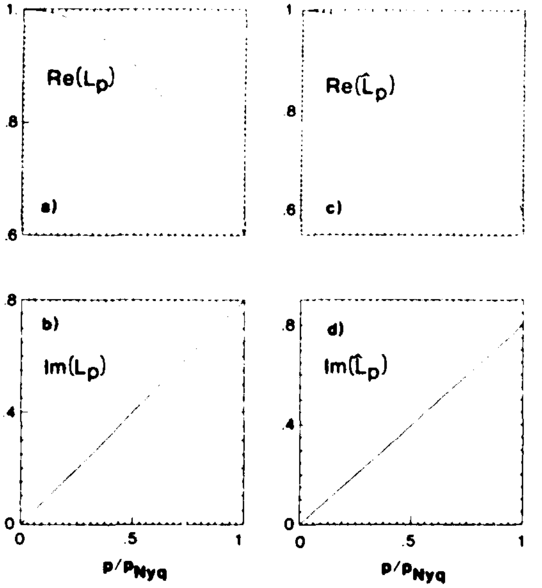


Fig. 2. Comparison of the analytically defined linear transfer function with the reconstructed ones: a) shows the real; and b) the imaginary component of the "true" values; c) and d) the estimated ones.

*The maximal frequency which can be resolved digitally is given by the Nyquist theorem. The theorem indicates that the smallest detectable period must contain at least two sampling points. Frequencies larger than p_{Nvq} produce an erroneous spectrum below p_{Nvq} due to aliasing.

†It is sufficient to compute and plot the transfer function for $p_1 \geq p_2$ (because $Q_p^{p_1,p_2} = Q_p^{p_2,p_1}$) and for $p > 0$ (because $Q_p^{p_1,p_2} = [Q_{p}^{p_1,\; p_2}]^*$, as the Fourier transform for real data $x(t)$ satisfies $X_p = X^*_{p}$).

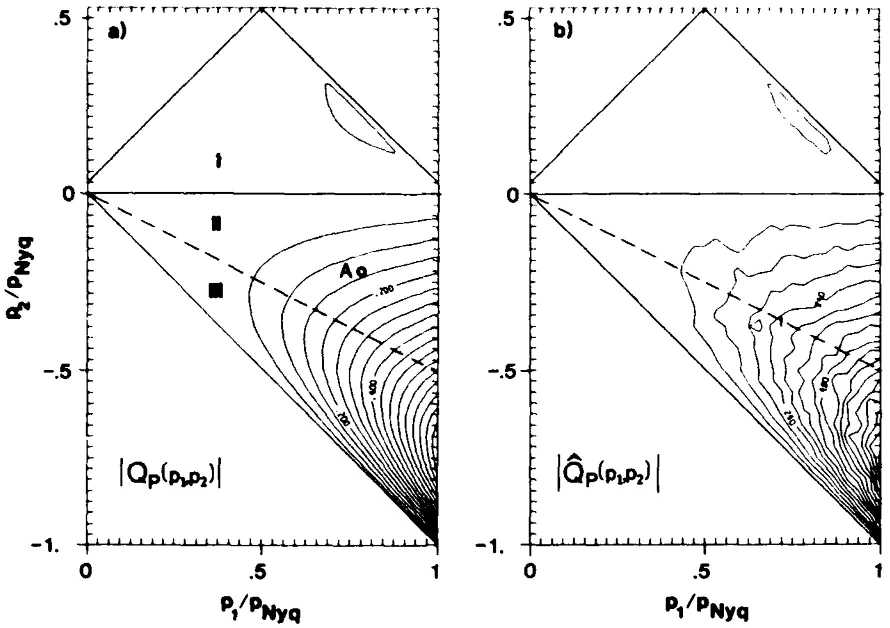


Fig. 3. Contour plot of the amplitude of a) the analytically defined quadratic transfer function. Point A illustrates $|Q_p^{p_1,p_2}|$ at $p_1/p_{\mathrm{Nyq}} = 0.78$ due to coupling with $p_2/p_{\mathrm{Nyq}} = 0.22$ and $p/p_{\mathrm{Nyq}} = 0.56$; b) reconstruction of $|Q_p^{p_1,p_2}|$ by the iterative method. The contour interval is 0.05 in a) and 0.06 ir b).

gives the strength of the coupling at the intermediate frequency p due to wave–wave coupling with a spectral component of larger frequency p_1 and one with a smaller frequency $p_2(p_1 > p > |p_2|)$. The triangular region (III) shows $|Q_p^{p_1,p_2}|$ at the lowest frequency component p due to coupling with the two others $(p < p_1, |p_2|)$. For a convenient graphical representation of the three coupling regions, we plot positive and negative values of p_2. Note, however, the positive and negative spectral components are related by $X_{-p} = X_p^*$. To facilitate the interpretation of fig. 3, we arbitrarily pick out one point, which is indicated as A. The number of contours at this point gives the amplitude of the transfer function $|Q_p^{p_1,p_2}|$ at frequency $p/p_{\mathrm{Nyq}} = 0.56$ as a result of its coupling with a higher frequency component $p_1/p_{\mathrm{Nyq}} = 0.78$ and with a lower frequency component $|p_2|/p_{\mathrm{Nyq}} = 0.22$.

To test the "goodness of fit" of this simulation, we compute the coherency from eq. (4b). The difference between the estimated total coherency and unity can be regarded as due to systematic errors of the estimation approach and due to the variance of the estimator of the statistical quantities (e.g. the cross-bispectrum). Fig. 4a shows the contribution of the linear, quadratic and mixed coherency to the total coherency of the model $\gamma^2(p)$. We have chosen relatively small values of the quadratic transfer function with respect to the linear component in order to simulate a realistic situation; consequently the contribution of $\gamma_Q^2(p)$ is small. Note that the total coherency of the model $\gamma^2(p)$ is close to unity indicating that the systematic error of the approach is small. We conclude that the iterative method is able to produce a good fit of the data.

To illustrate the necessity of applying the iterative method, we compute the transfer functions with the usual method by disregarding the non-Gaussianity of the input signal, X_p. The non-iterative method then leads to a quadratic

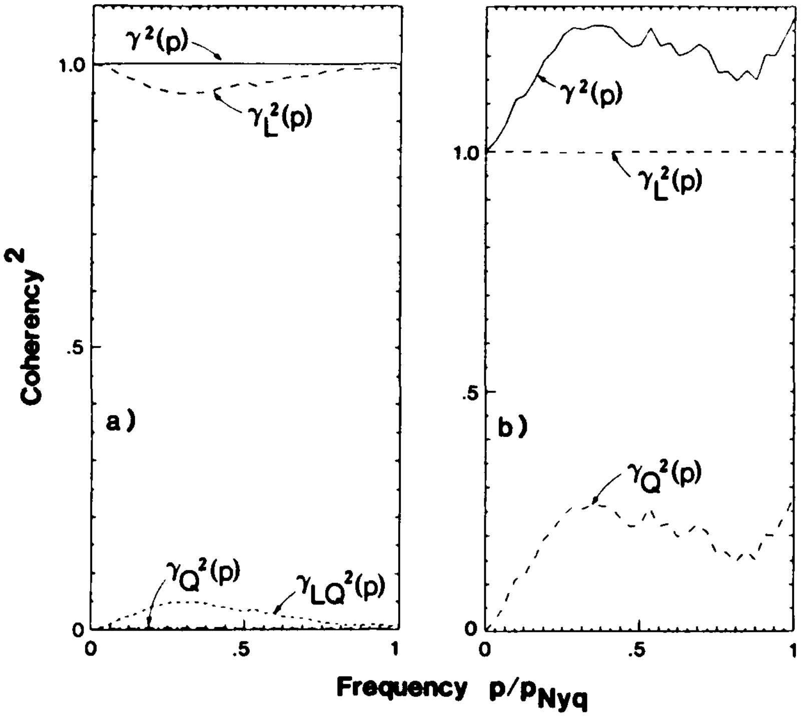


Fig. 4. Coherence spectra. $\gamma_L^2(p)$ linear component, $\gamma_Q^2(p)$ quadratic contribution, $\gamma_{LQ}^2(p)$ mixed term, $\gamma^2(p)$ summation over the three components. The difference between $\gamma^2(p)$ and 1.0 represents the error term $\gamma_n^2(p)$: a) gives the coherence for the iterative procedure (eq. (4a, b)); b) the one computed with the usual definition (eq.(6)).

transfer function which is much different from the true one. This erroneous result can easily be visualized by applying the usual definition of the coherency, given by eq. (6), which leads to values for the total coherency much larger than unity (fig. 4b). The systematic error introduced by assuming a Gaussian input signal can thus be considerable. Note, for the non-Gaussian case, the linear coherency is very close to unity. This is due to the very small change of the signal between the input and output as a result of the quadratic interaction (which is of order $\gamma_Q^2(p)$ in fig. 4a).

It is important to recognize that a good estimation of the quadratic transfer function requires a large number of realizations. For this simulation we have used 2000 realizations. This large number of realizations is necessary in order to estimate the auto- and cross-bispectra accurately and thus to assure a proper convergence of the iterative process. To illustrate this, we show in fig. 5 the convergence of both the power spectrum P_p and the amplitude of the auto- and cross-bispectra to a stable value as a function of an increasing amount of realizations. We have chosen for this demonstration the frequency components $p/p_{Nyq} = 0.47$, $p_1/p_{Nyq} = 0.31$, and $p_2/p_{Nyq} = 0.16$. While the auto-power spectrum stabilizes after 500 realizations, more than 1500 realizations for the auto- and cross-bispectra are needed. The large variance of the higher-order spectra can be attributed to the fact that all spectral components in a turbulent spectrum couple with each other and therefore the contribution of each triplet of waves is relatively small.

By testing the iterative method under different conditions, we have found another interesting feature which we briefly report. For a purely Gaussian input, X_p, the iterative nrocedure appears to

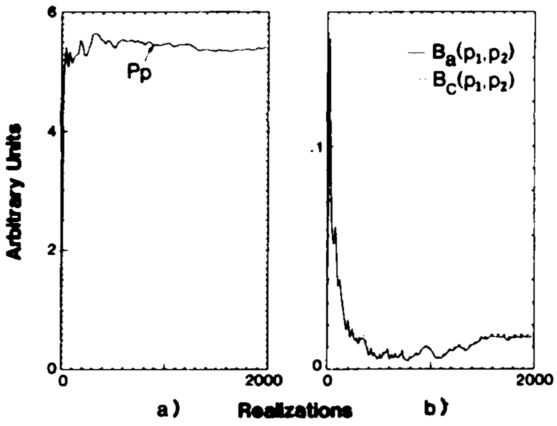


Fig. 5. Convergence of the spectral estimates to a stable value as function of the number of realizations. a) Convergence of the power spectrum at $p/p_{\mathrm{Nyq}} = 0.47$; and b) of the amplitude of the auto- and cross-bispectra at $p/p_{\mathrm{Nyq}} = 0.47$ due to wave–wave coupling with $p_2/p_{\mathrm{Nyq}} = 0.16$ and $p_1/p_{\mathrm{Nyq}} = 0.31$.

considerably reduce the number of realizations needed for a given statistical error of the quadratic transfer function. While the auto- and cross-power spectra are already estimated acceptably, the estimate of the bispectra is significantly poorer for the same amount of realizations. Because of a similar behavior of the auto- and cross-bispectra (see their convergence in fig. 5), the statistical errors therefore approximately cancel in eq. (3a) and the estimation of $Q_p^{p_1,p_2}$ is improved for the same amount of realizations. We conclude that the iterative procedure is valuable to increase the speed of the calculation for systems with a purely Gaussian input signal and also requires fewer realizations.

4. Nonlinear wave coupling in the edge of a tokamak

The plasma which connects the hot interior plasma of a tokamak with the cold wall is often called edge plasma. This region of the tokamak is characterized by steep density and temperature gradients and by a high density fluctuation level. Asymmetries induced by the limiter and an increased impurity level due to the plasma wall contact make this region even more complex. A good understanding of the physics of the edge plasma in a tokamak can, however, be crucial for fusion, as the fluctuations in the edge can substantially affect the global plasma confinement. The study of the linear and nonlinear behavior of the waves and instabilities in the edge plasma can add important information which, together with theoretical models, lead to optimized edge conditions with minimized particle and energy fluxes to the wall.

The linear features of fluctuations in the edge of tokamak plasmas have been studied for some time with probes [e.g. 1–4, 29–31]. In the following we extend this observation and give information about the nonlinear wave interaction by the estimation of the nonlinear transfer functions.

To make a direct comparison with theoretical works on turbulence in plasmas one is most interested in the transfer function for wavenumber spectra. Such spectra are, however, not easily obtainable in tokamak plasmas as this would necessitate a large number of spatial samples with probes or a simultaneous measurement of many spectral components with scattering technique. In a tokamak the installation of a Langmuir probe array has recently been realized [4]. A large probe array may substantially perturb the plasma. Multichannel scattering experiments are able to measure instantaneously different wave numbers. For the analysis a heterodyne system would be needed and the influence due to different scattering volumes has to be understood. To circumvent these technical questions we chose for this initial experiment a two probe technique.

Two spatially separated probes are capable of measuring temporal variations of the fluctuation. Instead of computing the bispectra for different wave numbers which fulfill the selection rule $k = k_1 + k_2$ we compute them for the frequency components $f = f_1 + f_2$. This approach enables us to detect only the resonantly coupled components, for which the frequency mismatch is zero. Those coupled spectral components with frequency mismatch, which are present in strong turbulence causing the broadening of the dispersion relation, contribute to the error term in our method.

We measured the signals with Langmuir probes in the turbulent edge region of the TEXT tokamak (major radius 100 cm, minor radius $a = 27$ cm). To measure the density fluctuations, the probes were biased into ion saturation current and the potential fluctuations were detected with floating probes. The signals are digitized with a 10 bit digitizer with 32k words storage capability per channel. Each time series is subdivided into 512 time segments of 64 data points. To estimate the transfer functions we have averaged over more than 1500 realizations gathered from 3 identical shots. The sampling interval used in the following presentation is 1 μs, which defines a temporal Nyquist frequency of 500 kHz, which is well above the dominant components of the turbulent power spectrum.

We will consider data obtained with the following tokamak parameters. The toroidal magnetic field is $B_0 = 1$ T, the plasma current is $I_0 = 100$ kA, the chord averaged density is 1.0×10^{13} cm^{-3} and we have a peak electron temperature of 600 eV. In the region behind the limiter we observed a broad turbulent spectrum with fluctuation levels $|\tilde{n}/n|$ of up to 50%. The edge plasma is characterized by a nonuniform radial electric field which changes sign just inside of the outermost closed flux surface. Behind this flux surface the radial electric field causes an $E_r \times B$ plasma rotation in the ion diamagnetic drift direction while it results in a rotation in the electron direction on the inside. The measurements also exhibit a localized instability which occurs in the region of maximum velocity shear and which is different from the turbulence structure outside of the velocity shear. The measured phase velocity of the turbulence can be described by an $E_r \times B$ drift superimposed on a pressure gradient drift [29]. In this work we include a description of nonlinear effects for the region outside the shear layer. Model equations [23, 24, 32, 33] applicable in the edge region are characterized by quadratic nonlinearities, hence, the three wave interaction is the relevant wave coupling process. Model equations with cubic nonlinearities [34, 35] giving rise to four wave interaction are generally not considered in the edge plasma context.

Our goal is to estimate the linear and quadratic transfer functions between the signals of two poloidally separated points. The experimental setup is as follows: two radially movable Langmuir probes are located at the top of the tokamak. They are separated by $\Delta x = 3.5$ mm. For the dominant power at low frequencies, this separation is small compared to the poloidal correlation length (several centimeters). In this paper, we analyze data taken at a radial position 1 cm behind the limiter. The density at this location is approximately 5.0×10^{11} cm^{-3}, the electron temperature $T_e = 10$ eV and $|\tilde{n}/n| = 40\%$. The power spectra of the density fluctuations monitored by both probes located at this radial position are shown in fig. 6. Both probes yield nearly the same spectrum.

The data, which have been digitized and stored in the computer, are processed using the procedure discussed in section 2. The propagation direction is first determined by observing the sign of the phase shift of the cross power spectrum between the two probes. The signal of the probe which first samples the turbulent structures (i.e. the "upstream" probe) is treated as the input signal. The "downstream" probe signal is considered as the output.

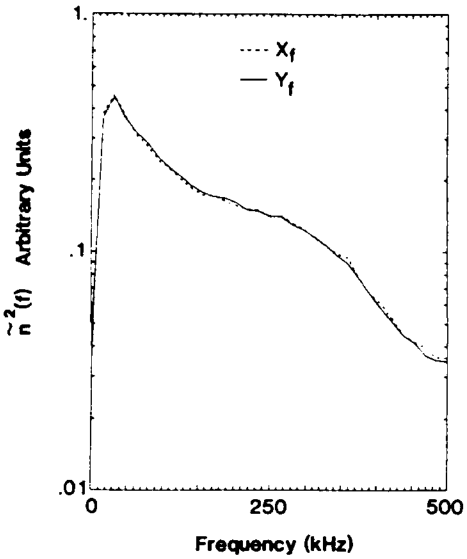


Fig. 6. Power spectrum of the density fluctuation 1.0 cm behind the limiter for the two Langmuir probes, separated by 3.5 mm in the poloidal direction.

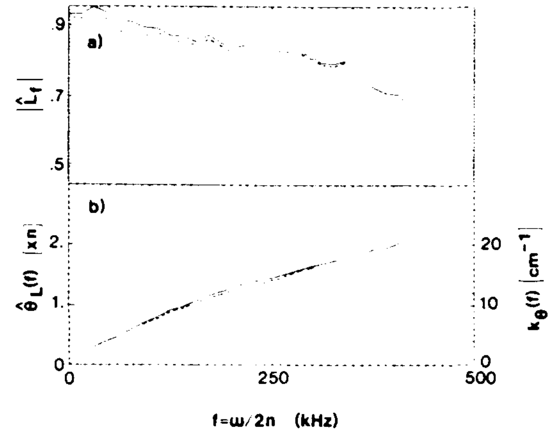


Fig. 7. Amplitude a) and phase b) of the measured linear transfer function. The dotted line represents the transfer function assuming Gaussian input.

The linear transfer function estimated from this data is presented, for easier physical interpretation in terms of amplitude and phase of $\hat{L}_f = |\hat{L}_f| \exp(i\hat{\theta}_L(f))$. Amplitude and phase are shown in figs. 7a, b. The phase of the linear transfer function $\hat{\theta}_L(f)$ gives the phase relationship between the input and the output signal due to linear effects. In our experiment, $\hat{\theta}_L(f)$ can be related to the mean dispersion relation, $\bar{k}_\theta(f)$, because a wave which propagates with wavenumber $k_\theta(f)$, will undergo a phase shift in propagating between the two probes equal to $k_\theta(f)\Delta x$. In fig. 7b we give the phase information as well as the related averaged wavenumber $\bar{k}_\theta(f) = \hat{\theta}_L(f)/\Delta x$. We find an approximately linear dispersion relation and thus a nearly constant phase velocity of the wave over the observed frequency range. At the radial position discussed here (1 cm behind the limiter), the propagation direction of the phase velocity is in the ion diamagnetic drift direction.

The magnitude, $|L_f|$, gives the change of the signal level between the input X_f and the output Y_f due to linear effects. An amplitude $|L_f|$ larger than unity indicates a wave which grows while propagating a distance Δx. An amplitude smaller than one can be due to a linear damping mechanism or a transfer of energy to other waves as a result of nonlinear wave-wave coupling. Other reasons for the change of the amplitude also exist, as for example, the neglect of higher-order nonlinear terms, an external source of noise or a multidimensional behavior of the fluctuation. In fig. 7a we observe that $|\hat{L}_f|$ is always smaller than unity and drops off rapidly for frequencies greater than 350 kHz. We conclude that all spectral components are damped; however, we would require an unphysically large damping coefficient to describe the damping at high frequencies. For more insight we look at the "goodness of fit" of the model by plotting the coherency (fig. 8). The total coherency of the model (eq. (5)), $\gamma^2(f)$, is high for spectral components up to the frequency where the amplitude of the linear transfer function starts to decrease rapidly. The fast increase of $\gamma_n^2(f) = 1.0 - \gamma^2(f)$ for frequencies larger than 350 kHz demonstrate that systematic errors become important at high frequencies. The effect of instrument noise can be neglected, as the power of the signal is significantly larger than the noise power for all frequencies. We believe that the deviation of the coherency from unity is mainly due to the one-dimensional approach we have chosen. In our

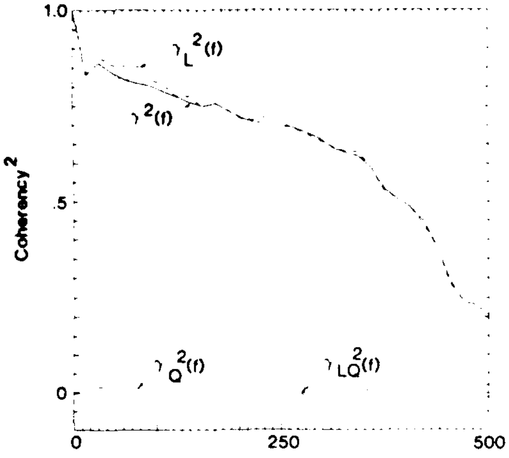


Fig. 8. Coherence between the two probes.

experimental set up, we look only at the poloidal component of the fluctuation. When a turbulent structure propagates at an angle with respect to the two probes it may be observed by only one of the probes and will cause a decorrelation between the signals. This effect will dominate when the scale length and thus the wave length of the structures becomes very small, as is the case for high frequencies (see fig. 7b).

We now turn our attention to the quadratic transfer function, for which the amplitude, $|\hat{Q}_f^{f_1,f_2}|$, is shown in a three-dimensional plot in fig. 9. During the propagation between the two probes, only a relatively small amount of the spectral amplitude of the output signal, Y_f, is generated at frequency $f=f_1+f_2>f_1, f_2$ due to wave-wave interaction with the frequency components f_1 and f_2 (see region (I) in Fig. 3a). The amount of the signal which is transferred to intermediate frequencies due to coupling with a higher frequency and a lower frequency component is larger. The largest values of the transfer function occur in the shaded region in fig. 9, corresponding to region (III) in fig. 3a. We conclude from this that the quadratic transfer function is strongest for the dominant low frequency components of the spectrum in fig. 6 (around 30 kHz) as a result of their interaction with two waves of similar frequency. The strength of this wave-wave coupling increases when f_1 and f_2 approach the Nyquist frequency. We must keep in mind, however, that because of the low power at high frequencies the absolute contribution of the high frequency components to the output signal is small, despite a large value of the transfer function.

When we examine the coherency due to quadratic interaction of the waves $\gamma_Q^2(f)$ (fig. 8a) we find very low values. The fraction of the total power at frequency f which originates from interaction with all frequency pairs which satisfy the selection rule $f=f_1+f_2$ is therefore small for a probe separation of 3.5 mm. The fact that the magnitude of the mixed term $\gamma_{LQ}^2(f)$ is larger in magnitude than $\gamma_Q^2(f)$ is not surprising. For small probe separations the turbulent structures reaching the first probe are still present when they reach the second probe (due to the linear transfer) and contribute even more to the output signal than that generated due to quadratic interaction between the two probes.

Having discussed the various sources of errors which affect the "goodness of fit" of the model, we have to ask if the measured quadratic transfer function gives a meaningful result. This question is critical, as we know from the coherency of the quadratic term as well as from theoretical models that the quadratic interaction is small. Also, the variance of the estimate is larger than that of the linear one, just as an estimate of a two-dimensional surface has more variance than an estimate of a one-dimensional curve, using the same number of data points. We have, however, good indication that the quadratic coefficient is qualitatively correct. When we look at the scaling of the coefficients with probe separation, we find: for an increased spacing of the probes, $\gamma_Q^2(f)$ increases and $\gamma_{LQ}^2(f)$ decreases. We can expect such a result for a correct measurement: for an increased probe separation, the interaction time for quadratic processes and linear damping is longer. This observation indicates that the quadratic transfer function is estimated correctly.

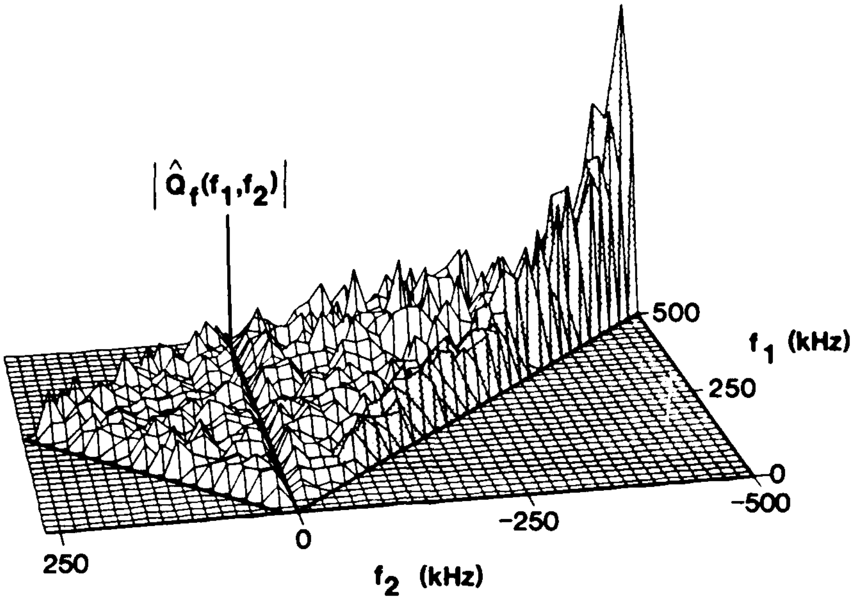


Fig. 9. Three-dimensional plot of the amplitude of the quadratic transfer function, obtained from the density fluctuations measured at the two probes.

How do these results compare with different theoretical predictions? Fundamental differences are predicted theoretically for two-dimensional and three-dimensional turbulence: in three-dimensional turbulence the energy cascades due to vortex stretching from large scale turbulent structures to small scale structures. In two-dimensional turbulence the energy is predicted to cascade, in contrast to the three-dimensional model, to smaller wave numbers and frequencies, as the vorticity is independent of the third direction. For tokamak plasmas with inherently strong, mainly toroidal, magnetic field a two-dimensional behavior of the turbulence is generally expected. The essentially two dimensional picture of the turbulence can be justified by the strong magnetic field which confines the charged particles in the perpendicular direction, while the electrons and ions can propagate relatively freely along the field lines. There are many different theoretical turbulence models for tokamak plasmas because a variety of linear instabilities and nonlinear mode coupling processes can be included [36]. Our experiment shows a dominant coupling of higher frequency spectral components in a way as to increase the amplitude at the lowest frequency component of the triplet. The strongest effect is found for coupling of two spectral components of comparable frequency to a very low frequency component. This result indicates an efficient "one step" process between relatively small scale turbulent structures ($k \geq 10$ cm^{-1}) to large scales ($k \leq 5$ cm^{-1}). The experiment indicates that the turbulence in the edge plasma of the TEXT tokamak is essentially of two-dimensional nature as expected theoretically for tokamak plasmas.

We mentioned in the introduction that much work dealing with turbulence is based on linear transfer functions [e.g. 1-4, 29]. The contribution of the nonlinear term in eq. (2) can then be neglected. The magnitude of the quadratic transfer function and, thus, the coherency $\gamma_Q^2(f)$ we observed suggest that such an assumption is, indeed, valid. The approximation can be confirmed quantitatively. When we neglect the quadratic terms and compute the linear transfer function (eq. (2a)) we find values which are very close to the ones for the non-Gaussian input case (dashed lines in figs. 7a and b). While our results are obtained for turbulence observed in the edge plasma of the TEXT tokamak with probes separated by $\Delta x = 3.5$ mm, they suggest that the quadratic terms can be neglected for similar experimental arrangements when one is interested only in the linear behavior of fully developed turbulence in the edge region of tokamaks. Note that the neglect of the nonlinear terms in the estimation of the linear transfer function can lead to an incorrect result for cases with small linear component L_p, as in this case the contribution of the quadratic term to eq. (2) can be considerable.

5. Conclusion

We have found that the usual method of estimating linear and quadratic transfer functions between two measurement points for Gaussian input signals can be extended to non-Gaussian input signals. This extension leads to a model with which a self-excited and fully turbulent system *can be* studied experimentally. The model is tested with simulated data and can reproduce the transfer functions. A check of the "goodness of fit" shows that the coherency is close to unity. The true output signal can thus be modeled nearly exactly from the input signal by using the estimated linear and quadratic transfer functions. Analysis of the same signals with the usual method, which assumes a Gaussian input signal, gives nonphysical values of the coherency with values larger than unity.

Using two Langmuir probes in the edge plasma of the TEXT tokamak we find: the linear transfer function yields a nearly linear dispersion relation of the waves with propagation in the ion diamagnetic drift direction. All spectral components are at least marginally damped. The amplitude of the quadratic transfer function is largest at the dominant low frequencies around 30 kHz due to coupling with two waves having comparable frequencies and lead to an increase of the amplitude at the low-frequency component. This result indicates an efficient "one step" process between small scale and large scale structures. The portion of the signal which is generated between the two probes due to quadratic interaction is small with respect to the linear contributions, but proves that many waves couple with each other in a fully turbulent spectrum. Our observation is consistent with two-dimensional turbulence models which predict an energy transfer to smaller scale structures due to three-wave coupling, which is in contrast to three dimensional models with energy cascading to larger scales.

We demonstrated in this paper the usefulness of the method to estimate the linear and quadratic transfer function for non-Gaussian inputs. Although we applied the method to plasma turbulence data, we emphasize that the method is equally applicable to analyze other data, as in fluid turbulence, for example. In a future work, we will extend this method to estimate the three-wave coupling coefficients and thus quantify energy cascading in turbulence. Such experimental measurements can be useful for a direct comparison with theoretical models.

Acknowledgements

We wish to thank Roger Bengtson for his encouragement and careful reading of the manuscript. We are pleased to thank Ken Gentle, Scott Levinson and Sung Bae Kim for useful discussions and comments on the manuscript. The consistent theoretical motivation and help of Pat Diamond and Paul Terry is gratefully acknowledged. We would further like to thank the TEXT group of the Fusion Research Center at U.T. Austin for their technical support.

This work was supported by the U.S. Department of Energy Contract No. DE-AC05-78ET53043, the Joint Services Electronics Program Contract No. F49620-82-C-0033, and in part by DOE Contract DE-AC05-85ER80251, and the

Texas Atomic Energy Research Foundation. Ch.P. Ritz was supported partially by the Swiss National Science Foundation.

References

[1] S.J. Zweben, P.C. Liewer and R.W. Gould, J. Nucl. Mater. 111-112 (1982) 39.

[2] T.A. Casper and J.L. Shohet, Nucl. Fusion 20 (1980) 1369.

[3] S.J Levinson, J.M. Beall, E.J. Powers and Roger D. Bengtson, Nucl. Fusion 24 (1984) 527.

[4] S.J. Zweben and R.W. Gould, Nucl. Fusion 25 (1985) 171.

[5] M.A. Flemming, F.H. Mullins and A.W.D. Watson, Proc. IEE Int. Conf. RADAR-77, Oct. 25–28 (1977) 552.

[6] E.J. Powers, J.Y. Hong and Y.C. Kim, IEEE Trans. Aerospace and Electronic Systems AES-17 (1981) 602.

[7] J.A. Pinkster, 1975 Society of Petroleum Engineering Journal, p. 487.

[8] F.H. Hsu and K.A. Blenkarn, 1972 Society of Petroleum Engineering Journal, p. 329.

[9] G.F.M. Remery and A.J. Hermans, 1972 Society of Petroleum Engineering Journal, p. 191

[10] D. Choi, J.H. Chang, R.O. Stearman and E.J. Powers, Proc. of the 2nd Int. Modal Analysis Conf., Orlando, Florida, 1984, pp. 602–609.

[11] N. Bloembergen, Nonlinear Optics (Benjamin, New York, 1965)

[12] R.H. Kraichnan, Fluid Mech. 5 (1958) 497.

[13] S. Chandrasekhar, J. Madras Univ. B 27 (1957) 251.

[14] B.B. Kadomtsev, Plasma Turbulence (Academic Press, New York, 1965); Phenomenes Collectifs dans les Plasmas (Edition MIR, Moscou, 1979).

[15] A. Hasegawa and K. Mima, Phys. Fluids 21 (1978) 87.

[16] P.H. Handel, Proc. Turbulence of Fluids and Plasmas, April 16–18, (1968) 381–395

[17] Ch.P. Ritz, S.J. Levinson, E.J. Powers, R.D. Bengtson and K.W. Gentle, Proc. Int. Conf. Plasma Phys. (1984) 73.

[18] O.M. Phillips, in: Nonlinear Waves, Leibovich and Seebas, eds. (Cornell Univ. Press. Ithaca, 1974) 186.

[19] Y.C. Kim, W.F. Wong, E.J. Powers and J.R. Roth, Proc. IEEE 67 (1979) 428.

[20] J.Y. Hong, Y.C. Kim and E.J. Powers, Proc. IEEE 68 (1980) 1026.

[21] D. Choi, R.W. Miksad, E.J. Powers and F.J. Fischer, J. of Sound and Vibr. 99 (1985) 309.

[22] M.D. Millionshchikov, Dokl. Akad. Nauk SSSR 32 (1941) 611.

[23] P.W. Terry and P.H. Diamond, Phys. Fluids 28 (1985) 1419.

[24] P.W. Terry, P.H. Diamond, K.C. Shaing, L. Garcia and B.A. Carreras, "The Spectrum of Resistivity Gradient Driven Turbulence", Institut for Fusion Studies, Report IFSR-174, Austin, Texas.

[25] R.E. Waltz and R.R. Dominguez, Phys. Fluids 26 (1983) 3338.

[26] D.R. Brillinger and M. Rosenblatt, in: Spectral Analysis of Time Series, B. Harris, ed. (Wiley, New York, 1967), p.153.

[27] Y.C. Kim and E.J. Powers, IEEE Trans. on Plasma Science, PS-7 (1979) 120.

[28] M.B. Priestley, Spectral Analysis and Time Series (Academic Press, London, 1981) chap. 6.

[29] Ch.P. Ritz, R.D. Bengtson, S.J. Levinson and E.J. Powers, Phys. Fluids 27 (1984) 2956.

[30] P.A. Duperrex, Ch. Hollenstein, B. Joye, R. Keller, J.B. Lister, F.B. Marcus, J.M. Moret, A. Pochelon and W. Simm, Phys. Lett. A. (1984) 133.

[31] S.J. Zweben, Phys. Fluids 28 (1985) 974.

[32] M. Wakatani and A. Hasegawa, Phys. Fluids 27 (1984) 611.

[33] R.E. Waltz, Phys. Fluids 28 (1985) 577.

[34] R.N. Franklin, Rep. Prog. Phys. 40 (1977) 1369.

[35] D.R. Nicholson, Phys. Rev. Lett. 52 (1984) 2152.

[36] P.C. Liewer, Nucl. Fusion 25 (1985) 543.

OFFSHORE ENGINEERING, VOLUME 5

Proceedings of the 5th International Symposium on Offshore Engineering held at COPPE, Federal University of Rio de Janeiro, Brazil, September, 1985.

Sponsored by:
FEDERAL UNIVERSITY OF RIO DE JANEIRO
PETROBRÁS S.A. (The Brazilian State Oil Company)
CPNq (The Brazilian Council for Scientific and Technological Development)
FINEP (The Brazilian Federal Agency for Supporting Studies and Projects)

Edited by:
F. L. L. B. Carneiro, R. C. Batista and A. J. Ferrante
COPPE/Federal University of Rio de Janeiro

PENTECH PRESS
London

MODELING OF NONLINEAR SYSTEMS IN OFFSHORE ENGINEERING FOR NON-GAUSSIAN INPUTS

K.I. Kim, E.J. Powers, Ch.P. Ritz and R.W. Miksad
College of Engineering
The University of Texas at Austin, Austin, Texas 78712, U.S.A.
F.J. Fischer
Shell Development Company, Houston, Texas 77001, U.S.A.

INTRODUCTION

When moored in random seas, ships and barges undergo large-amplitude oscillations at or near the undamped natural frequency of the vessel- mooring system. This phenomenon is known as low-frequency drift oscillation since the frequencies of such motions are well below those of the incident sea waves. Work by Pinkster (1979), and Remery and Hermans (1972) indicate that the mean and low-frequency wave drift forces are related to the amplitude of the incident wave by a second order nonlinearity. Thus the relationship between random wave excitation and the corresponding response of a moored vessel is nonlinear.

Recently, digital time series analysis techniques have been utilized to model the nonlinear relationship in the frequency domain by Choi et al. (1985), and to predict the low-frequency drift oscillation in time domain by Koh et al. (1984). A fundamental assumption underlying those approaches involves the fact that the "input" is assumed to be a random waveform which possesses Gaussian statistics, an assumption which results in a substantial simplification of the relevant mathematics. In many cases where the input excitation is under the control of the experimentalist it is often possible to satisfy the Gaussian assumption. In many practical cases, however, the input excitation is not under the direct control of the experimentalist, such as in the case of wave loading of offshore structures by large amplitude irregular waves. For example, if nonlinear wave-wave interactions have occurred, the resultant fluctuations in instantaneous sea wave height will not be Gaussian distributed. As discussed by Ritz and Powers(1985), when the input is weakly non-Gaussian, transfer functions of a nonlinear system can be estimated by an iterative approach, which also shows that considerable error occurs in estimating transfer functions of nonlinear systems with non-Gaussian inputs based on a Gaussian assumption.

It is the purpose of this paper to describe a new approach that we have recently developed for analyzing, modeling, and interpreting nonlinear systems, when the nonlinear system is excited by non-Gaussian inputs. The technique should have wide application to many areas of offshore engineering. In this paper, the validity of the approach has been demonstrated with data taken in model basin studies of nonlinear drift wave oscillations of moored vessels subjected to random sea wave excitation.

NONLINEAR SYSTEM MODELING WITH NON-GAUSSIAN INPUTS

We can extend nonlinear system concepts so as to investigate many nonlinear phenomena that occur in various engineering problems. In so doing, one approach is to express the input-output relationship in terms of a functional series such as Volterra series (see for example, Schetzen, 1980). Then the problem reduces to the measurement of the kernels involved in the expression.

In this paper it will be assumed that nonlinearities of order higher than the second are negligible and the input-output relationship can be reasonably described by the Volterra series up to second order. Our focus is on the frequency domain modeling where the key idea is to model the nonlinear system with a parallel combination of linear and quadratic transfer functions. Therefore our model can be expressed as follows;

$$Y(f_m) = H_1(f_m)X(f_m) + \sum_{i+j=m}\sum H_2(f_i,f_j)X(f_i)X(f_j) \qquad (1)$$

where $X(f_m)$ and $Y(f_m)$ are the N-point discrete Fourier transforms (DFT's) of the input (sea wave) and the output (barge sway response) respectively. $H_1(f_m)$ in Eq. (1) is referred to as the linear transfer function (LTF), and $H_2(f_i,f_j)$ the quadratic transfer function (QTF). In the discussion that follows, it will be assumed that the QTF is a symmetric function of its arguments, i.e., $H_2(f_i,f_j)=H_2(f_j,f_i)$ since it can be seen that the quadratic term in Eq. (1) will be identical if the arguments are interchanged.

The problem is to determine the LTF and the QTF by processing the input and the output data which are measured from experiments and stored in a digital computer. This can be done by solving the following set of equations which are obtained by multiplying Eq. (1) by $X^*(f_m)$ and $X^*(f_k)X^*(f_l)$, respectively.

$$\langle X^*(f_m)Y(f_m)\rangle = H_1(f_m)\langle|X(f_m)|^2\rangle + \sum_{i+j=m}\sum H_2(f_i,f_j)\langle X^*(f_m)X(f_i)X(f_j)\rangle \qquad (2)$$

$$\langle X^*(f_k)X^*(f_l)Y(f_m)\rangle = H_1(f_m)\langle X^*(f_k)X^*(f_l)X(f_m)\rangle + \sum_{i+j=m}\sum H_2(f_i,f_j)\langle X^*(f_k)X^*(f_l)X(f_i)X(f_j)\rangle (3)$$

The angle brackets denote statistical averagings, and * denotes the complex conjugate. Note that Eq. (3) is meaningful only when $m=i+j=k+l$ because of the properties of higher order spectra as summarized by Choi et al. (1985).

If the system input is a zero-mean Gaussian, the terms containing the third order moment in Eqs. (2) and (3) vanish, and $H_1(f_m)$ and $H_2(f_i,f_j)$ can be determined separately. It has been shown (Choi et al., 1985) that they are determined by the various spectra associated with the input and the output. Specifically, $H_1(f_m)$ and $H_2(f_i,f_j)$ are given in terms of the cross power spectrum and cross bispectrum between input and output respectively. However, in general, we have to solve Eqs. (2) and (3) simultaneously so that it is extremely difficult to find closed form solutions for LTF and QTF. As discussed later, analysis of the probability density function and the bispectrum characterizing the sea wave excitation used in our studies indicate that it is not appropriate to assume a Gaussian excitation. Accordingly, in this case, we need to find LTF and QTF by simultaneously solving Eqs. (2) and (3).

Now we will describe a practical method of solving Eqs. (2) and (3) which can be implemented on a digital computer. Then the linear and the quadratic transfer functions will be calculated at a discrete set of equally spaced frequency points.

Using vector notation, we can rewrite Eq. (1) as

$$Y(f_m) = \underline{h}^t\underline{x} = \underline{x}^t\underline{h} \qquad (4)$$

where t denotes transposition, and

$$\underline{h}^t = [H_1(f_m) \quad H_2(f_{m-M},f_M) \ \dots \ H_2(f_0,f_m) \ \dots \ H_2(f_M,f_{m-M})] \qquad (5)$$

$$\underline{x}^t = [X(f_m) \quad X(f_{m-M})X(f_M) \ \dots \ X(f_0)X(f_m) \ \dots \ X(f_M)X(f_{m-M})].(6)$$

In Eqs. (5) and (6), f_M signifies the Nyquist frequency associated with the sampling of time series data. Then solving Eqs. (2) and (3) is equivalent to solving the following matrix equation.

$$\langle \underline{x}^*Y(f_m)\rangle = \langle \underline{x}^*\underline{x}^t\rangle \underline{h} \tag{7}$$

Eq. (7) is linear in the transfer function vector $\underline{h}$, and so $\underline{h}$ is given by

$$\underline{h} = \langle \underline{x}^*\underline{x}^t\rangle^{-1}\langle \underline{x}^*Y(f_m)\rangle \tag{8}$$

if $\langle \underline{x}^*\underline{x}^t\rangle$ is not singular. Note that $\langle \underline{x}^*\underline{x}^t\rangle$ is a Hermitian matrix consisting of various spectral moments of the input signal. Except for the first element, the first row and column represent the bispectrum while the first element is the auto power spectrum of the input. The remaining elements are fourth order spectral moments of the input. The solution given by Eq. (8) can be considered as a result of multivariate linear regression analysis and thus the transfer functions obtained in this manner are optimum in the mean square sense when there is any additive noise present in the output.

From the standpoint of memory space and computation time, it is very useful to reduce the size of the matrix $\langle \underline{x}^*\underline{x}^t\rangle$ to be inverted. We can see that it depends on the number of terms on the R.H.S. of Eq. (1) (or the number of elements in $\underline{h}$ or $\underline{x}$), and again on the number of data points (N=2M) taken for the DFT's and the frequency index m. If we use the symmetricity of $H_2(f_i,f_j)$, we can rewrite Eq. (1) such that the number of terms in the quadratic part are reduced by about half. The reduction also enables us to decrease the number of elements in $\underline{h}$ or $\underline{x}$, which is important because it is directly related to the size of the matrix $\langle \underline{x}^*\underline{x}^t\rangle$. For example, the largest one is $(M+2)\times(M+2)$ when $m=0$, and the smallest one is $(\frac{M}{2}+2)\times(\frac{m}{2}+2)$ when $m=M-1$ or M.

It can be shown that we obtain, from Eq. (8), the same expressions for LTF and QTF as those of Choi et al., 1985, when the system input is a zero-mean Gaussian. Thus the solution given by Eq. (8) is a general one for the transfer functions of a quadratically nonlinear system with an arbitrary random input, i.e., it includes the zero-mean Gaussian input as a special case.

ANALYSIS OF EXPERIMENTAL DATA AND RESULTS

Data from a scaled (1:48) model wave basin test of a moored barge in an irregular sea are analyzed by the techniques developed in this paper. Expected quantities in Eq. (8) were obtained by dividing up the sample data record into 50 segments covering different time intervals and then averaging those quantities computed by the fast Fourier transform over different segments. Each segment contains 128 sample points of data.

Shown in Fig. 1 are the time traces of the random sea wave height and the barge sway response. Examination of Fig. 1 clearly indicates that the oscillation frequencies of the moored barge, which may be modeled as a dynamic system having a large mass and small restoring stiffness, are considerably lower than those of the sea wave input. We can see this more clearly from the plots of auto power spectra of the sea wave input and the barge sway response shown in Fig. 2. In particular, we note that most of the incident sea wave energy is contained in a frequency band ranging from 0.13 to 0.20 Hz while the sway response occurs at a low frequency equal to approximately 0.008 Hz.

We examined the amplitude statistics of the incident random sea wave input by computing the probability density function (PDF) shown in Fig. 3. Although bell-shaped, the PDF is not a pure Gaussian in that it is slightly skewed and has non-smooth tails. To further test the Gaussian nature of the signal we computed the auto bispectrum of the incident sea wave input. As described by Kim and Powers (1979), the auto bispectrum is a third order moment and should be identically zero for a Gaussian signal. In Fig. 4 (a) and (b) are shown perspective and contour plots of the bispectrum. The large peak occurring at approximately $f_i=f_j=0.165$ Hz, is indicative of a second harmonic component present in the incident sea waves which in turn results in non-Gaussian statistics.

Transfer functions are calculated based on Eq. (8) and amplitudes of the LTF and the QTF are shown in Fig. 5 and Fig. 6 respectively. Note that although the linear transfer function peaks near 0.008 Hz in Fig. 5, the corresponding response will be very small since the incident sea waves contain very little power at this frequency. On the other hand, the QTF shown Fig. 6 has its amplitude peaks along the line $f_i-f_j=0.008$ Hz which is the frequency of the barge sway response. This implies that all pairs of spectral components in the incident sea wave such that the differences are equal to 0.008 Hz contribute to the low frequency barge sway response. The narrow peaks also imply the sharp resonant behavior of the vessel-mooring system.

Next the actual auto power spectrum of the barge sway response is compared with that of the model output which is generated by applying the sea wave input signal to the computed transfer functions. Fig. 7 (a) was obtained by the new approach developed in this paper, and Fig. 7 (b) by the previous approach of Choi et al., 1985, which assumed a Gaussian input. Neglecting the third order moment terms in Eqs. (2) and (3) by

assuming that the sea wave input is Gaussian may lead to an erroneous result in that the model output power is larger than the actual output power (Fig. 7 (b)). Good agreement between the actual and the model output power in Fig. 7 (a) demonstrates superiority of the new approach.

The usefulness of the new technique has also been shown by comparing the actual and the model output signal in time domain. The results are shown in Fig. 8 where the model output signal has been recovered by inverse Fourier transforming the frequency domain model output. Note that the actual sway response shown in Fig. 8 (a) and the model sway response shown in Fig. 8 (b) are in good agreement. The small error (Fig. 8 (c)) indicates the validity of the computed transfer functions (both amplitude and phase), and also implies that the new approach can be used to "predict" the output successfully.

DISCUSSION AND CONCLUSION

In this paper, we have described a new approach for modeling a nonlinear system which is excited by a non-Gaussian input. It is shown that the linear and the quadratic transfer functions can be calculated, on a discrete set of frequencies, by solving matrix equations. The elements of the matrices consist of the second, the third and the fourth order spectral moments of the input and the output. Applying the technique developed in this paper to experimental low-frequency drift oscillation data, we have also demonstrated the feasibility of the method. In particular, the ability to estimate the phases of the transfer functions has been confirmed by comparing the actual and the model output in time domain.

As for the practical applications, we have found that the computation time required by our new approach is comparable to that for Gaussian input method. For example, in the case of analysis described in the previous section, it took about 100 seconds on a CDC Dual Cyber 170/750 with our new method while the other method required about 70 seconds. We also observed on the basis of simulation studies that the method developed here yielded better estimates of transfer functions even for Gaussian input cases.

ACKNOWLEDGEMENTS

The authors' work on nonlinear wave phenomena and applications of digital time series analysis techniques to such phenomena is supported by Air Force Office of Scientific Research (AFSC) Cont. F49620-82-C-0033, and NSF Grant MEA-8211205.

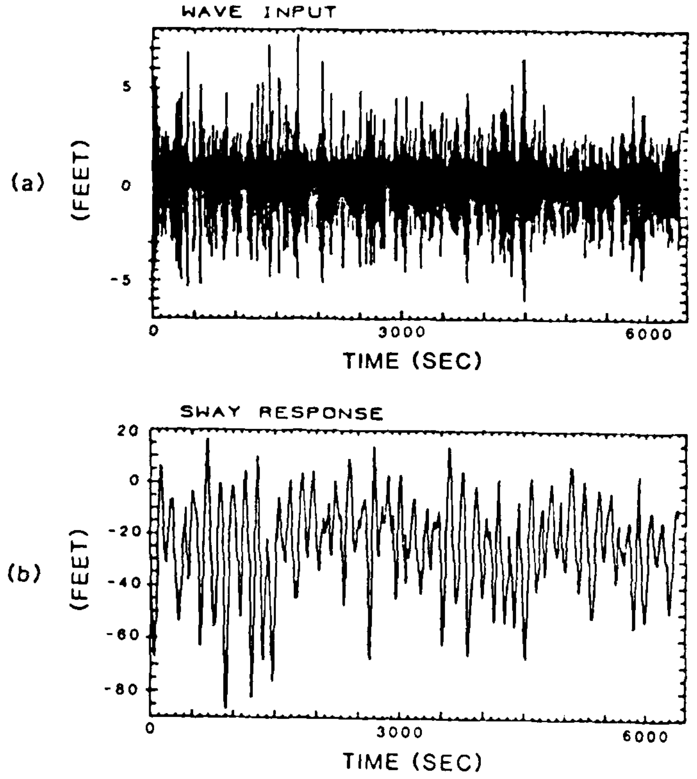


Fig. 1. The sway motion of a moored barge in response to an irregular sea: (a) irregular sea records, (b) sway response of barge.

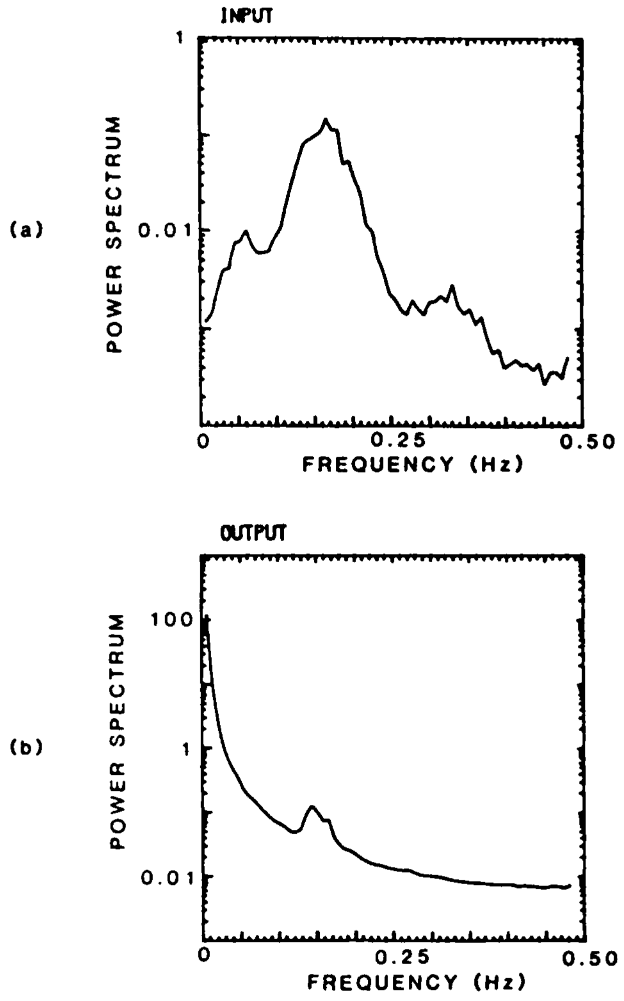


Fig. 2. Power spectrum of: (a) sea waves, (b) sway response of barge .

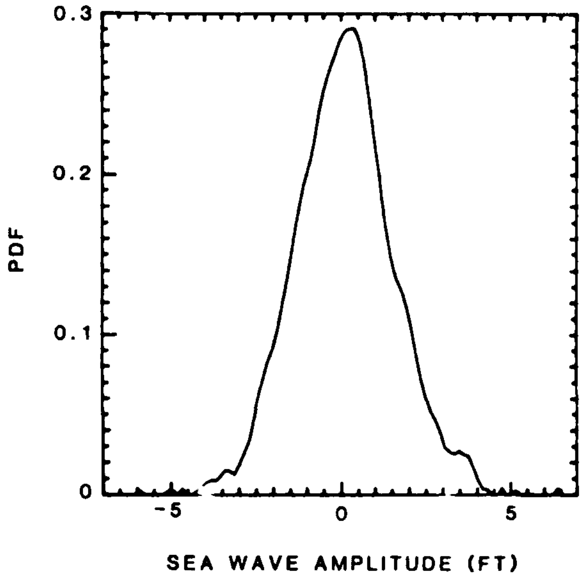


Fig. 3. Probability density function of the incident irregular sea wave amplitude.

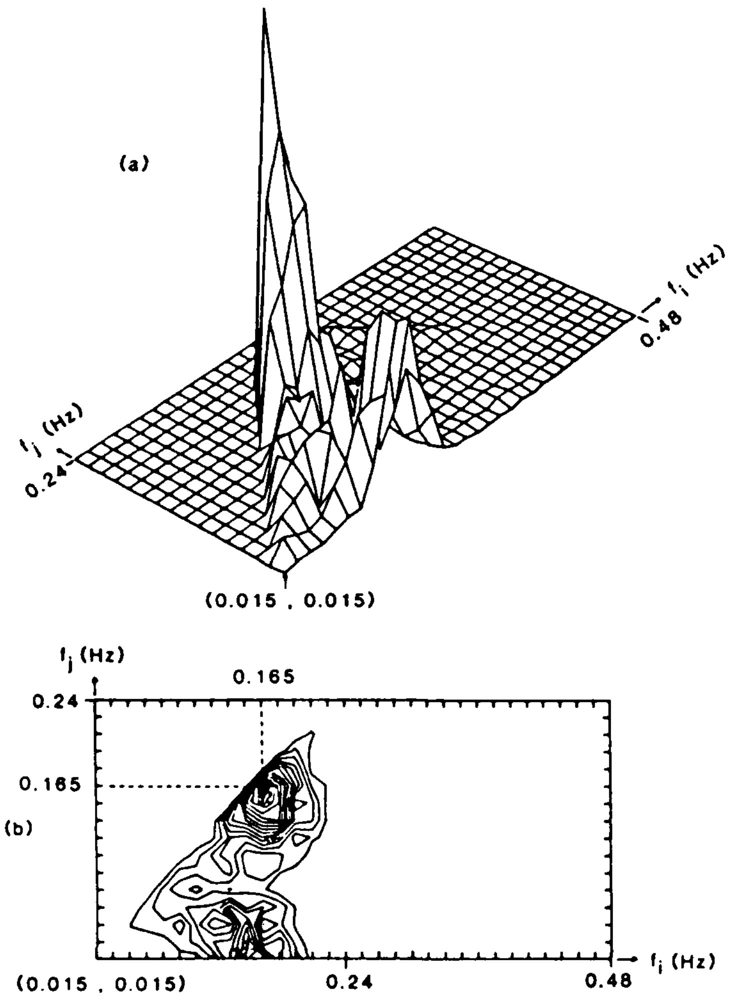


Fig. 4. Amplitude of auto bispectrum of the sea wave input:
(a) perspective view, (b) contour plot.

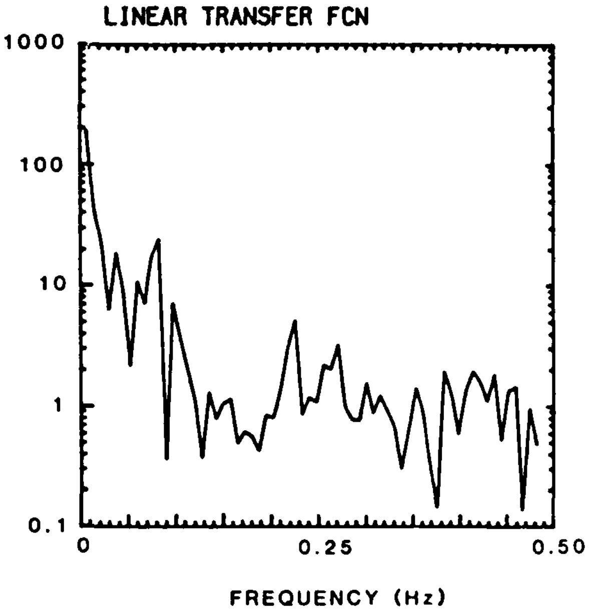


Fig. 5. Amplitude of linear transfer function.

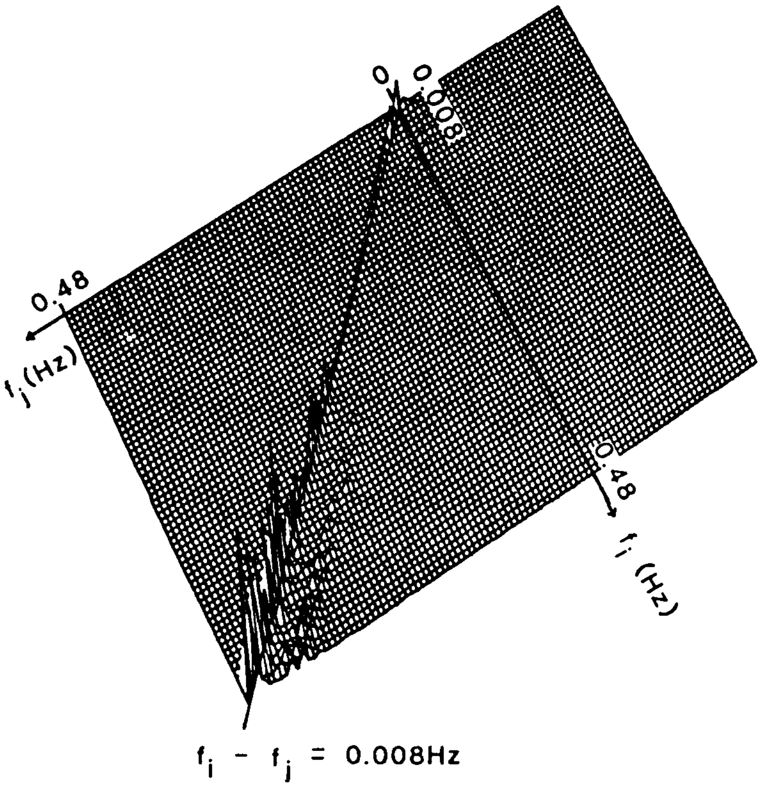


Fig. 6. Amplitude of quadratic transfer function.

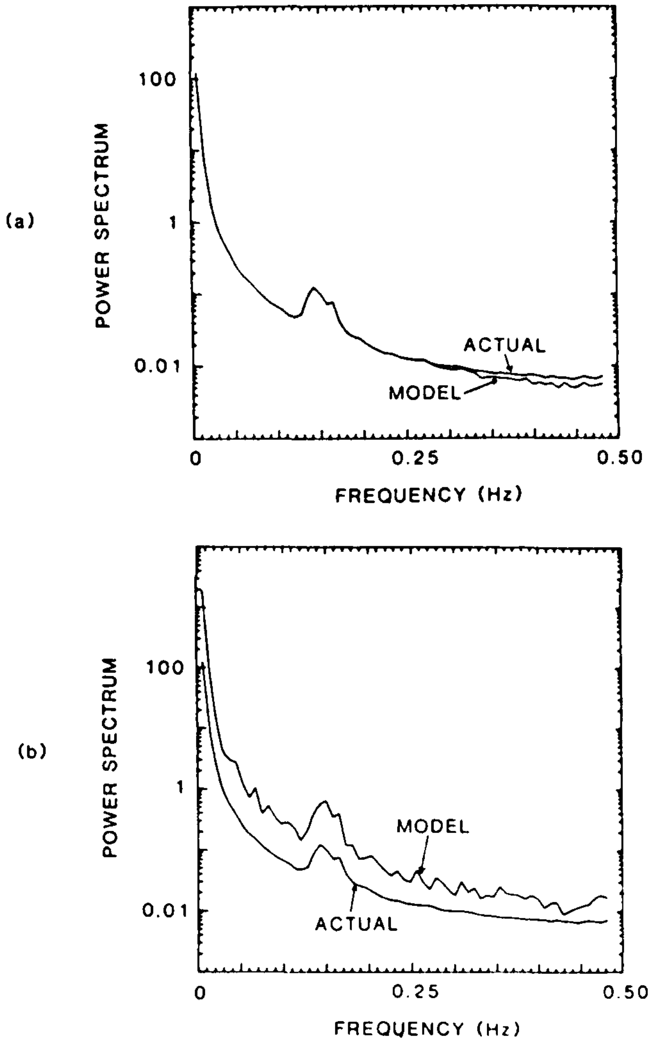


Fig. 7. *Power spectrum of actual and model sway responses:* (a) non-Gaussian method of this paper, (b) Gaussian method (Choi et al., 1985).

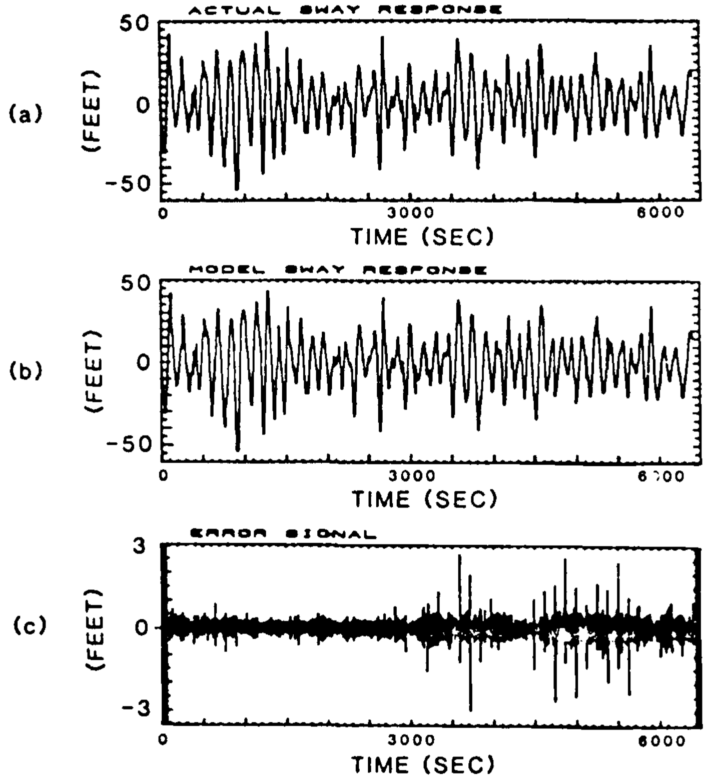


Fig. 8. Actual and model sway responses: (a) actual sway response, (b) output of model, (c) difference between actual and model output (note change in scale).

REFERENCES

Choi, D-W., Miksad, R.W., Powers, E.J. and Fischer, F.J. (1985) Application of Digital Cross-Bispectral Analysis Techniques to Model the Nonlinear Response of a Moored-Vessel System in Random Seas, Journal of Sound and Vibration, in press.

Kim, Y.C. and Powers, E.J. (1979) Digital Bispectral Analysis and Its Applications to Nonlinear Wave Interactions, IEEE Trans. Plasma Sci., Vol. PS-7, No. 2, June, 120-131.

Koh, T., Powers, E.J., Miksad, R.W. and Fischer, F.J. (1984) An Approach to Time Domain Modelling of Nonlinear Drift Oscillations in Random Seas, Offshore Engineering, Vol. 4, Pentech Press, London, 137-153.

Pinkster, J.A. (1979) Mean and Low Frequency Wave Drifting Forces on Floating Structures, Ocean Engineering, Vol. 6, 593-615.

Remery, G.F.M. and Hermans, A.J. (1972) The Slow Drift Oscillations of a Moored Object in Random Seas, Society of Petroleum Engineers Journal, June, 191-198.

Ritz, Ch.P. and Powers, E.J. (1985) Estimation of Nonlinear Transfer Functions for Fully Developed Turbulence, FRCR no. 273, Fusion Research Center, The University of Texas at Austin, March; submitted for publication in Physica D, 'Nonlinear Phenomena'.

Schetzen, M. (1980) The Volterra and Wiener Theories of Nonlinear Systems, New York, Wiley.

International Journal of Infrared and Millimeter Waves, Vol. 7, No. 3, 1986

THE APPLICATION OF HOMODYNE SPECTROSCOPY TO THE STUDY OF LOW-FREQUENCY MICROTURBULENCE IN THE TEXT TOKAMAK

D. L. Brower, N. C. Luhmann, Jr., and W. A. Peebles
University of California, Los Angeles
Los Angeles, CA 90024

and

Ch. P. Ritz and E. J. Powers
University of Texas, Austin
Austin, TX 78712

Received December 19, 1985

A new homodyne spectroscopy technique has been applied for the first time to tokamak microturbulence measurements in order to ascertain the frequency spectra and wave propagation direction of low-frequency density fluctuations. This method is employed in lieu of more expensive and complicated heterodyne detection schemes typically available for far-infrared laser scattering systems.

Introduction

Laser and millimeter-wave scattering techniques are commonly used to study the space-time statistics of electron density fluctuations in tokamak and other plasmas. Of particular importance is the determination of the direction of propagation of the fluctuations. Since the scattering geometry fixes the direction of the scattering wave vector $\underline{k}$, the direction of propagation information is carried by the sign of the fluctuation frequency ω. The fact that waves may be propagating both parallel and

0195-9271/86/0300-0447$05.00/0

antiparallel to **k** is manifested by the presence of blue and red sidebands centered around the incident wave frequency ω_o.

To recover the propagation direction information contained in the blue and red sidebands, heterodyne detection techniques are typically employed. This approach requires two coherent sources, with a frequency difference $\Delta\omega$, to be utilized as the incident and local oscillator beams. After mixing, the frequency range of the resultant signal is $\Delta\omega \pm \omega$, where ω is the frequency associated with the plasma fluctuations. As long as $\Delta\omega >> \omega$, the blue/red sidebands may be resolved. In contrast, if a classical homodyne approach is used, $\Delta\omega = 0$, and it is no longer possible to unambiguously determine the wave propagation direction.

Realization of a heterodyne detection system in the far-infrared is expensive and technically nontrivial. Utilization of a rotating grating to frequency shift a portion of the source beam is a feasible alternative although the frequency offset is limited to roughly $\Delta\nu \leqslant$ 150 kHz (insufficient for microturbulence measurements where fluctuations are observed up to 1 MHz). In addition, there is noise associated with the grating which limits resolution near zero frequency and fabrication can be costly.

Experimental Technique and Apparatus

A considerably simpler and inexpensive method proposed by Tsukishima[1] and Asada et al.,[2] permits detection of the wave propagation direction from the analysis of homodyne signals. The IF output of the scattered signal after the mixer is a real quantity described by

$$v(t) = \text{Re}\ \{\int_{-\infty}^{+\infty} \frac{d\omega}{2\pi} N(k_+,\omega)e^{i\omega t}\},$$

where $k_+ = k_s - k_o$ is the wave vector of the plasma fluctuation and $N(k_+,\omega) = \tilde{n}(k_+,\omega)$ with $\tilde{n}$ being the density fluctuation level. The sign of the frequency spectrum represents the propagation direction of the wave in the laboratory frame of reference and thus $N(k_+,\omega) \neq N(k_+,-\omega)$. However, from the real time signal $v(t)$, one is restricted to the reconstruction of the spectrum $N'(k_+,\omega)$ with the symmetry property $N'(k_+,\omega) = N'^*(k_+,-\omega)$, thereby losing wave propagation direction information. The key idea of the new homodyne spectroscopy technique proposed by Tsukishima[1] is to reconstruct the complex time signal $w(t)$ and recover the complete wave information by using two homodyne IF signals $v_1(t)$ and $v_2(t)$ which are phase shifted by 90° with respect to each other permitting one to write $w(t) = v_1(t) + i\, v_2(t)$.[1,2]

The blue and red sidebands of the scattered radiation, $S_+(\omega)$ and $S_-(\omega)$ where $S_\pm(\omega) \propto [\tilde{n}_\pm(\omega)]^2$, can be readily calculated from the two IF signals $v_1(t)$ and $v_2(t)$.

$$S_\pm(\omega) = \left(G_{11}(\omega) + G_{22}(\omega)\right) \pm 2\, \mathrm{Im}\left(G_{12}(\omega)\right), \quad \omega > 0,$$

where $G_{11}(\omega)$ and $G_{22}(\omega)$ are the auto-power and $G_{12}(\omega)$ the cross-power spectral densities of the two signals $v_1(t)$ and $v_2(t)$, and are given by

$$G_{ik}(\omega) = \langle V_i(\omega) V_k^*(\omega) \rangle , \qquad i,k = 1,2.$$

The spectral component $V_i(\omega) = \mathcal{F}(v_i(t))$ is the Fourier transform of the IF signal $v_i(t)$ and $< >$ denotes an ensemble average over many statistically similar realizations.

In the experimental results to be shown later, a time series of 32k data points (length of the time sample $T = 16$ ms at a sampling rate of 2 MHz) was subdivided into 128 realizations of 256 data points ($T' = 122\ \mu s$). By employing a fast Fourier transform algorithm, the frequency spectrum was obtained with a resolution of $\Delta\omega/2\pi = 1/T' = 8$ kHz. Mixer and amplifier noise contributions could be subtracted from the autopower spectra although signal-to-noise levels were sufficiently large so as to make it unnecessary.

A schematic of the experimental arrangement employed for application of the homodyne spectroscopy technique[3] to collective far-infrared scattering is shown in Fig. 1. The source beam utilized for the incident and local oscillator radiation is a $C^{13}H_3F$ far-infrared laser producing ≈ 20 mW of power at 245 GHz (1.22 mm). Detection is achieved by

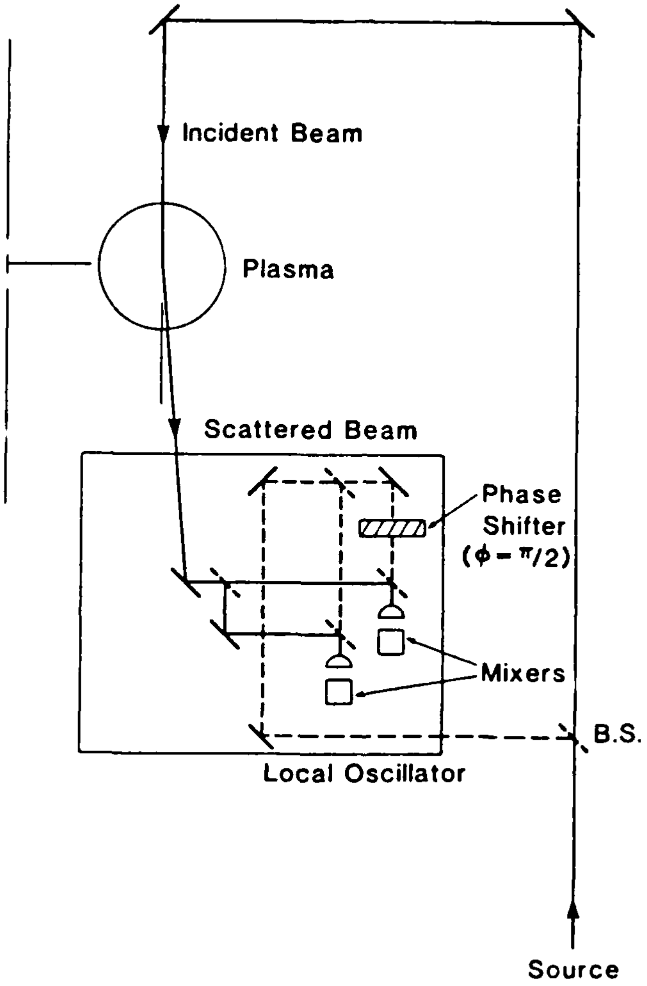


Figure 1 Experimental arrangement for homodyne spectroscopy measurement

the use of quasi-optical biconnical Schottky barrier diode mixers. The frequency shifted scattered radiation may be collected over a range of wave vectors from $0 < k_\perp < 15$ cm^{-1}. Detailed information on the scattering system and calibration procedures is described by Park et al.[4,5] The portion of the far-infrared laser utilized as the probe beam is weakly focussed along a vertical chord to a waist of ≈ 2 cm producing a measured wavenumber resolution $\Delta k_\perp = \pm 1$ cm^{-1}. The length of the scattering volume varies as a function of wavenumber and ranges from ±8 cm (e^{-1} point of scattered power) at $k_\perp = 12$ cm^{-1} to a chord average as $k_\perp \rightarrow 0$. Scattered radiation in the plane perpendicular to the toroidal magnetic field is examined.

At a particular wave vector $k_\perp$, the scattered radiation beam is divided equally into two components which are coupled into detectors 1 and 2 by 50 % reflectivity beam splitters. Similarly, the local oscillator beam is equally divided to provide rf drive for each mixer. In one leg of the local oscillator beam (det. 2), a phase shifter is inserted. This phase shifter consists of a piece of high density polyethylene (excellent transmission characteristics at 245 GHz) mounted on a rotation stage. By tuning the rotation angle, the path of the LO beam through the polyethylene is altered thereby changing its phase with respect to det. 1. This phase shifter is tuned such that there is a 90^o phase difference between dets. 1 and 2. The signal from each detector is then amplified and digitized so that the cross- and auto- power spectra may be *computed*.

Experimental Results

The above described technique may now be applied to density fluctuation measurements in a high temperature tokamak plasma. Microturbulence (low-frequency density fluctuation) is driven by the free energy associated with plasma inhomogeneities such as density and temperature gradients. For drift wave type fluctuations, it is predicted that the turbulence will exhibit a phase velocity $\underline{v}_{De} = \omega/k_\Theta = [k_B T_e/eB_T n_e]\ \nabla n_e \times \underline{B}_T/|\underline{B}_T|$, where $\underline{v}_{De}$ is the electron diamagnetic drift velocity, k_Θ and ω are the poloidal wave vector and frequency of the fluctuation, B_T is the toroidal magnetic field, T_e is the electron temperature, and n_e is the electron density. The geometry for scattering from electron drift waves is shown schematically in Fig. 2. The scattering system is positioned such that the incident beam impinges upon the plasma from the top of the torus (also see Fig. 1) along a vertical chord at the major radius R = 1 m. This provides for scattering from fluctuations with a poloidal wave vector, $\underline{k}_\Theta$. Depending upon the orientation of the collection optics with respect to the incident beam, the wavenumber matching condition (momentum conservation) gives $\underline{k}_s = \underline{k}_o \pm \underline{k}$ where $\underline{k}_s$, $\underline{k}_o$, $\underline{k}$ are the wave vectors of the scattered beam, incident beam, and plasma fluctuation, respectively. It is important to note that for a particular scattering geometry, a sign change will occur ($\pm\underline{k}$) when one switches from the plasma top to bottom (see Fig. 2 (a) and (c) or (b) and (d)).

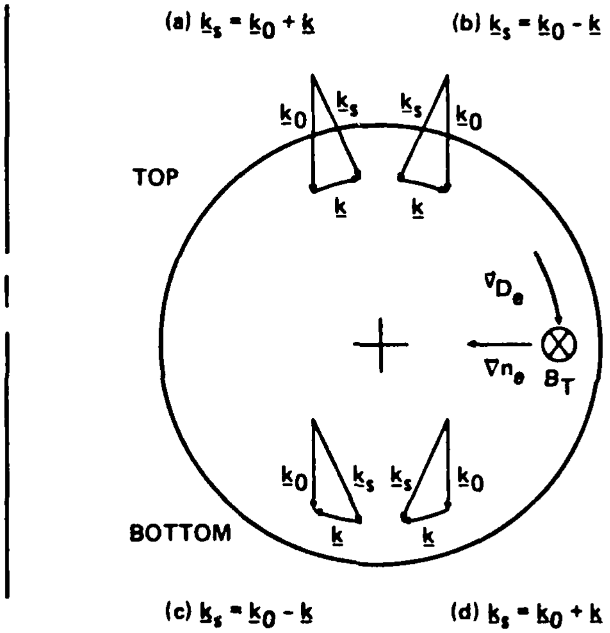


Figure 2. Tokamak scattering geometry.

Experimental results from the TEXT tokamak (major radius R = 1m and minor radius a = 27 cm) are shown in Fig. 3, for a scattering volume located at the plasma bottom

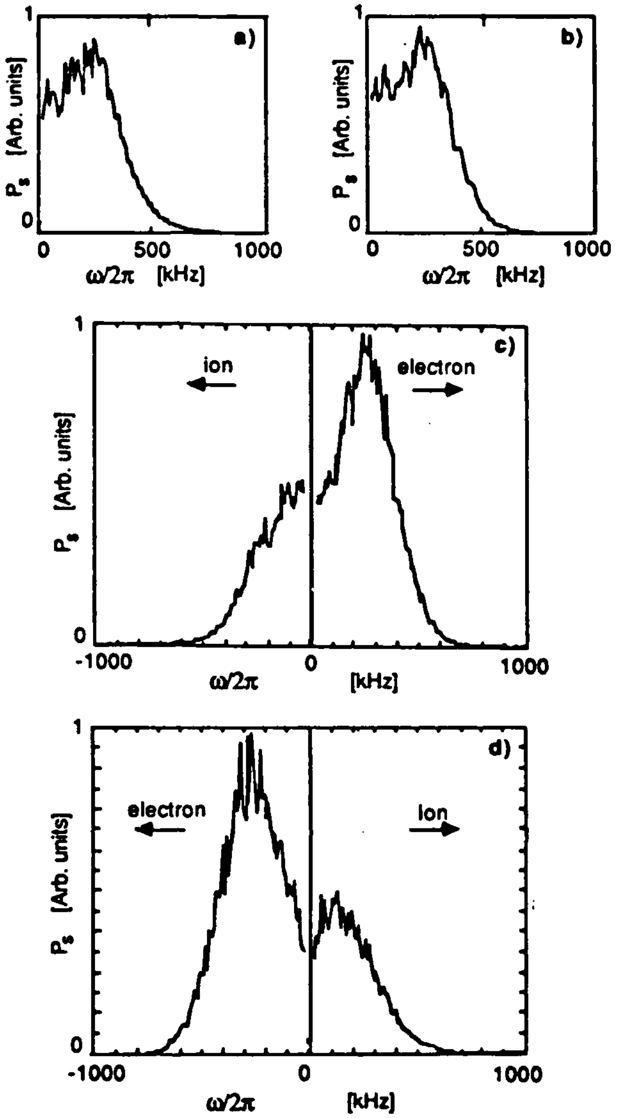


Figure 3. Application of homodyne spectroscopy technique to tokamak low-frequency microturbulence data, (a) homodyne signal from detector 1, (b) homodyne signal from detector 2, (c) frequency spectra using homodyne spectroscopy, $\mathbf{k}_s = \mathbf{k}_o - \mathbf{k}$, and (d) frequency spectra using homodyne spectroscopy, $\mathbf{k}_s = \mathbf{k}_o + \mathbf{k}$.

with poloidal wave vector $k_\Theta = 7$ cm^{-1}. The discharge parameters were $I_p = 400$ kA, $B_T = 28$ kG, and $\bar{n}_e = 2 \times 10^{13}$ cm^{-3}. In Figs. 3(a) and (b), the homodyne power spectra $S_k(\omega/2\pi)$ from the two mixers are illustrated. Each is characterized by a broad spectra which falls-off in power for $\omega/2\pi \geqslant 300$ kHz. The homodyne spectra provide no information regarding wave propagation direction as the $\pm\omega$ components are detected as $|\pm\omega|$. Now however, by implementing the homodyne spectroscopy technique of Tsukishima[1], wave propagation direction information can be ascertained as depicted in Fig. 3(c). Here, the fluctuations are observed to possess a clear peak at $+\omega/2\pi \approx 250 \pm 50$ kHz in the electron diamagnetic drift direction as measured in the laboratory frame of reference. This indicates a fluctuation phase velocity $v_{ph}(=\omega/k_\Theta) \approx 2 \times 10^5$ cm/sec which is in the drift wave region of velocities.[6] A substantial component is also observed at $\omega/2\pi < 0$, corresponding to the ion drift direction. The scattering geometry for this measurement is oriented according to Fig. 2(c), i.e. $\mathbf{k}_s = \mathbf{k}_o - \mathbf{k}$. If we reverse the geometry to that of Fig. 2(d), i.e. $\mathbf{k}_s = \mathbf{k}_o + \mathbf{k}$, one would expect a change in sign from the results of Fig. 3(c), which is indeed the case as shown in Fig. 3(d). The features of the scattered spectra are the same except that $-\omega$ now corresponds to the electron drift direction.

The component of the frequency spectra corresponding to the ion drift direction may result from factors other than a true ion drift feature of the plasma. On the TEXT tokamak, density fluctuations in the limiter shadow and scrape-off regions have been observed to propagate in the ion drift direction due to a strong radial electric field inducing a plasma rotation effect.[7] Another possibility is that the interaction volume may extend to the opposite side of the plasma thereby introducing components at $\pm\omega$ although both represent the same propagation direction. This effect will be described more thoroughly in the ensuing paragraph.

In Fig. 4, the frequency spectra at wave vector $k_\Theta = 7$ cm^{-1} are shown at three spatial positions along a vertical chord through plasma center, scattering volume ($L = \pm 14$ cm) positioned at the plasma top, midplane and bottom. The tokamak discharge conditions were $I_p = 300$ kA, $B_T = 28$ kG, and $\bar{n}_e = 3 \times 10^{13}$ cm^{-3}. At the plasma top (see Fig. 4(a)), the low frequency density fluctuations are observed to be propagating largely in the electron

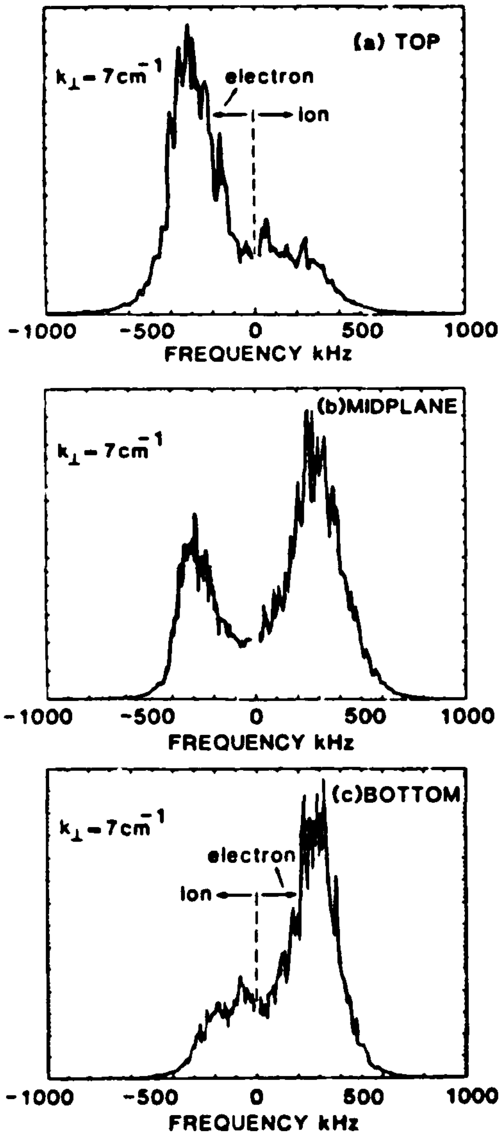


Figure 4 Homodyne spectroscopy frequency spectra for scattering volume centered at plasma (a) top, (b) midplane, and (c) bottom

diamagnetic drift direction with a peak at $\omega/2\pi \simeq -300$ kHz. The fluctuations traveling in the ion drift direction are typically at much lower frequency with no clear peak except at zero frequency. Likewise, at the plasma bottom (see Fig. 4(c)), the fluctuations are again observed to be propagating primarily in the electron drift direction with $\omega/2\pi \simeq +300$ kHz. The change in sign results from the reversal in direction of the fluctuations wave vector k_Θ with respect to the fixed wave vectors k_o and k_s of the incident and scattered radiation (see Fig. 2). When the scattering volume is centered on the midplane (see Fig. 4(b)), fluctuatons are detected both above and below the midplane resulting in peaks at $\pm$ $\omega/2\pi$. The wave propagation direction can only be resolved if the scattering volume is situated completely above or below the midplane. Similar observations have been made at other wavenumbers.

Summary

The wave propagation direction of microturbulence in a tokamak plasma has been accurately measured by application of a new homodyne spectroscopy technique. This method has been used in conjunction with a collective far-infrared laser scattering experiment on TEXT. The low-frequency density fluctuations are observed to propagate primarily in the electron diamagnetic drift direction, however, the broadband spectra also possess an appreciable level of fluctuations traveling in the ion drift direction. Application of the homodyne spectroscopy technique represents an inexpensive and easily implemented alternative to the more technically demanding heterodyne schemes available in the far-infrared.

Acknowledgements

The authors wish to express their appreciation to the Fusion Research Center, University of Texas at Austin, for generously providing machine time and operation for these measurements. This work is supported by the U.S. Department of Energy, Office of Fusion Energy under contract No. DE-AC05-78ET53043 and through subcontract no. UT-1-24120-52019 and the Department of Defense Joint Services Electronics Program under contract no. F49620-82-C-0033.

References

[1] Tsukishima, T., O. Asada, Japan J. Appl. Phys. 17,2059(1978).

[2] Asada, O., A. Inoue, T. Tsukishima, Rev. Sci. Instrum. 51,1308(1980).

[3] Tsukishima, T., et al., Int. Conf. on Infrared and Millimeter Waves, Takarazuka, Japan, 1984.

[4] Park, H., C.X. Yu, W.A. Peebles, N.C. Luhmann, Jr., R. Savage, Rev. Sci. Instrum. 53,1535(1982).

[5] Park, H., D.L. Brower, W.A. Peebles, N.C. Luhmann, Jr., R.L. Savage, Jr., C.X. Yu, Rev. Sci. Instrum. 56,1055(1985).

[6] Brower, D.L., W.A. Peebles, N.C. Luhmann, Jr., R.L. Savage, Jr, Phys. Rev. Lett. 54,689(1985).

[7] Ritz, C.P., R.D. Bengtson, S.J. Levinson, E.J. Powers, Phys. Fluids 27,2956(1984).

Digital complex demodulation applied to interferometry

D. W. Choi,[a] E. J. Powers, Roger D. Bengtson, and G. Joyce[b]

The University of Texas at Austin, Austin, Texas 78712

D. L. Brower, N. C. Luhmann, Jr., and W. A. Peebles

University of California, Los Angeles, California 90024

(Presented on 11 March 1986)

The objective of this paper is to describe the principles of digital complex demodulation, and to summarize its advantages with respect to rapid time response and insensitivity to noise. These advantages are demonstrated by application to interferometry data collected on the Texas experimental tokamak (TEXT).

INTRODUCTION

Interferometry involves the measurement of the line-integrated index of refraction of a medium. In any interferometer a beam of coherent radiation is split, passed along two paths (known as the reference and working arms), and remixed on a detector surface. Changes in the index of refraction of the medium along one of these paths, the working arm, alter the interference of the remixed beams, changing the detector output. *If* the variable medium is a plasma, measurement of the detected signal leads to an estimate of the line-averaged electron density along the working beam path.

In the simplest interferometer, the probe beam associated with the working arm is mixed with a reference beam whose phase is fixed. However, this simple scheme suffers from a severe limitation. It is not possible to distinguish between the cases of increasing and decreasing plasma density. This shortcoming may be overcome by introducing a small constant frequency shift ω_0 between the beams of the working and reference arms. The detector output becomes a phase-modulated sine wave at frequency ω_0 with the relative phase proportional to the line-integrated plasma density. The phase of the detector output is obtained by comparison of this signal with a reference signal oscillating at the frequency ω_0. In an electronic system, the relative phase of the detector signal with respect to the reference signal is obtained via sine and cosine producing phase comparators. Electronic systems are attractive for the following reasons: the comparator output signals require only simple processing, allowing for real-time data reduction and the computer signals need only be sampled at a low rate. There are also several disadvantages: the comparators are sensitive to variations in the amplitude of the detector and reference signals; the comparators may introduce large errors into the phase calculation; and electronic systems have difficulty coping with transient noise. As a result of this noise sensitivity, fringes are often skipped or added. To counter the effects of noise, the time response of the electronic system is lengthened (i.e., bandwidth reduced). Unfortunately, however, this has the undesirable effect of "washing out" rapid fluctuations in plasma density associated with phenomena such as sawtooth oscillations.

The objective of this paper is to describe the principles and application of digital complex demodulation, which is the equivalent of digital heterodyning. The advantages of the digital technique include an insensitivity to transient noise, faster time response, and improved accuracy of the phase measurements.

In the following sections, we describe the principles of complex demodulation, our approach to avoiding 2π phase ambiguities, the effects of noise, and an application to interferometry data collected on the Texas experimental tokamak (TEXT).

I. DIGITAL COMPLEX DEMODULATION

Digital complex demodulation is a digital version of analog heterodyne demodulation and allows the simultaneous measurement of the amplitude and phase modulations of a narrow-band modulated carrier signal as functions of time. This technique has been shown to be effective and flexible compared with other demodulation techniques.[1] In the following, we outline the key ideas of digital complex demodulation.

A sampled narrow-band signal $x(t_n)$ at a "carrier frequency" ω_0, whose amplitude and phase are modulated, may be described by

$$x(t_n) = A(t_n)\cos[\omega_0(t_n) + \theta(t_n)] \, , \tag{1}$$

where $A(t_n)$ and $\theta(t_n)$ are the respective amplitude and phase modulates at the time $t_n = n\Delta t$. Equation (1) may be rewritten in the form

$$x(t_n) = 1/2A(t_n)(\exp[i\omega_0 t_n + i\theta(t_n)] + \exp\{-i[\omega_0 t_n + \theta(t_n)]\}) \, . \tag{2}$$

As in analog demodulation, we must downshift the frequency by an amount $-\omega_0$. In digital demodulation this downshifting is accomplished by simply multiplying (in the computer) the time series data representing Eq. (1) or (2) by a "local oscillator" component $2\exp(-i\omega_0 t_n)$. This results in difference and sum frequency terms, $\omega_0 - \omega_0$ and $\omega_0 + \omega_0$, respectively. By choosing $\omega_0 = \omega_0$ the difference frequency is set to zero,

0034-6748/86/081989-03$01.30

$$x(t_n)2\exp(-i\omega_0 t_n) = A(t_n)\exp[i\theta(t_n)] + \exp\{-i[2\omega_0 t_n + \theta(t_n)]\}. \tag{3}$$

Also, as in analog demodulation, a low-pass filter is applied to the signal. In this case a digital low-pass filter is applied whose passband ω_F is chosen to block the sum frequency component, $\omega_F < \omega_0 + \omega_{l0}$, but to admit the inherent bandwidth of the carrier due to phase modulation, $\omega_F > |d\theta(t)/dt|_{\max}$. A linear digital low-pass filter, MAXFLAT,[2] is selected for this purpose. This filter has a variable passband and a variable transition region in which the filter response drops from 0.95 to 0.05. The amplitude modulation $A(t_n)$ and phase modulation $\theta(t_n)$ are easily recovered from the digital filter output, $y(t_n) = A(t_n)\exp[i\theta(t_n)]$,

$$A(t_n) = |y(t_n)|, \tag{4}$$

$$\theta(t_n) = \arctan\left(\frac{\mathrm{Im}[y(t_n)]}{\mathrm{Re}[y(t_n)]}\right) + 2N\pi, \tag{5}$$

where N is any integer.

II. PHASE AMBIGUITIES AND NOISE

As a result of the multivalued nature of the arctangent function, there is a 2π ambiguity in the phase demodulation; we make two assumptions: (1) The original phase demodulate $\theta(t)$ before sampling is continuous. This assumption is reasonable if $A(t)$ never crosses zero, since the bandlimitedness of the modulated signal $x(t)$ ensures the continuity of $\theta(t)$. (2) The change in the calculated phase modulate between any two consecutive samples is less than π. Thus, if one obtains a true, unambiguous value of $\theta(t_n)$, this condition guarantees an unambiguous value of $\theta(t_{n+1})$ by restricting the range of possible values in the interval

$$\theta(t_n) - \pi < \theta(t_{n+1}) < \theta(t_n) + \pi .$$

The length of this interval is 2π.

We now show that the proper choice of the sampling rate ensures that the phase shift between two samples is less than π. The function $x(t_n)$ is a product of $A(t_n)$ and $\exp\{i[\omega_0 t_n + \theta(t_n)]\}$. According to the convolution theorem, the bandwidth B_x of $x(t_n)$ is greater than the bandwidth of either $A(t_n)$ or $\exp\{i[\omega_0 t_n + \theta(t_n)]\}$. Thus, we have $|\Delta\theta(t_n)/2\pi\Delta t|_{\max} < B_x$. This inequality may be written in terms of the Nyquist frequency $f_N = (2\Delta t)^{-1}$, as $|\Delta\theta(t_n)|_{\max} < \pi(B_x/f_N)$. If the sampling frequency is sufficiently high to satisfy the sampling theorem ($B_x < f_N$), then $|\Delta\theta(t_n)|_{\max} < \pi$, and the problem of 2π phase ambiguities is removed.

A time series to be demodulated may be contaminated by noise which includes the quantization noise at the A/D conversion stage as well as electronic pickup and plasma noise. The multiplication of $x(t_n)$ by $\exp(-i\omega_{l0} t_n)$ translates the frequency contents of $x(t_n)$ by $-\omega_{l0}$ in the frequency domain, but has no effect on the signal-to-noise ratio. However, low-pass filtering increases the signal-to-noise ratio when the signal $A(t_n)\exp[i\theta(t_n)]$ is within the passband. If the noise is assumed to be white up to the Nyquist frequency f_N and the low-pass cutoff frequency is f_F, then the gain of the signal-to-noise ratio is f_N/f_F. In view of the fact that the low-pass filtering is being done digitally, it is a relatively simple matter to adjust the low-pass cutoff frequency to enhance the signal-to-noise ratio.

The error in the calculated phase modulate $\theta(t_n)$ depends on the signal-to-noise ratio $\mathrm{SNR}(t_n)$ of the filtered signal, $y(t_n)$. To avoid errors in the calculated phase due to a momentary low signal-to-noise ratio, a threshold value of $y(t_n)$, $y_{\min}$, is defined which depends on the amplitude of the noise. When the signal amplitude drops so that $|y(t_n)| < y_{\min}$, the calculated phase modulate $\theta(t_n)$ is ignored and the previously calculated phase modulate is carried over. If the noise is a transient, it is reasonable to neglect the phase measurement and assume the previous value of the phase as the current value.

In most interferometers the shift frequency ω_0 (referred to as the carrier frequency in this section) is generally not known with perfect precision and may also be unstable. The application of a constant digital local oscillator frequency shift, ω_{l0}, which is not precisely equal to ω_0 will therefore result in an accumulated phase error which increases linearly in time, $(\omega_0 - \omega_{l0})t_n$. Other sources of phase error include phase distortions introduced by the low-pass filter and deviations from a uniform sampling rate. To cancel these errors a reference signal $x_r(t)$, which equals the carrier signal without the plasma phase modulation $\theta(t_n)$, is also sampled,

$$x_r(t_n) = A_r(t_n)\exp(i\omega_0 t_n) .$$

Digital complex demodulation is then applied to this signal as well, to yield a reference phase modulate $\theta_r(t_n)$. The desired phase modulate $\theta_p(t_n)$, due to the plasma is simply the difference of these calculated phases,

$$\theta_p(t_n) = \theta(t_n) - \theta_r(t_n) .$$

III. APPLICATION TO INTERFEROMETER DATA

We have applied digital complex demodulation to data from a 1.2-mm far-infrared laser interferometer on TEXT. The detector and reference signals had carrier frequencies of 80 kHz. These signals were sampled at 200 kHz for 325 ms. The digital low-pass filter, MAXFLAT, was set to a bandpass frequency of 20% of the carrier, resulting in 200 filter

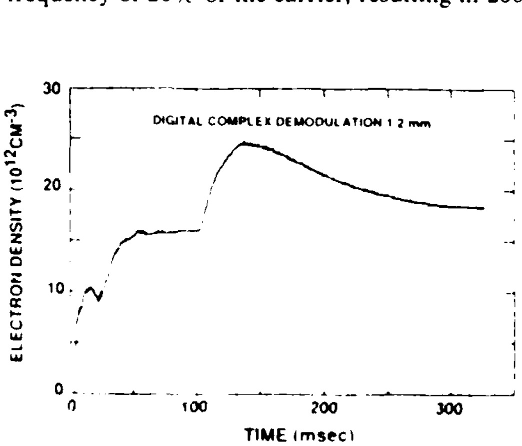


FIG. 1. Plasma density vs time deduced from digitally demodulated interferometer data taken on the TEXT.

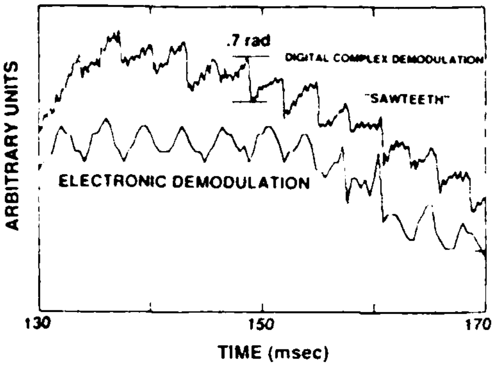


FIG. 2 Comparison of sawtooth oscillations observed by electronic demodulation and digital complex demodulation.

coefficients which characterize the filter impulse response. The amplitude threshold y_{min} for phase interpolation was set to 5% of the maximum signal level when free of noise. Figure 1 shows the calculated density for a typical shot. Note that the digital complex demodulation result faithfully tracks the density including the dip during the current rise at ~25 ms.

Figure 2 compares data reduced via complex demodulation with data from a standard interferometer utilizing electronic phase comparators. Both data are from a portion of a discharge during which "sawtooth" activity was occurring. The digitally produced traces display the sawtooth shape (generally observed via other faster diagnostics) while the electronically reduced data resemble a sine wave, due to the smoothing resulting from a derated time response, the latter being necessary to avoid skipped fringes due to noise effects. Note also that the time response of the complex demodulation approach is sufficiently fast to recover the Mirnov oscillations which are superimposed on the sawteeth oscillations.

These examples demonstrate that digital complex demodulation possesses a number of advantages with respect to insensitivity to noise, faster time response, and phase measurement accuracy. Also, the digital approach is inherently flexible, allowing relatively easy adjustment of the low-pass filter cutoff frequency and the amplitude threshold y_{min}. The limitations of digital complex demodulation stem from the memory and computation requirements. We typically sample the 80-kHz detector and reference signals at 200 kHz for 325 ms, yielding 130 000 samples. In addition, a 16-kHz digital low-pass filter (20%) contains 200 coefficients representing the filter impulse response. The resulting convolution integrals require 13 min on a VAX 11/780 with a nonoptimized code. The use of an array processor would probably reduce this time requirement considerably.

ACKNOWLEDGMENTS

This work is supported by U. S. DOE Contract DE-AC05-78ET53043. The phase-sensitive complex demodulation technique was developed under the auspices of the DOD Joint Services Electronics Program Contract F49620-82-C-0033.

[a] Present address: AT&T Bell Laboratories, N. Andover, Massachusetts.

[b] Present address: GTE, Inc., Waltham, Massachusetts.

[1] P. Y. Ktonas and N. Papp, Signal Process. 2, 373 (1980).

[2] J. F. Kaiser and W. A. Reed, Rev. Sci. Instrum. 48, 1447 (1977).

Resolving the propagation direction of tokamak microturbulence via homodyne spectroscopy

D. L. Brower, W. A. Peebles, and N. C. Luhmann, Jr.

University of California, Los Angeles, California 90024

Ch. P. Ritz and E. J. Powers

University of Texas, Austin, Texas 78712

(Presented on 11 March 1986)

A new homodyne spectroscopy technique [T. Tsukishima and O. Asada, Jpn. J. Appl. Phys. **17**, 2059 (1978)] has been applied to tokamak microturbulence measurements in order to resolve the frequency spectra and wave propagation direction of low-frequency density fluctuations. Application of this method provides a high-resolution, inexpensive, and easily implemented alternative to the more technically demanding heterodyne detection schemes typically available. Comparison of heterodyne and the new homodyne spectroscopy results will be made.

INTRODUCTION

Laser and millimeter-wave scattering techniques are commonly used to study the space-time statistics of electron density fluctuations in tokamak and other plasmas. Of particular importance is the determination of the direction of propagation of the fluctuations. Since the scattering geometry fixes the direction of the scattering wave vector **k**, the direction of propagation information is carried by the sign of the fluctuation frequency ω. The fact that waves may be propagating both parallel and antiparallel to **k** is manifested by the presence of blue and red sidebands centered around the incident wave frequency ω_0.

To recover the propagation direction information contained in the blue and red sidebands, heterodyne detection techniques are typically employed. This approach requires two coherent sources, with a frequency difference $\Delta\omega$, to be utilized as the incident and local oscillator beams. After mixing, the frequency range of the resultant signal is $\Delta\omega \pm \omega$, where ω is the frequency associated with the plasma fluctuations. As long as $\Delta\omega \gg \omega$, the blue/red sidebands may be resolved. In contrast, if a classical homodyne approach is used, $\Delta\omega = 0$, and it is no longer possible to unambiguously determine the wave propagation direction.

Realization of a heterodyne detection system in the far infrared is expensive (e.g., two lasers) and technically n[illegible] trivial [e.g., intermediate frequency (IF) stability [illegible] Implementation of a rotating grating to frequency shift [illegible] the source beam is a feasible alternative, alth[illegible] quency offset is limited to roughly Δ[illegible] 150 kH[illegible] cient for microturbulence measurements w[illegible] are observed up to 1 MHz). A considera[illegible] expensive method proposed by Tsukis[illegible] mits resolution of the wave propaga[illegible] analysis of homodyne signals. This new [illegible] copy technique utilizes advanced [illegible] while requiring minor modificati[illegible] tem.

I. EXPERIMENTAL TECHNIQUE AND APPARATUS

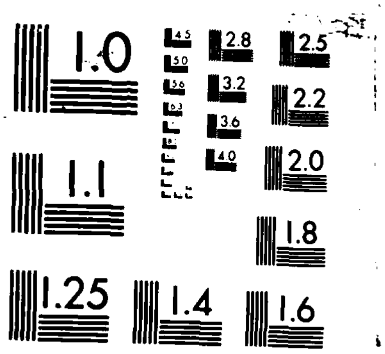
The IF output of the scattered signa[illegible] real quantity described by

$$v(t) = \mathrm{Re}\left(\int \frac{d\omega}{2\pi} N(k \ldots\right.$$

where $k_s = k_s - k_i$ is the wave vec[illegible] ation, and $N(k, \omega)$ [illegible] fluctuation level. The sign of th[illegible] sents the propagation dire[illegible] frame of reference and [illegible] However, from the rea[illegible] the reconstruction of th[illegible] metry property N [illegible] wave-propaga[illegible] new homodyne spe[illegible] shima and Asa[illegible] [illegible] and [illegible] tw[illegible] shifted b[illegible] wh[illegible]

[illegible]

AD-A177 348 JOINT SERVICES ELECTRONICS PROGRAM: BASIC RESEARCH IN 2/2
ELECRRONICS (JSEP) (U) TEXAS UNIV AT AUSTIN
ELECTRONICS RESEARCH CENTER E J POWERS 31 DEC 86
UNCLASSIFIED AFOSR-TR-87-0091-APP F49620-86-C-0045 F/G 9/5 NL



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963 A

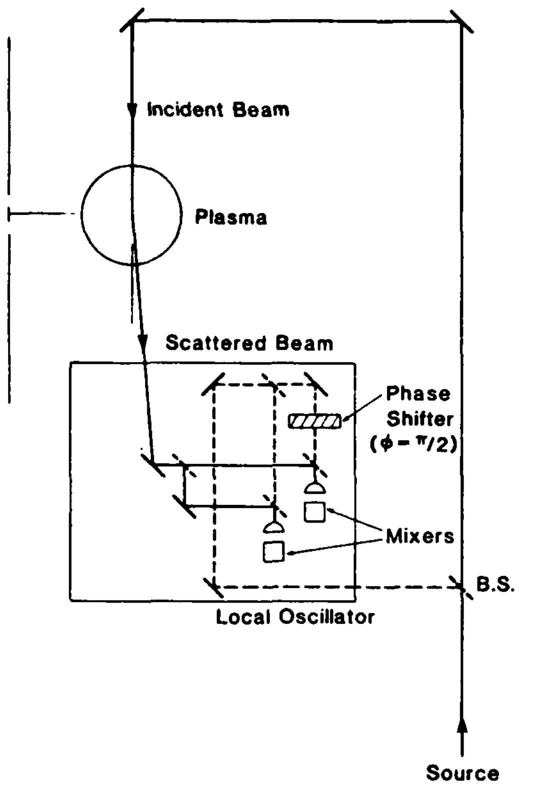


FIG. 1. Experimental arrangement for homodyne spectroscopy measurement.

transform algorithm, the frequency spectrum was obtained with a resolution of $\Delta\omega/2\pi = 1/T' \simeq 8$ kHz. Mixer and amplifier noise contributions could be subtracted from the autopower spectra although signal-to-noise levels were sufficiently large so as to make it unnecessary.

A schematic of the experimental arrangement employed for application of the new homodyne spectroscopy technique to collective far-infrared scattering is shown in Fig. 1. Detailed information on the scattering system is described by Park *et al.*[2]

At a particular wave vector $k_\perp$, the scattered radiation beam is divided equally into two components which are coupled into the detectors by 50% reflectivity beam splitters. Similarly, the local oscillator beam is equally divided to provide rf drive for each mixer. In one leg of the local oscillator beam a phase shifter is inserted. This phase shifter consists of a piece of high-density polyethylene (excellent transmission characteristics at 245 GHz) mounted on a rotation stage. By tuning the rotation angle, the path of the LO beam through the polyethylene is altered, thereby changing its phase. This phase shifter is tuned such that there is a 90° phase difference between the detected signals in the two channels. The signal from each detector is then amplified and digitized so that the cross- and auto-power spectra may be computed.

II. EXPERIMENTAL RESULTS

The new homodyne spectroscopy technique will now be applied to density fluctuation measurements in a high-temperature tokamak plasma. Microturbulence (low-frequency density fluctuations) is driven by the free energy associated with plasma inhomogeneities such as density and temperature gradients. For drift-wave-type fluctuations, it is predicted that the turbulence will exhibit a phase velocity of order $\mathbf{v}_{De} = \omega/\mathbf{k}_\Theta = (k_B T_e/eB_T n_e)\nabla n_e \times \mathbf{B}_T/|B_T|$, where $\mathbf{v}_{De}$ is the electron diamagnetic drift velocity, k_Θ and ω are the poloidal wave vector and frequency of the fluctuation, B_T is the toroidal magnetic field, T_e is the electron temperature, and n_e is the electron density. The scattering system is positioned such that the incident beam impinges upon the plasma from the top of the torus (see Fig. 1) along a vertical chord at the major radius $R = 1m$. This provides for scattering from fluctuations with a poloidal wave vector. $\mathbf{k}_\Theta$. Depending upon the orientation of the collection optics with respect to the incident beam, the wave-number matching condition (momentum conservation) gives $\mathbf{k}_s = \mathbf{k}_0 \pm \mathbf{k}$, where $\mathbf{k}_s$, $\mathbf{k}_0$, and $\mathbf{k}$ are the wave vectors of the scattered beam, incident beam, and plasma fluctuation, respectively. Similarly, energy conservation dictates $\omega_s = \omega_0 \pm \omega$, where the subscripts have the same meaning as above. In addition, it is important to note that for a specific scattering geometry, a sign change ($\pm \mathbf{k}$) will occur in the detected signal when one switches from the plasma top to bottom for fluctuations traveling in a particular poloidal direction.

Experimental results from the Texas Experimental Tokamak (TEXT) for a scattering volume located at the plasma bottom with poloidal wave vector $k_\Theta = 7$ cm^{-1} are shown in Fig. 2. The discharge parameters were $I_p = 400$ kA, $B_T = 28$ kG, and $\bar{n}_e = 2 \times 10^{13}$ cm^{-3}. In Fig. 2(a), the

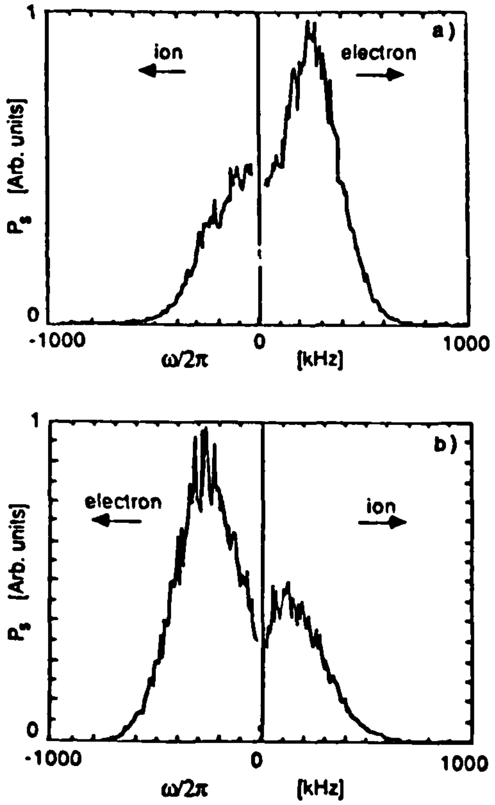


FIG. 2. Application of homodyne spectroscopy technique to tokamak low-frequency microturbulence data; (a) $\mathbf{k}_s = \mathbf{k}_0 - \mathbf{k}$, and (b) $\mathbf{k}_s = \mathbf{k}_0 + \mathbf{k}$.

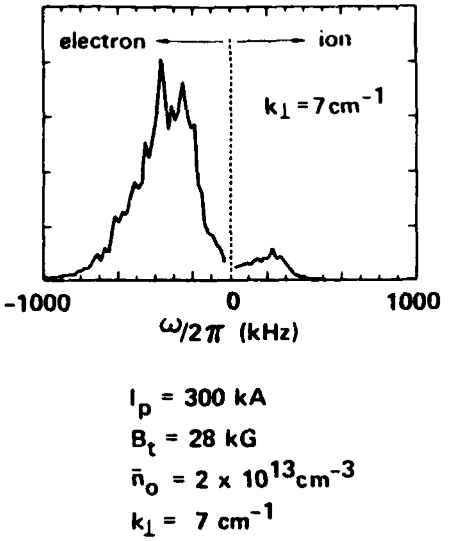


FIG. 3. Heterodyne detection frequency spectra.

fluctuations are observed to possess a clear peak at $+\omega/2\pi \simeq 275 \pm 50$ kHz in the electron diamagnetic drift direction as measured in the laboratory frame of reference. This indicates a fluctuation phase velocity $v_{ph} (= \omega/k_\Theta) \simeq 2\times10^5$ cm/s, which is in the drift wave region of velocities.[3] A substantial component is also observed at $\omega/2\pi < 0$, corresponding to the ion drift direction. If we reverse the geometry from that of Fig. 1, one would expect to see a change in sign, which is indeed the case as shown in Fig. 2(b). The features of the scattered spectra are the same except that $-\omega$ now corresponds to the electron drift direction.

The component of the frequency spectra corresponding to the ion drift direction may result from factors other than a true ion drift feature of the plasma. On the TEXT tokamak, density fluctuations in the limiter shadow and scrape-off regions have been observed to propagate in the ion drift direction because of a strong radial electric field inducing a plasma rotation effect.

Heterodyne scattering results for $k_\Theta = 7\,\mathrm{cm}^{-1}$ under plasma conditions of $I_p = 300$ kA, $B_T = 26$ kG, and $\bar n_e = 2\times10^{11}\,\mathrm{cm}^{-3}$ are shown in Fig. 3 for scattering volume positioned at the plasma bottom. The frequency difference between the local oscillator and incident beams (IF) is $\Delta\omega/2\pi \simeq 1100$ kHz. By subtracting the IF frequency contribution from the scattered spectra and defining it as the new zero, one can plot the fluctuation spectrum. The resolution near zero frequency is limited by the IF bandwidth which is ± 20 kHz. As with the homodyne spectroscopy technique, moving the scattering volume from plasma top to bottom results in a change of sign for the fluctuation spectrum. In comparing the results of Figs. 2 and 3, it is very evident that both the new homodyne spectroscopy and heterodyne detection methods produce similar spectra. Any differences can be attributed to plasma discharge conditions. The homodyne spectroscopy method provides improved resolution near zero frequency (± 8 kHz).

III. SUMMARY

The wave propagation direction of microturbulence in a tokamak plasma is measured by application of a new homodyne spectroscopy technique. The accuracy of this method is established by comparison with results from a heterodyne scattering system which are similar. Both techniques show the low-frequency density fluctuations to be propagating primarily in the electron diamagnetic drift direction: however, the broadband spectra also possess an appreciable level of fluctuations traveling in the ion drift direction. Application of the homodyne spectroscopy technique represents an inexpensive and easily implemented alternative to the more technically demanding heterodyne schemes available in the far infrared.

ACKNOWLEDGMENTS

The authors wish to express their appreciation to the Fusion Research Center, University of Texas at Austin, for generously providing machine time and operation for these measurements. This work is supported by the U.S. DOE, Office of Fusion Energy under contract No. DE-AC05-78ET53043 through subcontract No. UT-1-24120-52019 and the Department of Defense Joint Services Electronics Program under Contract No. F49620-82-C-0033.

[1] T. Tsukishima and O. Asada, Jpn. J. Appl. Phys. 17, 2059 (1978).
[2] H. Park, D. L. Brower, W. A. Peebles, N. C. Luhmann, Jr., R. L. Savage, Jr., and C. X. Yu, Rev. Sci. Instrum. 56, 1055 (1985).
[3] D. L. Brower, W. A. Peebles, N. C. Luhmann, Jr., and R. L. Savage, Jr. Phys. Rev. Lett. 54, 689 (1985).

DIGITAL ESTIMATION OF LINEAR/QUADRATIC TRANSFER FUNCTIONS WITH A GENERAL RANDOM INPUT

Kyoung Il Kim and Edward J. Powers
Department of Electrical and Computer Engineering
and Electronics Research Center
The University of Texas at Austin
Austin, Texas 78712 USA

Presented at

IEEE International Conference on Acoustics,
Speech, and Signal Processing 86

Tokyo, Japan

April 8-11, 1986

DIGITAL ESTIMATION OF LINEAR/QUADRATIC TRANSFER FUNCTIONS WITH A GENERAL RANDOM INPUT

Kyoung Il Kim and Edward J. Powers

Department of Electrical and Computer Engineering
and Electronics Research Center
The University of Texas at Austin
Austin, Texas 78712, U.S.A.

ABSTRACT

A new digital method of estimating linear and quadratic transfer functions of a quadratic system with a general random input is presented. The feasibility of the technique is demonstrated by analyzing simulated data. It is also shown that considerable error occurs in estimating the transfer functions based on a Gaussian input assumption, when in fact the input is non-Gaussian.

1. INTRODUCTION

A difficulty encountered when one attempts to apply the Volterra functional series to nonlinear problems is the measurement of the Volterra kernels [see, e.g. 1]. So far, a fundamental assumption underlying many approaches involves the fact that the "input" is assumed to be a stationary random process which possesses Gaussian statistics, an assumption which allows a substantial simplification of the relevant mathematics. In many practical cases, however, the input excitation is not under the direct control of the experimentalist which precludes the use of the so-called "probing" method, thus one must use the measured input data as they are observed.

It has been shown in [2] that, for a zero-mean Gaussian input, expressions for the linear and quadratic transfer functions are respectively given in terms of various spectral moments up to third order (i.e., the bispectrum). However, when a general random input is applied to the system, it is extremely difficult to find such closed form expressions for the transfer functions. Katzenelson and Gould [3] described an iterative method to solve this problem, and Eykhoff [4] considered a discrete time version. Also Ritz and Powers [5] showed that when the input is weakly non-Gaussian, transfer functions of a quadratic system can be estimated by an iterative approach in the discrete frequency domain.

It is the purpose of this paper to describe a new digital method of processing input and output signals in order to quantitatively measure the linear and the quadratic transfer functions even when we cannot assume a particular characteristic of the input. In the next section, it is shown that the linear and the quadratic transfer functions can be evaluated in the discrete frequency domain by solving matrix equations. In section 3, a known system is analyzed to show the feasibility of the analysis results. In addition, the results are compared with the transfer functions estimated by the "Gaussian input method" in order to illustrate the deleterious effects of assuming a Gaussian input when in reality it is not.

2. ESTIMATION OF TRANSFER FUNCTIONS

Since we will concentrate on frequency domain analysis and the objective is to find a digital method that can be practically implemented, we will start from the input-output relationship in the discrete frequency domain. In the following, we assume the unknown nonlinear system is of second order (i.e., quadratic), thus, higher order terms may be safely neglected. Then the model to be studied can be expressed as follows;

$$Y(f_m) = H_1(f_m)X(f_m) + \sum_{\substack{k\;l\\k+l=m}}\sum H_2(f_k,f_l)X(f_k)X(f_l) \quad (1)$$

where $X(f_m)$ and $Y(f_m)$ respectively represent the discrete Fourier transforms (DFT's) for a finite number (N) of observations of the input and the output signals of the nonlinear system described by Volterra series up to second order. On the other hand, $H_1(f_m)$ and $H_2(f_k,f_l)$ are linear and quadratic transfer functions which are given by the Fourier transforms of Volterra kernels at a discrete set of frequencies $\{f_n=n/N;\ n=-(N-1)/2,\ldots,-1,0,1,\ldots,N/2\}$. It will be assumed that the quadratic transfer function $H_2(f_k,f_l)$ is a symmetric function of its arguments, i.e., $H_2(f_k,f_l)=H_2(f_l,f_k)$.

Determination of the linear and quadratic transfer functions in terms of the input and output characteristics can be carried out by solving the following set of equations which are obtained by multiplying (1) by $X^*(f_m)$ and $X^*(f_i)X^*(f_j)$, respectively, and then taking an expected value of each side.

$$E[X^*(f_m)Y(f_m)] = H_1(f_m)E[|X(f_m)|^2] + \sum_{\substack{k\;l\\k+l=m}}\sum H_2(f_k,f_l)E[X^*(f_m)X(f_k)X(f_l)] \quad (2)$$

$$E[X^*(f_i)X^*(f_j)Y(f_m)] = H_1(f_m)E[X^*(f_i)X^*(f_j)X(f_m)] + \sum_{k}\sum_{\substack{l \\ k+l=m}} H_2(f_k,f_l)E[X^*(f_i)X^*(f_j)X(f_k)X(f_l)] \quad (3)$$

Note that (3) is meaningful only when $f_m=f_k+f_l=f_i+f_j$ because of the properties of higher order spectral moments [6].

When the system input is zero-mean Gaussian, the terms containing the third order moment of the input in (2) and (3) vanish. In this case, the linear and the quadratic transfer functions can be determined separately and expressed by the various spectra up to third order [2]. More specifically, they are given by

$$H_1(f_m) = \frac{E[X^*(f_m)Y(f_m)]}{E[|X(f_m)|^2]}, \quad (4)$$

$$H_2(f_k,f_l) = \frac{E[X^*(f_k)X^*(f_l)Y(f_k+f_l)]}{2E[|X(f_k)|^2]E[|X(f_l)|^2]}, \quad k+l=0. \quad (5)$$

However, for the case of a general input, we have to solve (2) and (3) simultaneously so that it is extremely difficult to find the closed form solutions like (4) and (5).

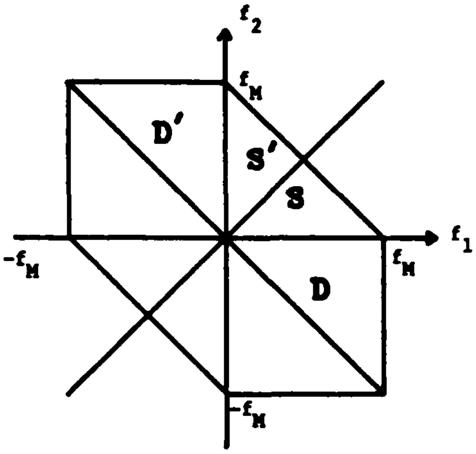


Fig. 1 Two dimensional frequency domain.

Next, we will describe a method of solving (2) and (3) which can be digitally implemented. Due to the symmetricity assumed for $H_2(f_k,f_l)$, we can express the output only in terms of the portions of the quadratic transfer function which are in the sum and difference interaction regions of the two dimensional frequency domain (regions S and D in Fig. 1). Using this fact and expanding the summation term in (1), we can rewrite (1) in the following vector form;

$$Y(f_m) = \underline{H}^t\underline{X} = \underline{X}^t\underline{H} \quad (6)$$

where t denotes transposition, and

$$\underline{H}^t = \begin{cases} [H_1(f_m),\ 2H_2(f_{\frac{m+1}{2}},f_{\frac{m-1}{2}}),\ldots,2H_2(f_m,f_0),\ldots,2H_2(f_M,f_{m-M})], \text{ for m odd} \\ [H_1(f_m),\ H_2(f_{\frac{m}{2}},f_{\frac{m}{2}})\ ,2H_2(f_{\frac{m}{2}+1},f_{\frac{m}{2}-1}),\ldots,2H_2(f_m,f_0),\ldots,2H_2(f_M,f_{m-M})], \text{ for m even,} \end{cases}$$

$$\underline{X}^t = \begin{cases} [X(f_m),\ X(f_{\frac{m+1}{2}})X(f_{\frac{m-1}{2}}),\ldots,X(f_m)X(f_0),\ldots,X(f_M)X(f_{m-M})], \text{ for m odd} \\ [X(f_m),\ X(f_{\frac{m}{2}})X(f_{\frac{m}{2}})\ ,X(f_{\frac{m}{2}+1})X(f_{\frac{m}{2}-1}),\ldots,X(f_m)X(f_0),\ldots,X(f_M)X(f_{m-M})], \text{ for m even.} \end{cases}$$

In (6), f_M signifies the Nyquist frequency associated with the sampling of the input and output signals. Then solving simultaneously (2) and (3) is equivalent to solving the following matrix equation;

$$E[\underline{X}^*Y(f_m)] = E[\underline{X}^*\underline{X}^t]\underline{H}. \quad (7)$$

Equation (7) is linear in the transfer function vector $\underline{H}$, and so $\underline{H}$ is given by

$$\underline{H} = \{E[\underline{X}^*\underline{X}^t]\}^{-1}E[\underline{X}^*Y(f_m)] \quad (8)$$

if $E[\underline{X}^*\underline{X}^t]$ is not singular. The solution given by (8) can also be considered as a result of multivariate linear regression analysis and thus the transfer functions obtained in this manner are optimum in the mean square sense when there is any additive noise present in the output. Note that $E[\underline{X}^*\underline{X}^t]$ is a Hermitian matrix consisting of various spectral moments of the input signal. Except for the first element, the first row and column of this matrix represent the bispectrum while the first element is the auto power spectrum of the input. The remaining elements are fourth order spectral moments of the input. The size of the matrix $E[\underline{X}^*\underline{X}^t]$ to be inverted depends on the number of data points (N=2M) taken for the DFT's and the frequency index m. For example, it is a $(M+2-m/2)\times(M+2-m/2)$ matrix if m is even, and it is $(M+2-\frac{m+1}{2})\times(M+2-\frac{m+1}{2})$ when m is odd. Therefore the size decreases as m increases. The largest one is $(M+2)\times(M+2)$ when m=0, and the smallest one is $(\frac{M}{2}+2)\times(\frac{M}{2}+2)$ when m=M-1 or M.

If the input is zero-mean Gaussian, it can be shown that the matrix $E[\underline{X}\,\underline{X}^t]$ becomes a diagonal matrix and thus the transfer functions given by (8) have the same expressions for the linear and quadratic transfer functions as those given by (4) and (5). Consequently the solution given by (8) is a general one for the transfer functions of a quadratically nonlinear system with an arbitrary random input, i.e., it includes the zero-mean Gaussian input as a special case.

3. ANALYSIS OF SIMULATED DATA

FORTRAN programs have been written on the basis of the analysis results described in the previous section and tested by analyzing simulated data generated by a known system.

The known system is formed as follows;

$$y(t) = -0.64x(t) + x(t-2) + 0.9x^2(t) + x^2(t-1). \quad (9)$$

Therefore the transfer functions are given by

$$H_1(f) = -0.64 + e^{-i4\pi f} \quad (10)$$

$$H_2(f_1,f_2) = 0.9 + e^{-i2\pi(f_1+f_2)}. \quad (11)$$

These actual transfer functions of the given system are shown in Fig. 2 and Fig. 3.

The approach developed in this paper has been tested by applying a zero-mean exponentially distributed input signal which is generated by IMSL routine. For both the input and output signals, 128000 sample points of data record have been generated, and they were divided up into 2000 segments of 64 data points each.

For a quantitative measure of the quality of the estimates, the normalized mean square errors involved in the computed transfer functions are defined as follows;

$$\mathrm{MSE}_1 = \frac{1}{M}\sum_{m=1}^{M}\frac{|H_1(f_m)-\hat{H}_1(f_m)|^2}{|H_1(f_m)|^2} \quad (12)$$

$$\mathrm{MSE}_q = \frac{1}{N_q}\sum_{k}\sum_{l}\frac{|H_2(f_k,f_l)-\hat{H}_2(f_k,f_l)|^2}{|H_2(f_k,f_l)|^2} \quad (13)$$

$$(f_k,f_l)\in SUD$$

where ^ signifies an estimated quantity, and N_q is the number of frequency points in the regions S and D.

By using the input and output signals generated above, the estimation of the linear and quadratic transfer functions have been carried out first by the Gaussian input method (Eqs. (4) and (5)), and then the results are compared with the estimates that have been obtained by the general input method developed in this paper (Eq. (8)). Fig. 4 and Fig. 5, respectively, show the linear and quadratic transfer functions estimated by the two methods. In this case, the mean square errors have been calculated as follows; by the Gaussian input method, $\mathrm{MSE}_1=9.5$, $\mathrm{MSE}_q=4.0$, and by the general input method, $\mathrm{MSE}_1=0.06$, $\mathrm{MSE}_q=0.04$. Comparing the mean square error values as well as the plots, one can clearly see the obvious differences between the transfer functions estimated by the two methods. This indicates that, in order to obtain useful estimates, one must use the new method developed in this paper when there is not any good a priori knowledge about the input signal statistics or when the input is not a zero-mean Gaussian signal.

4. CONCLUSION

In this paper, we have discussed the problem of estimating system transfer functions by processing random input and output data, and described a new digital estimation method which can be successfully applied to the modeling of a quadratically nonlinear system excited by a non-Gaussian input. Analyzing simulated data, we have also demonstrated the feasibility of the technique. Considering the computation time and the size of the processor memory available for practical applications, it is desirable that the number of data points in each segment of record data be not too large since it determines the dimensions of the matrices that must be inverted in order to solve the matrix equations.

ACKNOWLEDGEMENT

This work was supported by the Department of Defense's Joint Services Electronics Program through the Air Force Office of Scientific Research Contract F49620-82-C-0033.

REFERENCES

[1] G. Hung and L. Stark, "The Kernel Identification Method - Review of Theory, Calculation, Application, and Interpretation," Math Biosci., 37, pp. 135-190, 1977.

[2] L.J. Tick, "The Estimation of Transfer Functions of Quadratic Systems," Technometrics, Vol. 3, No. 4, pp. 563-567, Nov. 1961.

[3] J. Katzenelson and L.A. Gould, "The Design of Nonlinear Filters and Control Systems, Part I," Inf. and Control 5, pp. 108-143, 1962.

[4] P. Eykhoff, "Some Fundamental Aspects of Process Parameter Estimation," IEEE Trans. Auto. Cont., AC-8, pp. 347-357, Oct. 1963.

[5] Ch.P. Ritz and E.J. Powers, "Estimation of Nonlinear Transfer Functions for Fully Developed Turbulence," FRCR No. 273, Fusion Research Center, The University of Texas at Austin; accepted for publication in Physica D, 'Nonlinear Phenomena', 1985.

[6] D.R. Brillinger and M. Rosenblatt, "Asymptotic Theory of Estimates of K-th Order Spectra," in Spectral Analysis of Time Series Ed. by B. Harris, pp. 189-232, Wiley, New York, 1967.

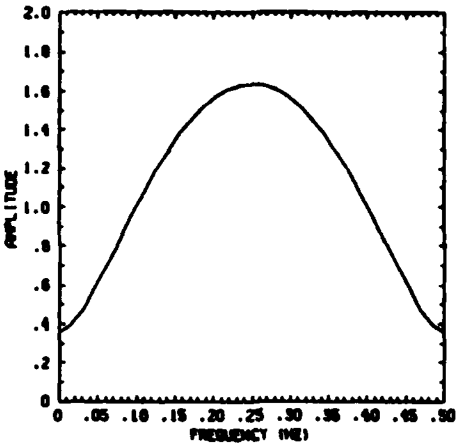


Fig. 2 Amplitude of the actual linear transfer function.

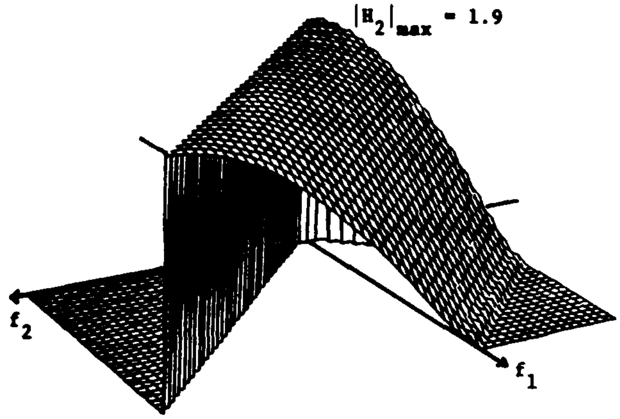


Fig. 3 Amplitude of the actual quadratic transfer function.

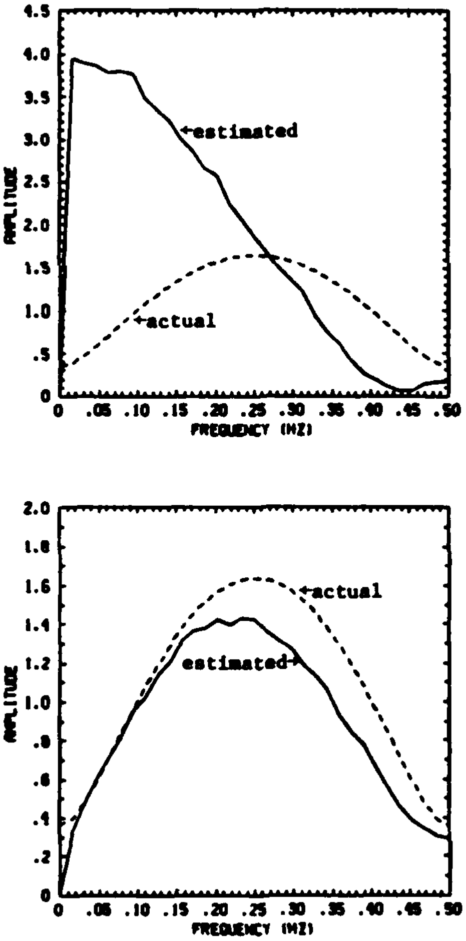


Fig. 4 Linear transfer function estimates: Top, Gaussian input assumption; bottom, method of this paper.

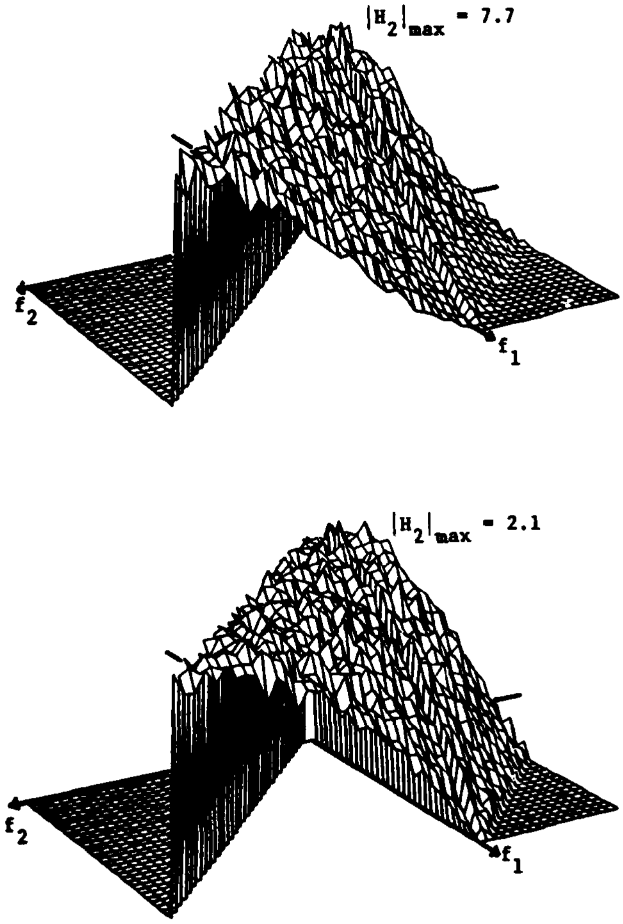


Fig. 5 Quadratic transfer function estimates: Top, Gaussian input assumption; bottom, method of this paper.

Correspondence

Identification of Nonlinear Systems in the Walsh Sequency Domain

KYOUNG IL KIM, JAE YOUNG HONG, MEMBER, IEEE, AND EDWARD J. POWERS, JR., FELLOW, IEEE

Abstract—**A method of measuring transfer functions for nonlinear dyadic-invariant (NLDI) systems described by Barrett's orthogonal model is discussed. In particular, this paper develops the expressions for the transfer functions up to third order, and the results show that the transfer functions can be obtained from the raw input and output data by computing the appropriate Walsh sequency power spectra.**

Key Words—**Nonlinear systems, Walsh sequency domain, transfer functions.**

Index Code—**P2d, K2d.**

I. Introduction

Ever since J. L. Walsh published a complete set of orthogonal functions [1], Walsh functions have been one of the most important examples of nonsinusoidal functions in engineering applications. The computational efficiency of the fast Walsh transform (FWT) and the feasibility of its software and hardware implementation provide the incentive to find useful applications. On the other hand, motivated by the fact that the Walsh functions form the natural basis for representing dyadic-invariant systems just as trigonometric functions do for time-invariant systems, some studies have been carried out to find useful properties of the dyadic-invariant systems. For example, an optimal linear dyadic-invariant (LDI) system was developed [2], the problem of modeling a multiple input/output LDI system in terms of sequency transfer functions was considered [3], and Walsh series expansions were utilized to develop a method of measuring the kernels up to order two in Wiener's nonlinear system model [4].

In this study, we present a method of measuring kernels for the identification of a class of nonlinear systems which can be described by Barrett's model. Since we let the model have dyadic convolution operations, we refer to these systems as nonlinear dyadic-invariant (NLDI) systems. In particular, we will obtain the expressions for kernels up to third order using Walsh transform techniques. Thus this study can be considered as an extension (through a different approach) of the results in [4], and also a parallel study to that which has been carried out by Hong *et al.* in the Fourier frequency domain [5].

By orthogonalizing the Volterra functional series for a white Gaussian input signal, Wiener provided a representation for nonlinear systems [6]. In many practical situations, however, the input to the system is usually nonwhite. In such cases, Barrett's orthogonal model [7] for nonlinear systems is a useful one since it is valid for Gaussian inputs with arbitrary spectral densities.

II. Development

Now we consider the space of real nonnegative-coordinate signals. Thus the input $x(t)$ is defined for $t \geq 0$. Following Barrett's orthogonal representation, we can express the output $y(t)$ of an unknown NLDI system for a zero-mean dyadic stationary [3], [8] Gaussian input process $x(t)$ as follows:

$$\begin{aligned} y(t) = h_0 &+ \int_0^\infty h_1(t_1)x(t \oplus t_1)\, dt_1 \\ &+ \int_0^\infty \int_0^\infty h_2(t_1, t_2)\{x(t \oplus t_1)x(t \oplus t_2) \\ &- E[x(t \oplus t_1)x(t \oplus t_2)]\}\, dt_1\, dt_2 \\ &+ \int_0^\infty \int_0^\infty \int_0^\infty h_3(t_1, t_2, t_3)\{x(t \oplus t_1)x(t \oplus t_2)x(t \oplus t_3) \\ &- x(t \oplus t_1)E[x(t \oplus t_2)x(t \oplus t_3)] \\ &- x(t \oplus t_2)E[x(t \oplus t_3)x(t \oplus t_1)] \\ &- x(t \oplus t_3)E[x(t \oplus t_1)x(t \oplus t_2)]\}\, dt_1\, dt_2\, dt_3 + \cdots \end{aligned} \tag{1}$$

where h_n is the nth order kernel, $E[\cdot]$ denotes the expectation operator, and the operator $\oplus$ signifies modulo-two addition (without carry) of the two real numbers involved. In the following context, and without the loss of generality, we assume that the kernels are symmetric functions in their arguments since it can be seen that the output $y(t)$ would be identical for any permutation of the arguments.

The Walsh transform of (1) is given by

$$\begin{aligned} Y(\sigma) = W\{y(t)\} &= \int_0^\infty y(t)\Psi(\sigma, t)\, dt \\ &= h_0\delta(\sigma) + \int_0^\infty H_1(\sigma_1)X(\sigma_1)\delta(\sigma \oplus \sigma_1)\, d\sigma_1 \\ &+ \int_0^\infty \int_0^\infty H_2(\sigma_1, \sigma_2)\{X(\sigma_1)X(\sigma_2) \\ &- E[X(\sigma_1)X(\sigma_2)]\}\delta(\sigma \oplus \sigma_1 \oplus \sigma_2)\, d\sigma_1\, d\sigma_2 \\ &+ \int_0^\infty \int_0^\infty \int_0^\infty H_3(\sigma_1, \sigma_2, \sigma_3)\{X(\sigma_1)X(\sigma_2)X(\sigma_3) \\ &- X(\sigma_1)E[X(\sigma_2)X(\sigma_3)] - X(\sigma_2)E[X(\sigma_3)X(\sigma_1)] \\ &- X(\sigma_3)E[X(\sigma_1)X(\sigma_2)]\} \\ &\cdot \delta(\sigma \oplus \sigma_1 \oplus \sigma_2 \oplus \sigma_3)\, d\sigma_1\, d\sigma_2\, d\sigma_3 + \cdots \end{aligned} \tag{2}$$

where $W\{\cdot\}$ signifies the Walsh transform operator, $\Psi(\sigma, t)$ denotes a generalized Walsh function [8], and $H_n(\sigma_1, \sigma_2, \cdots, \sigma_n)$ and $X(\sigma)$ are Walsh transforms of the nth order kernel and the input, respectively. We refer to $H_n(\sigma_1, \sigma_2, \cdots, \sigma_n)$ as the nth order transfer function in the Walsh sequency domain. Note that the nth order transfer function has the same symmetry property as that of the nth order kernel. In (2), the delta function $\delta(\sigma)$ is the Walsh transform of the unit step function, i.e., $\delta(\sigma) = \int_0^\infty \Psi(\sigma, t)\, dt$, and it has the

Manuscript received September 6, 1985; revised February 28, 1986. This work was supported by the Department of Defense Joint Services Electronics Program through the Air Force Office of Scientific Research Contract F49620-82-C-0033.

K. I. Kim and E. J. Powers, Jr. are with the Electronics Research Center and the Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, TX 78712. (512) 471-6179.

J. Y. Hong was with the Electronics Research Center and the Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, TX 78712. He is now with Intelligent Signal Processing, Austin, TX 78752. (512) 495-6729.

IEEE Log Number 8608940.

0018-9375/86/0800-0162$01.00 © 1986 IEEE

following properties [8] which are similar to those of the Dirac delta function:

$$\int_0^\infty f(t)\delta(t \oplus t_1)\, dt = f(t_1)$$

$$\int_0^\infty \delta(t \oplus t_1)\, dt = 1, \quad \text{for } t_1 > 0.$$

To get (2) from (1), we have utilized the product rules of generalized Walsh functions which are expressed as

$$\Psi(\sigma_1, t)\Psi(\sigma_2, t) = \Psi(\sigma_1 \oplus \sigma_2, t)$$

$$\Psi(\sigma, t_1)\Psi(\sigma, t_2) = \Psi(\sigma, t_1 \oplus t_2).$$

By averaging (2) and then integrating, h_0 is readily obtained such that

$$h_0 = \int_0^\infty E[Y(\sigma)]\, d\sigma.$$

Since we assume that the transfer functions are symmetric functions of their arguments, other transfer functions up to order n, i.e., $H_n(\sigma_1, \sigma_2, \cdots, \sigma_n)$, may be obtained by multiplying (2) by $X(\sigma_1')X(\sigma_2') \cdots X(\sigma_n')$, respectively, and then taking an expected value of each side. Because of the Gaussian nature of the input process and the orthogonality of the functional series in (2), only one term containing $H_n(\sigma_1, \cdots, \sigma_n)$ remains in calculating $E[Y(\sigma)X(\sigma_1') \cdots X(\sigma_n')]$. In particular, the transfer functions up to third order are given as follows:

$$H_1(\sigma) = \frac{\Gamma_{yx}(\sigma)}{\Gamma_x(\sigma)} \tag{3}$$

$$H_2(\sigma_1, \sigma_2) = \frac{1}{2!}\left\{\frac{\Gamma_{yxx}(\sigma_1, \sigma_2) - h_0\Gamma_x(\sigma_1)\delta(\sigma_1 \oplus \sigma_2)}{\Gamma_x(\sigma_1)\Gamma_x(\sigma_2)}\right\} \tag{4}$$

$$H_3(\sigma_1, \sigma_2, \sigma_3) = \frac{1}{3!}\left[\frac{\Gamma_{yxxx}(\sigma_1, \sigma_2, \sigma_3)}{\Gamma_x(\sigma_1)\Gamma_x(\sigma_2)\Gamma_x(\sigma_3)} - \left\{\frac{H_1(\sigma_1)}{\Gamma_x(\sigma_3)}\delta(\sigma_2 \oplus \sigma_3) + \frac{H_1(\sigma_2)}{\Gamma_x(\sigma_1)}\delta(\sigma_1 \oplus \sigma_3) + \frac{H_1(\sigma_3)}{\Gamma_x(\sigma_2)}\delta(\sigma_1 \oplus \sigma_2)\right\}\right]. \tag{5}$$

In (3)–(5), $\Gamma_x(\sigma)$ is the Walsh sequency power spectral density function of $x(t)$, and $\Gamma_{yx\cdots x}(\sigma_1, \sigma_2, \cdots, \sigma_n)$ is the nth order Walsh cross spectrum such that $\Gamma_x(\sigma)\delta(\sigma \oplus \sigma_1) = E[X(\sigma)X(\sigma_1)]$, and $\Gamma_{yx\cdots x}(\sigma_1, \sigma_2, \cdots, \sigma_n)\delta(\sigma_1 \oplus \sigma_2 \oplus \cdots \oplus \sigma_n \oplus \sigma) = E[Y(\sigma)X(\sigma_1)X(\sigma_2) \cdots X(\sigma_n)]$. Note that the nth order Walsh spectrum can be defined as the Walsh transform of the nth order logical correlation function, and, in general, it is not directly related to the Fourier spectrum. In the case of the second-order moment, the relationship between discrete Walsh and Fourier power spectra and their computation was discussed in [9], and the algorithmic properties of logical and arithmetic autocorrelation functions were investigated in [10].

We can see that the resulting expressions for the kernels up to third order have much the same form as those obtained in the Fourier frequency domain [4], except that they are expressed in terms of Walsh instead of Fourier spectra. This result implies that we may take advantage of the computation procedure with which we are familiar in Fourier analysis [11]; that is, the nth order transfer function $H_n(\sigma_1, \sigma_2, \cdots, \sigma_n)$ can be obtained from the raw input and output time-series data by computing the appropriate power spectra using the fast Walsh transform algorithm.

Finally, we note that the statistical approach used in this paper to obtain (3)–(5) is equivalent to the approach in which we minimize the mean-square error between the output of the actual system and the model [12], [13]. Therefore, considering $y(t)$ in (1) as the estimate of the desired signal and $x(t)$ as the signal immersed in noise, the transfer functions developed in this paper would specify the nonlinear filter which is optimum in the mean-square sense [13], [14]. Although, due to the lack of a simple relationship between arithmetic convolution and logical convolution, the existence of natural systems that can be modeled by (1) is open to question, we can extend the results developed in this paper so as to design an optimal NLDI filter which can be used as an alternative to a time-invariant filter. The implementation as well as the design of such filters in the Walsh sequency domain would be very efficient.

REFERENCES

[1] J. L. Walsh, "A closed set of orthogonal functions," *Amer. J. Math.*, vol. 44, pp. 5-24, 1923.

[2] F. Pichler, "Walsh functions and optimal linear systems," in *Proc. 1970 Symp. Applications of Walsh Functions* (Washington, DC), Mar. 31-Apr. 3, 1970, pp. 17-22.

[3] S. Cohn-Sfetcu and S. T. Nichols, "On the identification of linear dyadic invariant systems," *IEEE Trans. Electromagn. Compat.*, vol. EMC-17, pp. 111-117, May 1975.

[4] A. S. French and E. G. Butz, "The use of Walsh functions in the Wiener analysis of nonlinear systems," *IEEE Trans. Comput.*, vol. C-23, pp. 225-232, Mar. 1974.

[5] J. Y. Hong, Y. C. Kim, and E. J. Powers, "On modeling the nonlinear relationship between fluctuations with nonlinear transfer functions," *Proc. IEEE*, vol. 68, pp. 1026-1027, Aug. 1980.

[6] N. Wiener, *Nonlinear Problems in Random Theory*. Cambridge, MA: MIT Press, 1958.

[7] J. F. Barrett, "The use of functionals in the analysis of nonlinear physical systems," *J. Electron. Contr.*, vol. 15, pp. 567-615, 1963.

[8] M. Maqusi, *Applied Walsh Analysis*. London, U.K.: Heyden, 1981.

[9] G. S. Robinson, "Discrete Walsh and Fourier power spectra," in *Proc. 1972 Symp. Applications of Walsh Functions* (Washington, DC), Mar. 27-29, 1972, pp. 298-309.

[10] N. Ahmed and T. Natarajan, "On logical and arithmetic autocorrelation functions," *IEEE Trans. Electromagn. Compat.*, vol. EMC-16, pp. 177-183, Aug. 1974.

[11] J. S. Bendat and A. G. Piersol, *Engineering Applications of Correlation and Spectral Analysis*. New York: Wiley, 1980.

[12] P. Eykoff, "Some fundamental aspects of process-parameter estimation," *IEEE Trans. Automat. Contr.*, vol. AC-8, pp. 347-357, Oct. 1963.

[13] T. Koh and E. J. Powers, "Second-order Volterra filtering and its application to nonlinear system identification," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 1445-1455, Dec. 1985.

[14] J. Katzenelson and L. A. Gould, "The design of nonlinear filters and control systems, Part I," *Inform. Contr. 5*, pp. 108-143, 1962.