Martin Hromčík

Advanced systems and control algorithms and their applications

CZECH TECHNICAL UNIVERSITY IN PRAGUE
FACULTY OF ELECTRICAL ENGINEERING

Habilitation thesis

Advanced systems and control algorithms and their applications

Praha, 2012

Martin Hromčík

Copyright

Abstract

This habilitation thesis is focused on advanced algorithms for control and systems design problems and their application in various industries. Namely, two new polynomial factorization solvers are presented first, with direct impact towards audio compensation devices design. In addition, results related to biomedical research and the aerospace area are presented.

Anotace

Předkládaná habilitace pojednává o pokročilých algoritmech z oblasti teorie systémů a návrhu řízení a jejich aplikacích. Dva nové algoritmy pro polynomiální faktorizace s aplikacemi v audio oblasti, jsou prezentovány v prvním přiloženém článku. Další články se pak týkají aplikací metod řízení a systémů v oblastech biomedicíny a leteckého a kosmického výzkumu.

Acknowledgments

My thanks come here to my colleagues and friends who directly and significantly contributed to particular papers collected in this habilitation thesis.

- Michael Šebek, professor at the Faculty of Electrical Engineering, CTU in Prague, is the co-author of the first and second presented papers. Additionally, it happened to a great extent thanks to his outstanding activities and respect in the controls and systems community that the European project ACFA 2020 was awarded to our department and to our aerospace group in the year 2008, giving rise - among others - to the research results presented in the fifth attached paper.

- Petr Kujan, Vladimír Tichý, Tomáš Haniš and Jana Nováková, my finished or finishing PhD students, all worked really great in the recent years and managed to achieve exquisite research results under my supervision, worth publishing in prestigious SCI journals.

I am aware of the fact that I missed to name many people who helped me substantially in my career on the way to this habilitation, and they all fully deserve and have my deep gratitude and respect. I apologize to all of you, my dear colleagues and friends, for not mentioning you explicitly at this place, and I beg for your understanding.

The research presented in this habilitation thesis has been supported by the following grants.

- *Centrum aplikované kybernetiky, 2001-2011, MŠMT projekt číslo 1M0567*

- *ACFA 2020 - Active Control for Flexible 2020 Aircraft, projekt 7. RP EU, AAT-2007-RTD-1, 2008-2011*

- *AKTION - Structural Vibration Control, vědecká spolupráce v rámci dvoustranných a mnohostranných mezivládních dohod, 51p4, 2008-2009*

- *Aktivní tlumení pro mobilní prostředky, program MŠMT KONTAKT, ME10094, 2010-2012*

- *Kolaborativní algoritmy pro řízení formací bezpilotních letounů COLA4CON, 2010-2012, Číslo: ME10010*

- *Algoritmy pro analýzu a návrh řízení složitých systémů, 2008-2010, Číslo: GA102/08/0186*

- *Struktura a universalita v návrhu řízení složitých systémů, 2005-2007, Číslo: GA102/05/0011*

- *Moderní numerické metody pro systémy, signály a řízení, 2002-2004, Číslo: GA102/02/0709*

Comments to the published works

The presented habilitation thesis takes the form of a collection of selected SCI Expanded journal papers published by the applicant in recent five years. In accordance with the multidisciplinary nature of the controls and systems research area, and thanks to the variety of project opportunities the applicant has been facing in the past years, the covered topics range from algorithms development through aircraft and spacecraft systems to functional brain modelling approaches. Though, the unifying theme of the systems theory and control fundamentals can be identified in the whole collection upon a closer look.

The structured list of the presented papers follows.

A. Hromčík, M. - Šebek, M.: Numerical Algorithms for Polynomial Plus/Minus Factorization. *International Journal of Robust and Nonlinear Control*. 2007, vol. 17, no. 8, p. 786-802. ISSN 1049-8923. *(IF 1.495)*

B. Kujan, P. - Hromčík, M. - Šebek, M.: Complete Fast Analytical Solution of the Optimal Odd Single-Phase Multilevel Problem. *IEEE Transactions on Industrial Electronics*. 2010, vol. 2010, no. 57(7), p. 2382-2397. ISSN 0278-0046. *(IF 3.439)*

C. Tichý, V. - Barbera, M. - Collura, A. - Hromčík, M. - Hudec, R. - et al.: Tests of Lobster Eye Optics for Small Space X-ray Telescope. *Nuclear Instruments and Methods in Physics Research, Section A, Accelerators, Spectrometers, Detectors and Associated Equipment*. 2011, vol. 633, no. 1, p. S169-S171. ISSN 0168-9002. *(IF 1.142)*

D. Nováková, J. - Hromčík,M. – Jech, R.: Dynamic Causal Modeling and subspace identification methods, Biomedical Signal Processing and Control, available online August 9, 2011 DOI: 10.1016/j.physletb.2003.10.071, August 2011 *(IF 0.734)*

E. Haniš, T. - Hromčík M.: Optimal Sensor placement and Spillover Suppression, Mechanical Systems and Signal Processing (in press, January 2012) *(IF 1.762)*

The *paper A*, co-authored by Michael Šebek, was published in 2007 in the Journal of Robust and Nonlinear Control as a direct follow-up of the applicant's dissertation. Two new numerical routines for plus-minus factorization, based on FFT and the matrix factorization approach respectively, are presented. Applications are targeted at the audio processing area, and an acoustics compenasation case study, related to research done at the University of Uppsala, is processed.

The remaining papers are results of the applicant's cooperation with his Ph.D. students he has supervised in the years 2004-2011. They are namely Petr Kujan, Vladimír Tichý, Tomáš Haniš and Jana Nováková.

The *paper B*, co-authored by Petr Kujan and Michael Šebek, was published in the IEEE TIE in 2010. Analytical solution of the optimal pulse-width-modulation problem for odd-symmetry waveforms is resolved. Performance of the proposed algorithm is assessed and compared to exiting routines. An active-filter case study is included.

Paper C, co-authored by Vladimír Tichý and our colleagues from the Astronomical Institute, Faculty of Nuclear Sciences and Physical Engineering at CTU, Institute of Experimental and Applied Physics at CTU, INAF-OAPA in Palermo, and the RIGAKU company, is devoted to assessment of X-ray optical characteristics of a small space X-ray telescope prototype, developed within Vladimir's dissertation. Measurements were conducted at the INAF-OAPA experimental X-ray 35 meters long parallel beam generator.

Paper D, co-authored by Jana Nováková and Robert Jech from the First Faculty of Medicine, Charles University in Prague, is a result of our cooperation with the General Teaching Hospital, Department of Neurology, Prague. Based on our experience gained during our fMRI data-processing job related to a clinical study performed at the hospital, we propose an alternative procedure for determination of substantial functional connections among selected brain areas. Our method, compared to currently dominating DCM approach (Dynamic Causal Modelling), does not require a-priori hypotheses related to existence of important connections. Instead, the structure is actually given as an output.

Finally, the *paper E* co-authored by Tomáš Haniš is devoted to control algorithms developed at the Department of Control Engineering, FEE CTU, in the years 2008-2011 within the European R&D FP7 project Active Control for Flexible Aircraft ACFA 2020 that I had the privilege to lead on behalf of CTU as a project consortium member. High capacity airliner blended-wing-body near-future concept was elaborated by our consortium partners and resulting delivered high-fidelity models, comprising both the rigid-body flight dynamics part and highly flexible nature of the construction, were used to devise advanced, yet low-complexity, robust and optimal MIMO (multiple-input multiple-output) control laws.

Paper A.

Hromčík, M. - Šebek, M.: Numerical Algorithms for Polynomial Plus/Minus Factorization. *International Journal of Robust and Nonlinear Control*. 2007, vol. 17, no. 8, p. 786-802. ISSN 1049-8923. *(IF 1.495)*

INTERNATIONAL JOURNAL OF ROBUST AND NONLINEAR CONTROL
Int. J. Robust Nonlinear Control 2007; **17**:786–802
Published online 30 October 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/rnc.1132

Numerical algorithms for polynomial plus/minus factorization

M. Hromčík[1,*,†] and M. Šebek[2]

[1] *Department of Control Engineering, Faculty of Electrical Engineering, Center for Applied Cybernetics, Czech Technical University in Prague, Czech Republic*
[2] *Department of Control Engineering, Faculty of Electrical Engineering, Czech Technical University in Prague, Czech Republic*

SUMMARY

Two new algorithms are presented in the paper for the plus/minus factorization of a scalar discrete-time polynomial. The first method is based on the discrete Fourier transform theory (DFT) and its relationship to the Z-transform. Involving DFT computational techniques and the famous fast Fourier transform routine brings high computational efficiency and reliability. The method is applied in the case study of H_2-optimal inverse dynamic filter to an audio equipment. The second numerical procedure originates in a symmetric spectral factorization routine, namely the Bauer's method of the 1950s. As a by-product, a recursive *LU* factorization procedure for Toeplitz matrices is devised that is of more general impact and can be of use in other areas of applied mathematics as well. Performance of the method is demonstrated by an l_1 optimal controller design example. Copyright © 2006 John Wiley & Sons, Ltd.

Received 30 August 2006; Accepted 30 August 2006

KEY WORDS: polynomial design methods; numerical algorithms; polynomial factorizations

1. INTRODUCTION

This paper describes a new method for the plus–minus factorization of a discrete-time polynomial. Given a polynomial in the z variable,

$$p(z) = p_0 + p_1 z + p_2 z^2 + \cdots + p_n z^n$$

without any roots on the unit circle, its plus/minus factorization is defined as

$$p(z) = p^+(z) p^-(z) \tag{1}$$

*Correspondence to: M. Hromčik, Center for Applied Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague, Karlovo namesti 13-G, Prague, Czech Republic.
† E-mail: hromcik@utia.cas.cz

Contract/grant sponsor: The Ministry of Education of the Czech Republic (MSMT); contract/grant number: 1M0567
Contract/grant sponsor: The Grant Agency of the Czech Republic (GACR); contract/grant number: 102/05/0011

where $p^{+}(z)$ has all roots inside and $p^{-}(z)$ outside the unit disc. Clearly, the scalar plus/minus factorization is unique up to a scaling factor.

Polynomial plus/minus factorization has many applications in control and signal processing problems. For instance, efficient algebraic design methods for time-optimal controllers [1], quadratically optimal filters for mobile phones [2, 3], and l_1 optimal regulators [4], to name just a few, all recall the $+/-$ factorization as a crucial computational step.

2. EXISTING METHODS

In any case, the plus/minus factors for $n \geqslant 5$ cannot be achieved by a finite number of algebraic operations. This conclusion is due to the Galois's theorem stating that the roots of a polynomial of degree greater or equal to five cannot be expressed in a closed form. Therefore, all numerical algorithms for plus/minus factorization are iterative in nature and give just an approximation to the genuine factors. Some existing approaches to this problem are mentioned in this section.

The most natural way is based on the computation of polynomial roots. Having determined the roots $r_1, r_2, \ldots, r_n$ of $p(z)$ via any standard procedure for polynomial roots [5] and considering that $p(z) \neq 0$ for all $|z| = 1$ by assumption, one can divide the roots into two groups $R^{+} = \{r_i : m^{\bigstar}(r_i) = 0, |r_i| < 1\}, R^{-} = \{r_i : m^{\bigstar}(r_i) = 0, |r_i| > 1\}$. Clearly, R^{+} and R^{-} are the sets of roots of $p^{+}(z)$ and $p^{-}(z)$, respectively.

Performance of this procedure heavily hinges on the accuracy of the computed polynomial roots. If these roots are distinct and separated enough, standard numerical routines [5] can determine them with good precision. However, it is well known that the relative accuracy of a computed root decreases as its multiplicity grows [5], and so does the accuracy of the spectral factor thus obtained.

In addition, if the degree of the involved polynomial is high, say over 50, the very computation of the spectral factor coefficients is problematic due to rounding errors. It means that even if the desired *roots* of the spectral factor are evaluated with good accuracy, its particular *coefficients*, which are typically required in applications, are not accurate.

An alternative algorithm relies on polynomial spectral factorization and greatest polynomial divisor computation. If $q(z)$ is the spectral factor of the symmetric product $p(z)p(z^{-1})$ then the greatest common divisor of $p(z)$ and $q(z)$ is obviously the plus factor of $p(z)$. The minus factor can be derived similarly from $p(z^{-1})$ and $q(z^{-1})$. As opposed to the previous approach based on direct roots computation which typically makes problems for higher degrees and/or roots multiplicities, this procedure relies on numerically reliable algorithms for polynomial spectral factorization [6, 7]. Unfortunately, the polynomial greatest common divisor computation is much more sensitive. As a result, both these techniques do not work properly for high degrees (say over 50).

In this report, we will introduce a completely new approach to the problem, inspired by our work on efficient algorithms for polynomial spectral factorization, see [7]. It is based on the discrete Fourier transform theory (DFT) theory and provides both a fruitful view on the relationship between DFT and the Z-transform theory, and a powerful computational tool in the form of the fast Fourier transform (FFT) algorithm.

DOI: 10.1002/rnc

3. DISCRETE FOURIER TRANSFORM

If $\mathbf{p} = [p_0, p_1, \ldots, p_N]$ is a vector of complex numbers, then its *direct* DFT is given by the vector $\mathbf{y} = [y_0, y_1, \ldots, y_N]$, where

$$y_k = \sum_{i=0}^{N} p_i \mathrm{e}^{-\mathrm{j}(2\pi k/N+1)i} \tag{2}$$

The vector $\mathbf{y}$ is called the image of vector $\mathbf{p}$. Conversely, if $\mathbf{y} = [y_0, y_1, \ldots, y_N]$ is given, then its inverse DFT recovers the original vector $\mathbf{p} = [p_0, p_1, \ldots, p_N]$, where

$$p_i = \frac{1}{N+1} \sum_{k=0}^{N} y_k \mathrm{e}^{\mathrm{j}(2\pi i/N+1)k} \tag{3}$$

DFT is of great interest in various engineering fields. For its relationship to Fourier series of sampled signals, DFT is frequently used in signal processing. One of the experimental identification methods employs DFT as well [8]. The close relationship of DFT to interpolation is also well known and was used recently to solve some tasks of the polynomial control theory [9] and to treat robustness analysis problems of certain kind [10].

For numerical computation of DFT, the efficient recursive FFT algorithm was developed by Cooley and Tukey in 1965 [11]. If the length of the input is a power of two, a faster version of FFT (sometimes called *radix-2 FFT*) can be employed [11]. In general, the FFT routine features a highly beneficial computational complexity and involves $\mathcal{O}(N\log(N))$ multiplications and additions for a vector of length N.

Thanks to the importance of DFT mentioned above, the FFT algorithms are naturally available as built-in functions of many computing packages (MATLAB™, MATHEMATICA™, etc.). This is another good reason for employing the procedure proposed in this paper.

4. PLUS/MINUS FACTORIZATION AND DFT

4.1. Theory

Given a polynomial

$$p(z) = p_0 + p_1 z + \cdots + p_d z^d$$

non-zero for $|z| = 1$, we first apply a direct degree shift to arrive at a two-sided polynomial

$$\tilde{p}(z) = p_0 z^{-\delta} + \cdots + p_d z^{d-\delta}$$

where δ is the number of roots of $p(z)$ lying inside the unit circle. Now, instead of solving Equation (1), we look for $\tilde{p}^+(z) = \tilde{p}_0^+ + \tilde{p}_1^+ z^{-1} + \cdots + \tilde{p}_\delta^+ z^{-\delta}$ and $\tilde{p}^-(z) = \tilde{p}_0^- + \tilde{p}_1^- z + \cdots + \tilde{p}_{d-\delta}^- z^{d-\delta}$ such that

$$\tilde{p}(z) = \tilde{p}^+(z)\tilde{p}^-(z) \tag{4}$$

Relationship between the pairs $\tilde{p}^+, \tilde{p}^-$ and p^+, p^- are obvious.

In order to solve Equation (4), logarithm is applied. As $\tilde{p}(z)$, $\tilde{p}^+(z)$ and $\tilde{p}^-(z)$ are all analytic and non-zero in $1-\varepsilon<|z|<1+\varepsilon$ the logarithms exist. Let us denote them as $\ln \tilde{p}(z) = n(z), \ln \tilde{p}^+(z) = x^+(z), \ln \tilde{p}^-(z) = x^-(z)$. Here $n(z)$, obtained from the given $\tilde{p}(z)$, is a Laurent

DOI: 10.1002/rnc

infinite power series

$$n(z) = \cdots + n_1 z + n_0 + n_{-1} z^{-1} + \cdots$$

It can be directly decomposed,

$$n(z) = x^{+}(z) + x^{-}(z^{-1})$$

with power series

$$x^{+}(z) = x_0^{+} + x_1^{+} z^{-1} + \cdots = \frac{n_0}{2} + n_{-1} z^{-1} + \cdots, \quad x^{-}(z) = x_0^{-} + x_1^{-} z + \cdots = \frac{n_0}{2} + n_1 z + \cdots \tag{5}$$

analytic for $1 - \varepsilon < |z|$ and $1 + \varepsilon > |z|$, respectively.

At this time, the necessity of the degree shift yielding the two-sided polynomial $\tilde{p}$ can be explained. According to the Cauchy's theorem of argument [12], the curve $p(z)$ for $|z| = 1$ encircles the origin in the complex plane as many times as is the number of roots of $p(z)$ lying in the complex unit disc. Hence, the logarithms cannot be applied directly as its imaginary part, reading the phase of $p(z)$, would not be continuous. An easy solution to avoid this situation is to move the desired number of roots of $p(z)$ from infinity to zero by performing proper degree shift.

Once $x^{+}(z)$ and $x^{-}(z)$ are computed, the plus/minus factors $\tilde{p}^{+}, \tilde{p}^{-}$ are recovered as

$$\tilde{p}^{+} = \mathrm{e}^{x^{+}(z)} = \tilde{p}_0^{+} + \tilde{p}_1^{+} z^{-1} + \cdots, \quad \tilde{p}^{-} = \mathrm{e}^{x^{-}(z)} = \tilde{p}_0^{-} + \tilde{p}_1^{-} z + \cdots$$

Since $x^{+}(z)$ is analytic in $1 - \varepsilon < |z|$, so is $\tilde{p}^{+}(z)$ and hence it can be expanded according to (3). Moreover, as a result of exponential function, $\tilde{p}^{+}(z)$ is non-zero in $1 - \varepsilon < |z|$. In other words, it has all its zeros inside the unit disc and is therefore Schur stable. Note also that $\tilde{p}^{+}(z)$ has to be a (finite) polynomial of degree d (due to the uniqueness of the solution to the problem which is known to be a polynomial) though $n(z)$ is an infinite power series. Similar reasoning proves the $\tilde{p}^{-}$ factor desired properties.

4.2. Numerical algorithm

Numerical implementation follows the ideas considered above. A polynomial $p(z)$ is represented by its coefficients p_i, $i = 0 \ldots r$ or, equivalently, by function values P_k in the Fourier interpolating points g^k, $k = -R \ldots 0 \ldots R$, where $R \geqslant d$, $g = \mathrm{e}^{\mathrm{j}(2\pi/(2R+1))}$. Accordingly, a power series can be approximated by a finite set of its coefficients or by its values in a finite number of interpolation points on the unit circle. Some operations of the procedure, namely the decomposition of $n(z)$ into $x^{+}(z)$ and $x^{-}(z)$, are performed in the time domain (operations on coefficients), while the others (evaluation of logarithmic and exponential functions) are executed in the frequency domain (operations with values over $|z| = 1$). Mutual conversion between the two domains is mediated by the shifted discrete Fourier transform operator defined as

$$X_k = \sum_{i=-R}^{R} x_i g^{-ki}, \quad x_i = \frac{1}{2R+1} \sum_{k=-R}^{R} X_k g^{ki}$$

which approximates the Z-transform by dealing with $-R \leqslant i \leqslant +R$ instead of infinite $-\infty < i < +\infty$ and with $z = g^k, -R \leqslant k \leqslant +R$ instead of continuum $z = \mathrm{e}^{\mathrm{j}\phi}, -\pi \leqslant \phi \leqslant +\pi$.

The accuracy of the results depends on the number of interpolation points $2R + 1$ involved in the computation. This number can be considered as a simple tuning knob of the computational process.

Then, the resulting numerical routine looks as follows:

Algorithm 1: Scalar discrete-time plus–minus factorization
Input: Scalar polynomial

$$p(z) = p_0 + p_1 z + \cdots + p_d z^d, \quad \text{non-zero for } |z| = 1$$

Output: Polynomials $p^+(z)$ and $p^-(z)$, the plus and minus factors of $p(z)$.

Step 1: *Choice of the number of interpolation points*. Decide about the number R. R approximately 10–50 times larger than d is recommended up to our practical experience.

Step 2: *Degree shift*. Find out the number δ of zeros of $p(z)$ inside the unit disc. A modification of well-known Schur stability criterion can be employed, see [13] for instance.

Having δ at hand, construct a two-sided polynomial $\tilde{p}(z)$ as

$$\tilde{p}(z) = p(z)z^{-\delta} = p_0 z^{-\delta} + \cdots + p_d z^{d-\delta} = \tilde{p}_{-\delta} z^{-\delta} + \cdots + \tilde{p}_0 + \cdots + \tilde{p}_{d-\delta} z^{d-\delta}$$

Step 3: *Direct FFT (I)*. Using the FFT algorithm, perform direct DFT, defined by (2), on the vector

$$\mathbf{p} = [\underbrace{\tilde{p}_0, \tilde{p}_1, \ldots, \tilde{p}_{d-\delta}, 0, 0, \ldots, 0, \tilde{p}_{-\delta}, \ldots, \tilde{p}_{-1}}_{2R+1}]$$

In this way, the set $\mathbf{P} = [P_0, P_1, \ldots, P_{2R}]$ of the values of $\tilde{p}(z)$ at the Fourier points is obtained.

Step 4: *Logarithmization*. Compute the logarithms $N_i = \ln(P_i)$ of all particular P_i's and form the vector $\mathbf{N} = [N_0, N_1, \ldots, N_{2R}]$ of them. N_i's thus obtained are the values of the function $n(z) = \ln(\tilde{p}(z))$ at related Fourier points on the unit complex circle.

Step 5: *Inverse FFT (I)*. To get the vector $\mathbf{n} = [n_0, n_1, \ldots, n_R, n_{-R}, \ldots, n_{-1}]$, containing the coefficients of the two-sided polynomial $n(z) = n_{-R}z^{-R} + \cdots + n_{-1}z^{-1} + n_0 + n_1 z + \cdots + n_R z^R$ approximating the power series of $\ln(m(z))$ for the given R, perform inverse DFT, defined by (3), on the vector $\mathbf{N}$ using the FFT algorithm.

Step 6: *Decomposition*. Take the 'causal part' $\mathbf{x}^+$ of $\mathbf{n} : \mathbf{x}^+ = [n_0/2, n_1, \ldots, n_R]$. Similarly, construct $\mathbf{x}^-$ as $\mathbf{x}^- = [n_0/2, n_{-1}, \ldots, n_{-R}]$.

Step 7: *Direct FFT (II)*. Evaluate $x^+(z) = n_0/2 + n_1 z^{-1} + \cdots + n_R z^{-R}$ at the Fourier points by applying direct FFT on the set $\mathbf{x}^+$ and get $\mathbf{X}^+ = [X_0^+, \ldots, X_R^+]$. Proceed with $x^-(z)$ in obvious way.

Step 8: *Exponential function*. To get the plus/minus factors, the exponential functions $\tilde{p}^+(z) = e^{x^+(z)}$ and $\tilde{p}^-(z) = e^{x^-(z)}$ remain to be evaluated. First, we compute the values of $\tilde{p}^+(z)$ and $\tilde{p}^-(z)$ at the Fourier points: $\tilde{\mathbf{P}}^+ = [e^{X_0^+}, \ldots, e^{X_R^+}]$. Similar steps apply for the minus part.

Step 9: *Inverse FFT (II)*. Finally, the coefficients $\tilde{\mathbf{p}}^+ = [\tilde{p}_0^+, \ldots, \tilde{p}_R^+]$ of $\tilde{p}^+(z)$ are recovered by inverse FFT performed on the vector $\tilde{\mathbf{P}}^+$. The resulting approximation to the plus factor $\tilde{p}^+(z)$ then equals $\tilde{p}^+(z) = \tilde{p}_0^+ + \tilde{p}_{-1}^+ z^{-1} + \cdots + \tilde{p}_{-\delta}^+ z^{-\delta}$. Proceed with the minus part accordingly.

Step 10: *Finalization*. Convert the plus–minus factors $\tilde{p}^+(z)$ and $\tilde{p}^-(z)$ of $\tilde{p}(z)$ into the desired factors of $p(z)$ using the following formulas:

$$p^- = \tilde{p}^-, \quad p^+ = \tilde{p}^{+\bigstar}$$

where the star stands for discrete-time conjugate, $z \to z^{-1}$.

Note that one obtains R coefficients of $\tilde{p}^+$ and $\tilde{p}^-$ in Step 9. However, $p^+(z)$ being the plus factor of $p(z)$ is known to be of degree δ only and only the first $\delta + 1$ coefficients of $\tilde{p}^+(z)$ should be significant as a result while the remaining ones should be negligible. As the number R increases, these values theoretically converge to zero indeed since the formulas of DFT become better approximations to the Z-transform definitions.

5. RADIX-2 MODIFICATION OF THE ALGORITHM

The basic version of the routine proposed above is based on the shifted discrete Fourier transform. This modification of DFT appears useful during the derivation of the Algorithm 1 due to its more transparent relationship to the spectral theory. It can be easily transformed to the standard DFT as it is defined in the Section 3, simply by reordering related vector entries (see Steps 2 and 4 of Algorithm 1). However, $2R+1$ interpolation points are used for the FFT algorithm and unfortunately this number is always odd and cannot equal any power of two. Therefore, the radix-2 fast version of the FFT routine cannot be addressed. Nevertheless, this slight drawback can be easily avoided if the periodicity of direct and inverse DFT formulas is taken into account. Basically, one can construct the initial set as

$$[\underbrace{\tilde{p}_0, \tilde{p}_1, \dots, \tilde{p}_{d-\delta}, 0, 0, \dots, 0, \tilde{p}_{-\delta}, \dots, \tilde{p}_{-1}}_{2^R}]$$

which has a power-of-two entries in total. The Algorithm 1 remains valid also in this case with $2R+1$ replaced by 2^R and $R+1$ by 2^{R-1}, respectively, up to one point: in Step 6, the decomposition reads $\mathbf{x}^+ = [n_0/2, n_1, \dots, n_R/2]$ instead of $\mathbf{x}^+ = [n_0/2, n_1, \dots, n_R]$. This minor modification of the proposed method further increases its efficiency since the powerful radix-2 FFT can be called.

6. COMPUTATIONAL COMPLEXITY

Thanks to the fact that the FFT algorithm is extensively used during the computation, the overall routine features an expedient computational complexity.

Provided that the above modifications of the computational procedure are considered, namely if the resulting number of interpolation points is taken as a power of two, the fast radix-2 FFT can be employed. In this case, $(R\log_2 R)/2$ multiplications and $R\log_2 R$ additions are needed to evaluate either direct or inverse DFT of a vector of length R [11]. Let us suppose in addition that computing the logarithm or exponential of a scalar constant takes at most k multiplications and l additions. Then the particular steps of the modified Algorithm 1 involve $(R\log_2 R)/2$ multiplications and $R\log_2 R$ additions (Steps 3, 5, 7 and 9), and kR multiplications and lR additions (Steps 4 and 8), respectively. Hence the overall procedure consumes

$$4\frac{R\log R}{2} + 2kR = 2R\log R + 2lR$$

complex multiplications, and

$$4R\log R + 2lR$$

complex additions. By inspecting the above formulas one can see that asymptotically the proposed method features $\mathcal{O}(R\log R)$ complex multiplications and additions.

DOI: 10.1002/rnc

7. UPGRADING LOUDSPEAKERS DYNAMICS

An original approach has been published by Sternad *et al.* in [14] on how to improve the performance of an audio equipment at low additional costs. The authors use the linear-quadratic-gaussian (LQG) optimal feedforward compensator technique to receive an inverse dynamic filter for a moderate quality loudspeaker. By attaching a signal processor implementing this filter prior to the loudspeaker, the dynamical imperfections of the original device are eliminated and the overall equipment behaves as an apparatus of a much higher class. To learn more about this research and to get some working examples, refer [15] (Figure 1).

Unlike their predecessors, the authors try to modify the sound over the whole range of frequencies. Such a complex compensation fully employs the increasing performance of signal hardware dedicated to CD-quality audio signals, and at the same time calls for fast and reliable factorization solvers [14]. We believe our new algorithm will significantly contribute to this goal.

The loudspeaker dynamics is considered in the form of an auto regressive model with external input (ARX) model

$$y(t) = z^{-k}\frac{B(z)}{A(z)}u(t)$$

Since the impulse response is rather long for a high sampling frequency (CD-quality standard of 44 kHz was used), both the numerator and denominator of the model are of high orders, say 1–500.

The model has an unstable inverse in general since some of its zeros may lie outside the unit disc. Hence, a stable approximation has to be calculated to be used in the feedforward structure. The authors recall the LQG theory and seek for a compensating filter

$$u(t) = \frac{Q(z)}{P(z)}w(t)$$

such that the criterion $J = E[|y(t) - w(t-d)|^2 + \rho|u(t)|^2]$ is minimized.

For broadband audio signals, the optimal filter is given in the form

$$u(t) = \frac{Q_1(z)A(z)}{\beta(z)}w(t)$$

where β results from the spectral factorization

$$\beta\beta^* = BB^* + \rho AA^*$$

and Q_1 is the solution of a subsequent Diophantine equation

$$z^{k-d}B^*(z) = r\beta^*(z)Q_1(z^{-1}) + zL^*(z)$$

see [14] (Figure 2).

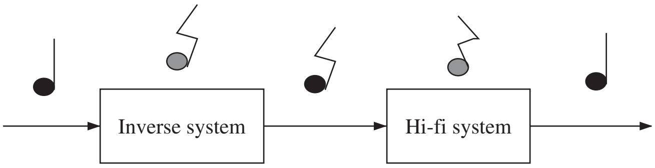


Figure 1. Pre-filter compensation scheme (adopted from [14]).

As for the spectral factor computation, the authors employ the Newton–Raphson iterative scheme [6] in the cited work [14]. According to their results and our experience, this method has been probably the best available procedure for scalar polynomial spectral factorization so far [15, 16]. This method works quite well also for high degrees of involved polynomials in contrast to the straightforward way of computing and distributing the roots of $BB^* + \rho AA^*$.

Let us perform a benchmark experiment to compare the existing approach and our newly proposed algorithm for particular numerical data kindly provided by Mikael Sternad and colleagues from the University of Uppsala. Up to now, two models of the loudspeakers dynamics have been sent to us for testing purposes and the results related to the more complex one are presented in the following.

The data in concern are given as follows. The numerator $B(z) = B_0 + B_1 z^{-1} + \cdots + B_{250} z^{-250}$ is an unstable polynomial of degree 250, $A(z)$ is stable of degree 90, and $k = 160$. Taking $\rho = 0$, the spectral factorization of $m(z) = B(z)B^*(z) = m_{250} z^{-250} + \cdots + m_0 + \cdots + m_{250} z^{250}$ is to be performed. In this special case, the spectral factor $x(z)$ of $m(z)$ can be effectively constructed as

$$x(z) = B^+(z)(B^-(z))^* z^{-k}$$

where B^+, B^- are the plus and minus factors of B, respectively, and k is the degree of B^-.

All presented experiments were realized on a PC computer with Pentium III/1.2 GHz processor and 512 MB RAM, under MS Windows 2000 in MATLAB, version 6.1.

Results of this experiment for various values of the parameters N are summarized and related in Table I. Namely, the computational time and accuracy of results are of interest. To obtain the former characteristic, the MATLAB abilities were employed (the built-in functions `tic/toc`). The computational error is defined here as the largest coefficient of the expression $B^+B^- - B$, evaluated in the MATLAB workspace, divided by the largest coefficient of B (all in absolute value).

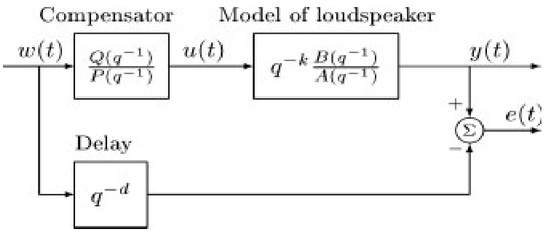


Figure 2. Optimal filtering problem setup (adopted from [14]).

Table I. Accuracy and efficiency of compared algorithms.

	Time (s)	Accuracy
FFT(14)	0.23	6.28×10^{-3}
FFT(15)	0.45	2.52×10^{-8}
FFT(16)	0.89	4.65×10^{-11}
FFT(17)	1.75	2.40×10^{-12}

DOI: 10.1002/rnc

These tests prove the power of the new algorithm in such tough examples. Neither of the two procedures described in Section 2 can factor this large polynomial. Direct roots evaluation method, based on the standard MATLAB function `roots`, gives totally meaningless results (accuracy of 10^{43}) while the routine based on spectral factorization fails due to numerical problems with greatest common polynomial divisor evaluation (polynomial toolbox function `rdiv` was used [16]).

8. DISCUSSION

The success in modifying a selected numerical procedure, originally developed for polynomial spectral factorization, to handle the non-symmetric plus/minus decomposition suggests that other well-known spectral factorization routines might work well in the non-symmetric context as well. Actually, we decided to go this way and succeeded in adapting a classical polynomial spectral factorization routine, the Bauer's algorithm [17], for the plus/minus case. Our results are presented in subsequent paragraphs.

9. BAUER'S METHOD FOR SPECTRAL FACTORIZATION

F. I. Bauer published his method for spectral factorization of a discrete-time scalar polynomial as early as in 1955, see [17, 18]. The procedure is based on the relationship between polynomials and related infinite Toeplitz-type Sylvester matrices.

9.1. Algebra of Sylvester matrices

Given a two-sided polynomial $p(z) = p_{-m}z^{-m} + \cdots + p_0 + \cdots + p_n z^n$, we define its Sylvester companion matrix T_p^N of order N,

$$N \geqslant \max(n, m)$$

as an $N \times N$ square matrix constructed according to the following scheme:

$$T_p^N = \begin{pmatrix} p_0 & p_1 & \dots & p_n & 0 & \dots & 0 \\ p_{-1} & p_0 & p_1 & \dots & p_n & \ddots & \vdots \\ \vdots & p_{-1} & \ddots & \ddots & & \ddots & 0 \\ p_{-m} & \vdots & \ddots & & & & p_n \\ 0 & p_{-m} & & & & \ddots & \vdots \\ \vdots & \ddots & \ddots & & \ddots & \ddots & p_1 \\ 0 & \cdots & 0 & p_{-m} & \dots & p_{-1} & p_0 \end{pmatrix}$$

To show the relationship between the polynomial algebra and the algebra of Sylvester matrices, let us consider two simple polynomials $p_1(z) = 3z^{-1} + 2 + z$ and $p_2(z) = z^{-1} + 3$. Their companion matrices of order four read, respectively

$$T^4_{p_1} = \begin{pmatrix} 2 & 1 & 0 & 0 \\ 3 & 2 & 1 & 0 \\ 0 & 3 & 2 & 1 \\ 0 & 0 & 3 & 2 \end{pmatrix}$$

$$T^4_{p_2} = \begin{pmatrix} 3 & 0 & 0 & 0 \\ 1 & 3 & 0 & 0 \\ 0 & 1 & 3 & 0 \\ 0 & 0 & 1 & 3 \end{pmatrix}$$

Their sum $p_3(z) = p_1(z) + p_2(z)$ equals

$$p_3(z) = 4z^{-1} + 5 + z$$

and its companion matrix can be computed as direct sum of related companion matrices $T^4_{p_1}, T^4_{p_2}$:

$$T^4_{p_3} = \begin{pmatrix} 5 & 1 & 0 & 0 \\ 4 & 5 & 1 & 0 \\ 0 & 4 & 5 & 1 \\ 0 & 0 & 4 & 5 \end{pmatrix}$$

Similarly, their product $p_4 = p_1 p_2 = 3z^{-2} + 11z^{-1} + 7 + 3z$ has a companion matrix

$$T^4_{p_4} = T^4_{p_1} T^4_{p_2} = \begin{pmatrix} 7 & 3 & 0 & 0 \\ 11 & 7 & 3 & 0 \\ 3 & 11 & 7 & 3 \\ 0 & 3 & 11 & 6 \end{pmatrix}$$

9.2. *Bauer's method for spectral factorization*

As we have illustrated above, finite-dimensional matrices are sufficient to accommodate 'finite' algebraic problems. On the other hand, if we do not restrict to finite dimensionality of related matrices, transcendent problems, including spectral factorization, involving polynomials can be resolved by this approach as well.

We will illustrate the Bauer's spectral factorization method by means of a simple example. An interested reader can find detailed description in the original work [17] or, alternatively, in the survey paper [19].

DOI: 10.1002/rnc

Given $p(z) = 2z^{-1} + 5 + 2z$, its companion matrix of order five reads

$$T_p = \begin{pmatrix} 5 & 2 & 0 & 0 & 0 \\ 2 & 5 & 2 & 0 & 0 \\ 0 & 2 & 5 & 2 & 0 \\ 0 & 0 & 2 & 5 & 2 \\ 0 & 0 & 0 & 2 & 5 \end{pmatrix}$$

As p is symmetric and positive definite on the unit circle its spectral factor x exists such that

$$x^{\bigstar}x = p$$

holds and x is stable. The spectral factor coefficients can be approximated using the Cholesky factorization of T_p:

$$T_x = \begin{pmatrix} 2.236 & 0.8944 & 0 & 0 & 0 \\ 0 & 2.049 & 0.9759 & 0 & 0 \\ 0 & 0 & 2.012 & 0.9941 & 0 \\ 0 & 0 & 0 & 2.003 & 0.9985 \\ 0 & 0 & 0 & 0 & 2.001 \end{pmatrix}$$

The diagonals of T_x obviously converge to the genuine spectral factor coefficients: $x(z) = 1 + 2z$.

An interesting feature of this routine is that particular columns of T_x can be computed iteratively, using only latest preceding column and the coefficients of $p(z)$, see [19] for details. As a result, the final algorithm is favourably memory efficient. Mainly for this reason the method is still quite popular in spite of the fact that some later approaches, see e.g. [20, 21], provide a faster rate of convergence.

10. PLUS/MINUS FACTORIZATION VIA BAUER'S APPROACH

A modification of the Bauer's method for the non-symmetric polynomial plus/minus factorization is worked out in this section.

10.1. LU factorization

As we have shown in Section 9.1, algebra of companion matrices is not limited to the symmetric case. Also, the matrix theory provides useful factorization techniques for non-symmetric matrices along with stable and efficient procedures for their computation.

Bauer's method calls for the Cholesky factorization to get the desired spectral factor. This routine assumes the input matrix to be symmetric and positive definite which is the case of the spectral factorization problem. However, if we aim at modifying the method in order to capture the non-symmetric plus/minus factorization case, we need to leave this concept and employ another technique since the companion matrix is no longer symmetric.

The Cholesky factorization decomposes the input matrix into a product of two matrices basically that are upper and lower triangular, respectively. Considering this observation, the most natural alternative for the non-symmetric plus/minus case seems to be the *LU*-factorization concept.

Definition 10.1 (*general LU-factorization*)
LU factorization expresses any square matrix A as the product of a permutation of a lower triangular matrix and an upper triangular matrix,

$$A = LU$$

where L is a permutation of a lower triangular matrix with ones on its diagonal and U is an upper triangular matrix.

The permutations are necessary for theoretical reasons in the general case. For instance, the matrix

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

cannot be expressed as the product of triangular matrices without interchanging the two rows. However, the special band structure of the companion matrices can be exploited to show that the permutations are not necessary and the factorization can be expressed simply as a product of a lower and an upper triangular matrix.

Lemma 10.1
Given a scalar discrete-time two-sided polynomial $p(z)$ with roots not lying on the unit circle, its companion matrix can be factored in the form $T_p = LU$, where L and U are lower and upper triangular matrices, respectively.

Proof
If a (possibly two-sided) polynomial p is non-zero at the unit circle then the principal minors of its companion matrix are known to be non-zero, see the reasoning in [19]. Further, according to [22], Theorem 3.2.1, a matrix A has the desired lower–upper triangular factorization if its all principal minors are non-zero. Combining these two observations, we arrive at the statement of the lemma. □

Following Lemma 10.1, a new algorithm for polynomial plus/minus factorization is suggested in the next subsection.

10.2. Plus/minus factorization algorithm

Given a (scalar, one-sided) polynomial

$$p(z) = p_0 + p_1 z + \cdots + p_d z^d$$

non-zero for $|z| = 1$, we first apply a direct degree shift to arrive at a two-sided polynomial

$$\tilde{p}(z) = p_0 z^{-\delta} + \cdots + p_d z^{d-\delta}$$

DOI: 10.1002/rnc

where δ is the number of roots of $p(z)$ lying inside the unit circle. Now, instead of solving Equation (1), we look for $\tilde{p}^+(z) = \tilde{p}_0^+ + \tilde{p}_1^+ z^{-1} + \cdots + \tilde{p}_\delta^+ z^{-\delta}$ and $\tilde{p}^-(z) = \tilde{p}_0^- + \tilde{p}_1^- z + \cdots + \tilde{p}_{d-\delta}^- z^{d-\delta}$ such that

$$\tilde{p}(z) = \tilde{p}^+(z)\tilde{p}^-(z) \tag{6}$$

Relationship between the pairs $\tilde{p}^+, \tilde{p}^-$ and p^+, p^- are obvious.

Having composed the companion matrix $T_{\tilde{p}}^N$ of sufficiently high order N, its LU factorization is performed. An approximation to the plus and minus factors of $\tilde{p}$ can then be read from the last column of the L and U factors, respectively, similarly to the spectral factorization case.

The degree shift yielding the two-sided polynomial $\tilde{p}$ is necessary to assure the correct decomposition of $\tilde{p}$ into stable and antistable parts. If the shift were not performed or were different from δ, the decomposition would still work in principle, however, the strict stability and antistability of particular factors would be lost.

Detailed description of the resulting algorithm follows.

Algorithm 2: Scalar discrete-time plus/minus factorization

Input: Scalar polynomial
$p(z) = p_0 + p_1 z + \cdots + p_d z^d$, non-zero for $|z| = 1$.

Output: Polynomials $p^+(z)$ and $p^-(z)$, the plus and minus factors of $p(z)$.

Step 1: *Choice of the companion matrix size*. Decide about the number N. N approximately 10–50 times larger than d is recommended up to our practical experience.

Step 2: *Degree shift*. Find out the number δ of zeros of $p(z)$ inside the unit disc. A modification of the well-known Schur stability criterion can be employed, see [13] for instance.

Having δ at hand, construct a two-sided polynomial $\tilde{p}(z)$ as

$$\begin{aligned}\tilde{p}(z) &= p(z)z^{-\delta} = p_0 z^{-\delta} + \cdots + p_d z^{d-\delta}\\ &= \tilde{p}_{-\delta} z^{-\delta} + \cdots + \tilde{p}_0 + \cdots + \tilde{p}_{d-\delta} z^{d-\delta}\end{aligned}$$

Step 3: *Construction of $T_{\tilde{p}}^N$*. Following the Section 4.8.1, construct the Sylvester companion matrix related to $\tilde{p}$ of order N.

Step 4: *LU decomposition of $T_{\tilde{p}}^N$*. Perform the LU decomposition of $T_{\tilde{p}}^N$:

$$T_{\tilde{p}}^N = LU$$

L and U are lower and upper triangular matrices, respectively.

Step 5: *Construction of polynomial factors*. Columns of the L and U matrices contains a non-zero vector l, u of length $\delta + 1$ and $d - \delta + 1$ lying under and above the main diagonal, respectively. Take the last full column $l = [l_0, l_1, \ldots, l_\delta]$ to create the plus factor of $p(z)$ as

$$p^+(z) = l_0 + l_1 z + \cdots + l_\delta z^\delta$$

The minus factor is constructed in a similar way using the last vector u.

11. EXAMPLE

To illustrate the usefulness of polynomial plus/minus factorization and to demonstrate the power of the proposed algorithm at the same time, we will discuss the l_1 optimal control problem.

DOI: 10.1002/rnc

l_1 optimization is a modern design technique, see [23] for a survey. The design goal lies in minimizing the l_1 norm of a closed-loop transfer function. Such a way, the magnitude of measured output signal is minimized with respect to bounded, yet persistent input disturbances. l_1 optimal controllers have already found an application in some irrigation channel regulation problem, see [24] for instance.

Quite recently a new method has been suggested by Z. Hurák *et al.* for the computation of an l_1 optimal discrete-time single-input single-output compensator, see [4]. Unlike their predecessors, the authors rely on the transfer function description purely and carefully exploit the algebraic structure of the problem. The resulting algorithm is given in [4] along with the following example.

Let us compute a feedback controller that minimizes ℓ_1 norm of the sensitivity function for a plant described by

$$G(z^{-1}) = \frac{b(z)}{a(z)} = \frac{-45 - 132z^{-1} + 9z^{-2}}{-20 - 48z^{-1} + 5z^{-2}}$$

The solution consists of the following computational steps:

1. plus/minus factorization of $a(z^{-1}) = a^+(z^{-1})a^-(z^{-1})$ and $b(z^{-1}) = b^+(z^{-1})b^-(z^{-1})$.
2. Find the minimum degree solution to $a(z^{-1})x_0(z^{-1}) + b(z^{-1})y_0(z^{-1}) = 1$.
3. Find a solution to $a^-(z^{-1})b^-(z^{-1})x(z^{-1}) + y(z^{-1}) = a(z^{-1})x_0(z^{-1})$ of given degree of $y(z^{-1})$ and with minimum $\|\,.\,\|_1$ norm.
4. The optimal controller is given by

$$C(z^{-1}) = \frac{a^+(z^{-1})b^+(z^{-1})y_0(z^{-1}) + a(z^{-1})x(z^{-1})}{a^+(z^{-1})b^+(z^{-1})x_0(z^{-1}) - b(z^{-1})x(z^{-1})}$$

The first step can be efficiently and reliably performed using Algorithm 2. We take small-size Sylvester matrices first for illustrative purposes, say N equal to 4. T_a and T_b read, respectively

$$T_a = \begin{bmatrix} -48 & 5 & 0 & 0 \\ -20 & -48 & 5 & 0 \\ 0 & -20 & -48 & 5 \\ 0 & 0 & -20 & -48 \end{bmatrix}$$

$$T_b = \begin{bmatrix} -132 & -45 & 0 & 0 \\ 9 & -132 & -45 & 0 \\ 0 & 9 & -132 & -45 \\ 0 & 0 & 9 & -132 \end{bmatrix}$$

DOI: 10.1002/rnc

and their LU factorization gives rise to

$$T_a^+ = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0.4167 & 1 & 0 & 0 \\ 0 & 0.3993 & 1 & 0 \\ 0 & 0 & 0.4003 & 1 \end{bmatrix}$$

$$T_a^- = \begin{bmatrix} -48 & 5 & 0 & 0 \\ 0 & -50.083 & 5 & 0 \\ 0 & 0 & -49.997 & 5 \\ 0 & 0 & 0 & -50 \end{bmatrix}$$

and

$$T_b^+ = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -0.06818 & 1 & 0 & 0 \\ 0 & -0.06663 & 1 & 0 \\ 0 & 0 & -0.06667 & 1 \end{bmatrix}$$

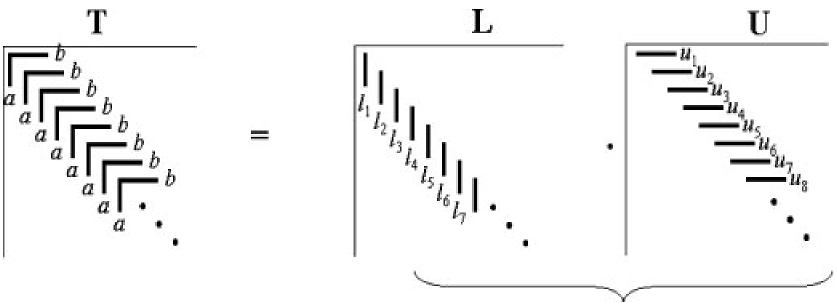
$$T_b^- = \begin{bmatrix} -132 & -45 & 0 & 0 \\ 0 & -135.1 & -45 & 0 \\ 0 & 0 & -135 & -45 \\ 0 & 0 & 0 & -135 \end{bmatrix}$$

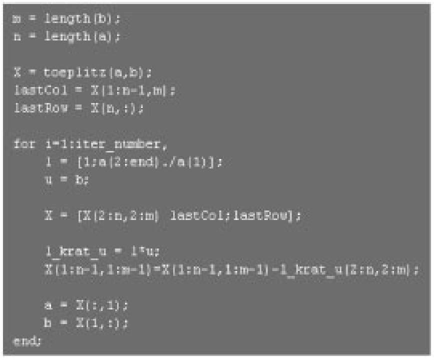
These matrix factors give a fair approximation to a^+, a^-, b^+, b^- polynomials:

$$a^+ = 0.40003z^{-1} + 1, \quad a^- = -49.997z^{-1} + 5$$
$$b^+ = -0.067z^{-1} + 1, \quad b^- = -135z^{-1} - 45$$

To get more accurate results, N is increased. Taking $N = 20$ yields perfectly accurate results,

$$a^+ = 2/5z^{-1} + 1, \quad a^- = -50z^{-1} + 5$$
$$b^+ = -1/15z^{-1} + 1, \quad b^- = -135z^{-1} - 45$$





```
m = length(b);
n = length(a);

X = toeplitz(a,b);
lastCol = X(1:n-1,m);
lastRow = X(n,:);

for i=1:iter_number,
    l = [1;a(2:end)./a(1)];
    u = b;

    X = [X(2:n,2:m) lastCol;lastRow];

    l_krat_u = l*u;
    X(1:n-1,1:m-1)=X(1:n-1,1:m-1)-l_krat_u(2:n,2:m);

    a = X(:,1);
    b = X(1,:);
end;
```

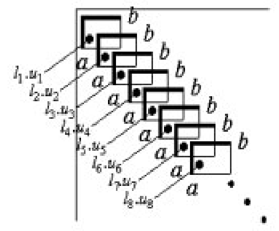


Figure 3. Algorithm diagram.

12. *LU* FACTORIZATION OF TOEPLITZ MATRICES

The *LU* decomposition can be performed *via* standard routines, see [22] for instance, implemented in standard packages such as LAPACK or commercial MATLAB. Nevertheless, thanks to the strong structurality of involved Toeplitz matrices, dedicated efficient routines for their *LU* factorization can be developed.

We assume the *L* and *U* factors in special forms depicted in Figure 3. Analysing the product *LU*, the procedure given in the figure in the form of a MATLAB pseudocode can be developed to receive subsequent iterations of *l* and *u* vectors.

13. CONCLUSIONS

A new method for the discrete-time plus/minus factorization problem in the scalar case has been proposed. The new method relies on numerically stable and efficient FFT algorithm. Besides its good numerical properties, the derivation of the routine also provides an interesting view into the related mathematics, combining the results of the theory of functions of complex variable, the theory of sampled signals, and the discrete Fourier transform techniques. The suggested method is employed in a practical application of improving the quality of a hi-fi system.

Encouraged by the success in modifying a spectral factorization algorithm for the plus/minus factorization case, we decided to re-visit another classical spectral factorization routine, namely

DOI: 10.1002/rnc

the Bauer's method of the 1950s. This idea has proved fruitful and our efforts resulted in another plus/minus factorization routine. As a by-product, a recursive LU factorization procedure for Toeplitz matrices has been developed that has a more general impact and can be of use in other areas of applied mathematics as well. Performance of the method was demonstrated by an l_1 optimal control system design example.

REFERENCES

1. Kučera V. *Analysis and Design of Discrete Linear Control Systems*. Academia: Prague, 1991.
2. Sternad M, Ahlén A. Robust filtering and feedforward control based on probabilistic descriptions of model errors. *Automatica* 1993; **29**:661–679.
3. Ohrn K, Ahlén A, Sternad M. A probabilistic approach to multivariable robust filtering and open-loop control. *IEEE Transactions on Automatic Control* 1995; **40**:405–417.
4. Hurák Z, Böttcher A, Šebek M. Minimum distance to the range of a lower triangular Toeplitz operator in l_1 norm and application in l_1 optimal control. *SIAM Journal on Control and Optimization* 2006; **45**(1):107–122.
5. Higham NJ. *Accuracy and Stability of Numerical Algorithms*. SIAM: Philadelphia, PA, 1996.
6. Ježek J, Kučera V. Efficient algorithm for matrix spectral factorization. *Automatica* 1985; **29**:663–669.
7. Hromcik M, Jezek J, Sebek M. New algorithm for spectral factorization and its practical application. *Proceedings of the European Control Conference ECC'2001*, Porto, Portugal, 1–5 September 2001.
8. Ljung L. *System Identification*: *Theory for the User*. Prentice-Hall Information and Systems Sciences Series. Prentice-Hall: Englewood Cliffs, NJ, 1987.
9. Hromčík M, Šebek M. Numerical and symbolic computation of polynomial matrix determinant. *Proceedings of the 38th Conference on Decision and Control CDC'99*, Phoenix, AZ, U.S.A., 7–10 December 1999.
10. Hromčík M, Šebek M. Fast Fourier transform and robustness analysis with respect to parametric uncertainties. *Proceedings of the 3rd IFAC Symposium on Robust Control Design ROCOND 2000*, Prague, CZ, 21–23 June 2000.
11. Bini D, Pan V. *Polynomial and Matrix Computations*, *Volume 1*: *Fundamental Algorithms*. Birkhäuser: Boston, 1994.
12. Needham T. *Visual Complex Analysis*. Oxford University Press: Oxford, 1997.
13. Barnett S. *Polynomials and Linear Control*. Marcel Dekker: New York, Basel, 1983.
14. Sternad M, Johansson M, Rutstrom J. Inversion of loudspeaker dynamics by polynomial LQ feedforward control. *Proceedings of the 3rd IFAC Symposium on Robust Control Design ROCOND 2000*, Prague, CZ, 21–23 June 2000.
15. University of Uppsala, Signals and Systems Department. *Adaptive Signal Processing*, *Course Homepage*, http://www.signal.uu.se/Courses/CourseDirs/AdaptSignTF/Adapt00.html
16. Kwakernaak H, Šebek M. *PolyX Home Page*, http://www.polyx.cz/, http://www.polyx.com/
17. Bauer FL. Ein direktes iterations verfahren zur Hurwitz-zerlegung eines polynoms (in German). *Archiv der Elektrischen Uebertragung* 1955; **9**:285–290.
18. Bauer FL. Beitrage zur entwicklung numerischer verfahren fur programmgesteuerte rechenanlagen, II. Direkte faktorisierung eines polynoms (in German). *Sitzungsherichte Bayerishen Akademie Wissenschaften* 1956; 163–203.
19. Wu SP, Boyd S, Vandenberghe L. FIR filter design via spectral factorization and convex optimization. In *Applied Computational Control*, *Signal and Communications*, Datta B (ed.). Birkhauser: Basel, 1997.
20. Ježek J, Kučera V. Efficient algorithm for matrix spectral factorization. *Automatica* 1985; **29**:663–669.
21. Hromčík M, Ježek J, Šebek M. New algorithm for spectral factorization and its practical application. *6th European Control Conference ECC 2001*, Porto, Portugal, 4–7 September 2001; 3104–3109.
22. Golub GH, Van Loan CF. *Matrix Computations*. The Johns Hopkins University Press: Baltimore, London, 1990.
23. Dahleh MA, Diaz-Bobillo IJ. *Control of Uncertain Systems*: *A Linear Programming Approach*. Prentice-Hall: Englewood Cliffs, NJ, 1995.
24. Malaterre P-O, Khammash M. l-1 Controller design for a high-order 5-pool irrigation canal system. *IEEE-CDC Conference*, Sydney, Australia, December 2000.

DOI: 10.1002/rnc

Paper B.

Kujan, P. - Hromčík, M. - Šebek, M.: Complete Fast Analytical Solution of the Optimal Odd Single-Phase Multilevel Problem. *IEEE Transactions on Industrial Electronics*. 2010, vol. 2010, no. 57(7), p. 2382-2397.
ISSN 0278-0046. *(IF 3.439)*

Complete Fast Analytical Solution of the Optimal Odd Single-Phase Multilevel Problem

Petr Kujan, *Member, IEEE*, Martin Hromčík, and Michael Šebek, *Senior Member, IEEE*

***Abstract*—In this paper, we focus on the computation of optimal switching angles for general multilevel (ML) odd symmetry waveforms. We show that this problem is similar to (but more general than) the optimal pulsewidth modulation (PWM) problem, which is an established method of generating PWM waveforms with low baseband distortion. We introduce a new general modulation strategy for ML inverters, which takes an analytic form and is very fast, with a complexity of only $\mathcal{O}(n \log^2 n)$ arithmetic operations, where n is the number of controlled harmonics. This algorithm is based on a transformation of appropriate trigonometric equations for each controlled harmonics to a polynomial system of equations that is further transformed to a special system of composite sum of powers. The solution of this system is carried out by a modification of the Newton's identity via Padé approximation, formal orthogonal polynomials (FOPs) theory, and properties of symmetric polynomials. Finally, the optimal switching sequence is obtained by computing zeros of two FOP polynomials in one variable or, alternatively, by a special recurrence formula and eigenvalues computation.**

***Index Terms*—Composite sum of powers, formal orthogonal polynomials (FOPs), multilevel (ML) inverters, Newton's identities, optimal pulsewidth modulation (PWM) problem, Padé approximation, polynomial methods, selected harmonics elimination.**

I. Introduction

The optimal multilevel (ML) or pulsewidth modulation (PWM) problem, sometimes called the selected harmonic elimination (SHE) problem, is an established method for generating ML waveforms with low baseband distortion. The principal problem is to determine the switching times (angles) to produce the baseband and to not generate specific higher order harmonics. This way, it is possible to separate the undesirable highest harmonics.

The optimal ML problem offers several advantages compared to traditional modulation methods [1]–[4]. This approach allows better performance with low switching frequency, direct control over output waveform harmonics, and the ability to leave untouched harmonics divisible by three for three-phase systems.

Up to now, a lot of different perspectives were proposed. All the methods assume quarter symmetry, and all formulations result in the Fourier series representation for different waveforms. The principal problem lies in solving a multivariate trigonometric system of equations or, after substitution for Chebyshev polynomials, in solving a multivariate polynomial system of equations. There are several techniques of how to solve them.

The most effective method for single-phase quarter-symmetric inverter is described in [5]–[7]. This method is based on trigonometric identity for cosine function where the original trigonometric system is transformed to a polynomial system of specific structure leading to the polynomial system of sum of odd powers. The problem results in the construction of a special set of one variable polynomials and computation of their zeros. These polynomials are formal orthogonal (FOPs), and a recurrence formula is derived for them. The solution is based on diagonal Padé approximation. In the case of single-phase inverter for a given modulation index,[1] one or no solution exists. An exact algorithm with a small complexity $\mathcal{O}(n \log^2 n)$ was found. The main result of this paper is, in fact, a generalization of this work for general odd symmetry ML waveforms.

Three-phase inverter systems pose a very interesting topic with many industrial applications. In the three-phase connection, all harmonics divisible by three are ignored as they are automatically canceled in the electric system. This is a more complicated problem because a special structure of the system of equations is damaged. One unique, several different, or no solution exists for a given modulation index. From these, only one solution is selected—the one that minimizes other undesirable and uncontrolled higher harmonics. For more details, see [8]–[11]. These papers also rely on the conversion to a system of polynomials using trigonometric identities. This system of polynomials is solved by the Gröbner basis theory or by the elimination method based on computation of resultants [10]. In addition, a substitution for elementary symmetric polynomials or power sums is applied in [9] and [10]. Applicability of this method is restricted to say five odd harmonics because an appropriate system for a higher number of eliminated harmonics is too large, and its solution is extraordinarily time consuming. Nevertheless, fast analytical methods similar to the algorithms

Manuscript received April 24, 2009; revised July 20, 2009 and September 10, 2009; accepted October 5, 2009. Date of publication October 20, 2009; date of current version June 11, 2010. The work of P. Kujan was supported by the Grant Agency of the Czech Republic under Grant P103/10/P323. The work of M. Hromčík was supported by the Ministry of Education of the Czech Republic under Project 1M0567. The work of M. Šebek was supported by the Grant Agency of the Czech Republic under Grant 102/08/0186.

P. Kujan and M. Šebek are with the Department of Control Engineering, Faculty of Electrical Engineering, Czech Technical University in Prague, 16627 Prague 6, Czech Republic, and also with the Department of Control Theory, Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, 18208 Prague 8, Czech Republic (e-mail: kujanp@fel.cvut.cz).

M. Hromčík is with the Center for Applied Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague, 16627 Prague 6, Czech Republic, and also with the Department of Control Theory, Institute of Information Theory and Automation of the Academy of Sciences of the Czech Republic, 18208 Prague 8, Czech Republic.

Digital Object Identifier 10.1109/TIE.2009.2034677

[1] This is basically the ratio of the first harmonic to the amplitude of one level of an ML waveform.

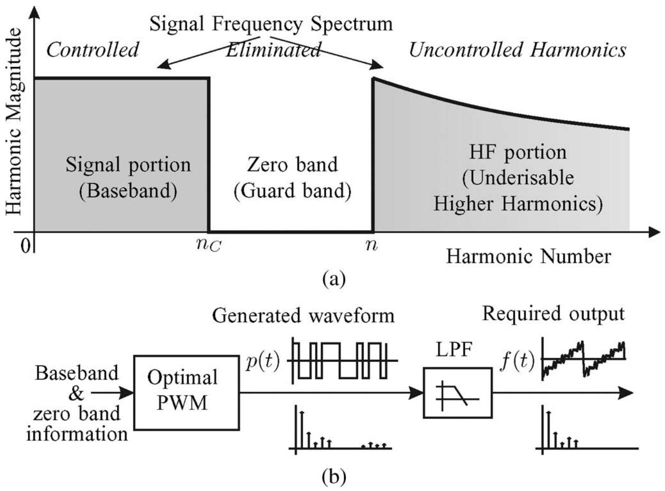


Fig. 1. (a) Frequency spectrum of a separated baseband signal. The baseband can be recovered by an LPF. (b) Principal scheme for the optimal PWM or ML problem.

for single-phase systems presented further in this paper seem to appear soon (see [12] for some first results).

Other methods presented in the literature dealing with the system of polynomial equations are numerical iterative routines [13], genetic algorithms [14], optimization theory [15]–[18], homotopy and continuation [19], or a predictive control algorithm [20].

Applications of the optimal ML or PWM problem cover the control of large electric drives, power electronics converters, active harmonic filters, control of (micro) electromechanical systems, or digital audio amplifier. Implementation of fast and efficient algorithms proposed in this paper on dedicated hardware, e.g., digital signal processors, opens a possibility of a more effective on-the-fly realizations and more accurate and faster solutions. It can result in increasing fuel or power efficiency and better performance (see [21]).

II. Optimal ML Problem

A key issue in the optimal ML problem is the determination of the switching times (angles) to produce the signal portion (baseband) and to not generate specific higher order harmonics (guard band or zero band). This spectral gap separates the baseband, which has to be identical to the required output waveform, from an uncontrolled higher frequency portion. The required output signal can be recovered by means of an analog low-pass filter (LPF) with a cutoff frequency in the guard band. The procedure is depicted in Fig. 1.

Methods described in this section are based on exploiting appropriate trigonometric transcendental equations that define the harmonic content of the generated periodic ML waveform $p(t)$, which is equal to the required finite frequency spectrum of $f(t)$. The main problem lies in solving these systems of equations.

The solution of the optimal ML problem is a sequence of switching times $\overline{\alpha}^\star = (\alpha_1, \ldots, \alpha_n)$. This sequence is obtained from the solution of the following system of equations:

$$a_{p_0}(\overline{\alpha}) = a_{f_0} \tag{1a}$$

$$\left.\begin{aligned} a_{p_k}(\overline{\alpha}) &= a_{f_k} \\ b_{p_k}(\overline{\alpha}) &= b_{f_k} \end{aligned}\right\} \text{ for all } \quad k \in \mathcal{H}_C \tag{1b}$$

$$\left.\begin{aligned} a_{p_k}(\overline{\alpha}) &= 0 \\ b_{p_k}(\overline{\alpha}) &= 0 \end{aligned}\right\} \text{ for all } \quad k \in \mathcal{H}_E \tag{1c}$$

$$\text{subject to} \quad 0 < \alpha_i < T \tag{1d}$$

where $\overline{\alpha} = (\alpha_1, \ldots, \alpha_n)$ are unknown variables, a_{p_0} and a_{p_k}, b_{p_k} are the zeroth and kth cosine and sine Fourier coefficients of the generated waveform $p(t)$, respectively, and a_{f_0} and a_{f_k}, b_{f_k} are the zeroth and kth cosine and sine Fourier coefficients of the required output waveform $f(t)$. $\mathcal{H}_C$ is the set of controlled harmonics, and the number of elements is n_C. $\mathcal{H}_E$ is the set of eliminated harmonics, and the number of elements is n_E. The number of equations is $n = 1 + 2(n_C + n_E)$.

If only one solution $\overline{\alpha}$ of (1) exists, then it is the optimal solution, and $\overline{\alpha}^\star = \overline{\alpha}$. If the solutions of (1) are $\overline{\alpha}_1, \ldots, \overline{\alpha}_m$, $m > 1$, then the optimal solution $\overline{\alpha}^\star$ is chosen as the minimizer of the total harmonic distortion (THD), i.e.,

$$\overline{\alpha}^\star = \arg \min_{\overline{\alpha} = \{\overline{\alpha}_1, \ldots, \overline{\alpha}_m\}} \text{THD}(\overline{\alpha}) \tag{2}$$

where

$$\text{THD}(\overline{\alpha}) \text{ (in percent)} = 100 \sqrt{\frac{\sum_{i=n_c+1}^{n+N} \left(\frac{a_{p_i}(\overline{\alpha}) + b_{p_i}(\overline{\alpha})}{i}\right)^2}{\sum_{i=1}^{n_c} \left(\frac{a_{p_i}(\overline{\alpha}) + b_{p_i}(\overline{\alpha})}{i}\right)^2}}. \tag{3}$$

If no solution of (1) is found, then the optimal solution $\overline{\alpha}^\star$ is computed as a general minimization problem, i.e.,

$$\overline{\alpha}^\star = \arg \min_{\overline{\alpha}} \sqrt{\sum_{k \in \mathcal{H}_E} \left(a_{p_k}(\overline{\alpha}) + b_{p_k}(\overline{\alpha})\right)^2}$$

$$\text{subject to} \quad \text{(1a) and (1b).} \tag{4}$$

In the rest of this paper, we focus on single-phase odd ML and bilevel PWM waveforms, which lead to a special structure of (1), with only one solution satisfying the condition (1d). The solution of (1) is then found by an analytical procedure.

III. Switching Waveforms

We will show by analysis of different ML waveforms (general, odd, even, half-wave, quarter-wave and bilevel, three-level) that an effective (analytical) solution is possible for waveforms with odd and quarter-wave symmetry only. The Fourier series of these waveforms are odd and, therefore, contain sine coefficients only (the zeroth harmonic and cosine coefficients are equal to zero). The sine and cosine Fourier coefficients are included in other cases, and therefore, it is not possible to make simplifying arrangements for an effective solution.

The optimal PWM problem for a quarter-symmetric three-level inverter is solved in [5] and [6]. This waveform generates only odd sine harmonics, and only the first harmonic is controlled. In this paper, we present solutions for more general odd ML waveforms, which generate all (odd as well as even)

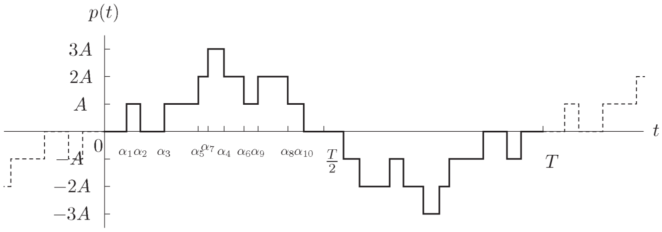


Fig. 2. General odd multilevel (seven-level) waveform.

sine harmonics. Therefore, our approach covers the solution of the quarter-symmetric PWM and ML problem, and it is more general. Furthermore, the first few n_c harmonics are controlled.

A. General Odd ML Waveform

The Fourier series of a T periodic general odd ML waveform $p(t)$ with amplitude A (see Fig. 2) is sine, i.e.,

$$p(t) \sim \sum_{k=1}^{\infty} b_k \sin \omega k t \tag{5}$$

where

$$b_k = \frac{2A}{k\pi}\left((-1)^{k+1} o_n - \sum_{i=1}^{n}(-1)^i \cos \omega k \alpha_i\right), \qquad k = 1, 2, 3, \ldots. \tag{6}$$

The unknown switching times $\overline{\alpha} = (\alpha_1, \ldots, \alpha_n)$ are subject to $0 < \alpha_1 < \alpha_3 < \cdots < \alpha_{2\lceil n/2 \rceil - 1} < T/2$ ($\lceil n/2 \rceil$ rising edges) and $0 < \alpha_2 < \alpha_4 < \cdots < \alpha_{2\lfloor n/2 \rfloor} < T/2$ ($\lfloor n/2 \rfloor$ falling edges), and $\omega = 2\pi/T$ is the angular frequency. The integer n is the number of switching times in the half period, and o_n is the odd parity test described by

$$o_n = \frac{1-(-1)^n}{2} = \begin{cases} 0, & \text{for even } n, \\ 1, & \text{for odd } n. \end{cases} \tag{7}$$

The number of levels is equal to

$$2 \max_{i=1,\ldots,n} |\Lambda_i| + 1 \tag{8}$$

where

$$\Lambda_1 = M(a_1) \quad \Lambda_{i+1} = \Lambda_i + M(a_{i+1}) \quad i=1,\ldots,n-1$$
$$(a_1, \ldots, a_n) = \text{sort}_<(\alpha_1, \alpha_2, \ldots, \alpha_n) \tag{9}$$
$$M(a_i) = \begin{cases} 1, & a_i \in \alpha_{2j-1} \\ -1, & a_i \in \alpha_{2j}. \end{cases} \tag{10}$$

In the following, we describe some special cases:

1) *proper odd ML waveform:* $(2\lceil n/2 \rceil + 1)$-level waveform with n switching times in the half period, satisfying the condition $\alpha_{2\lceil n/2 \rceil - 1} < \alpha_2$ (see Fig. 3);
2) *proper three-level waveform:* only 0 and $+A$ levels in the half period, satisfying the condition $0 < \alpha_1 < \alpha_2 < \alpha_3 < \cdots < \alpha_n$ (see Fig. 4);

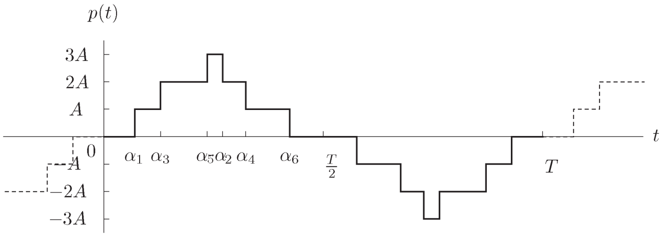


Fig. 3. Odd proper multilevel waveform.

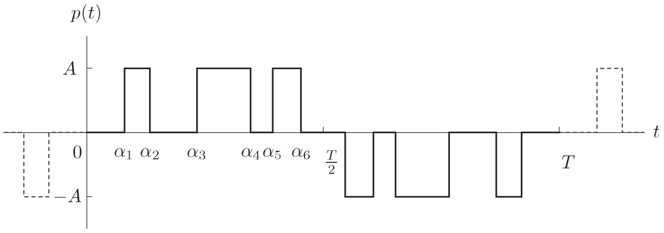


Fig. 4. Odd proper three-level waveform.

3) *bilevel waveform:* has a slightly different Fourier series expansion and is therefore described separately in Section III-B.

For the sequel, we put $T = 2\pi$ and $\omega = 1$ for simplicity. Then, all solutions α_i are transformed back to the original period by a substitution $\alpha_i \mapsto \alpha_i T/(2\pi)$.

For further generalization and simplification of the notation, we introduce (6) and (27) for the bilevel waveform (see Section III-B) in the following form:

$$b_k(\overline{\alpha}) = A_k\left(B_k + C_k \sum_{i=1}^{n}(-1)^i \cos(k\alpha_i)\right), \qquad k = 1, 2, \ldots. \tag{11}$$

The parameters for 2π periodic odd ML waveform are

$$A_k = \frac{2A}{k\pi} \quad B_k = (-1)^{k+1} o_n \quad C_k = -1. \tag{12}$$

According to the previous analysis of the optimal ML problem then, for a single-phase system, the controlled harmonics of the output ML waveform $p(t)$ are b_{p_k}, $k \in \mathcal{H}_C = \{1, 2, \ldots, n_C\}$, and the eliminated harmonics are b_{p_k}, $k \in \mathcal{H}_E = \{n_C + 1, n_C + 2, \ldots, n_C + n_E\}$. Thus, we have

$$b_{p_k}(\overline{\alpha}) = A_k\left(B_k + C_k \sum_{i=1}^{n}(-1)^i \cos(k\alpha_i)\right) = b_{f_k}, \qquad k = 1, 2, \ldots, n_c \tag{13a}$$

$$b_{p_k}(\overline{\alpha}) = B_k + C_k \sum_{i=1}^{n}(-1)^i \cos(k\alpha_i) = 0, \qquad k = n_c + 1, n_c + 2, \ldots, n \tag{13b}$$

subject to

$$0 < \alpha_1 < \alpha_3 < \cdots < \alpha_{2\lceil n/2 \rceil - 1} < \pi \tag{13c}$$

$$0 < \alpha_2 < \alpha_4 < \cdots < \alpha_{2\lfloor n/2 \rfloor} < \pi \tag{13d}$$

where $\overline{\alpha} = (\alpha_1, \ldots, \alpha_n)$ are unknown variables (switching times), $n = n_C + n_E$, A_k, B_k, and C_k are set according to (12), and b_{f_k}, $k = 1, 2, \ldots, n_C$, on the right-hand side (RHS) of the equations are real numbers defining the required signal $f(t)$ (baseband frequency spectrum). The integer n_E defines the number of zero harmonics in the guard band.

1) Polynomial Equations: In this section, we convert the trigonometric equations in (13) to polynomial equations and simplify them. According to the trigonometric identity for multiple angles of cosine

$$\cos(k\alpha_i) = T_k(\cos\alpha_i). \tag{14}$$

We substitute by Chebyshev polynomial T_k of the first kind (see, e.g., [22, p.771] or [5]) and convert the kth harmonic of (11) to multivariate polynomials, i.e.,

$$b_{p_k}(\overline{x}) = A_k\left(B_k + C_k\sum_{i=1}^{n}(-1)^i T_k(x_i)\right) \tag{15}$$

in variables $(x_1, \ldots, x_n) = \overline{x}$. The dependence between x_i and α_i is given by

$$\alpha_i = \arccos x_i, \qquad i = 1, \ldots, n. \tag{16}$$

According to (13c) and (13d)

$$\begin{aligned} &-1 < x_n < \cdots < x_4 < x_2 < 1 \\ &-1 < x_{n-1} < \cdots < x_3 < x_1 < 1. \end{aligned} \tag{17}$$

Thus, the trigonometric system (13) is transformed to a polynomial system, i.e.,

$$b_{p_k}(\overline{x}) = A_k\left(B_k + C_k\sum_{i=1}^{n}(-1)^i T_k(x_i)\right) = b_{f_k}, \qquad k = 1, 2, \ldots, n_c \tag{18a}$$

$$b_{p_k}(\overline{x}) = B_k + C_k\sum_{i=1}^{n}(-1)^i T_k(x_i) = 0, \qquad k = n_c + 1, n_c + 2, \ldots, n \qquad \text{subject to (17)} \tag{18b}$$

where the variables are $(x_1, \ldots, x_n) = \overline{x}$. This polynomial system (18) can be re-solved using existing methods, such as the Gröbner basis approach, elimination based on resultants, and other algorithms (see [23] and [24]). Note that the polynomials in this system are partially symmetric. It means that we can arbitrarily permutate variables x_{2i} or x_{2i-1} and the function $b_{p_k}(\overline{x})$ is left unchanged.

However, the following steps show how the system of equations in (18) [respectively (15)] can be further simplified by conversion to a new linear system in new variables. These new variables are composite sums of powers and create new polynomial system of equations. We present new effective algorithm for this system, which is much more effective compared to direct application of standard polynomial methods to (18).

From (15), the expression $\sum_{i=1}^{n}(-1)^i T_k(x_i)$ for odd k reads

$$\begin{aligned} \sum_{i=1}^{n}(-1)^i T_k(x_i) &= -\sum_{j=1}^{\frac{k+1}{2}} t_{k,2j-1}\sum_{i=1}^{n}(-1)^{i+1}x_i^{2j-1} \\ &= -\sum_{j=1}^{\frac{k+1}{2}} t_{k,2j-1}p_{2j-1}, \qquad k \text{ is odd} \end{aligned}$$

where $t_{k,2j-1}$ is the $(2j-1)$th coefficient of x^{2j-1} in the Chebyshev polynomial of degree k, and p_{2j-1} are composite sums of powers (new unknown variables) for which the following identity holds:

$$\begin{aligned} p_{2j-1} &= \sum_{i=1}^{n}(-1)^{i+1}x_i^{2j-1} \\ &= x_1^{2j-1} - x_2^{2j-1} + \cdots + (-1)^{n+1}x_n^{2j-1}, \\ &\quad j = 1, 2, \ldots. \end{aligned} \tag{19}$$

Then, one can write (15) in the following form:

$$\begin{aligned} &b_{p_{2i-1}}(p_1, p_3, \ldots, p_{2i-1}) \\ &\quad = A_{2i-1}\left(B_{2i-1} - C_{2i-1}\sum_{j=1}^{i} t_{2i-1,2j-1}p_{2j-1}\right), \\ &\quad i = 1, \ldots, \lceil n/2 \rceil. \end{aligned} \tag{20}$$

Similarly, for even k, we have

$$\begin{aligned} &b_{p_{2i}}(p_2, p_4, \ldots, p_{2i}) \\ &\quad = A_{2i}\left[B_{2i} - C_{2i}\left((-1)^i o_n + \sum_{j=1}^{i} t_{2i,2j}p_{2j}\right)\right], \\ &\quad i = 1, \ldots, \lfloor n/2 \rfloor \end{aligned} \tag{21}$$

where

$$\begin{aligned} p_{2j} &= \sum_{i=1}^{n}(-1)^{i+1}x_i^{2j} \\ &= x_1^{2j} - x_2^{2j} + \cdots + (-1)^{n+1}x_n^{2j}, \qquad j = 1, 2, \ldots. \end{aligned} \tag{22}$$

Finally, we apply back substitution to (18) having the following polynomial system of equations:

$$\begin{aligned} b_{p_{2i-1}}(\overline{p}) &= A_{2i-1}\left(B_{2i-1} - C_{2i-1}\sum_{j=1}^{i} t_{2i-1,2j-1}p_{2j-1}\right) \\ &= b_{f_{2i-1}}, \qquad i = 1, 2, \ldots, \left\lceil \frac{n_c}{2} \right\rceil \end{aligned} \tag{23a}$$

$$\begin{aligned} b_{p_{2i}}(\overline{p}) &= A_{2i}\left[B_{2i} - C_{2i}\left((-1)^i o_n + \sum_{j=1}^{i} t_{2i,2j}p_{2j}\right)\right] \\ &= b_{f_{2i}}, \qquad i = 1, 2, \ldots, \left\lfloor \frac{n_c}{2} \right\rfloor \end{aligned} \tag{23b}$$

$$b_{p_{2i-1}}(\overline{p}) = B_{2i-1} - C_{2i-1}\sum_{j=1}^{i} t_{2i-1,2j-1}p_{2j-1} = 0, \qquad i = \left\lceil \frac{n_c}{2} \right\rceil + 1, \ldots, \left\lceil \frac{n}{2} \right\rceil \tag{23c}$$

$$b_{p_{2i}}(\overline{p}) = B_{2i} - C_{2i}\left((-1)^i o_n + \sum_{j=1}^{i} t_{2i,2j}p_{2j}\right) = 0, \qquad i = \left\lfloor \frac{n_c}{2} \right\rfloor + 1, \ldots, \left\lfloor \frac{n}{2} \right\rfloor \tag{23d}$$

where $\overline{p} = (p_1, p_2, \ldots, p_n)$ are unknown variables. Because $\mathcal{H}_C = \{1, 2, \ldots, n_C\}$ and $\mathcal{H}_E = \{n_C + 1, n_C + 2, \ldots, n_C + n_E\}$, $n = n_C + n_E$, the previous system is linear and of n equations with n unknown variables $p_1, \ldots, p_n$. Now, if we separate unknowns $p_1, \ldots, p_n$ on the left-hand side (LHS) of (23), then the new RHS for b'_{f_i} are

$$-\frac{1}{C_{2i-1}}\left(\frac{b_{f_{2i-1}}}{A_{2i-1}} - B_{2i-1}\right)$$

$$-(-1)^i o_n - \frac{1}{C_{2i}}\left(\frac{b_{f_{2i}}}{A_{2i}} - B_{2i}\right) \quad \frac{B_{2i-1}}{C_{2i-1}}$$

$$-(-1)^i o_n + \frac{B_{2i}}{C_{2i}}.$$

The itemized form of (23) for an ML waveform [parameters A_i, B_i, and C_i are (12)] for $n = 6$ and $n_C = 3$ reads

$$\begin{bmatrix} t_{1,1} & & & & & \\ 0 & t_{2,2} & & & & 0 \\ t_{3,1} & 0 & t_{3,3} & & & \\ 0 & t_{4,2} & 0 & t_{4,4} & & \\ t_{5,1} & 0 & t_{5,3} & 0 & t_{5,5} & \\ 0 & t_{6,2} & 0 & t_{6,4} & 0 & t_{6,6} \end{bmatrix} \cdot \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \\ p_6 \end{bmatrix} = \begin{bmatrix} \frac{\pi b_{f_1}}{2A} \\ \frac{\pi b_{f_2}}{A} \\ \frac{3\pi b_{f_3}}{2A} \\ 0 \\ 0 \\ 0 \end{bmatrix}. \tag{24}$$

The system of equations in (23) is a special linear system where $t_{k,i}$ on the LHS are the ith coefficients of the k degree Chebyshev polynomial of the first kind, and on the RHS, there are b'_{f_i} and zeros. According to Gauss–Banachiewitz decomposition for orthogonal Chebyshev polynomials (for more details, see [25]), the solution of (23) for the general ML waveform is

$$p_{2i} = o_n + 2^{-2i+1}\frac{\pi}{A}\sum_{j=1}^{\underline{K}}\binom{2i}{i-j} j\, b_{f_{2j}}, \tag{25a}$$

$$\underline{K} := \begin{cases} i, & \ldots i < \lfloor n_c/2 \rfloor \\ \lfloor n_c/2 \rfloor, & \ldots i \geq \lfloor n_c/2 \rfloor \end{cases}; \qquad i = 1, 2, \ldots, \lfloor n/2 \rfloor \tag{25b}$$

$$p_{2i-1} = -\, o_n + 2^{-2i+1}\frac{\pi}{A}\sum_{j=1}^{\overline{K}}\binom{2i-1}{i-j}(2j-1)\, b_{f_{2j-1}}, \tag{25c}$$

$$\overline{K} := \begin{cases} i, & \ldots i < \lceil n_c/2 \rceil \\ \lceil n_c/2 \rceil, & \ldots i \geq \lceil n_c/2 \rceil \end{cases}; \qquad i = 1, 2, \ldots, \lceil n/2 \rceil. \tag{25d}$$

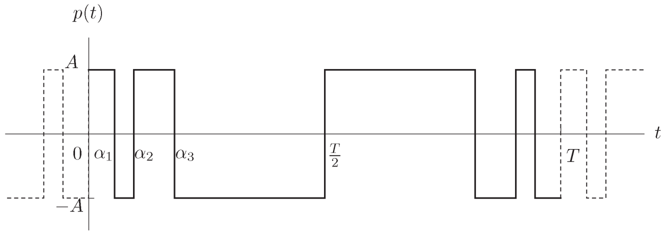


Fig. 5. Odd bilevel PWM waveform.

The number of operations is $\mathcal{O}(nn_C)$ only, instead of the standard recursive procedure for solution of triangular linear system (23), which takes $\mathcal{O}(n^2)$ operations, and it is moreover not necessary to generate and store in memory the coefficients of Chebyshev polynomials $t_{i,j}$. For example, in the converter problem, where $n_C = 1$ (only the first harmonic is controlled), the number of operations is linear compared to quadratic.

To sum up, the problem of optimal ML, namely, the solution of trigonometric system (13) or polynomial system (18), was converted to a more simple solution of system of composite sum of powers (19) and (22), which is in compact form, i.e.,

$$p_j = x_1^j - x_2^j + \cdots + (-1)^{n+1}x_n^j, \qquad j = 1, 2, \ldots, n \quad \text{subject to (17)} \tag{26}$$

where p_j are easily solved according to (25), and unknowns are $\overline{x} = (x_1, x_2, \ldots, x_n)$. The effective solution for this special polynomial system of composite sum of powers is described in Section IV. The unknown switching times α_i are then obtained according to (16).

B. Odd Bilevel PWM Waveform

The Fourier series of T periodic odd bilevel PWM waveform $p(t)$ with amplitude A (see Fig. 5) is sine with the following coefficients:

$$b_k = \frac{4A}{k\pi}\left(o_{n+k} + \sum_{i=1}^{n}(-1)^i \cos(\omega k \alpha_i)\right), \qquad k = 1, 2, \ldots, \tag{27}$$

where $0 < \alpha_1 < \alpha_2 < \cdots < \alpha_n < T/2$ are the unknown switching times.

The parameters according to (11) are

$$A_k = \frac{4A}{k\pi} \quad B_k = o_{n+k} \quad C_k = 1 \tag{28}$$

and the composite sum of powers is

$$p_{2i} = o_n - 2^{-2i}\frac{\pi}{A}\sum_{j=1}^{\underline{K}}\binom{2i}{i-j} j\, b_{f_{2j}}, \qquad i = 1, 2, \ldots, \lfloor n/2 \rfloor \tag{29a}$$

$$p_{2i-1} = o_{n+1} - 2^{-2i}\frac{\pi}{A}\sum_{j=1}^{\overline{K}}\binom{2i-1}{i-j}(2j-1)\, b_{f_{2j-1}}, \qquad i = 1, 2, \ldots, \lceil n/2 \rceil \tag{29b}$$

where $\overline{K}$ and $\underline{K}$ are according to (25d) and (25b). The inequality condition for variables x_i is

$$-1 < x_n < x_{n-1} < \cdots < x_2 < x_1 < 1. \tag{30}$$

IV. Composite Sum of Powers

As shown in Section III-A1, the solution of the optimal odd ML problem depends only on computation of the composite sum of powers (26). The itemized form is

$$\begin{aligned} &x_1 - x_2 + \cdots + (-1)^{n+1} x_n = p_1 && \text{(31a)}\\ &x_1^2 - x_2^2 + \cdots + (-1)^{n+1} x_n^2 = p_2 &&\\ &\qquad\qquad\vdots &&\\ &x_1^n - x_2^n + \cdots + (-1)^{n+1} x_n^n = p_n && \text{(31b)}\\ &\quad \text{subject to (17) for optimal ML problem or} &&\\ &\quad \text{subject to (30) for optimal bilevel PWM problem} &&\\ & && \text{(31c)} \end{aligned}$$

where the RHS are real numbers according to (25) for the general odd ML waveform, or (29) for the odd bilevel PWM waveform. Note that this system is very similar to standard power sums $\sum_{i=1}^{n} x_i^k = p_k$, $k = 1, \dots, n$, that are easily solvable by the Newton's identity (see [24] and [26]).

For the following steps, it is better to focus on the following configuration of the power sums:

$$p_j(y_1, \dots, y_n) = \sum_{i=1}^{k} y_i^j - \sum_{i=k+1}^{n} y_i^j, \qquad j = 1, \dots, n \tag{32}$$

where $k \leq \lfloor n/2 \rfloor$. When $k > \lfloor n/2 \rfloor$, we can multiply the equation system in (32) by -1 and convert it to the case $k < \lfloor n/2 \rfloor$. This form in (32) can be obtained by resorting variables in (31). The polynomials $p_j(y_1, y_2, \dots, y_n)$ in (32) are partially symmetric because the power sums $\sum_{i=1}^{k} y_i^j$ and $\sum_{i=k+1}^{n} y_i^j$ are symmetric polynomials (see [24]) in variables $\overline{y}^{\,+} = (y_1, \dots, y_k)$ and $\overline{y}^{\,-} = (y_{k+1}, \dots, y_n)$ separately. Then, we have

$$\begin{aligned} &p_j(y_1, \dots, y_k, y_{k+1}, \dots, y_n)\\ &\qquad = p_j\left(y_{\pi_1(1)}, \dots, y_{\pi_1(k)}, y_{\pi_2(k+1)}, \dots, y_{\pi_2(n)}\right) \end{aligned} \tag{33}$$

where $(y_{\pi_1(1)}, \dots, y_{\pi_1(k)})$ and $(y_{\pi_2(k+1)}, \dots, y_{\pi_2(n)})$ are arbitrary permutations of $\overline{y}^{\,+}$ and $\overline{y}^{\,-}$, respectively. Therefore, the total number of solutions is $k!(n-k)!$. All of them are combinations of two sets coming from permutations of elements of vectors $\overline{y}^{\,+}$ and $\overline{y}^{\,-}$.

Equation (31) is converted to (32) in the following way. If n is an even integer, then $n/2$ variables with positive sign and the same number with negative sign are in (31). Therefore, converting to (32) is accomplished by introducing the following new variables:

$$\overline{y}^{+} = (y_1, y_2, \dots, y_k) = (x_1, x_3, \dots, x_{2k-1}) \tag{34a}$$

$$\overline{y}^{-} = (y_{k+1}, y_{k+2}, \dots, y_{2k}) = (x_2, x_4, \dots, x_{2k}) \tag{34b}$$

where $k = n/2$. If n is odd, then $\lfloor n/2 \rfloor + 1$ variables with positive sign and $\lfloor n/2 \rfloor$ variables with negative sign are in (31). Therefore, conversion similar to the case with n even leads to $k > \lfloor n/2 \rfloor$, which is not in agreement with condition $k \leq \lfloor n/2 \rfloor$ of (32). Therefore, each equation in (31) must be multiplied by -1, and for that reason, the signs of RHS of (32) must be changed, i.e., $p_i \mapsto -p_i$. Then, the following substitution can be done:

$$\overline{y}^{\,+} = (y_1, y_2, \dots, y_k) = (x_2, x_4, \dots, x_{2k}) \tag{35a}$$

$$\overline{y}^{\,-} = (y_{k+1}, \dots, y_{2k+1}) = (x_1, x_3, \dots, x_{2k+1}) \tag{35b}$$

where $k = \lfloor n/2 \rfloor$.

The solution $x_1, \dots, x_n$ of the optimal odd ML problem is obtained as follows. From all solutions of (32), only one is chosen—the one that is in agreement with (31c), which means that all elements $\overline{y}^{\,+}$ and $\overline{y}^{\,-}$ are real numbers strictly inside the interval $(-1, 1)$. When no such solution exists, then none of the switching sequences allows us to generate the required harmonics (e.g., this situation arises when we require high first harmonic for low amplitude of ML waveform for a given n). As all elements $\overline{y}^{\,+}$ and $\overline{y}^{\,-}$ can be permuted, the elements of $\overline{y}^{\,+}$ and $\overline{y}^{\,-}$ are reindexed so that for $\overline{y}^{\,+}$, $-1 < y_k < \cdots < y_1 < 1$ holds, and for y^{-}, $-1 < y_n < \cdots < y_{k+1} < 1$ holds. Therefore, according to (34) for even n and (35) for odd n, we have $(x_1, \dots, x_n) = (y_1, y_{k+1}, y_2, y_{k+2}, \dots, y_n, y_k)$ and $(x_1, \dots, x_n) = (y_{k+1}, y_1, y_{k+2}, y_2, \dots, y_k, y_n)$, respectively. Finally, the condition (31c) for $\overline{x}$ must hold.

A. *Solving Composite Sum of Powers*

In this section, the algorithm for solving the composite sum of powers in (32) is described. The solution is inspired by [5] and [6], where a special case of quarter-symmetric three-level inverter problem is studied. The problem was also tackled in [27] and [28], the authors, however, did not use the Padé approximation and the theory of FOPs that play a crucial role in the analytical solution of the whole problem. The other applications of solving composite sum of powers are in coding theory and geometric optics. We will find the exact solution as the set of roots of the following two polynomials:

$$\begin{aligned} V_k(y) &= \prod_{i=1}^{k} (y - y_i)\\ &= y^k + v_{k,k-1} y^{k-1} + \cdots + v_{k,0} \end{aligned} \tag{36}$$

$$\begin{aligned} W_{n-k}(y) &= \prod_{i=1}^{n-k} (y - y_{i+k})\\ &= y^{n-k} + w_{n-k,n-k-1} y^{n-k-1} + \cdots + w_{n-k,0}. \end{aligned} \tag{37}$$

Then, let us do a logarithmic derivative of

$$\frac{V_k(y)}{W_{n-k}(y)} = \frac{\prod_{i=1}^{k} (y - y_i)}{\prod_{i=1}^{n-k} (y - y_{i+k})} \tag{38}$$

to get

$$\frac{V_k'(y)}{V_k(y)} - \frac{W_{n-k}'(y)}{W_{n-k}(y)} = \sum_{i=1}^{k} \frac{1}{y-y_i} - \sum_{i=1}^{n-k} \frac{1}{y-y_{i+k}}. \tag{39}$$

The expansion of $1/(y-z)$ at $y=\infty$ is the series $\sum_{j=0}^{\infty} z^j/y^{j+1}$. Then, we have

$$\frac{V_k'(y)}{V_k(y)} - \frac{W_{n-k}'(y)}{W_{n-k}(y)} = \sum_{j=0}^{\infty} \frac{p_j^+}{y^{j+1}} - \sum_{j=0}^{\infty} \frac{p_j^-}{y^{j+1}}. \tag{40}$$

where $p_j^+ = \sum_{i=1}^{k} y_i^j$, $p_j^- = \sum_{i=1}^{n-k} y_{i+k}^j$ and $p_j = p_j^+ - p_j^-$. Thus, we get

$$\frac{V_k'(y)}{V_k(y)} - \frac{W_{n-k}'(y)}{W_{n-k}(y)} = \sum_{j=0}^{\infty} \frac{p_j}{y^{j+1}}. \tag{41}$$

By integrating, (41) we get

$$\frac{V_k(y)}{W_{n-k}(y)} = y^{2k-n} e^{\left(-\sum_{j=1}^{\infty} \frac{p_j}{j y^j}\right)} = f(y). \tag{42}$$

The series expansion of $f(y)$ leads to the Padé approximation.

B. Padé Approximation

In this section, we will find the unknown coefficients of polynomials $V_k(y)$ and $W_{n-k}(y)$ according to the theory of Padé approximation (for more details, see [29] and [30]). We rewrite (42) in the following way:

$$\begin{aligned} &\frac{V_k(y)}{W_{n-k}(y)} + O(y^{-n+k-2}) \\ &= \left(\frac{1}{y}\right)^{n-2k} \left(\mu_0 + \mu_1 \frac{1}{y} + \mu_2 \left(\frac{1}{y}\right)^2 + \cdots\right) \\ &= f(y), \qquad y \to \infty \end{aligned} \tag{43}$$

where the RHS of (43) is the series expansion of $f(y)$ at infinity. In this case, the expansion of function $f(y)$ contains the negative powers of y.

We consider the following form:

$$\begin{aligned} \frac{\widetilde{V}_k(y)}{\widetilde{W}_{n-k}(y)} + O(y^{n+1}) &= y^{2k-n} f(y^{-1}) \\ &= e^{\left(-\sum_{j=1}^{\infty} \frac{p_j}{j} y^j\right)} = F(y), \qquad y \to 0 \end{aligned} \tag{44}$$

where $\widetilde{V}_k(y) = y^k V_k(y^{-1})$ and $\widetilde{W}_{n-k}(y) = y^{n-k} V_{n-k}(y^{-1})$ (this is only reversion of polynomial coefficients). Therefore, we solve (44) [instead of solving (43)] as the problem of Padé approximation with the following notation:

$$[k/n-k]_F(y) = \frac{\widetilde{V}_k(y)}{\widetilde{W}_{n-k}(y)} = \frac{\widetilde{V}_k^{[k,n-k]}(y)}{\widetilde{W}_{n-k}^{[k,n-k]}(y)} \tag{45}$$

of the function

$$F(y) = e^{\left(-\sum_{j=1}^{\infty} \frac{p_j}{j} y^j\right)} = e^{\sum_{j=1}^{\infty} c_j y^j} \qquad \text{at } y \to 0, \quad \text{where } c_j = -\frac{p_j}{j}. \tag{46}$$

The solution of the original problem in (43) is then obtained by reversing the coefficients of polynomials $\widetilde{V}_k(y)$ and $\widetilde{W}_{n-k}(y)$.

Now, it is necessary to solve the series expansion of the function $F(y)$ at $y=0$ in the form

$$F(y) = \sum_{i=0}^{\infty} \mu_i y^i = \mu_0 + \mu_1 y + \mu_2 y^2 + \cdots. \tag{47}$$

The direct solution is carried out according to [31, Ch. 4.7, exercise 4] and reads

$$\mu_0 = 1, \mu_k = -\frac{1}{k} \sum_{j=1}^{k} p_j \mu_{k-j}, \qquad k = 1, 2, \ldots. \tag{48}$$

In the case of the optimal odd ML problem (or odd bi-level PWM problem), two eventualities can occur (see Section IV). The first is for odd n and $k = \lfloor n/2 \rfloor$, and the second is for even n and $k = n/2$. Both cases will be described separately.

Equation (44), after cross multiplication, gives

$$\widetilde{V}_k^{[k,n-k]}(y) = \widetilde{W}_{n-k}^{[k,n-k]}(y) F(y) + O(y^{n+1}) \tag{49}$$

and a detailed form of the previous equation, considering (47), leads to

$$\begin{aligned} &(\widetilde{v}_{k,k} y^k + \widetilde{v}_{k,k-1} y^{k-1} + \cdots + \widetilde{v}_{k,0}) - O(y^{n+1}) \\ &= (\widetilde{w}_{n-k,n-k} y^{n-k} + \widetilde{w}_{n-k,n-k-1} y^{n-k-1} + \cdots + \widetilde{w}_{n-k,0}) \\ &\quad \times (\mu_0 + \mu_1 y + \mu_2 y^2 + \cdots). \end{aligned} \tag{50}$$

First, let us consider the following cases.

n Is an Odd Number and $k = \lfloor n/2 \rfloor$: The problem of the shifted diagonal Padé approximation, i.e.,

$$[k, k+1]_F(y) = \frac{\widetilde{V}_k^{[k,k+1]}(y)}{\widetilde{W}_{k+1}^{[k,k+1]}(y)} \tag{51}$$

is solved. Equating the coefficients of $y^{k+1}, \ldots, y^{2(k+1)+1}$ in (50) leads to the following linear system:

$$\left[\begin{array}{cccc|c} \mu_0 & \mu_1 & \cdots & & \mu_{k+1} \\ \mu_1 & & \iddots & & \vdots \\ \vdots & \iddots & & & \mu_{2k+1} \\ \hline \mu_{k+1} & \cdots & \mu_{2k+1} & & \mu_{2(k+1)} \end{array}\right] \cdot \left[\begin{array}{c} \widetilde{w}_{k+1,k+1} \\ \vdots \\ \widetilde{w}_{k+1,1} \\ \hline \widetilde{w}_{k+1,0} \end{array}\right] = \left[\begin{array}{c} 0 \\ \vdots \\ 0 \\ \hline \widetilde{K}_k \end{array}\right] \tag{52}$$

where $\widetilde{w}_{k+1,0}$ is coefficient of y^0 of polynomial $\widetilde{W}_{k+1}^{[k,k+1]}(y)$, and due to definiteness and the condition that $w_{k+1,k+1} = 1$, we put $\widetilde{w}_{k+1,0} = 1$, and $\widetilde{K}_k$ will be a nonzero constant. The last equation of the system in (52) is reduced. Therefore, we solve the linear system with a Toeplitz structure (Hankel matrix) of size $(k+1)\times(k+1)$ as follows:

$$\begin{bmatrix} \mu_0 & \mu_1 & \cdots & \mu_k \\ \mu_1 & & \iddots & \vdots \\ \vdots & \iddots & & \mu_{2k-1} \\ \mu_k & \cdots & \mu_{2k-1} & \mu_{2k} \end{bmatrix} \cdot \begin{bmatrix} \widetilde{w}_{k+1,k+1} \\ \widetilde{w}_{k+1,k} \\ \vdots \\ \widetilde{w}_{k+1,1} \end{bmatrix} = \begin{bmatrix} -\mu_{k+1} \\ -\mu_{k+2} \\ \vdots \\ -\mu_{2k+1} \end{bmatrix}. \tag{53}$$

From the found solution $\widetilde{W}_{k+1}^{[k,k+1]}(y)$, the polynomial $W_{k+1}^{[k,k+1]}(y)$ is recovered by reversing the coefficients. Alternatively, the solution can be obtained as the solution of the following linear system:

$$\left[\begin{array}{cccc|c} \mu_0 & \mu_1 & \cdots & & \mu_{k+1} \\ \mu_1 & & \iddots & & \vdots \\ \vdots & \iddots & & & \mu_{2k+1} \\ \hline \mu_{k+1} & \cdots & \mu_{2k+1} & & \mu_{2(k+1)} \end{array}\right] \cdot \left[\begin{array}{c} \widetilde{w}_{k+1,0} \\ \vdots \\ \widetilde{w}_{k+1,k} \\ \hline \widetilde{w}_{k+1,k+1} \end{array}\right] = \left[\begin{array}{c} 0 \\ \vdots \\ 0 \\ \hline K_k \end{array}\right] \tag{54}$$

where $w_{k+1,k+1}$ is equal to 1.

Unknown polynomial coefficients of $V_k^{[k,k+1]}(y)$ are obtained from the known polynomial coefficients of $\widetilde{W}_{k+1}^{[k,k+1]}(y)$ as follows:

$$\begin{bmatrix} \widetilde{v}_{k,0} \\ \widetilde{v}_{k,1} \\ \vdots \\ \widetilde{v}_{k,k} \end{bmatrix} = \begin{bmatrix} 0 & 0 & \cdots & 0 & \mu_0 \\ \vdots & \vdots & \iddots & \iddots & \mu_1 \\ 0 & 0 & \iddots & \iddots & \vdots \\ 0 & \mu_0 & \mu_1 & \cdots & \mu_k \end{bmatrix} \cdot \begin{bmatrix} \widetilde{w}_{k+1,k+1} \\ \widetilde{w}_{k+1,k} \\ \vdots \\ \widetilde{w}_{k+1,1} \\ \widetilde{w}_{k+1,0} \end{bmatrix} \tag{55}$$

equating coefficients of $x^0, x^1, \ldots, x^k$ in (50). Obviously, $\widetilde{w}_{k+1,0} = 1$, $\mu_0 = 1$, and $\widetilde{v}_{k,0} = 1$. Therefore, the previous matrix equation is simplified to

$$\begin{bmatrix} \widetilde{v}_{k,1} \\ \widetilde{v}_{k,2} \\ \vdots \\ \widetilde{v}_{k,k} \end{bmatrix} = \begin{bmatrix} 0 & \cdots & 0 & \mu_0 \\ \vdots & \iddots & \iddots & \mu_1 \\ 0 & \iddots & \iddots & \vdots \\ \mu_0 & \mu_1 & \cdots & \mu_{k-1} \end{bmatrix} \cdot \begin{bmatrix} \widetilde{w}_{k+1,k} \\ \widetilde{w}_{k+1,k-1} \\ \vdots \\ \widetilde{w}_{k+1,1} \end{bmatrix} + \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{bmatrix}. \tag{56}$$

The polynomial $V_k^{[k,k+1]}(y)$ can be constructed analogously from the found solution $\widetilde{V}_k^{[k,k+1]}(y)$ by reversing coefficients or by the following linear system:

$$\begin{bmatrix} v_{k,k-1} \\ v_{k,k-2} \\ \vdots \\ v_{k,0} \end{bmatrix} = \begin{bmatrix} 0 & \cdots & 0 & \mu_0 \\ \vdots & \iddots & \iddots & \mu_1 \\ 0 & \iddots & \iddots & \vdots \\ \mu_0 & \mu_1 & \cdots & \mu_{k-1} \end{bmatrix} \cdot \begin{bmatrix} w_{k+1,1} \\ w_{k+1,2} \\ \vdots \\ w_{k+1,k} \end{bmatrix} + \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{bmatrix}. \tag{57}$$

n Is an Even Number and $k = n/2$: The procedure is similar to the previous case. The diagonal Padé approximation, i.e.,

$$[k,k]_F(y) = \frac{\widetilde{V}_k^{[k,k]}(y)}{\widetilde{W}_k^{[k,k]}(y)} \tag{58}$$

is solved. The coefficients of $\widetilde{W}_k^{[k,k]}(y)$ are due to

$$\begin{bmatrix} \mu_1 & \mu_2 & \cdots & \mu_k \\ \mu_2 & & \iddots & \vdots \\ \vdots & \iddots & & \mu_{2k-2} \\ \mu_k & \cdots & \mu_{2k-2} & \mu_{2k-1} \end{bmatrix} \cdot \begin{bmatrix} \widetilde{w}_{k,k} \\ \widetilde{w}_{k,k-1} \\ \vdots \\ \widetilde{w}_{k,1} \end{bmatrix} = \begin{bmatrix} -\mu_{k+1} \\ -\mu_{k+2} \\ \vdots \\ -\mu_{2k} \end{bmatrix} \tag{59}$$

equating the coefficients of $y^{k+1}, y^{k+2}, \ldots, y^{2k+1}$ in (50). The coefficients of $\widetilde{V}_k^{[k,k]}(y)$ are obtained as follows:

$$\begin{bmatrix} \widetilde{v}_{k,1} \\ \widetilde{v}_{k,2} \\ \vdots \\ \widetilde{v}_{k,k} \end{bmatrix} = \begin{bmatrix} 0 & \cdots & 0 & \mu_0 \\ \vdots & \iddots & \iddots & \mu_1 \\ 0 & \iddots & \iddots & \vdots \\ \mu_0 & \mu_1 & \cdots & \mu_{k-1} \end{bmatrix} \cdot \begin{bmatrix} \widetilde{w}_{k,k} \\ \widetilde{w}_{k,k-1} \\ \vdots \\ \widetilde{w}_{k,1} \end{bmatrix} + \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{bmatrix}. \tag{60}$$

C. Family of FOPs

According to the theory of Padé approximation, $V(y)$ and $W(y)$ are FOPs, and therefore, related formulas and theorems can be applied (see, e.g., [30], [32], and [33] for references).

1) Three-Term Recurrence Formula for $W(y)$ and $V(y)$:

n is an odd number and $k = \lfloor n/2 \rfloor$: According to [32, p. 101] with (51) we have

$$[k/k+1]_F(y) = \frac{\widetilde{V}_k^{[k,k+1]}(y)}{\widetilde{W}_{k+1}^{[k,k+1]}(y)} = \frac{\widetilde{Q}_{k+1}^{(0)}(y)}{\widetilde{P}_{k+1}^{(0)}(y)} \tag{61}$$

where $\widetilde{P}_{k+1}^{(0)}(y) = y^{k+1} P_{k+1}^{(0)}(y^{-1})$, and $\widetilde{Q}_{k+1}^{(0)}(y) = y^k Q_{k+1}^{(0)}(y^{-1})$. The polynomial $P_{k+1}^{(0)}(y)$ is an FOP of the first kind with respect to the linear functional $\mathcal{L}^{(0)}[y^i] = \mu_i$, where μ_i is generated according to (48). The polynomial $Q_{k+1}^{(0)}(y)$ is the associated FOP (sometimes called the polynomial of the second kind) to $P_{k+1}^{(0)}(y)$. Thus, according to (61), $W_{k+1}^{[k,k+1]}(y) = P_{k+1}^{(0)}(y), V_k^{[k,k+1]}(y) = Q_{k+1}^{(0)}(y)$, and we can write the following three-term recurrence formulas:

$$W_{-1}^{[-2,-1]}(y) = 0 \quad W_0^{[-1,0]}(y) = 1 \tag{62a}$$

$$W_i^{[i-1,i]}(y) = (y + B_i) W_{i-1}^{[i-2,i-1]}(y) - C_i W_{i-2}^{[i-3,i-2]}(y)$$
$$i = 1, 2, \ldots, k+1, \ldots \tag{62b}$$

where

$$B_i = -\frac{\mathcal{L}^{(0)}\left[y\left(W_{i-1}^{[i-2,i-1]}(y)\right)^2\right]}{K_{i-1}} \quad C_i = \frac{K_{i-1}}{K_{i-2}} \tag{63a}$$

$$K_i = \sum_{j=0}^{i} \mu_{i+j} w_{i,j}. \tag{63b}$$

The linear moment functional $\mathcal{L}^{(0)}[\cdot]$ in (63a) of arbitrary polynomial $Z(y) = \sum_{i=0}^{n} z_i y^i$ is solved according to $\mathcal{L}^{(0)}[Z(y)] = \sum_{i=0}^{n} z_i \mu_i$, where $\mathcal{L}^{(0)}[y^i] = \mu_i$ and $w_{i,j}$ are the coefficients of

$W_i^{[i-1,i]}(y) = \sum_{j=0}^{i} w_{i,j} y^j$. Note that the constant K_i in (63b) is the same as the constant in (54).

The polynomial $V_k^{[k,k+1]}(y)$ is associated FOP to $W_{k+1}^{[k,k+1]}(y)$, and therefore, we have

$$V_{-1}^{[-1,0]}(y) = -1 \quad V_0^{[0,1]}(y) = 0 \tag{64a}$$

$$V_i^{[i,i+1]}(y) = (y+B_i)V_{i-1}^{[i-1,i]}(y) - C_i V_{i-2}^{[i-2,i-1]}(y), \quad i = 1, 2, \ldots, k, \ldots \tag{64b}$$

where B_i and C_i are identical to (63).

n is an even number and $k = n/2$: Similarly as above, we have the following equation for (58):

$$[k/k]_f(y) = \frac{\widetilde{V}_k^{[k,k]}(y)}{\widetilde{W}_k^{[k,k]}(y)} = \mu_0 + y\frac{\widetilde{Q}_k^{(1)}(y)}{\widetilde{P}_k^{(1)}(y)} \tag{65}$$

where $\widetilde{P}_k^{(1)}(y) = y^k P_k^{(1)}(y^{-1})$, and $\widetilde{Q}_k^{(1)}(y) = y^{k-1} Q_k^{(1)}(y^{-1})$. The polynomial $P_k^{(1)}(y)$ is the adjacent FOP of the first kind with respect to the linear functional $\mathcal{L}^{(1)}[y^i] = \mathcal{L}^{(0)}[y^{i+1}] = \mu_{i+1}$, where μ_i is generated according to (48). The polynomial $Q_k^{(1)}(y)$ is the associated adjacent FOP to $P_k^{(1)}(y)$. Thus, according to (65) $W_k^{[k,k]}(y) = P_k^{(1)}(y)$, $\widetilde{V}_k^{[k,k]}(y) = \mu_0 \widetilde{P}_k^{(1)}(y) + y\widetilde{Q}_k^{(1)}(y)$, and therefore, we can write the following three-term recurrence formula for $W_i^{[i,i]}(y)$:

$$W_{-1}^{[-1,-1]}(y) = 0 \quad W_0^{[0,0]}(y) = 1 \tag{66a}$$

$$W_i^{[i,i]}(y) = (y+B_i)W_{i-1}^{[i-1,i-1]}(y) - C_i W_{i-2}^{[i-2,i-2]}(y), \quad i = 1, 2, \ldots, k, \ldots \tag{66b}$$

where

$$B_i = -\frac{\mathcal{L}^{(1)}\left[y\left(W_{i-1}^{[i-1,i-1]}(y)\right)^2\right]}{K_{i-1}} \qquad C_i = \frac{K_{i-1}}{K_{i-2}} \tag{67a}$$

$$K_i = \sum_{j=0}^{i} \mu_{i+j+1} w_{i,j} \tag{67b}$$

where the linear moment functional $\mathcal{L}^{(1)}[\cdot]$ in (67a) of arbitrary polynomial $Z(y) = \sum_{i=0}^{n} z_i y^i$ is solved according to $\mathcal{L}^{(1)}[Z(y)] = \sum_{i=0}^{n} z_i \mu_{i+1}$, and $w_{i,j}$ are the coefficients of $W_i^{[i,i]}(y) = \sum_{j=0}^{i} w_{i,j} y^j$.

Finding a recurrent formula for the polynomial $V_k(y)$ is more difficult due to the fact that $V_k^{[k,k]}(y)$ is not an associated FOP to $W_k^{[k,k]}(y)$. From (65), we know, however, that $\widetilde{V}_k(y) = \mu_0 \widetilde{P}_k^{(1)}(y) + y\widetilde{Q}_k^{(1)}(y)$. We apply "tilde notation" (reversion of coefficients) on both sides of the equation $\widetilde{\widetilde{V}}_k^{[k,k]}(y) = \mu_0 \widetilde{\widetilde{P}}_k^{(1)}(y) + \widetilde{y\widetilde{Q}_k}^{(1)}(y)$ and get $V_k^{[k,k]}(y) = \mu_0 P_k^{(1)}(y) + Q_k^{(1)}(y)$. Thus, the recursion for $V_k^{[k,k]}(y)$ is a composition of $P_k^{(1)}(y)$ and $Q_k^{(1)}(y)$, where $Q_k^{(1)}(y)$ is the associated FOP to $P_k^{(1)}(y)$, with the following three-term recurrence formula:

$$Q_{-1}^{(1)}(y) = -1 \quad Q_0^{(1)}(y) = 0 \tag{68a}$$

$$Q_i^{(1)}(y) = (y+B_i)Q_{i-1}^{(1)}(y) - C_i Q_{i-2}^{(1)}(y), \quad i = 1, 2, \ldots, k, \ldots \tag{68b}$$

where B_i and C_i are due to (67). The recurrence formula for $P_k^{(1)}(y)$ is given by (66), where $P_k^{(1)}(y) = W_k^{[k,k]}(y)$. Therefore, we have

$$\begin{aligned} V_i^{[i,i]}(y) &= \mu_0\left((y+B_i)P_{i-1}^{(1)}(y) - C_i P_{i-2}^{(1)}(y)\right) \\ &\quad + (y+B_i)Q_{i-1}^{(1)}(y) - C_i Q_{i-2}^{(1)}(y) \\ &= (y+B_i)\left(\mu_0 P_{i-1}^{(1)}(y) + Q_{i-1}^{(1)}(y)\right) \\ &\quad - C_i\left(\mu_0 P_{i-2}^{(1)}(y) + Q_{i-2}^{(1)}(y)\right) \\ &= (y+B_i)V_{i-1}^{[i-1,i-1]}(y) - C_i V_{i-2}^{[i-2,i-2]}(y), \\ &\quad i = 1, 2, \ldots, k, \ldots \end{aligned} \tag{69}$$

where B_i and C_i are according to (67), and the initial conditions are

$$V_{-1}^{[-1,-1]}(y) = \mu_0 P_{-1}^{(1)}(y) + Q_{-1}^{(1)}(y) = 1 \cdot 0 + (-1) = -1$$

$$V_0^{[0,0]}(y) = \mu_0 P_0^{(1)}(y) + Q_0(y)^{(1)} = 1 \cdot 1 + 0 = 1.$$

2) *Determinantal Formulas for $W(y)$ and $V(y)$:*

According to [32, Ch. 2], one can write the following determinantal formulas for polynomials $W(y)$ and $V(y)$.

n is an odd number and $k = \lfloor n/2 \rfloor$: We have

$$W_{k+1}^{[k,k+1]}(y) = D_{w_{k+1}} \det\begin{bmatrix} \mu_0 & \mu_1 & \cdots & \mu_k & \mu_{k+1} \\ \mu_1 & & \iddots & \iddots & \mu_{k+2} \\ \vdots & \iddots & \iddots & & \vdots \\ \mu_k & \mu_{k+1} & \cdots & \mu_{2k-1} & \mu_{2k} \\ 1 & y & \ldots & y^k & y^{k+1} \end{bmatrix} \tag{70}$$

$$V_k^{[k,k+1]}(y) = D_{v_k} \det\begin{bmatrix} \mu_0 & \mu_1 & \cdots & \mu_{k-1} & \mu_k \\ \mu_1 & & \iddots & \iddots & \mu_{k+1} \\ \vdots & \iddots & \iddots & & \vdots \\ \mu_{k-1} & \mu_k & \cdots & \mu_{2k-2} & \mu_{2k-1} \\ 0 & 1 & \ldots & \sum_{i=0}^{k-1}\mu_i y^{k-i-1} & \sum_{i=0}^{k}\mu_i y^{k-i} \end{bmatrix} \tag{71}$$

where $D_{w_{k+1}}$ and D_{v_k} are normalization factors so that $W_{k+1}^{[k,k+1]}(y)$ and $V_k^{[k,k+1]}(y)$ are monomials, and the moments μ_i are generated according to (48).

n is an even number and $k = n/2$: We have (72)–(73), shown at the bottom of the next page, where D_{w_k} and D_{v_k} are normalization factors so that $W_k^{[k,k]}(y)$ and $V_k^{[k,k]}(y)$ are monomials, and the moments μ_i are generated according to (48).

TABLE I

PARTIAL RESULTS FOR AN ILLUSTRATIVE EXAMPLE WHERE $n_C = 3, n_E = 13, A = 2.3$, AND $(b_{f_1}, b_{f_2}, b_{f_3}) = (-2, 0.5, 1)$

p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}	p_{11}	p_{12}	p_{13}	p_{14}	p_{15}	p_{16}	p_{17}
-1.3659	0.3415	-0.5122	0.3415	-0.2134	0.3201	-0.0747	0.2988	0.	0.2801	0.044	0.2641	0.0715	0.2504	0.0894	0.2384	0.1013
μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_7	μ_8	μ_9	μ_{10}	μ_{11}	μ_{12}	μ_{13}	μ_{14}	μ_{15}	μ_{16}	μ_{17}
1.3659	0.7621	0.3623	0.1482	0.0432	-0.0158	-0.0406	-0.0552	-0.0578	-0.0594	-0.0561	-0.0541	-0.0497	-0.047	-0.0428	-0.0403	-0.0367
$w_{8,0}$	$w_{8,1}$	$w_{8,2}$	$w_{8,3}$	$w_{8,4}$	$w_{8,5}$	$w_{8,6}$	$w_{8,7}$		$v_{8,0}$	$v_{8,1}$	$v_{8,2}$	$v_{8,3}$	$v_{8,4}$	$v_{8,5}$	$v_{8,6}$	$v_{8,7}$
0.001	-0.0512	-0.2215	0.3118	1.2376	-0.4483	-2.0023	0.1779		-0.0023	0.0209	0.1091	-0.5804	0.1302	1.6544	-1.1417	-1.188
y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8		y_9	y_{10}	y_{11}	y_{12}	y_{13}	y_{14}	y_{15}	y_{16}
-0.9222	-0.7985	-0.6458	-0.2468	0.0178	0.5245	0.9095	0.9836		-0.9109	-0.7287	-0.1693	0.0865	0.4478	0.6023	0.8841	0.9762
α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8	α_9	α_{10}	α_{11}	α_{12}	α_{13}	α_{14}	α_{15}	α_{16}	
0.1813	0.2186	0.4286	0.4863	1.0187	0.9244	1.553	1.1065	1.8202	1.4842	2.2729	1.7409	2.4956	2.3873	2.7446	2.7162	
b_{p_1}	b_{p_2}	b_{p_3}	b_{p_4}	...	$b_{p_{16}}$	$b_{p_{17}}$	$b_{p_{18}}$	$b_{p_{19}}$	$b_{p_{20}}$	$b_{p_{21}}$	$b_{p_{22}}$	$b_{p_{23}}$	$b_{p_{24}}$	$b_{p_{25}}$		THD
-2	0.5	1	0	...	0	0.2171	-0.0469	0.0158	0.3334	-0.3591	-0.2791	-0.0791	-0.0003	0.1343		1.81%

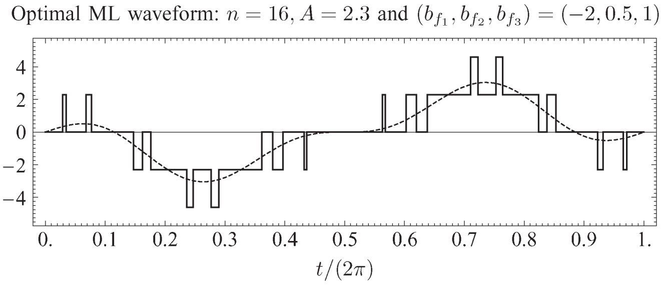


Fig. 6. Solution of an illustrative example.

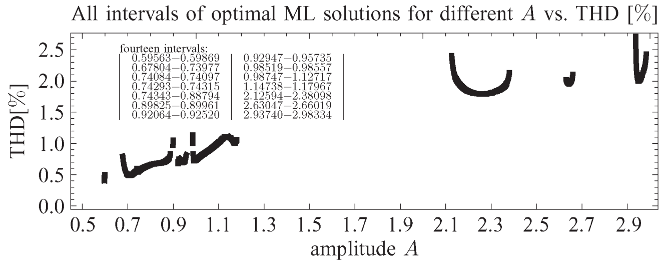


Fig. 7. All intervals of optimal ML solutions for increasing A versus THD (in percent): $n = 16$ and $(b_{f_1}, b_{f_2}, b_{f_3}) = (-2, 0.5, 1)$.

3) Eigenvalues Formulation: The solution of composite sum of powers is the set of zeros of polynomials $W(y)$ and $V(y)$. As these are FOPs, it is possible to obtain these zeros as eigenvalues of a special matrix (see [32, p. 79]) by

$$J_{k+1} = \begin{bmatrix} -B_1 & 1 & 0 & \dots & 0 \\ C_2 & -B_2 & 1 & \ddots & \vdots \\ 0 & C_3 & -B_3 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & 1 \\ 0 & \dots & 0 & C_{k+1} & -B_{k+1} \end{bmatrix} \tag{74}$$

where B_i and C_i are computed according to (63). Thus, for odd n, we have

$$\begin{aligned} W_{k+1}^{[k,k+1]}(y) &= \det(yI_{k+1} - J_{k+1}) \\ V_k^{[k,k+1]}(y) &= \det(yI_k - J'_k) \end{aligned} \tag{75}$$

where J'_k is the matrix obtained by suppressing the first row and the first column of J_{k+1}. Therefore, the zeros of $W_{k+1}^{[k,k+1]}(y)$ are the eigenvalues of J_{k+1}, and the zeros of $V_k^{[k,k+1]}(y)$ are the eigenvalues of J'_k.

4) Other Orthogonal Properties—The Zeros: The position of zeros of (classical) orthogonal polynomials has very important properties. Each n-degree polynomial in an orthogonal sequence has all n of its roots real from interval (a, b), distinct, and strictly inside the interval of orthogonality. The roots of each polynomial lie strictly between the roots of the next higher degree polynomial in the sequence. This interesting property can be partially employed in a numerical iterative search algorithms for the zeros in recurrence algorithm—for the choice of the initial iteration in Newton's method.

Not all nice properties extend to FOPs nevertheless. In particular, the zeros of FOPs need not be simple or even real. For

$$W_k^{[k,k]}(y) = D_{w_k} \det \begin{bmatrix} \mu_1 & \mu_2 & \cdots & \mu_k & \mu_{k+1} \\ \mu_2 & & \iddots & \iddots & \mu_{k+2} \\ \vdots & \iddots & \iddots & & \vdots \\ \mu_{k-1} & \mu_{k+1} & \cdots & \mu_{2k-1} & \mu_{2k} \\ 1 & y & \dots & y^{k-1} & y^k \end{bmatrix} \tag{72}$$

$$V_k^{[k,k]}(y) = D_{v_k} \det \begin{bmatrix} \mu_1 & \mu_2 & \cdots & \mu_k & \mu_{k+1} \\ \mu_2 & & \iddots & \iddots & \mu_{k+2} \\ \vdots & \iddots & \iddots & & \vdots \\ \mu_{k-1} & \mu_{k+1} & \cdots & \mu_{2k-1} & \mu_{2k} \\ 1 & y + \mu_1 & \dots & \sum_{i=1}^{k-1} \mu_i y^{k-i-1} & \sum_{i=1}^{k} \mu_i y^{k-i} \end{bmatrix} \tag{73}$$

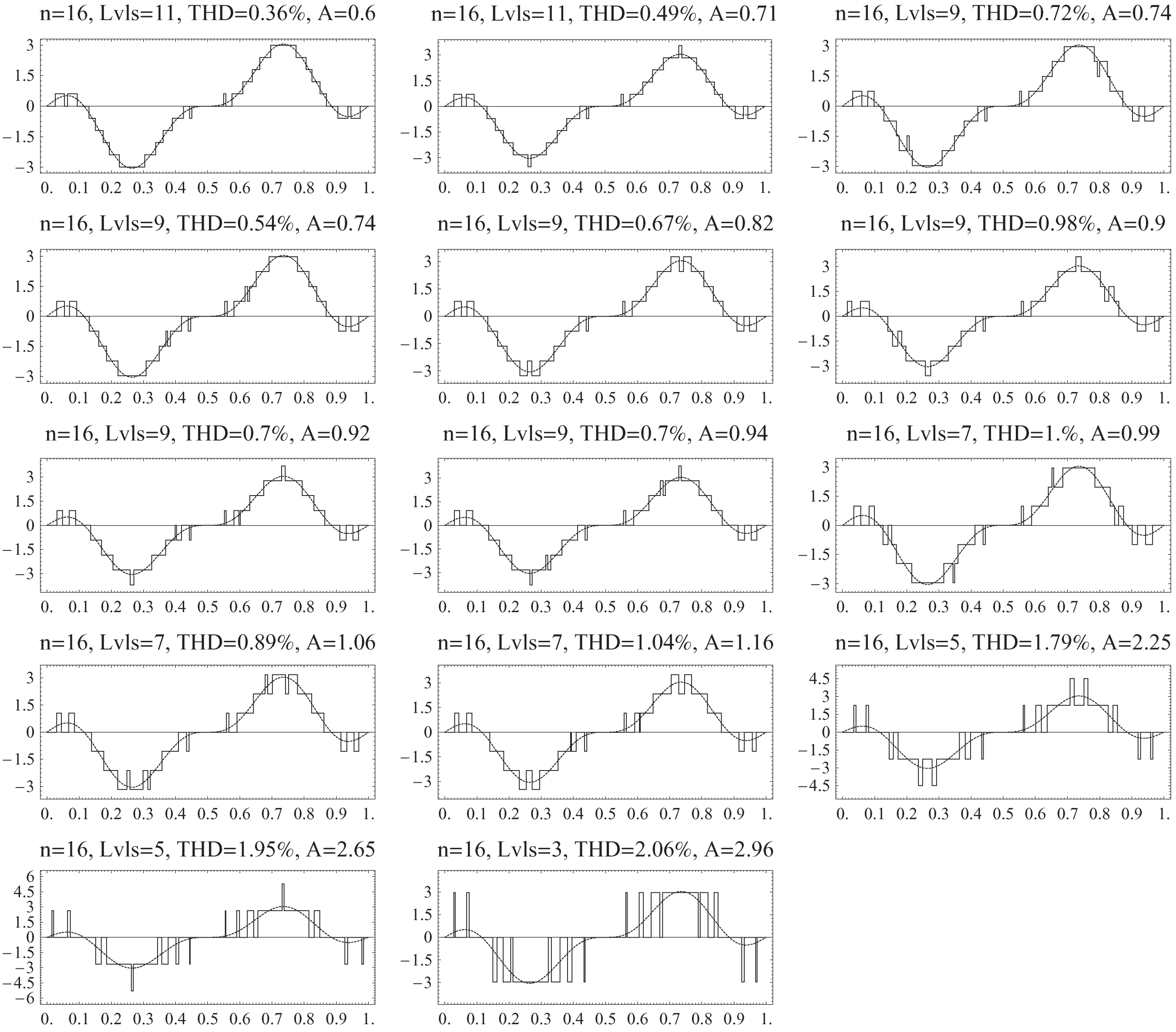


Fig. 8. All possible configurations of optimal ML waveforms for increasing amplitude: $A, n = 16$ and $(b_{f_1}, b_{f_2}, b_{f_3}) = (-2, 0.5, 1)$.

FOPs, the following holds nevertheless: if $\mathcal{L}[\cdot]$ is defined, then for all $k \geq 0$, 1) P_k and P_{k+1} have no common zeros, 2) Q_k and Q_{k+1} have no common zeros, and 3) P_k and Q_k have no common zeros.

V. Illustrative Numerical Example

Let us consider the optimal ML problem with controlled harmonics $(b_{f_1}, b_{f_2}, b_{f_3}) = (-2, 0.5, 1)$, fixed $n = 16$, and amplitude $A = 2.3$. The partial results of computation for this specific n and A are shown in Table I (the line 2: power sums p_i, 4: moments μ_i, 6: the coefficients of FOPs W and V, 8: the zeros W and V, 10: result—switching times α_i, 12: test—the required frequency spectrum of the ML waveform b_{p_i} computed from α_i and THD). Fig. 6 depicts the obtained solution for the ML problem.

The following figures illustrate complete solution of ML problem where n and A are varying. Fig. 7 depicts increasing amplitude A (in steps of 10^{-4}) and fixed $n = 16$ versus THD (in percent) $(N = 20)$. The solution is in 14 intervals for the amplitude A, where the ML problem has a solution (no other amplitude A solves this ML problem for $n = 16$), and Fig. 8 shows all switching configurations for all these intervals.

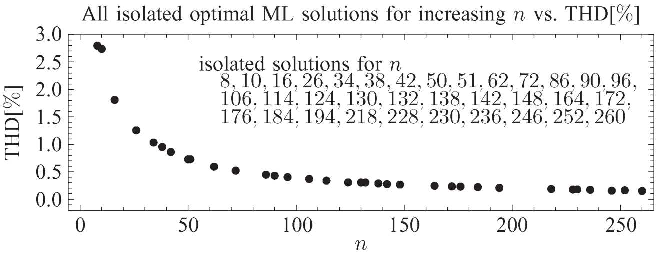


Fig. 9. All isolated optimal ML (five-level) solutions for increasing n versus THD (in percent): $A = 2.3$ and $(b_{f_1}, b_{f_2}, b_{f_3}) = (-2, 0.5, 1)$.

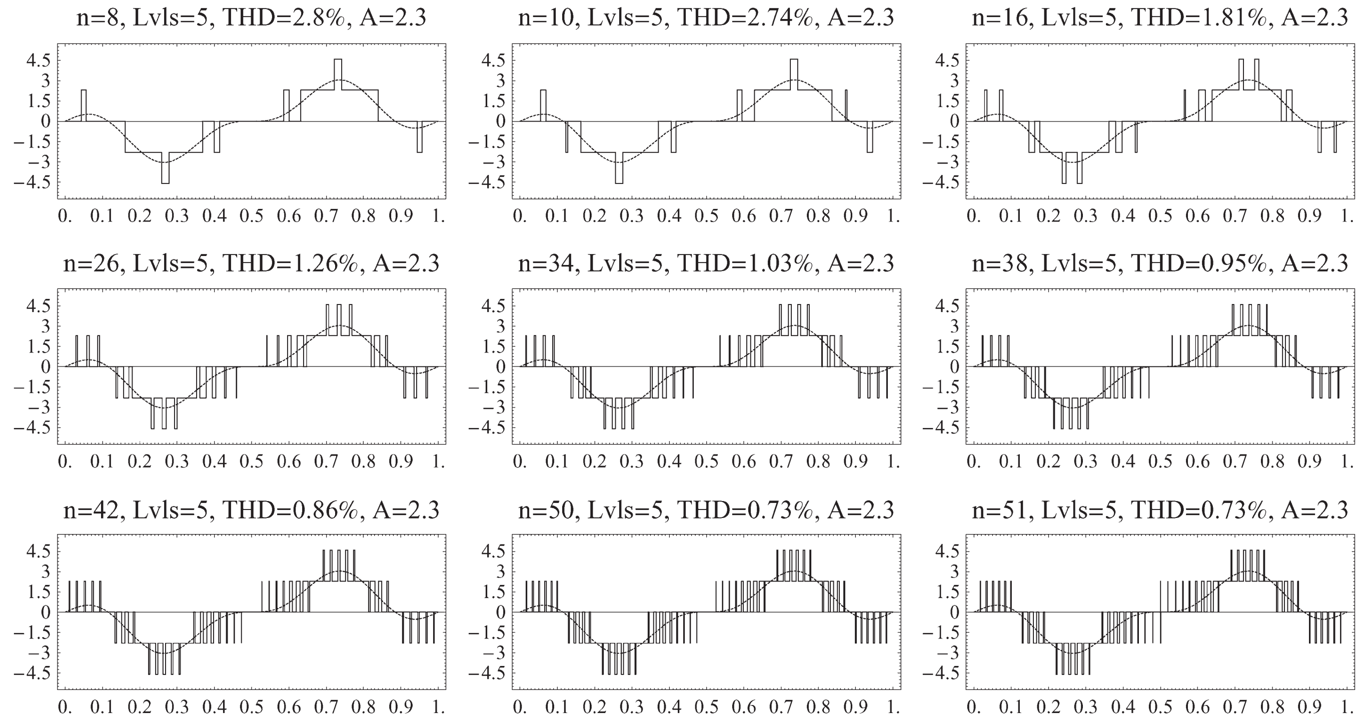


Fig. 10. All possible configurations of optimal ML waveforms for different n : $A = 2.3$ and $(b_{f_1}, b_{f_2}, b_{f_3}) = (-2, 0.5, 1)$.

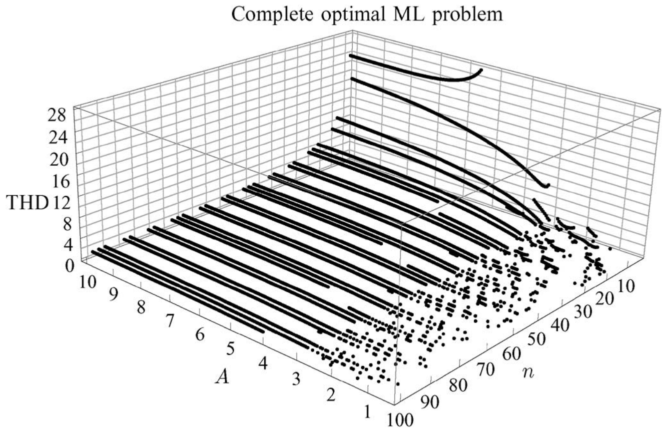


Fig. 11. Complete optimal ML solutions $(b_{f_1}, b_{f_2}, b_{f_3}) = (-2, 0.5, 1)$ for varying n and A versus THD (in percent).

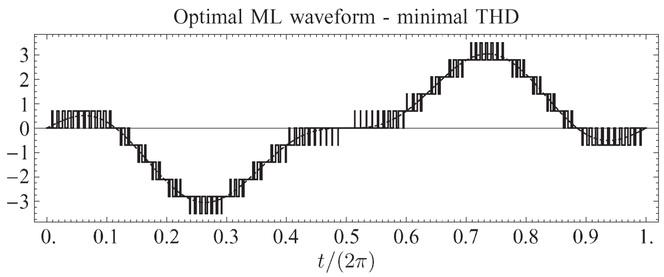


Fig. 12. Optimal ML with minimal THD (in percent): $A = 0.7$, $n = 96$, number of levels $= 11$, THD $= 0.125\%$, and $(b_{f_1}, b_{f_2}, b_{f_3}) = (-2, 0.5, 1)$.

Fig. 9 depicts increasing number of switching n and fixed $A = 2.3$ versus THD. The first nine isolated solutions are given in Fig. 10.

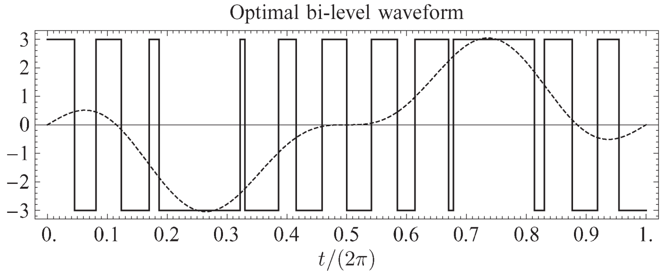


Fig. 13. Optimal bilevel waveform: $A = 3$, $n = 10$, THD $= 11.96\%$, and $(b_{f_1}, b_{f_2}, b_{f_3}) = (-2, 0.5, 1)$.

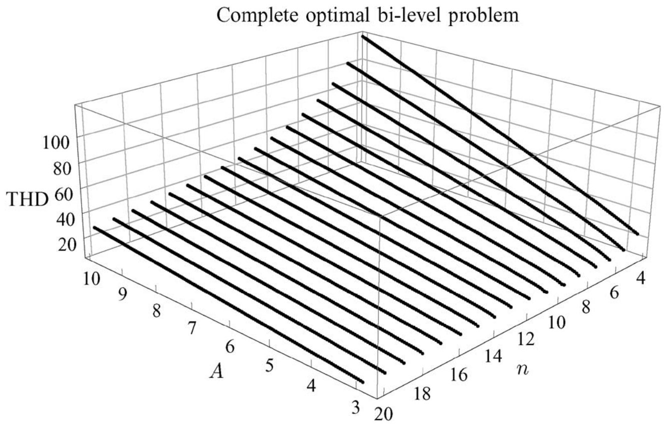


Fig. 14. Complete optimal bilevel solutions, varying n and A versus THD (in percent) and $(b_{f_1}, b_{f_2}, b_{f_3}) = (-2, 0.5, 1)$.

The complete solution (n is from 4 to 100, and A is from 0.05 to 10, with step 0.05) is visualized in Fig. 11, where a varying amplitude A and number of switching n versus THD are visualized. Fig. 12 show the ML signal with minimal THD.

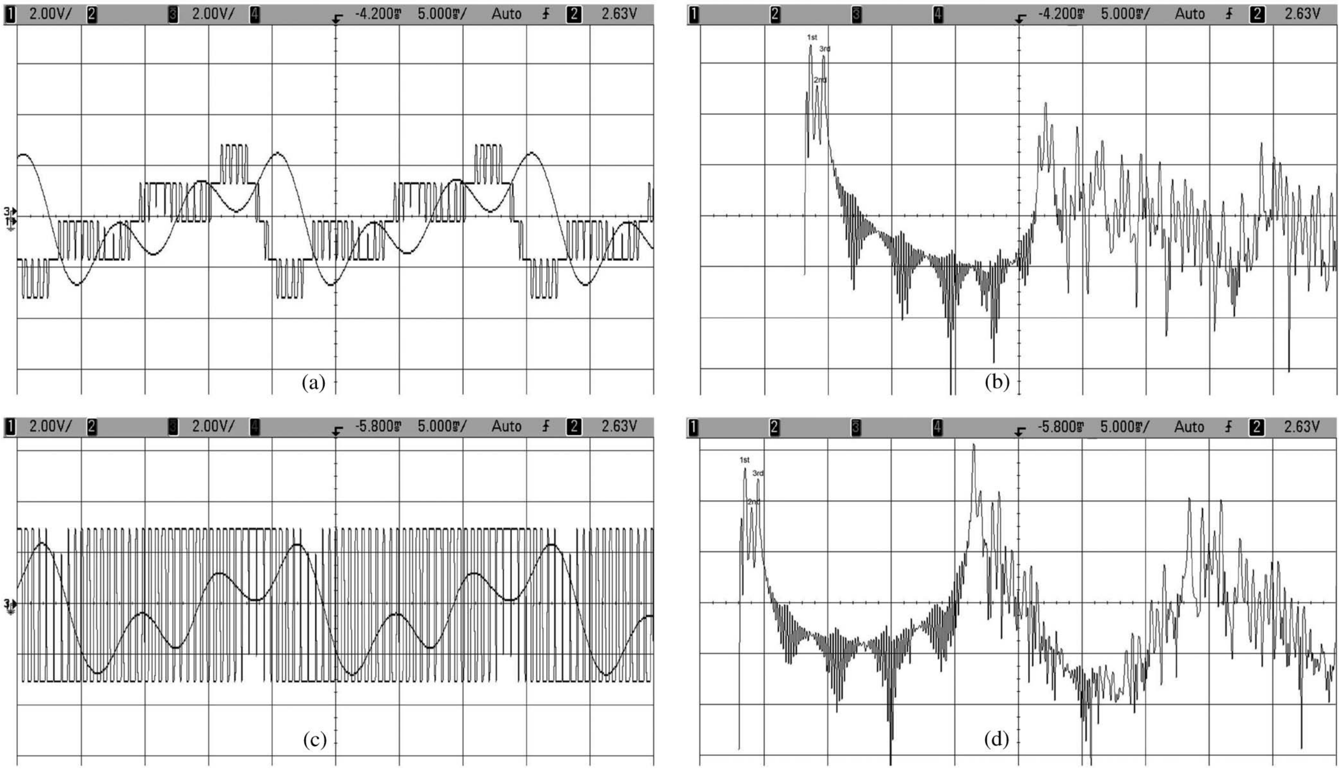


Fig. 15. Experimental results: output voltage (switching waveforms and filtered waveforms) and its spectrum. Baseband portion is $b_{f_{1,2,3}} = (1.5, -0.6, 1.2)$ and eliminated harmonics (zero band) are 4, 5, 6, . . ., 36. (a) and (b) Five-level waveform. (c) and (d) Bilevel waveform.

The results for the bilevel waveform (there is different solving procedure, see Section III-B) are depicted in Fig. 13, and the complete solution is in Fig. 14. We can see that there exist solutions in all cases (unlike ML), but THD is much worse in the ML case.

The *Mathematica*[2] package (all algorithms described in this paper) with other simulations and demo examples can be downloaded from the authors' webpages [35].

VI. Experimental Results

To verify the performance of the proposed algorithms, an experimental setup was built in the laboratory. It is composed of the Agilent 33120A waveform generator with related software Agilent IntuiLink WaveForm Editor installed on a laboratory personal computer.

In the experimental example, we solve the optimal five-level and bilevel problems for $b_{f_{1,2,3}} = (1.5, -0.6, 1.2)$ and $n = 36$ with a frequency of 50 Hz and $A = 1.5$ V and $A = 3$ V, respectively. According to proposed algorithms, we obtain the switching times $\overline{\alpha} = (0.000373, 0.000533, \ldots, 0.009668, 0.009784)$ and $\overline{\alpha} = (0.000279, 0.000502, \ldots, 0.009533, 0.009725)$, respectively. The offline fast Fourier transform (FFT) analysis of the experimental data shows that the THDs are 1.25% and 5.43.%, respectively, which are slightly larger than the theoretical values of 1.08% and 5.21.%, respectively, for given A. The solution is depicted in Fig. 15. Subsequently, the switching output waveform is filtered by the low-pass Butterworth filter (switched capacitor filter Maxim MAX291, eighth order), and the filtered output corresponds to the required baseband.

VII. Case Study: Active Filters

The main goal of active filters is the cancellation of noise or distortion of harmonic signals. These undesirable effects are consequences of disturbances or nonlinearities of load (see [36] and [37] for more details).

Let us consider the simplified principal scheme according to Fig. 16(a). The basic principle of active filters is based on generating harmonic signals with an amplitude opposite that of the undesirable harmonics so that they are canceled in total. This suitable signal is then generated as a filtered PWM or ML waveform that is easily and efficiently realizable.

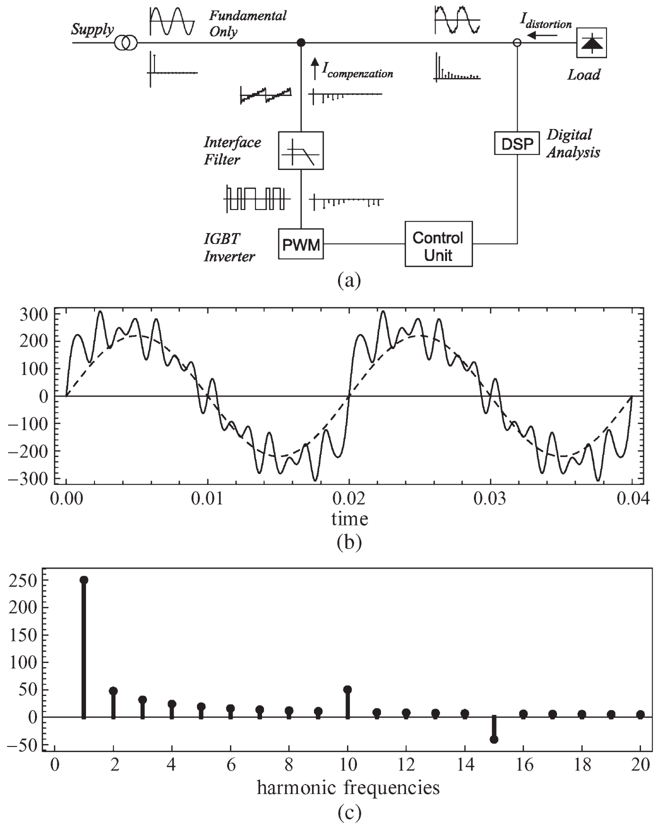
Active filters are installed in a wide range of industrial and nonindustrial applications (pulp and paper facilities, chemical plants, steel plants, car industry, and banks or telecommunication centers due to the large number of computers and Uninterruptible Power Supply (UPS) systems).

Numerical Example

Let us consider electrical power grid $f = 50$ Hz and compensate the harmonic distortion caused by a set of drives. The fundamental harmonic in a power grid is deviated strongly by the odd[3] saw signal and in addition amplified tenth and fifteenth harmonics. The signal, which

[2] The *Mathematica* Web pages are in [34].

[3] If the analyzed signal is not odd, we can make odd extension and use our approach.



(a)

(b)

(c)

Fig. 16. (a) Diagram illustrating components of the connected active filter with waveforms showing cancellation of harmonics from load. (b) Fundamental harmonic and deviated fund. Harmonic in a power grid. (c) Spectrum of a deviated fundamental waveform.

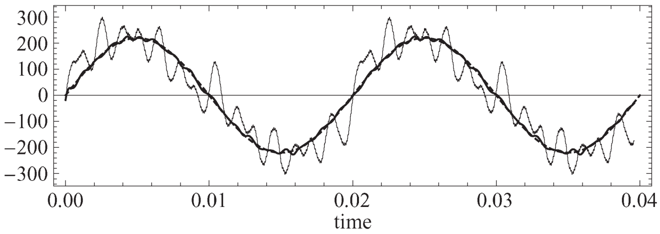


Fig. 17. Restored fundamental harmonic, filtered optimal odd, and filtered quarter-symmetric PWM waveform.

is biased, is depicted in Fig. 16(b). Its frequency amplitude spectrum $a_1, \ldots, a_{20}$ is depicted in Fig. 16(c), and it is $(250.1, 47.7, 31.8, 23.9, 19.1, 15.9, 13.6, 11.9, 10.6, 50.3, 8.7, 8.0, 7.3, 6.8, -40.6, 6.0, 5.6, 5.3, 5.0, 4.8, \ldots)$.

It is desirable to suggest appropriate switching $(\alpha_1, \ldots, \alpha_{220})$ of the odd bilevel PWM waveform so that after its filtration, we get harmonic signal with reverse amplitude spectrum $(b_1, b_2, \ldots, b_{20}, b_{21}, \ldots, b_{220}) = (-30.1, -47.7, \ldots, -4.8, 0, \ldots, 0)$. In this operation, we restrict the first 20 harmonics only, and the following 200 harmonics are zeroed. The nullity of higher harmonics is given because of consequent filtering (we use the Chebyshev filter of the fourth order with cutoff frequency $f_c = 23f$) of the odd bilevel waveform. The solution is depicted in Fig. 17. The solution obtained by a numerical algorithm for quarter-wave signals (see [5]) is also displayed in the figure. Apparently, the improvement in quality of filtration due to the results for odd harmonics presented in this paper is considerable compared to [5], where only quarter-symmetric waveforms are studied. The THD of odd symmetric waveform is 2.75% compared to the quarter symmetric 18.3% (the even harmonics are uncontrolled).

VIII. Complexity of the Optimal Odd ML Problem

The complexity analysis of the optimal odd ML problem follows. Solving the RHS of the system of composite sum of powers p_i [see (23)] takes $\mathcal{O}(nn_C)$ number of operations. The moments μ_i are computed in $\mathcal{O}(n^2)$ operations according to (47), but a significantly faster algorithm can be found. We can, for instance, use the fast Newton iteration method that takes only $\mathcal{O}(n \log n)$ operations (this method employs an FFT technique for polynomial multiplication) (see [38] and [39]). The computation of Hankel linear system takes $\mathcal{O}(n \log^2 n)$ number of operations (superfast algorithm; see [40] and [41]) or we can use the well-known Levinson–Durbin algorithm with complexity $\mathcal{O}(n^2)$ operations. The calculation of matrix equation with a triangular Hankel matrix takes $\mathcal{O}(n \log n)$ operations (see [40]). It is somewhat more intricate to establish the complexity for computations of the zeros of polynomials $V(y)$ and $W(y)$ because many algorithms of different complexity are available. For example, the algorithm based on computing the eigenvalues of the companion matrix takes $\mathcal{O}(n^3)$ operations. In contrast, the combination of three-term recurrence algorithm (which takes $\mathcal{O}(n^2)$ operations), employing the property of interlacing the zeros (if it is possible, but this property is not always guaranteed for FOPs), and the iterative Newton algorithm leads to a linear number of operations—we easily compute the zeroes in every step. Hence, the highest possible number of operations is considered during the computation of the recurrence formula.

It is important to mention that the solution of the Hankel system is ill-conditioned for high n, which restricts the computation in double precision real arithmetic. Therefore, either of the polynomials $V(y)$ and $W(y)$ is also ill-conditioned, and computation of their roots is difficult from numerical point of view. By using extended precision arithmetic, the range of n can be enlarged. However, we show that the solution can also be expressed as the solution to a Padé approximation problem and, consequently, introduce FOPs. Numerically stable algorithms using properties of FOPs should therefore exist and are subject to research now.

For a special case of the quarter-symmetric waveforms [5], it is possible to adopt these results and devise the solution of system of sums of odd powers that is needed for the solution of this problem. It is sufficient to put the odd harmonics equal to zero and compute the polynomial $W(y)$ only. Such a solution was described in [5] and [6], and our procedures cover their solution for $n_C = 1$ as a special case.

IX. Conclusion

Efficient algorithms for the optimal odd ML problem in the single-phase connection have been developed and studied in this paper. In Section III, we revealed that an efficient analytical solution can be found only for odd and quarter-wave symmetric waveforms with arbitrary number of levels. The quarter-wave symmetric case is solved in [5] and [6]. Therefore, we concentrated on more general odd symmetry waveforms, including all harmonics.

Both cases lead to the solution of special systems of composite sum of powers that are derived from generalization of the Newton's identity. We formulated and solved the problem via Padé approximation. The optimal switching times are the zeros of shifted diagonal Padé approximation polynomials $[k, k+1]_F(y) = V_k^{[k,k+1]}(y)/W_{k+1}^{[k,k+1]}(y)$ for an odd number of switching n and diagonal Padé approximation $[k, k]_F(y) = V_k^{[k,k]}(y)/W_k^{[k,k]}(y)$ for an even n. Due to the connection between the theory of Padé approximation and FOPs, we demonstrated that $V(y)$ and $W(y)$ are FOPs, and we formulated other methods for the solution of the optimal odd ML problem. Namely, we derived an appropriate three-term recurrence formula, a determinantal formula, and a formulation via eigenvalue computation. The obtained polynomials are FOPs.

The results are summarized as follow.

1) After variable transformations, the solution of the optimal odd ML problem is given by the zeros of two polynomials $W(y)$ and $V(y)$ that are suitably sorted.
2) The polynomials $W(y)$ and $V(y)$ are given by the shifted diagonal Padé approximation

$$[k, k+1]_f(y) = V_k^{[k,k+1]}(y)/W_{k+1}^{[k,k+1]}(y) = \exp\left(-\sum_{j=1}^{\infty} \frac{p_j}{j} y^j\right) = F(y) \qquad (76)$$

for odd n and by the diagonal Padé approximation

$$[k, k]_f(y) = V_k^{[k,k]}(y)/W_k^{[k,k]}(y) = F(y)$$

for even n, where $p_j = \sum_{i=1}^{k} y_i^j - \sum_{i=k+1}^{n} y_i^j,\ j = 1, \ldots, n$, is computed according to (25) for ML and (29) for the bilevel odd waveform.
3) The polynomials $V(y)$ and $W(y)$ also give the solution of a Padé approximation and therefore constitute a set of FOPs, where the polynomial $V_k^{[k,k+1]}(y)$ is the associated polynomial (or polynomial of the second kind) to $W_{k+1}^{[k,k+1]}(y)$ (polynomial of the first kind) for odd n. In the case of even n, the polynomials $V_k^{[k,k]}(y)$ and $W_k^{[k,k]}(y)$ are deduced from the adjacent family of FOPs $V_k^{(1)[k,k+1]}(y)$ and $W_{k+1}^{(1)[k,k+1]}(y)$.
4) The solution to the optimal ML problem can be obtained through the following:
   a) the Hankel system in (53) and (56) for odd n and in (59) and (60) for even n: the complexity of a fast algorithm being $\mathcal{O}(n \log n^2)$;
   b) the simple three-term recurrence relationship in (62) and (64) for odd n and in (66) and (69): the complexity being $\mathcal{O}(n^2)$ operations;
   c) the determinants of special polynomial matrices in (70) and (71) for odd n and in (72) and (73) for even n;
   d) the eigenvalues of special matrices in (74) and (75) for odd n.

It is also important to stress that our solution is consistent with the solution of [5] in the case of waveforms with quarter symmetry.

At the end of this paper, a numerical example and experimental verification results are presented. The numerical example illustrates a complete solution of the ML and bilevel PWM problem and the presented exact results could not be obtained without our fast analytical methods. Experimental results verified our expected behavior of optimal ML and bilevel PWM problem. An active filter case study then illustrates an advantage of our approach compared to an existing analytical scheme for quarter-symmetric waveforms.

Acknowledgment

The authors would like to thank M. Špiller, Faculty of Electrical Engineering, Czech Technical University in Prague, Czech Republic, for his help and support with the experimental setup.

References

[1] D. G. Holmes and T. A. Lipo, *Pulse Width Modulation for Power Converters: Principles and Practice*, 1st ed. Hoboken, NJ: Wiley, Oct. 2003, ser. Power Engineering (IEEE).

[2] J. R. Wells, X. Geng, P. L. Chapman, P. T. Krein, and B. M. Nee, "Modulation-based harmonic elimination," *IEEE Trans. Power Electron.*, vol. 22, no. 1, pp. 336–340, Jan. 2007.

[3] R. Ray, D. Chatterjee, and S. Goswamie, "A modified reference approach for harmonic elimination in pulse-width modulation inverter suitable for distributed generations," *Electron. Power Compon. Syst.*, vol. 36, no. 8, pp. 815–827, Aug. 2008.

[4] A. Khaligh, J. Wells, P. Chapman, and P. Krein, "Dead-time distortion in generalized selective harmonic control," *IEEE Trans. Power Electron.*, vol. 23, no. 3, pp. 1511–1517, May 2008.

[5] D. Czarkowski, D. V. Chudnovsky, G. V. Chudnovsky, and I. W. Selesnick, "Solving the optimal PWM problem for single-phase inverters," *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.*, vol. 49, no. 4, pp. 465–475, Apr. 2002.

[6] D. V. Chudnovsky and G. V. Chudnovsky, "Solution of the pulse width modulation problem using orthogonal polynomials and Korteweg–de Vries equations," *Proc. Nat. Acad. Sci.*, vol. 96, no. 22, pp. 12 263–12 268, Oct. 1999.

[7] H. Huang, S. Hu, and D. Czarkowski, "Solving the ill-conditioned polynomial for the optimal PWM," in *Proc. 11th Int. Conf. Harmonics Quality Power*, 2004, pp. 555–558.

[8] Z. Du, L. Tolbert, J. Chiasson, and B. Ozpineci, "Reduced switching-frequency active harmonic elimination for multilevel converters," *IEEE Trans. Ind. Electron.*, vol. 55, no. 4, pp. 1761–1770, Apr. 2008.

[9] J. N. Chiasson, L. M. Tolbert, Z. Du, and K. J. McKenzie, "The use of power sums to solve the harmonic elimination equations for multilevel converters," *Eur. Power Electron. Drives J.*, vol. 15, no. 1, pp. 19–27, Feb. 2005.

[10] J. N. Chiasson, L. Tolbert, K. McKenzie, and Z. Du, "Elimination of harmonics in a multilevel converter using the theory of symmetric polynomials and resultants," *IEEE Trans. Control Syst. Technol.*, vol. 13, no. 2, pp. 216–223, Mar. 2005.

[11] J. N. Chiasson, L. M. Tolbert, K. McKenzie, Z. Du, and K. Wang, *Harmonic Elimination Technique and Multilevel Converters*, Knoxville, TN: Power Eng. Lab., Univ. Tennessee. [Online]. Available: http://powerelec.ece.utk.edu/research multilevel converters.html

[12] P. Kujan, M. Sebek, and M. Hromčík, "Construction of new system of polynomial equations for 3-phase multilevel voltage converter," in *Proc. Eur. Control Conf.*, Kos, Greece, 2007.

[13] J. Sun and I. Grotstollen, "Pulsewidth modulation based on real-time solution of algebraic harmonic elimination equations," in *Proc. 20th Int. Conf. Ind. Electron., Control Instrum.*, 1994, vol. 1, pp. 79–84.

[14] B. Ozpineci, L. M. Tolbert, and J. N. Chiasson, "Harmonic optimization of multilevel converters using genetic algorithms," *IEEE Power Electron. Lett.*, vol. 3, no. 3, pp. 92–95, Sep. 2005.

[15] Y. Liu, H. Hong, and A. Huang, "Real-time calculation of switching angles minimizing THD for multilevel inverters with step modulation," *IEEE Trans. Ind. Electron.*, vol. 56, no. 2, pp. 285–293, Feb. 2009.

[16] V. Agelidis, A. Balouktsis, and C. Cossar, "On attaining the multiple solutions of selective harmonic elimination PWM three-level waveforms through function minimization," *IEEE Trans. Ind. Electron.*, vol. 55, no. 3, pp. 996–1004, Mar. 2008.

[17] V. Agelidis, A. Balouktsis, and M. Dahidah, "A five-level symmetrically defined selective harmonic elimination PWM strategy: Analysis and experimental validation," *IEEE Trans. Power Electron.*, vol. 23, no. 1, pp. 19–26, Jan. 2008.

[18] V. Agelidis, A. Balouktsis, I. Balouktsis, and C. Cossar, "Multiple sets of solutions for harmonic elimination PWM bipolar waveforms: Analysis and experimental verification," *IEEE Trans. Ind. Electron.*, vol. 21, no. 2, pp. 415–421, Mar. 2006.

[19] T. Kato, "Sequential homotopy-based computation of multiple solutions for selected harmonic elimination in PWM inverters," *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.*, vol. 46, no. 5, pp. 586–593, May 1999.

[20] M. Perez, P. Cortes, and J. Rodriguez, "Predictive control algorithm technique for multilevel asymmetric cascaded H-Bridge inverters," *IEEE Trans. Ind. Electron.*, vol. 55, no. 12, pp. 4354–4361, Dec. 2008.

[21] L. G. Franquelo, J. Rodriguez, J. I. Leon, S. Kouro, R. Portillo, and M. A. M. Prats, "The age of multilevel converters arrives," *IEEE Ind. Electron. Mag.*, vol. 2, no. 2, pp. 28–39, Jun. 2008.

[22] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*, 1st ed. New York: Dover, Jun. 1965.

[23] A. Dickenstein and I. Z. Emiris, *Solving Polynomial Equations: Foundations, Algorithms, and Applications*. New York: Springer-Verlag, Jun. 2005, ser. Algorithms and Computation in Mathematics.

[24] D. A. Cox, J. Little, and D. O'Shea, *Using Algebraic Geometry*, 2nd ed. New York: Springer-Verlag, Mar. 2005.

[25] P. Kujan, M. Hromčík, and M. Šebek, "Effective solution of a linear system with Chebyshev coefficients," *Integral Transforms Spec. Funct.*, vol. 20, no. 8, pp. 619–628, Aug. 2009.

[26] R. Zippel, *Effective Polynomial Computation*. Norwell, MA: Kluwer, Jul. 1993, ser. International Series Engineering and Computer Science.

[27] Y. Wu and C. N. Hadjicostis, "On solving composite power polynomial equations," *Math. Comput.*, vol. 74, no. 250, pp. 853–868, Aug. 2004.

[28] L. Gosse and O. Runborg, "Existence, uniqueness and a constructive solution algorithm for a class of finite Markov moment problems," *SIAM J. Appl. Math.*, vol. 68, no. 6, pp. 1618–1640, Sep. 2008.

[29] G. A. Baker and P. Graves-Morris, *Padé Approximants*, vol. 1. Cambridge, U.K.: Cambridge Univ. Press, Jan. 1996.

[30] L. Lorentzen and H. Waadeland, *Continued Fractions With Applications*. Amsterdam, The Netherlands: North Holland, Apr. 1992.

[31] D. E. Knuth, *The Art of Computer Programming*, vol. 2, *Seminumerical Algorithms*, 3rd ed. Boston, MA: Addison-Wesley Longman, 1997.

[32] C. Brezinski, *Computational Aspects of Linear Control*, 1st ed. New York: Springer-Verlag, Jun. 2002.

[33] W. Gautschi, *Orthogonal Polynomials: Computation and Approximation*. London, U.K.: Oxford Univ. Press, Jun. 2004.

[34] S. Wolfram, Champaign, IL: Wolfram Research, Inc. [Online]. Available: http://www.mathematica.com

[35] P. Kujan, Optimal Odd Single-Phase Multilevel Problem. [Online]. Available: http://support.dce.felk.cvut.cz/pub/kujanp/software/optimalpwm/index.html

[36] V. Bhavaraju and P. Enjeti, "Analysis and design of an active power filter for balancing unbalanced loads," *IEEE Trans. Power Electron.*, vol. 8, no. 4, pp. 640–647, Oct. 1993.

[37] L. Asiminoaei, F. Blaabjerg, and S. Hansen, "Evaluation of harmonic detection methods for active power filter applications," in *Proc. 20th Annu. IEEE Appl. Power Electron. Conf. Expo.*, Mar. 6–10, 2005, vol. 1, pp. 635–641.

[38] A. Bostan and É. Schost, "A simple and fast algorithm for computing exponentials of power series," *Inf. Process. Lett.*, vol. 109, no. 13, pp. 754–756, Jun. 2009.

[39] Tech. Rep. P. Zimmermann and G. Hanrot, Newton iteration revised, 2004. [Online]. Available: http://webloria.loria.fr/~zimmerma/papers/fastnewton.ps.gz

[40] D. Bini and V. Pan, *Polynomial and Matrix Computations*, vol. 1, *Fundamental Algorithms*. New York: Springer-Verlag, Aug. 1994.

[41] A. W. Bojanczyk, R. P. Brent, and F. R. de Hoog, A weakly stable algorithms for general Toeplitz systems, Comput. Sci. Lab., ANU, Canberra, Australia. Tech. Rep. TR-CS-93-15. [Online]. Available: http://citeseer.ist.psu.edu/bojanczyk95weakly.html

Petr Kujan (M'10) received the M.Sc. and Ph.D. degrees from the Czech Technical University (CTU), Prague, Czech Republic, in 2004 and 2009, respectively.

He is currently a Researcher with the Department of Control Engineering, Faculty of Electrical Engineering. His research interests include the area of optimal ML/PWM problem, selective harmonic elimination, optimal class-D amplifiers.

Martin Hromčík received the M.Sc. and Ph.D. degrees from the Czech Technical University (CTU), Prague, Czech Republic, in 1999 and 2005, respectively.

He is currently an Assistant Professor with the Department of Control Engineering and a Researcher with the Center for Applied Cybernetics, Faculty of Electrical Engineering, CTU. He is also with the Department of Control Theory, Institute of Information Theory and Automation of the Academy of Sciences of the Czech Republic, Prague. His research interests include robust control, aircraft control systems, and power electronics.

Michael Šebek (M'90–SM'91) received the M.Sc. and Ph.D. degrees from the Czech Technical University (CTU), Prague, Czech Republic, and the Czech Academy of Science, Prague, in 1978 and 1981, respectively.

Since 2004, he has been a Professor with the Department of Control Engineering, Faculty of Electrical Engineering. He is currently the Head of the Department of Control Engineering, Faculty of Electrical Engineering, CTU. He is also with the Department of Control Theory, Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, Prague. He achieved important results in polynomial approach to linear control, contributed to robust H-infinity design and robust stability with parametric uncertainties, pioneered the theory of n-D polynomial equations, and developed a bunch of efficient computational algorithms for operations, functions, and equations with polynomial matrices. He is the author of over 260 research papers published in international scientific journals and conference proceedings. The Science Citation Index registers more than 250 citations of his works. His research interests include linear control and systems theory, robust control, n-D systems and filters, numerical methods and software for control and signal processing.

Paper C.

Tichý, V. - Barbera, M. - Collura, A. - Hromčík, M. - Hudec, R. - et al.: Tests of Lobster Eye Optics for Small Space X-ray Telescope. *Nuclear Instruments and Methods in Physics Research, Section A, Accelerators, Spectrometers, Detectors and Associated Equipment*. 2011, vol. 633, no. 1, p. S169-S171. ISSN 0168-9002. *(IF 1.142)*

Contents lists available at ScienceDirect

Nuclear Instruments and Methods in Physics Research A

journal homepage: www.elsevier.com/locate/nima

Tests of lobster eye optics for small space X-ray telescope

Vladimir Tichý [a,*], Marco Barbera [b,c], Alfonso Collura [c], Martin Hromčík [d], René Hudec [e,f], Adolf Inneman [g], Jan Jakůbek [h], Jiří Maršík [i], Veronika Maršíková [g], Ladislav Pína [j], Salvatore Varisco [c]

[a] *Department of Control Engineering, Faculty of Electrical Engineering, Czech Technical University in Prague, Karlovo náměstí 13, CZ-12135, Prague 2, Czech Republic*
[b] *Universitá degli Studi di Palermo, Dipartimento di Scienze Fisiche ad Astronomiche, Via Archirafi 36, IT-90123 Palermo, Italy*
[c] *INAF-Osservatorio Astronomico di Palermo, Piazza del Parlamento 1, IT-90134 Palermo, Italy*
[d] *Centre for Applied Cybernbetics, Faculty of Electrical Engineering, Czech Technical University in Prague, Karlovo náměstí 13, CZ-12135, Prague 2, Czech Republic*
[e] *Astronomical Institute, Academy of Sciences of the Czech Republic, Fričova 298, CZ-25165 Ondřejov, Czech Repubic*
[f] *Department of Radioelectronics, Faculty of Electrical Engineering, Czech Technical University in Prahue, Technická 4, CZ-16607, Prague 6, Czech Republic*
[g] *Rigaku Innovative Technologies Europe s.r.o., Novodvorská 994, CZ-14221, Prague 4, Czech Republic*
[h] *Institute of Experimental and Applied Physics, Czech Technical University, Horská 3a/22, CZ-12800, Prague 2, Czech Republic*
[i] *Division of Precision Mechanics and Optics, Department of Instrumentation and Control Engineering, Faculty of Mechanical Engineering, Czech Technical University in Prague, Technická 4, CZ-16607, Prague 6, Czech Republic*
[j] *Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University in Prague, Břehová 7, CZ-11519, Prague 1, Czech Republic*

ARTICLE INFO

Available online 18 June 2010

Keywords:
Lobster eye
X-ray optics
X-ray telescope
All-sky monitor
X-ray imaging
X-ray

ABSTRACT

The Lobster eye design for a grazing incidence X-ray optics provides wide field of view of the order of many degrees, for this reason it can be a convenient approach for the construction of space all-sky X-ray monitors. We present preliminary results of tests of prototype lobster eye X-ray optics in quasi parallel beam full imaging mode conducted using the 35 m long X-ray beam-line of INAF-OAPA in Palermo (Italy). X-ray images at the focal plane have been taken with a microchannel plate (MCP) detector at several energy values from 0.3 to 8 keV. The gain, the field of view and the angular resolution have been measured and compared with theoretical values.

1. Introduction

The lobster eye (LE) [1,2] is a grazing incidence reflective optics. In particular, the Schmidt design [2] uses two orthogonal sets of reflecting surfaces, each set focusing in one direction. One potential area of application is on future space all-sky monitors [3–5].

2. Tested lobster eye

In the experiments reported here, a prototype lobster eye called P-25 (Fig. 1) has been used. This LE manufactured in Rigaku Innovative Techonologies Europe s.r.o., Prague, Czech Republic consists of 2 × 60 flat reflecting plates coated by gold of RMS microroughness 1 nm. The plates have dimensions 24 × 24 mm^2, thickness 0.1 mm. Average spacing between the plates is 0.3 mm. The LE has 250 mm focal length and it is designed for optimal efficiency at 1 keV photon energy.

3. Experimental setup

The X-ray imaging tests were performed in the 35 m long X-ray beam-line in INAF-OAPA, Palermo, Italy [6,7]. The X-ray tube with exchangeable targets and filters is installed at one end of the vacuum pipe. The LE P-25 and the microchannel plate (MCP) detector [8] were installed on remote controlled positioning devices inside a vacuum test chamber located at the opposite side of the pipe (Fig. 2). Used MCP detector was in Chevron configuration with resistive anode encoder and front plate coated with KBr (Quantar Technology Inc., Santa Cruz, CA, USA) with 40 mm active diameter, and 100 μm FWHM spatial resolution. LE could be rotated around vertical (yaw angle) and horizontal (pitch angle) axis. The MCP could be translated in all basic three directions *X*, *Y* and *Z* (focus adjustment). For the measurements, six fluorescent energy lines have been chosen: 0.28, 0.93, 1.5, 2.9, 4.5 and 8.0 keV.

4. Results

Image, as focal cross with bright center typical for LE has been obtained (Fig. 3).

[*] Corresponding author. Tel.: +420 2 2435 5706; fax: +420 2 2491 8646.
E-mail address: tichyvl1@fel.cvut.cz (V. Tichý).

0168-9002/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.nima.2010.06.157

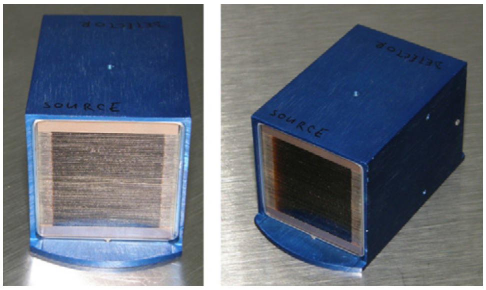


Fig. 1. Lobster eye P-25.

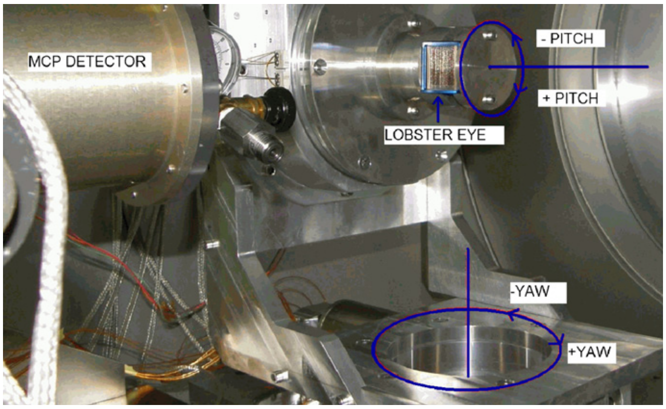


Fig. 2. Lobster eye and MCP on positioning devices.

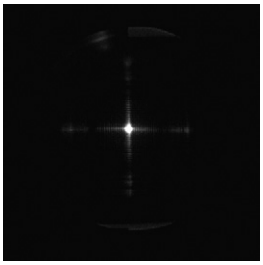


Fig. 3. Image with centered optics.

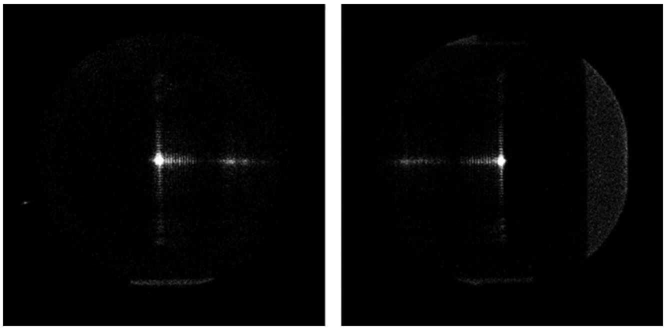


Fig. 4. Limiting positions in "yaw" angle.

Table 1
Gain at various energies.

Energy (keV)	0.28	0.93	1.5	2.9	4.5	8.0
Gain	78 ± 2	58 ± 1	75 ± 1	26 ± 1	27 ± 1	11 ± 1

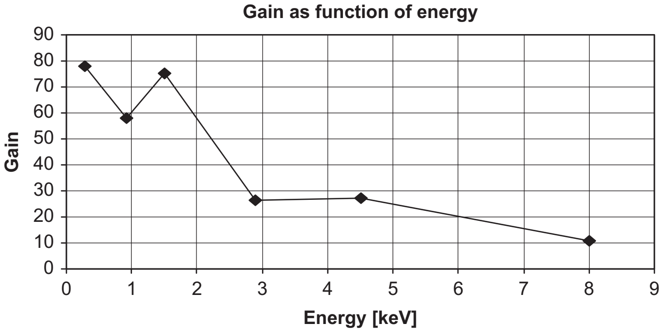


Fig. 5. Gain as function of incoming beam energy.

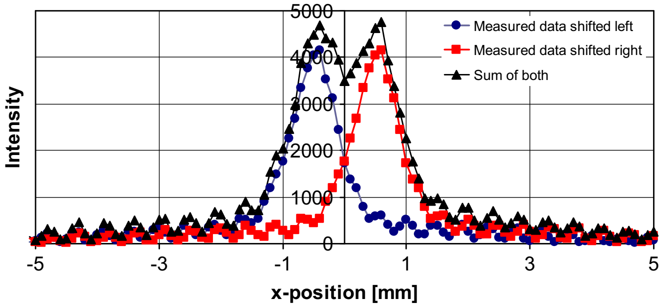


Fig. 6. Estimation of spatial resolution at energy 930 eV.

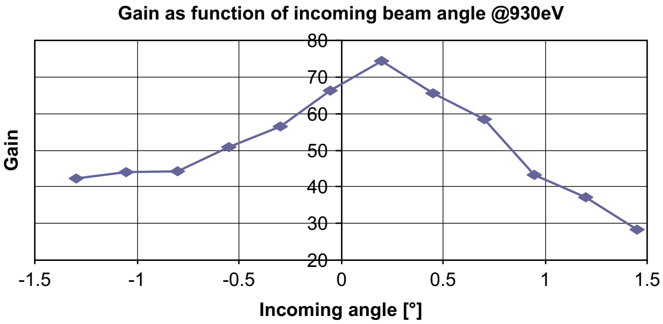


Fig. 7. Gain as function of incoming beam angle at energy 930 eV.

Scanning in "yaw" angle, we have found limiting positions when one arm of the focal cross vanishes (Fig. 4). The measured angular width between these positions is 2.9 ± 0.1°, this value can be considered as the field of view (FOV). It is determined only by LE geometry and it does not depend on the energy.

The optical axis of the LE was considered as the middle position between the mentioned limiting values. In this position, FWHM of the central peak is always in the range 12–13′ (arcmin). The measured gain, estimated by the ratio of the flux in the square of size 300 μm (size of chambers of LE) around peak, and incident flux onto the LE, is shown in Table 1 and Fig. 5 for the different investigated photon energies.

The angular resolution can be estimated summing profile of focal cross along the one line with the same data shifted to simulate resulting image of two point sources. Searching for position, when intensity between peaks of the sum falls to 80% of intensity of lower peak, value 13 ± 1′ was estimated as spatial resolution of tested LE at energy 930 eV (Fig. 6). The gain has been measured at various yaw angles within the FOV. Results are shown in Fig. 7.

In the paper [1], relations for estimation of gain and spatial resolution are published. There, it is supposed LE is accurately manufactured from ideally thin and ideally reflecting mirrors (a photon energy is not taken into account there). The relations give spatial resolution of this LE as 5.5′ and gain as 1736.

5. Conclusions

A prototype lobster eye P-25 was tested in the X-ray vacuum beam-line at INAF-OAPA. The measured FWHM spatial resolution is approximately 2 times larger than theoretical value. This is caused probably by small deformations of plates (they are not ideally flat) and some manufacturing aberrations. The gain vs. yaw angle shown in Fig. 7 is indicative that the LE is assembled a little asymmetric. Estimated gain is approx. 20 × smaller than mentioned value 1736, however this value is estimated for ideal LE assembled from ideally reflecting and ideally thin mirrors. Preliminary computations based on simulations, give gain around 760 for photons of energy 930 eV. In that simulations, thickness of mirrors and their reflectivity for this energy have been taken into account. Other decrease of gain is caused mainly by unflatness of mirrors and other manufacturing aberrations. More detailed analysis of the problem will be the subject of an another paper.

Acknowledgments

Authors would like to thank Prof. Salvatore Sciortino, director of INAF-Osservatorio Astronomico di Palermo for providing the use of the X-ray beam-line for these tests. Thanks to Rigaku Innovative Technologies Europe s.r.o. for providing of the lobster eye. This research was supported from grants MSMT ME 09028, GA CR 102/09/0997, GA AV CR IAAX01220701 and Ministry of Education of the Czech Republic Grant no. 1M0567.

References

[1] W.H.K. Schmidt, Nucl. Instr. and Methods 127 (1975) 285.
[2] J.R.P. Angel, ApJ 233 (1979) 364.
[3] A. Inneman, et al., Nucl. Phys. B—Proc. Supl 166 (2007) 229.
[4] R. Hudec, et al., Proc. SPIE 4851 (2003) 578.
[5] V. Tichý et al., Proc. ICSO (2008).
[6] M. Barbera, R. Candia, A. Collura, G. Di Cicca, C. Pelliciari, S. Sciortino, S. Varisco, Proc. SPIE (2006), 6266, 62663F.
[7] 〈http://www.astropa.unipa.it/XACT/index.html〉.
[8] J.L. Wiza, Nucl. Instr. and Methods 162 (1979) 587.

Paper D.

Nováková, J. - Hromčík,M. – Jech, R.: Dynamic Causal Modeling and subspace identification methods, Biomedical Signal Processing and Control, available online August 9, 2011 DOI: 10.1016/j.physletb.2003.10.071. *(IF 0.734)*

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Dynamic Causal Modeling and subspace identification methods

J. Nováková [a,b,*], M. Hromčík [a,b], R. Jech [c]

[a] Department of Control Engineering, Czech Technical University in Prague, Faculty of Electrical Engineering, Prague, Czech Republic
[b] Institute of Information Theory and Automation, Prague, Czech Republic
[c] Department of Neurology, Charles University in Prague, First Faculty of Medicine and General Teaching Hospital, Prague, Czech Republic

ARTICLE INFO

Article history:
Received 21 October 2010
Received in revised form 2 July 2011
Accepted 8 July 2011
Available online xxx

Keywords:
System identification
Subspace identification methods
Dynamic Causal Modeling
fMRI
DCM

ABSTRACT

The main contribution of the paper is in formulating the problem of detection of brain regions structure within the framework of dynamic system theory. The motivation is to see if the mature domain of experimental identification of dynamic systems can provide a methodology alternative to Dynamic Causal Modeling (DCM) which is currently used as an exclusive tool to estimate the structure of interconnections among a given set of brain regions using the measured data from functional magnetic resonance imaging (fMRI). The key tool proposed for modeling the structure of brain interconnections in this paper is subspace identification methods which produce linear state-space model, thus neglecting the bilinear term from DCM. The procedure is illustrated using a simple two-region model with maximally simplified linearized hemodynamics. We assume that the underlying system can be modeled by a set of linear differential equations, and identify the parameters (in terms of state space matrices), without any a priori constraints. We then transform the hidden states so that the implicit state matrix has a form or structure that is consistent with the generation of (region-specific) hemodynamic signals by coupled neuronal states.

1. Introduction

Nowadays, inference about connectivity or coupling among brain regions, using fMRI, usually rests upon some form of Dynamic Causal Modeling (DCM). DCM uses Bayesian techniques to identify the underlying neuronal system in terms of coupling parameters. Crucially, one has to specify prior constraints on the sparsity or form of the connections and then test different models (forms) of connectivity. In this paper, we work on a more efficient, direct approach. DCM is used to compare mathematical models with and without specific connections which entails fitting or inverting different models and then comparing their evidence. It is a methodology which enumerates possible models first, and then tests their validity using the conventional tools for testing statistic hypotheses [5]. The identification can take a considerable amount of time, especially when one compares large numbers of models.

The main contribution of our work is to estimate the full connectivity of any DCM (under linear and first order assumptions) in a way that is extremely efficient. This may be especially useful in the context of DCM, because recent developments in model comparison allow one to evaluate the evidence of reduced models (in which some connections are omitted) given the estimates of a full model [9].

The basic idea behind our approach is to estimate the significance of interconnections among the brain regions by identifying the coupling among the states of an underlying linear state-space model. This is done by finding the state matrices describing the dynamics of neuronal states through the measured hemodynamic responses, using conventional linear system identification techniques. However, these methods do not apply constraints on the form of the state matrix. We finesse this problem by modeling the data with a number of hidden states that is greater than the number of observed brain regions. We then find a transformation of the hidden states that conforms to the known expected block structure of the state matrix appropriate for our problem. This transformation relies on the numerically reliable Schur decomposition of the original state matrix and related eigen decompositions. We can then interpret the transformed states in terms of neuronal and hemodynamic states. The transformed state matrix gives direct information on couplings between particular neuronal states, and also defines the mapping from neuronal to hemodynamic subsystems.

The structure of the paper is a follows. In the next section we give some overview of DCM analysis. In the third section we present our alternative methodology. The fourth section brings a simple

Abbreviations: BOLD, blood oxygenation level-dependent; DCM, Dynamic Causal Modeling; MRI, magnetic resonance imaging; fMRI, functional MRI; SPM, Statistical Parametric Mapping.

[*] Corresponding author at: Department of Control Engineering, Faculty of Electrical Engineering, Czech Technical University in Prague, Technická 2, 166 27 Prague 6, Czech Republic. Tel.: +420 2 2435 7648.
E-mail address: tauchjan@fel.cvut.cz (J. Nováková).

1746-8094/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bspc.2011.07.002

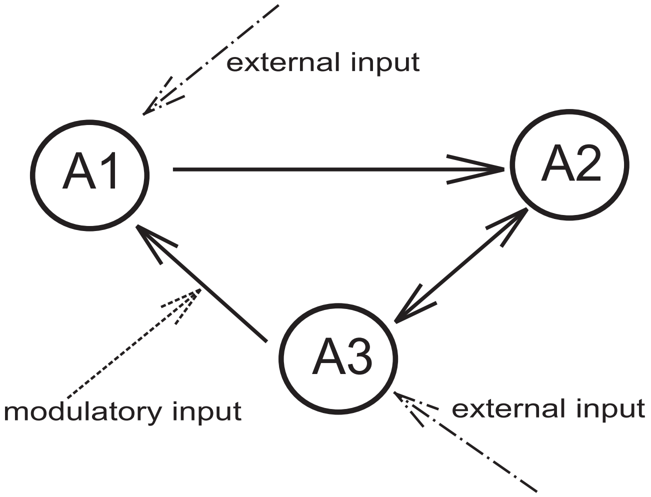


Fig. 1. Brain regions A1, A2, A3, intrinsic connections among them and two types of input signals – a representative of a predefined model.

computational example. The paper is concluded with a summary and a list of open problems.

1.1. Overview of DCM analysis

Dynamic Causal Modeling (DCM) is a statistical technique for detection of interconnections among selected brain regions [2,5,12,14]. DCM assumes a bilinear model in the form (1) and the interconnections among brain regions are qualitatively and quantitatively characterized by its parameters (note the presence of so-called modulatory inputs that modulate interconnections directly).

$$\dot{x}(t) = \left(A + \sum_{j=1}^{M} u_j(t) B_m^j \right) x(t) + Bu(t), \quad y(t) = Cx(t) + Du(t) \qquad (1)$$

where A is effective connectivity matrix for interconnections among regions, B_j is effective connectivity matrix encoding the changes in intrinsic connections induced by jth modulatory input u_j, and matrix B representing strength of extrinsic inputs leading directly to brain regions, see Fig. 1. DCM procedure then combines the bilinear neuronal model (1) of interacting regions with the biophysical model by Friston, based on principles of the *hemodynamic model* and *balloon model* [7] which describes how the unmeasured neuronal activity in a given brain region is transformed into the hemodynamic responses measured by the fMRI. The input signal is generally a deterministic on-off function representing stimulation process

The first two steps of DCM analysis are selection of several brain regions, described by coordinates of voxel (volumetric pixel) cluster, and definition of inferences (hypotheses) about the region interactions which will be confronted with real fMRI data in the statistical hypothesis-testing manner. The hypotheses usually result from clinical experience and empirical knowledge of a neurologist, trained in functional brain organization. The necessity to rely on an expert in this step can be regarded as a drawback of this method and full enumeration of all possible interaction structures is combinatorially prohibitive. The final step – testing statistical hypotheses – can be computationally very demanding, it can easily take up to a few days on a regular PC.

The DCM procedure is implemented in the SPM (Statistical Parametric Mapping) toolbox for Matlab [20], a popular tool for processing fMRI data. In addition, the toolbox also contains a simulator for characteristic time series generated by several activated brain regions.

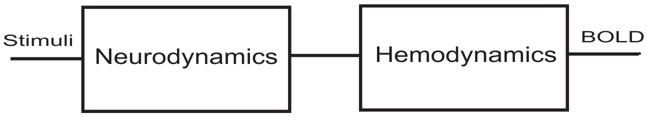


Fig. 2. Brain dynamics system structure – two types of dynamics, at first faster dynamics, slower dynamics forms output BOLD signal in each activated brain region.

2. Systems identification approach to detection of brain region interconnection

2.1. Definition of the dynamic system

The main goal of this section is to define the complex dynamic system represented by selected brain regions within the context of the systems theory, and to cast it as a task for system identification procedures.

The *system*, as interpreted by the systems theory, is a complex object consisting of *interconnected* subsystems and components which transforms *inputs* into *outputs* and this transformation can be characterized by a mathematical model, usually in the form of differential equations. The input stimulus signals that enter into the *brain system* reflect the particular fMRI based neurological experiment, and can be modeled as rectangular signals (on/off or active/inactive) as they correspond to hand motion, pictures projection, electrical stimulation, etc. The measured outputs are BOLD signals which are usually visualized as volumetric 3D plots. They can also be viewed as rectangular for which at every time instance the measured value assumes a shape of a 3-dimensional array (cube). Hence the input–output behavior of the *brain system* can be measured experimentally. However the system is characterized by specific intrinsic structure comprised of two different parts called neurodynamics and hemodynamics, see Fig. 2. The input (stimulus) signals enter the faster dynamics (neurodynamics) representing the intrinsic interconnections among brain regions. Neurodynamics could be modeled by several first order systems, each corresponding to a given brain region and their intrinsic connections as is done by DCM in fact [5]. The neuronal response of every brain region is only observed in the fMRI data after passing through the slower hemodynamics part, which can be modeled as a simple system (filter) for each brain region separately. In contrast to the nonlinear balloon model used within DCM, higher order hemodynamic linear filters (at least order two) are necessary to capture the oscillatory behavior as shown in the next section. The structure is revealed in Fig. 3 for a simple case.

The task of brain regions structure detection has now been formulated as a system identification problem. All the input and output signals are measured, we can therefore apply classical black-box identification methods used commonly in diverse industries [1,8,10,15,18].

2.2. fMRI data fitting by subspace methods

We now proceed to application of subspace identification methods for simulated fMRI data. First we model the hemodynamic response only. The function *spm_dcm_create* from SPM toolbox [20] can be used to generate such a realistic data set. Crucially, the method requires having more states corresponding to the neuronal dynamics than there are outputs corresponding to the hemodynamics, see Fig. 4. For the purpose of numerical simulations we generated data by convolving boxcar stimulus functions with canonical hemodynamic response functions and added white

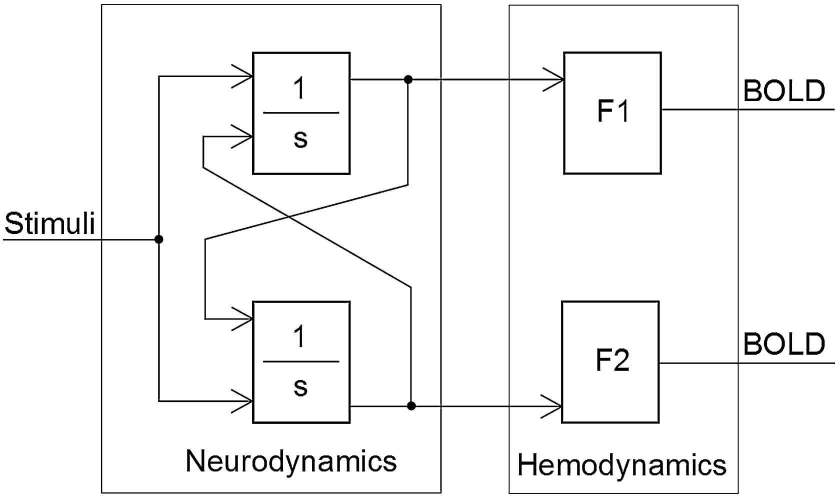


Fig. 3. The detailed structure of brain system – neurodynamics is modeled by reciprocally connected first order systems. Each region has also own hemodynamics represented by higher order system.

Gaussian observation noise of various amplitudes. The details of these simulations are provided in the figure legend.

We focused on subspace N4SID identification methods implemented in System Identification Toolbox for Matlab (version 2007b) [19]. Subspace methods combine results of systems theory, geometry and numerical linear algebra [4,11]. They seem suitable for our task especially for their fine numerical reliability for MIMO system identification. In addition, they give rise to models in the state-space form directly. We have used this identification method for hemodynamics modeling of a single region by fitting to fMRI data simulated by SPM toolbox. The simulation experiments were carried out for various combinations of important data parameters (number of samples, signal-to-noise ratio) and we proved applicability of subspace identification methods for fitting simulated fMRI data by linear dynamic higher-order models (discrete-time domain, sampling period 1.7 s, see [17] for details), see Fig. 4. The

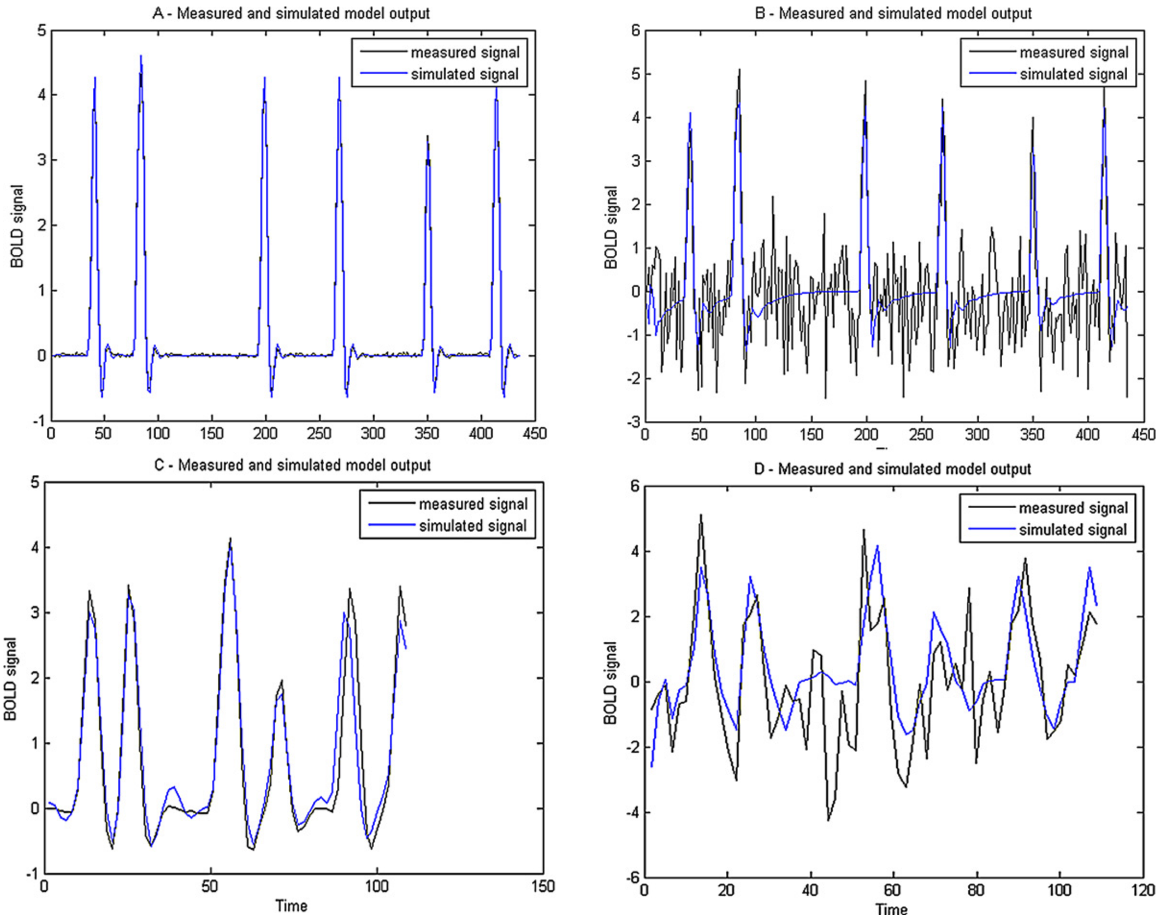


Fig. 4. fMRI data fitting for data with different parameters – A (SNR = 50, number of samples = 256) – 3rd order model, B (SNR = 1, number of samples = 256) – 6th order model, C (SNR = 50, number of samples = 64) – 6th order model, D (SNR = 1, number of samples = 64) – 8th order model. For more details see [17].

subspace identification proves useful here and fits successfully the simulated data by the identified linear model as is shown at the first triplet of pictures; just for the last data set with smaller signal-to-noise ratio and number of samples the model is not able to fit data sufficiently. We can summarize that subspace identification methods are a promising technique for hemodynamic response fitting. We attempt to extend the identification procedure to the whole system including neurodynamics too in the next section.

It should be also noted that an alternative approach to the conventional identification methods is Volterra series approach, which expresses the output signal as a nonlinear convolution of the inputs. In fact, it is an input–output description of a system without the necessity of characterizing the state variables. In spite of this, the Volterra series are able to characterize the effective connectivity by the constituent Volterra kernels, see [6] for details. In comparison with Volterra series modeling, subspace identification methods rely completely on the linear systems theory and methods and produce linear state space description, featuring state variables as crucial elements for our procedure of brain regions structure detection.

2.3. Identification procedure for brain system structure

Subspace identification methods return a linear state space model (2). The matrix A represents the dynamics, B is related to the inputs and C characterizes the outputs. The matrix D indicates direct connection from input to output in general. Choosing the linear model (2) instead of the bilinear model (1) for brain region system description is intentional, ignoring so-called modulatory inputs illustrated in Fig. 1 is motivated by simplicity. The hidden states x include certain transformation of all the neuronal and hemodynamic states in our model. This means the number of hidden states is much greater than the number of observations y (and that C is not a square matrix).

$$\begin{aligned} x(t) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) + Du(t) \end{aligned} \tag{2}$$

If we had the state space description in suitable form we could see intrinsic connections among selected brain regions directly. Unfortunately matrices A, B and C as a result of subspace methods are usually full and inappropriate to the specific structure of the *brain system*. Apparently it is necessary to transform the state space model into a realization reflecting separation of neurodynamics and hemodynamics. Matrix D of identified state space description is zero because there is no direct connection from input to output.

One way to enforce this structure into the state space realization is a similarity transformation with a suitable transformation matrix T as in (3).

$$A_{new} = T^{-1}AT, \quad B_{new} = T^{-1}B, \quad C_{new} = CT, \quad D_{new} = D. \tag{3}$$

These transformed matrices correspond to a transformation of variables in the form $x = Tx_{new}$, where x_{new} become our new desirable states that can be interpreted directly as neuronal and hemodynamic ones. The next section illustrates construction of the T matrix in a simple case which corresponds to the special brain structure according to Fig. 3.

3. Example

We consider a system including one input (stimulus) signal, two brain regions and two output (BOLD) signals, see Fig. 3. The output filters for hemodynamics modeling are considered as first order systems only for this moment (note that it does not fully correspond to orders necessary to model accurately hemodynamic filters as identified in the Section 2.2, so it is not possible to use SPM toolbox as the data generator, and we use the generator (5) instead).

The subspace identification methods yield the full matrices A, B, and C, see (4). The matrix D is zero (no direct throughputs are present in the system considered).

$$A = \begin{pmatrix} a_1 & a_2 & a_3 & a_4 \\ a_5 & a_6 & a_7 & a_8 \\ a_9 & a_{10} & a_{11} & a_{12} \\ a_{13} & a_{14} & a_{15} & a_{16} \end{pmatrix} \quad B = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix}$$

$$C = \begin{pmatrix} c_1 & c_2 & c_3 & c_4 \\ c_5 & c_6 & c_7 & c_8 \end{pmatrix} \tag{4}$$

$$A = \begin{pmatrix} e_1 & 0 & g_1 & 0 \\ 0 & e_2 & 0 & g_2 \\ 0 & 0 & e_3 & c_{12} \\ 0 & 0 & c_{21} & e_4 \end{pmatrix} \quad B = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix} \quad C = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \tag{5}$$

However, the desired form is in (5). This form reveals the specific structure of the *brain system* with the neuronal dynamics affected directly by the inputs and the hemodynamics projected immediately into the measured outputs. Matrix A contains the eigenvalues e_1, e_2 and the gain coefficients g_1, g_2 defining the hemodynamic SISO filters associated to a particular brain region. The lower right submatrix represents the (much faster) neurodynamics. The coefficients c_{12} and c_{21} are the crucial parameters which determine the intrinsic neuronal interconnections between the two modeled brain regions. The matrix B represents the structure of inputs and matrix C corresponds to the structure of outputs, in agreement with Fig. 3.

3.1. Similarity transformations

This section describes the sequence of similarity transformations steps leading from the full state-space model (4) to the structured form realization (5) from which the coupling parameters c_{12} and c_{21} can be detected. We consider a system with one input and two brain regions, each modeled by first order dynamics and with corresponding two output BOLD signals. Each similarity transformation follows the conventional rule (3).

The first step is Schur decomposition applied to the identified dynamic matrix A. It yields zero elements under the main diagonal on which the eigenvalues are displayed. These are then ordered to separate the eigenvalues of hemodynamics (slow) and neurodynamics (fast). The subsequent steps are devised to impact the remaining parts of state space description and to preserve the effect of the previous transformation steps. In this way, the eigenvectors of a selected submatrix of the new dynamic matrix A are calculated and used for diagonalization of the submatrix representing hemodynamics filters, and the null space of output matrix C is used for zeroing its selected elements. We also use inverse submatrix for adjustment of parts concerning gain coefficients of output (hemodynamic) filters. All steps are detailed in a Matlab pseudocode-form below, and are illustrated by a numerical example in the next section.

```
»[T1,A1] = schur(A)
»[T2,A2] = ordschur(T1,A1,[1,2,3,4])
»G2 = ss(T2\A*T2, T2\B, C*T2, 0);

»[t1,aj1] = eig(G2.a(1:2,1:2));
»C2 = G2.c*blkdiag(t1,eye(2));
»T3 = T2*blkdiag(t1,eye(2))*[eye(4,2), null(C2)];
»G3 = ss(T3\G2.a*T3, T3\G2.b, G2.c*T3, 0);

»t2 = inv(G3.a(1:2,3:4));
»T4 = [eye(2) zeros(2);zeros(2) t2];
»G4 = ss(T4\G3.a*T4, T4\G3.b, G3.c*T4, 0);
```

Note that the transformation matrix $T1$ resulting from the Schur decomposition is applied to identified dynamic matrix A

Please cite this article in press as: J. Nováková, et al., Dynamic Causal Modeling and subspace identification methods. Biomed. Signal Process. Control (2011), doi:10.1016/j.bspc.2011.07.002

and the Matlab function *ordschur* is able to sort eigenvalues on the main diagonal, so we obtain new state space description *G*2 with dynamic matrix *G*2.*a* containing separated hemodynamic and neurodynamic eigenvalues.

The next transformation with the matrix *T*3 includes null space of output matrix *C*2, and the eigenvectors of the hemodynamic part of matrix *G*2.*a* are used as well. The hemodynamic part of the dynamic matrix *G*2.*a* is now diagonal due to appropriate eigenvectors application, the input matrix *G*2.*b* is also modified (zeroing of values representing input to the hemodynamic filters). The transformation also ensures zeroing of values representing output from neurodynamic part in the matrix *G*2.*c*. We can see that there is no external input into the hemodynamic filters, only from the neurodynamic part, and there is no measured output from the neurodynamic part, which is desirable.

The last important transformation step is *G*3.*a* adjustment, and we namely want to impact the submatrix related to the hemodynamic filters. So we apply transformation matrix *T*4 including inverse of a submatrix of *G*3.*a* to diagonalize appropriate submatrix. So the almost final result of these transformation steps is the dynamic matrix *G*4.*a* with eigenvalues and gains of hemodynamic filters in the upper part, and the submatrix concerning neurodynamics at the bottom which contains interconnections between two regions on the next diagonal. Matrices *G*4.*b* and *G*4.*c* are also modified according to the form in (5) and they reflect *brain system* structure (no external input into hemodynamic filters, no external output from the neurodynamic part).

3.2. Numerical example

Data for the identification procedure were generated using the system (6) with structure according to Fig. 3.

$$A = \begin{pmatrix} -1 & 0 & 1 & 0 \\ 0 & -2 & 0 & 2 \\ 0 & 0 & -10 & 0 \\ 0 & 0 & 5 & -10 \end{pmatrix} \quad B = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}$$

$$C = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \tag{6}$$

Matrices (7) illustrate the state space description of a system with two regions modeled by first order systems, one input and two outputs, as identified by a subspace identification method [3,4,11] implemented in the functions of System Identification Toolbox for Matlab.

$$A = \begin{pmatrix} -2.36 & 12.22 & -7.52 & -11.40 \\ -4.91 & -5.06 & 7.07 & 4.15 \\ 3.14 & -0.13 & -2.18 & -3.02 \\ -0.09 & 12.61 & -8.61 & -13.39 \end{pmatrix} \quad B = \begin{pmatrix} 3.15 \\ -1.22 \\ 0.38 \\ 3.44 \end{pmatrix}$$

$$C = \begin{pmatrix} -1.16 & 0.20 & -0.27 & 1.16 \\ -1.79 & 0.50 & 0.14 & 1.80 \end{pmatrix} \tag{7}$$

Particular similarity transformation steps described in Section 3.1 leading to state space description (5) were applied. The final result reflecting the desired structure is in (8). We can also see at Fig. 5 that step response of transformed system is the very same as step response of the data generator. Therefore we did not change the input–output response of the originally identified system by similarity transformation and we found one of the equivalent state realization that reveal coupling structure between neurodynamics and hemodynamics.

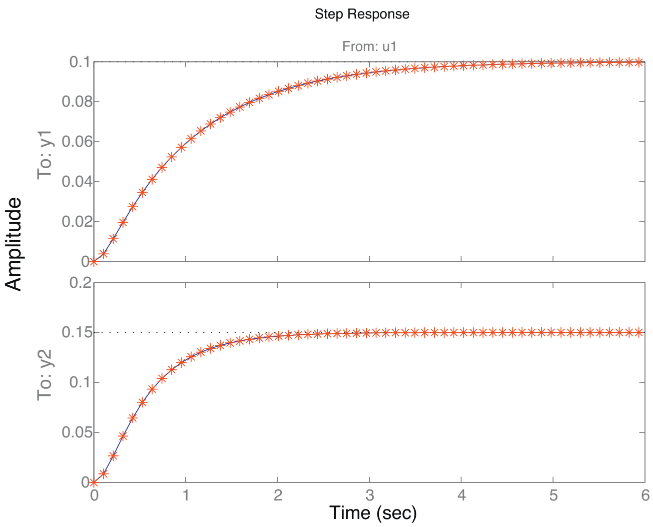


Fig. 5. Step response of identified transformed system and original system for the data generation are the same.

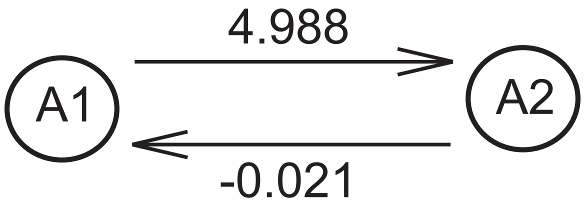


Fig. 6. Diagram of the final detected connections.

$$A = \begin{pmatrix} -0.999 & 0 & 0.999 & 0 \\ 0 & -2.001 & 0 & 2.001 \\ 0 & 0 & -10.071 & -0.021 \\ 0 & 0 & 4.988 & -9.986 \end{pmatrix} \quad B = \begin{pmatrix} 0 \\ 0 \\ 1.001 \\ 0.999 \end{pmatrix}$$

$$C = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \tag{8}$$

Here the (4, 3) element of the matrix indicates a significant connection between the two regions, that would be visualized in the DCM diagrams style as in Fig. 6.

4. Conclusions and open problems

In this paper we proposed to formulate the task of detection of brain regions structure within the well-established and mature framework of system identification as a promising alternative to Dynamic Causal Modeling which is based on statistical hypothesis-testing. The motivation for developing this alternative approach comes from the need to reduce the computational burden so that the fMRI data can be processed in real-time. We proposed a concrete computational procedure based on the popular subspace identification techniques applied to the measured (simulated respectively) fMRI data combined with a similarity transformation which enforces the structure into the problem (this structure accounts for the separation of dynamics into the neuronal and hemodynamic part). The procedure was demonstrated by a simple simulation example.

Surely the proposed and demonstrated method simplifies the problem a lot by the assumption of a linear model: the bilinear terms in the neuronal dynamics considered within the DCM framework are neglected here because modification of subspace identification techniques for bilinear models does not appear to be straightforward and is subject to further research.

The model of hemodynamics is also considered as an LTI model although currently some nonlinear models (such as the balloon model) are used within the fMRI community. In addition, the procedure is fully functional for first order hemodynamic filters only. It was observed though that when using system identification techniques to some fMRI data generated by the SPM toolbox, every output hemodynamics filter should be modeled as at least a second order system with complex conjugate eigenvalues, reflecting the oscillatory response [17]. The similarity transformations then become more complicated and the procedure proposed in this paper cannot handle it at this moment

Acknowledgements

We would like to express our deep gratitude to the anonymous reviewers for their inspiring comments and recommendations leading to considerable improvements of our paper. We would also like to thank our colleague Zdeněk Hurák for his careful proofreading.

The study was supported by the Czech Ministry of Education (grants MSM6840770038 and 1M0567) and from the Czech Ministry of Health (grants IGA MZČR NR8937-4 and 1A/8629-5).

References

[1] M. Basseville, A. Benveniste, M. Goursat, L. Mevel, Subspace-based algorithms for structural identification, damage detection and sensor data fusion, Journal of Advance in Signal Processing – Special issue on Advances in Subspace-based Techniques for Signal Processing and Communications 1 (2007) 200.

[2] T. Ethofer, S. Anders, M. Erb, C. Herbert, S. Wiethoff, J. Kissler, W. Grodd, D. Wildgruber, Cerebral pathways in processing of affective prosody: a dynamic causal modeling study, NeuroImage 30 (2005) 580–587.

[3] W. Favoreel, B.D. Moor, P.V. Overschee, Subspace identification of bilinear systems subject to white inputs, IEEE Transactions on Automatic Control 44 (1999) 1156–1165.

[4] W. Favoreel, B.D. Moor, P.V. Overschee, Subspace state space system identification for industrial processes, Journal of Process Control 10 (2000) 149–155.

[5] K.J. Friston, L. Harrison, W. Penny, Dynamic causal modeling, NeuroImage 19 (2003) 1273–1302.

[6] K.J. Friston, Human Brain Function, Volterra Kernels and Effective Connectivity, Elsevier Press, London, 2004.

[7] K.J. Friston, A. Mechelli, R. Turner, C.J. Price, Nonlinear responses in fMRI: The Balloon Model, Volterra Kernels, and other hemodynamics, NeuroImage 12 (2000) 466–477.

[8] S. Gannot, M. Moonen, Subspace methods for multimicrophone speech dereverberation, Journal of Applied Signal Processing 11 (2003) 1074–1090.

[9] K.J. Friston, B. Li, J. Daunizeau, K.E. Stephan, Network discovery with DCM, NeuroImage (2010).

[10] H. Garnier, W. Liuping, Identification of Continuous-Time Models from Sampled Data, Springer, London, 2008.

[11] T. Katayama, Subspace Methods for System Identification, Springer, London, 2005.

[12] S.J. Kiebel, O. David, K.J. Friston, Dynamic causal modeling of evoked responses in EEG/MEG with lead field parametrization, NeuroImage 30 (2006) 1273–1284.

[14] W.D. Penny, K.E. Stephan, A. Mechelli, K.J. Friston, Comparing dynamic causal models, NeuroImage 22 (2004) 1157–1172.

[15] J.K. Rice, M. Verhaegen, Distributed Control: a sequentially semi-separable approach, in: Proceedings of the 47th IEEE Conference on Decision and Control, Mexico, 2008.

[17] J. Tauchmanova, M. Hromcik, Subspace identification methods and fMRI analysis, in: IEEE EMBS, Vancouver, 2008.

[18] Y. Yao, F. Gao, Subspace identification for two-dimensional dynamic batch process statistical monitoring, Chemical Engineering Science 63 (2008) 3411–3418.

[19] Websites of Matlab Mathworks. http://www.mathworks.com/ (accessed August 3, 2011).

[20] Websites of Statistical Parametric Mapping toolbox. http://www.fil.ion.ucl.ac.uk/spm/ (accessed August 3, 2011).

Paper E.

Haniš, T. - Hromčík M.: Optimal Sensor placement and Spillover Suppression, Mechanical Systems and Signal Processing
(in press, January 2012) *(IF 1.762)*

Mechanical Systems and Signal Processing ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at SciVerse ScienceDirect

Mechanical Systems and Signal Processing

journal homepage: www.elsevier.com/locate/ymssp

Optimal sensors placement and spillover suppression

Q1 Tomas Hanis [a,*], Martin Hromcik [b]

[a] Department of Control Engineering, Czech Technical University in Prague, Faculty of Electrical Engineering, Technicka 2, 166 27 Prague, Czech Republic
[b] Center for Applied Cybernetics, Czech Technical University in Prague, Faculty of Electrical Engineering, Karlovo namesti 13-G, 121 35 Prague, Czech Republic

ARTICLE INFO

Article history:
Received 20 July 2010
Received in revised form
8 September 2011
Accepted 1 December 2011

Keywords:
Optimal sensor placement
Effective independence (EFI)
Spillover suppression

ABSTRACT

A new approach to optimal placement of sensors (OSP) in mechanical structures is presented. In contrast to existing methods, the presented procedure enables a designer to seek for a trade-off between the presence of desirable modes in captured measurements and the elimination of influence of those mode shapes that are not of interest in a given situation. An efficient numerical algorithm is presented, developed from an existing routine based on the Fischer information matrix analysis. We consider two requirements in the optimal sensor placement procedure. On top of the classical EFI approach, the sensors configuration should also minimize spillover of unwanted higher modes. We use the information approach to OSP, based on the effective independent method (EFI), and modify the underlying criterion to meet both of our requirements—to maximize useful signals and minimize spillover of unwanted modes at the same time. Performance of our approach is demonstrated by means of examples, and a flexible Blended Wing Body (BWB) aircraft case study related to a running European-level FP7 research project 'ACFA 2020—Active Control for Flexible Aircraft'.

1. Introduction

Optimal sensor placement (OSP) in mechanical systems and structures has become a popular and frequently discussed research topic during last 10 years. Applications cover modeling, identification, fault detection, and active control of such systems as bridges [9,8], rail wagons [14], large space structures [15]. The goal is to tell the designers of the whole mechanical system where displacement, force, inertial acceleration, or other sensors are to be installed so that they are as informative as possible.

Various approaches have been developed. We will mention two in brief. The former, information based approach, is based on the analysis of the output shape matrix. An iterative elimination algorithm, denoted as EFI (for "Effective Independence") has been developed that repeatedly deletes the lines of the initial, full output shape matrix with lowest amount of information, measured by either the trace or determinant of an underlying Fischer information matrix. See [4] for more detailed treatments and [9,8,15] for some case studies.

An alternative approach is based on the idea of maximizing the energy of the underlying modes in the optimally placed sensors. Related procedures lead to optimization problems over output Gramians of the system [16].

Both these approaches are applied on pre-selected modes of interest. For instance, in an active damping application for a transport vehicle, see a recent report [14], the bandwidth and thus implied modes are defined according to some comfort standards and considerations regarding impact of particular modes on the loads induced in the structure. Typically, a few

[*] Corresponding author. Tel.: +420 2 2435 7612; fax: +420 224 918 646.
E-mail addresses: han.tom@seznam.cz, hanistom@fel.cvut.cz (T. Hanis), xhromcik@control.felk.cvut.cz (M. Hromcik).

0888-3270/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ymssp.2011.12.007

lower modes are selected as a result of such analysis. Resulting optimal sensors selection is subsequently called, with only those pre-selected modes in mind.

However, also those not-considered, typically mid- or high-frequency modes are still present in the process and, if excited by disturbances or the control action, they can influence the active damping system behavior in an unexpected manner. This phenomenon, denoted as spillover, cannot be captured directly by the two existing approaches mentioned above. Although some procedures have been developed that address these issues, see e.g. [17], they are based on advanced signal processing (filtering) of the measured signals and do not suggest how to modify the sensor positions themselves accordingly.

And it is exactly the problem that this paper is focused on. The aforementioned information approach is taken as the starting point. The underlying criterion is modified so that the influence of desirable modes is maximized, and those unwanted modes are minimized in the observations at the same time, see Section 2. The result is a compromise where suitably chosen simple weights serve as a tuning knob for the designer. A related numerical procedure is then developed, based on the EFI approach, in Section 3. Two examples are presented in Section 4 where one can appreciate the intuitively expected placements and study the influence of tuning. Further, a case study related to a large flexible BWB aircraft and its active vibration control system is presented in Section 5. Conclusions and suggestions for further research then follow in Section 6.

2. The effective independence method (EFI)

Optimal sensors placement techniques are extensively discussed in papers [2–9]. A short overview of the EFI method follows in this section, adopted from [9,8].

The aim of the EFI method is to select measurement positions that make the mode shapes readings of interest as linearly independent as possible. The method originates from the estimation theory and is based on maximization of related Fisher information matrix, measured by its determinant or trace. That is in fact equivalent to minimization of the condition number of the information matrix related to selected sensors. The number of sensors is iteratively reduced from an initially large candidate set by removing those sensors which contribute least of all the candidate position to the linear independence of the target mode readings. In the end, the remaining sensors are delivered as the optimal sensor set. As a useful guideline to stop the iterative removing process, the determinant of the Fisher information matrix can be plotted with respect to the number of sensors; if a considerable drop is identified, further reduction should be considered with care.

2.1. Structural model

The sensor placement problem can be investigated from uncoupled modal coordinates of governing structural equations as follows:

$$\ddot{q}_i+M_i^{-1}\cdot C_i\cdot\dot{q}_i+M_i^{-1}\cdot K_i\cdot q_i=M_i^{-1}\cdot\Phi_i^T\cdot B_0\cdot u \tag{1}$$

$$y=\Phi\cdot q+\epsilon=\sum_{i=1}^{N}q_i\cdot\Phi_i+\epsilon \tag{2}$$

where q_i is the ith modal coordinate and is also the ith element of the vector, q, in the 2nd equation, M_i, K_i and C_i are the corresponding ith modal mass, stiffness and damping matrix, respectively, Φ is the mode shape matrix with its ith column as the ith mass-normalized mode shape, B_0 is simply a location matrix formed by ones (corresponding to actuators) and zeros (no load), specifying the positions of the force vector u. y is a measurement column vector indicating which positions of the structure are measured, and ϵ is a stationary Gaussian white noise with zero mean and a variance of σ^2.

2.2. Method principle

From the output measurement, the EFI algorithm analyzes the covariance matrix of the estimate error for an efficient unbiased estimate of the modal coordinates as follows [5,6,2,3,7,9]:

$$E[(q-\hat{q})\cdot(q-\hat{q})^T]=\left[\left(\frac{\partial y}{\partial q}\right)^T\cdot[\sigma^2]^{-1}\cdot\left(\frac{\partial y}{\partial q}\right)\right]^{-1}=Q^{-1} \tag{3}$$

where Q is the Fisher information matrix, σ^2 represents the variance of the stationary Gaussian measurement white noise ϵ in (2), E denote the mean value, and $\hat{q}$ is the efficient unbiased estimate of q. Maximizing Q over all sensor positions will result in the best state estimate of q. Ψ denotes the eigenvectors matrix of Q and λ is related diagonal eigenvalue matrix. The EFI coefficients of the candidate sensors are computed by the following formula:

$$E_D=[\Phi\cdot\Psi]\otimes[\Phi\cdot\Psi]\cdot\lambda^{-1}\cdot 1 \tag{4}$$

where $\otimes$ represents a term-by-term matrix multiplication, and **1** is an $n\times 1$ column vector with all elements of **1**. E_D's entries are the EFI indices, which evaluate the contribution of all candidate sensor locations to the linear independence of

Please cite this article as: T. Hanis, M. Hromcik, Optimal sensors placement and spillover suppression, Mech. Syst. Signal Process. (2011), doi:10.1016/j.ymssp.2011.12.007

the target modes measurement. Simple selection procedure is then employed to sort the elements of the E_D vector, and to remove its smallest entry at a time and also related candidate sensor, giving rise to a reduced mode shape matrix Φ. The E_D coefficients are then updated according to the new modal shape matrix, and the process is repeated iteratively until the number of remaining sensors equals a preset value. The remaining lines of the Φ matrix (or related EFI indices) define the optimal measurement locations.

3. The effective independence method with modified criterion

The main result of the paper is presented in this section. We develop a numerical scheme for OSP, based on the EFI method, such that the spillover [10–13] of unwanted higher modes is minimized.

3.1. Method principle

The modified criterion is based on the EFI reasoning presented above. Main task of the pure EFI is just to maximize information on desired modes through optimal configuration of sensors (measurements) expressed by the Fisher information matrix (FIM), or its trace or determinant respectively. The modified criterion we propose reads

$$J_{MEFI} = \alpha J_{EFI} + (1-\alpha) J_{SNR} \tag{5}$$

with optimum

$$J^*_{MEFI}(\alpha_0) = \max_{\substack{[i,j,k]\in\Omega \\ \alpha=\alpha_0}[\alpha J_{EFI}+(1-\alpha)J_{SNR}]} \tag{6}$$

where

$$J_{EFI} = \mathrm{tr}(Q^m_{[i,j,k]}) \tag{7}$$

with optimum

$$J^*_{EFI} = \max_{[i,j,k]\in\Omega} \mathrm{tr}(Q^m_{[i,j,k]}) \tag{8}$$

stands for the standard EFI part (maximize the information content for those desirable modes), and

$$J_{SNR} = \frac{\mathrm{tr}(Q^m_{[i,j,k]})}{\mathrm{tr}(Q^n_{[i,j,k]})} \tag{9}$$

with optimum

$$J^*_{SNR} = \max_{[i,j,k]\in\Omega} \left[\frac{\mathrm{tr}(Q^m_{[i,j,k]})}{\mathrm{tr}(Q^n_{[i,j,k]})} \right] \tag{10}$$

is a newly added term to penalize the unwanted mode shapes in sensor readings. Ω is the set of all candidate triples of sensors (we are considering three sensors to be selected to simplify indexing). $Q^m_{[i,j,k]}$ is the Fisher information matrix (see (3)) for mth modes (those to be captured), where $Q^n_{[i,j,k]}$ is the Fisher information matrix for the unwanted modes. Note that maximizing (9) increases information about the desirable modes in the measurements (maximizing numerator of (9) and simultaneously suppresses the unwanted modes influence (minimizing denominator of (9).

The coefficient $\alpha \in (0,1)$ serves as a tuning parameter and defines the relative importance of each part of the criterion. Selection of the parameter α is problem-dependent. However, although it is not possible to give a generally valid value for α, its influence for particular data can be investigated by means of related SNR-plots as explained in Example 1 in detail, see Section 4).

The ratio part in J_{SNR} however becomes problematic as both terms in $\mathrm{tr}(Q^m_{[i,j,k]})/\mathrm{tr}(Q^n_{[i,j,k]})$ approach zero (near the nodes of both desirable and unwanted mode shapes) which leads to irrelevant results. This unintended behavior is suppressed by applying a suitable mapping function on $\mathrm{tr}(Q^m_{[i,j,k]})$ and $\mathrm{tr}(Q^n_{[i,j,k]})$ to assure for reasonably high information content (those degenerated, almost $\frac{0}{0}$ candidates, are effectively discriminated). A suitable mapping function can take the following form, for example (see also Fig. 1):

$$f(t) = \sqrt[n]{(1+t^n)} \tag{11}$$

3.2. Modified EFI algorithm

Now we have an accordingly modified criterion. Next task is to modify the EFI heuristic in a very similar manner, to arrive at a tractable numerical scheme for the problem. Critical part of EFI method is in evaluation of E_D vector (see (4)), so the modified evaluation takes the following shape:

$$E_{DM}(\alpha) = \alpha E_D + (1-\alpha) E_{DSNR}$$

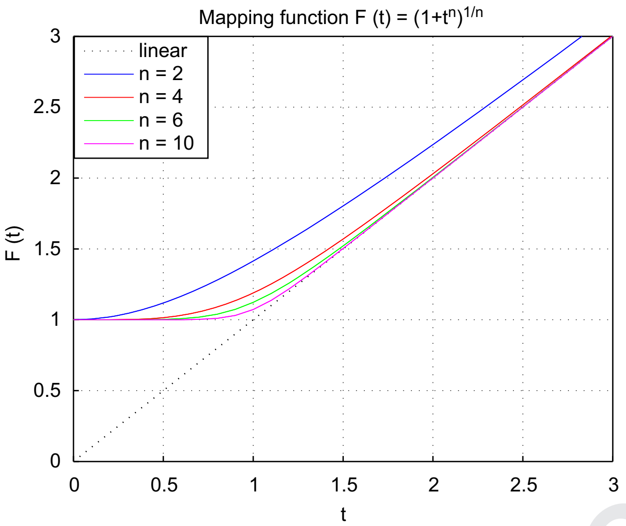


Fig. 1. Mapping function.

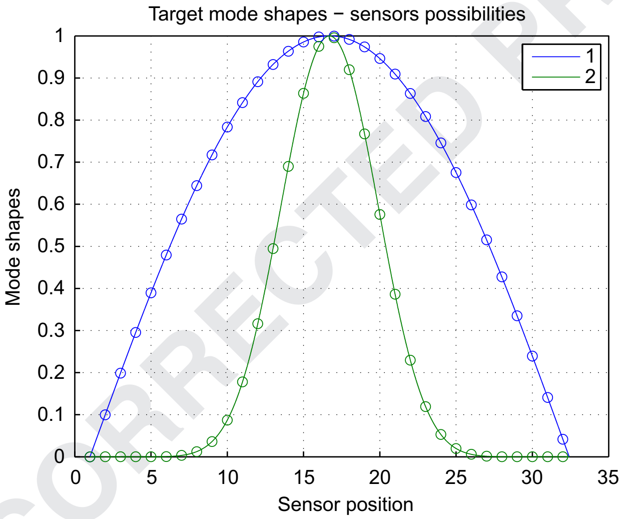


Fig. 2. Mode shapes and candidate sensor positions.

$$E_D = [\Phi \cdot \Psi] \otimes [\Phi \cdot \Psi] \cdot \lambda^{-1} \cdot 1$$

$$E_{DSNR} = \frac{[\Phi^m \cdot \Psi^m] \otimes [\Phi^m \cdot \Psi^m] \cdot \lambda^{m-1} \cdot 1}{[\Phi^n \cdot \Psi^n] \otimes [\Phi^n \cdot \Psi^n] \cdot \lambda^{n-1} \cdot 1} \tag{12}$$

Note that potential numerical issues near the node points are covered by the mapping function (11) applied on E_D and E_{DSNR} vector.

4. Example

Let us consider a flexible system with two modes of interest depicted in Fig. 2. Its structural equations read

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \cdot \ddot{q} + \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix} \cdot \dot{q} + \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix} \cdot q = \Phi^T \cdot I_{33\times 33} \cdot u \tag{13}$$

$$y = \Phi \cdot q \tag{14}$$

$$\Phi^T = \left[\begin{matrix} 0 & 0.0998 & 0.1987 & 0.2955 & 0.3894 & 0.4794 & 0.5646 & 0.6442 & \dots \\ 0 & 0.0000 & 0.0000 & 0.0000 & 0.0001 & 0.0006 & 0.0033 & 0.0123 & \dots \end{matrix}\right.$$

$$\begin{matrix} 0.7174 & 0.7833 & 0.8415 & 0.8912 & 0.9320 & 0.9636 & 0.9854 & 0.9975 & \dots \\ 0.0361 & 0.0870 & 0.1780 & 0.3161 & 0.4947 & 0.6899 & 0.8637 & 0.9752 & \dots \end{matrix}$$

$$\begin{matrix} 0.9996 & 0.9917 & 0.9738 & 0.9463 & 0.9093 & 0.8632 & 0.8085 & 0.7457 & \dots \\ 0.9957 & 0.9197 & 0.7672 & 0.5758 & 0.3864 & 0.2297 & 0.1193 & 0.0532 & \dots \end{matrix}$$

$$\left.\begin{matrix} 0.6755 & 0.5985 & 0.5155 & 0.4274 & 0.3350 & 0.2392 & 0.1411 & 0.0416 \\ 0.0198 & 0.0059 & 0.0013 & 0.0002 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \end{matrix}\right]$$

In this case it is fairly intuitive to decide by common sense where sensors should be placed if we want to maximize measurement of the first mode and reduce the second one. One can see results of the classical EFI approach in Fig. 6, related to the EFI criterion (7). It is clear that the EFI approach gives rise to sensors configuration optimal to fit the desired mode (first one), but spillover of the second one is huge. Measured energy of both modes (required E_{RQ} and not required E_{NOTRQ}) is printed in upward (Fig. 6). The signal to noise ratio coefficient (defined in dB units) was evaluated to represent

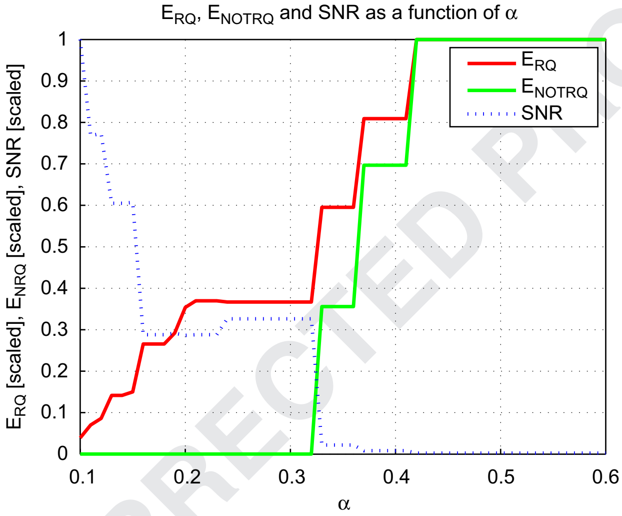


Fig. 3. The α-dependency of SNR coefficient, captured energy of required (J_{RQ}) and not required (J_{NOTRQ}) modes.

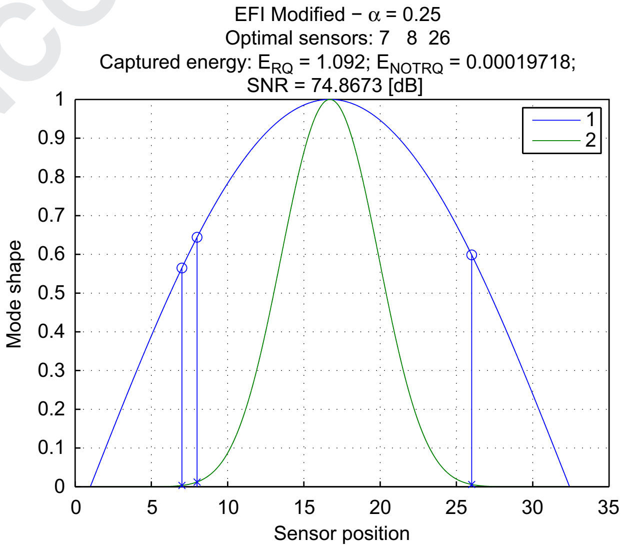


Fig. 4. OSP by the modified EFI algorithm.

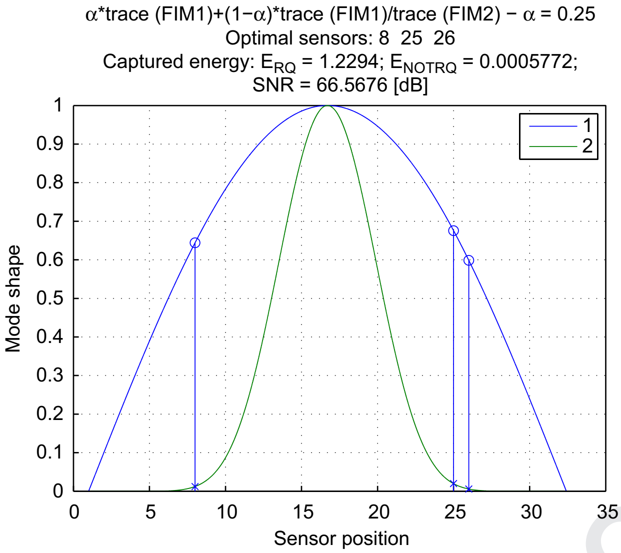


Fig. 5. OSP by direct maximization of J_{MEFI}.

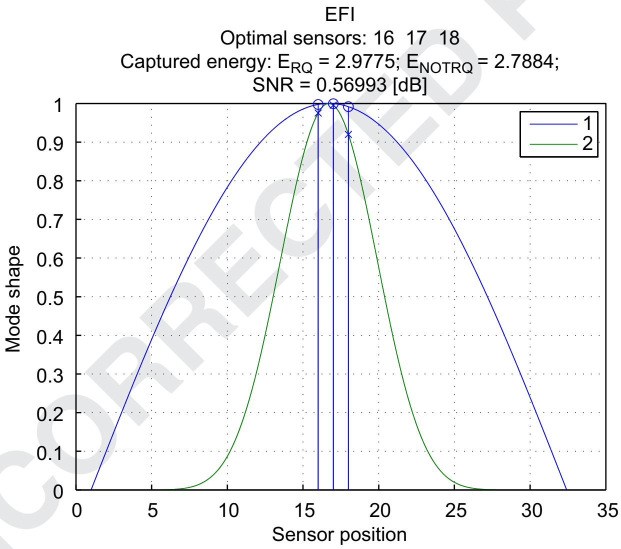


Fig. 6. OSP by classical EFI.

Fig. 7. BWB visualization.

Please cite this article as: T. Hanis, M. Hromcik, Optimal sensors placement and spillover suppression, Mech. Syst. Signal Process. (2011), doi:10.1016/j.ymssp.2011.12.007

spillover. SNR is defined by the following form:

$$SNR = 20 \cdot \log_{10}\left(\frac{E_{RQ}}{E_{NOTRQ}}\right) \tag{15}$$

Spillover reduction of the unwanted mode can be achieved by our modified criterion (see (5)). First, one has to select the α-value properly in the modified criterion (5). The dependencies of the captured energy of wanted modes (E_{RQ}) and of the captured energy of unwanted modes (E_{NOTRQ}) on α are depicted in Fig. 3. The optimal selection of the α value is at the point where E_{RQ} is large and E_{NOTRQ} is still sufficiently small. In our case, the suitable range for α is apparently the 0.2–0.3 interval, and the value of 0.25 is therefore selected.

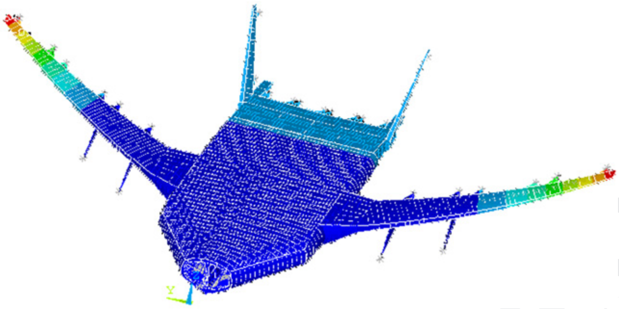


Fig. 8. Shape of first mode.

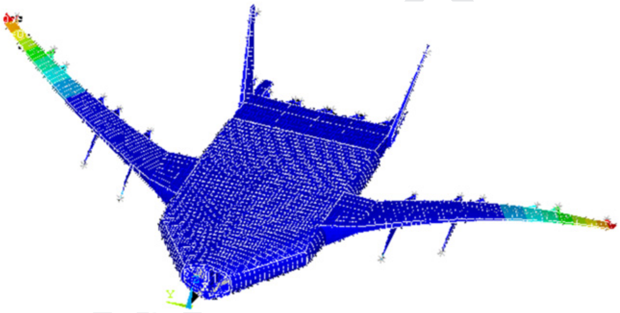


Fig. 9. Shape of second mode.

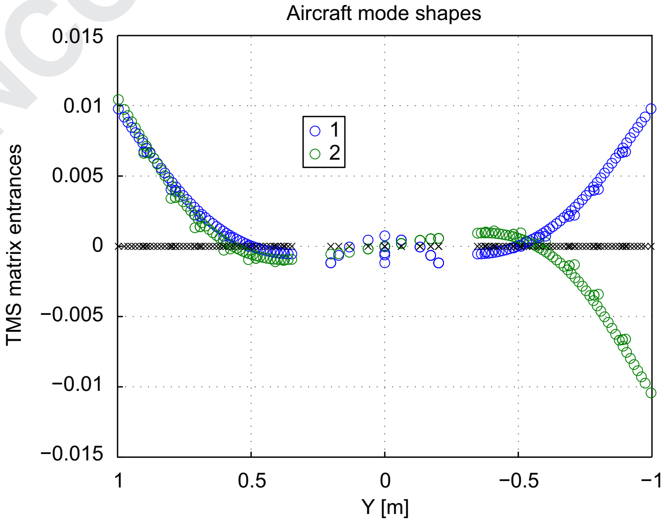


Fig. 10. Shape of first (blue o) and second (green o) modes and sensors reference positions with zero deflection (black x). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Q2

Having α, we can proceed with the modified criterion (5) and related modified EFI algorithm of Section 3.2. Results are presented in Fig. 4. One can see that spillover of the second mode with respect to the first mode is reduced if the sensors are selected according to the proposed criterion (5), and that the measurement of the useful mode is still at a good level. In addition, the suggested modified EFI algorithm appears to be an efficient approach to solve the problem (5)—mind the modes symmetry and compare (4) (modified EFI algorithm) and Fig. 5 (optimum of (5) found by “brute force”—in this particular very simple case it is feasible to exploit all the sensors combinations and select the true optimum, at the cost of high computational burden though).

For completeness, the standard EFI approach results for three sensors are given in Fig. 6. Obviously, first mode is captured very well (which is good), nevertheless, the second mode is not attenuated at all (it is not a part of the problem formulation for the standard EFI approach).

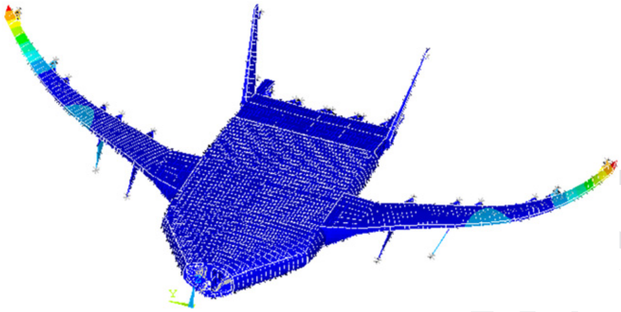


Fig. 11. Shape of third mode.

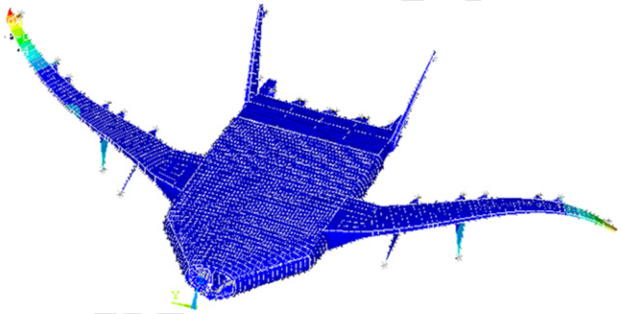


Fig. 12. Shape of fourth mode.

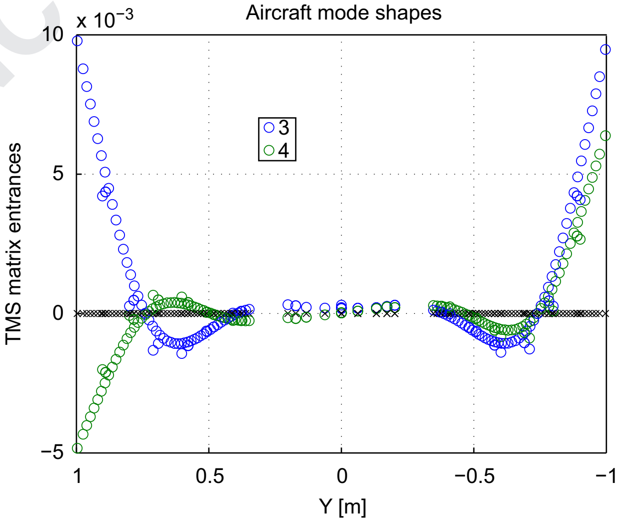


Fig. 13. Shape of third and fourth modes and sensors reference positions with zero deflection (black x).

Please cite this article as: T. Hanis, M. Hromcik, Optimal sensors placement and spillover suppression, Mech. Syst. Signal Process. (2011), doi:10.1016/j.ymssp.2011.12.007

5. Case study

ACFA 2020 is a collaborative research project funded by the European Commission under the seventh research framework programme (FP7). The project deals with innovative active control concepts for ultra efficient 2020 aircraft configurations like the blended wing body (BWB) aircraft (see Fig. 7). The Advisory Council for Aeronautics Research in Europe (ACARE) formulated the "ACARE vision 2020", which aims for 50% reduced fuel consumption and related CO_2 emissions per passenger-kilometer and reduction of external noise. To meet these goals is very important to minimize the environmental impact of air traffic but also of vital interest for the aircraft industry to enable future growth. Blended Wing Body type aircraft configurations are seen as the most promising future concept to fulfill the ACARE vision 2020 goals because aircraft efficiency can be dramatically increased through minimization of the wetted area and reducing of structural load and vibration by active damping in an integrated control law design (adopted from [1]).

The ability to distinguish between particular modes in measurement simply by optimization of appropriate sensor configuration is critical in this application due to the presence of more flexible modes in a narrow frequency range of 0–10 Hz. We cannot therefore rely on signal processing (filtering), and we have to think of a smart sensors configuration instead.

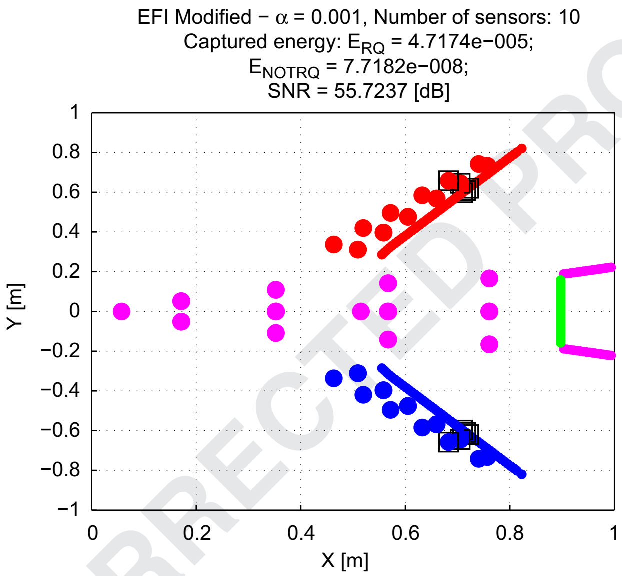


Fig. 14. Optimal sensor positions.

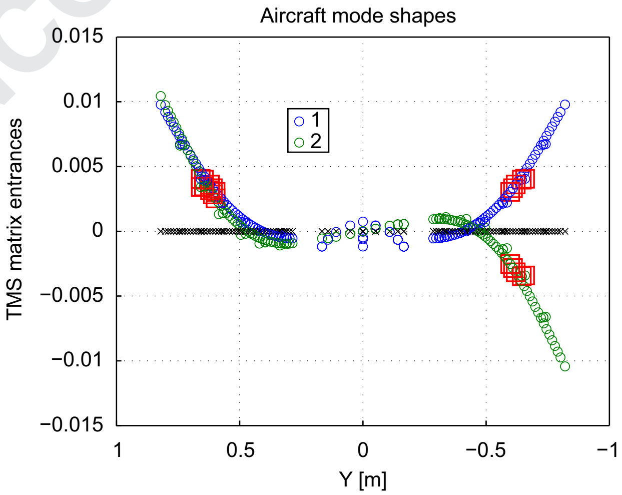


Fig. 15. Optimal sensor positions (red squares) plotted in required modes shapes (first mode shape—blue o and second mode shape—green o) and sensors reference positions with zero deflection (black x). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The most significant modes of the aircraft are first symmetrical and anti-symmetrical wing bending modes (in frequency first and second modes). Shape of the first and second aircraft mode modeled in ANSYS can be seen from Figs. 8 and 9. The target mode shapes of these modes are plotted in Fig. 10. For all next considerations we will assume these modes to be controlled and then we need to maximize information content of these modes in measurement.

The second symmetrical and anti-symmetrical modes, also called engine modes (in frequency thirrd and fourth modes) are considered as a non-controlled modes and we need to minimize information content of these modes in measurement. Shape of the third and fourth aircraft modes modeled in ANSYS can be seen from Figs. 11 and 12 and the target mode shapes are plotted in Fig. 13.

Results of optimization for case of first and second modes as required versus third and fourth modes to be rejected are plotted in Fig. 14. One can see that information content of required modes captured by this configuration of sensors is thousand times higher than information content of not-required modes (SNR approach 56 dB).

Selected sensors are superimposed into target mode shapes. One can see from Fig. 15 that higher deflections of wings during first symmetrical and anti-symmetrical bending modes are at more outboard positions. On the other hand, the nodes (zero deflection of wings due to particular mode) of the second symmetrical and anti-symmetrical wing bending modes are situated in the second third of wing lengths as can be seen from Fig. 16. Sensors location optimization therefore

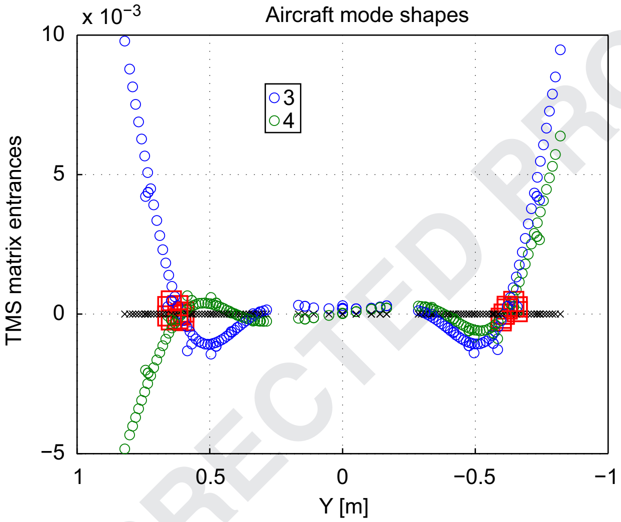


Fig. 16. Optimal sensor positions (red squares) plotted in not-required modes shapes (third mode shape—blue o and fourth mode shape—green o) and sensors reference positions with zero deflection (black x). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

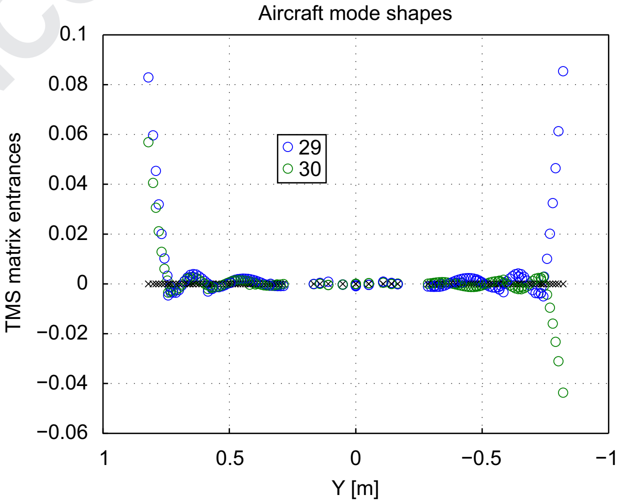


Fig. 17. Shape of 29th (blue o) and 30th (green o) modes and sensors reference positions with zero deflection (black x). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Please cite this article as: T. Hanis, M. Hromcik, Optimal sensors placement and spillover suppression, Mech. Syst. Signal Process. (2011), doi:10.1016/j.ymssp.2011.12.007

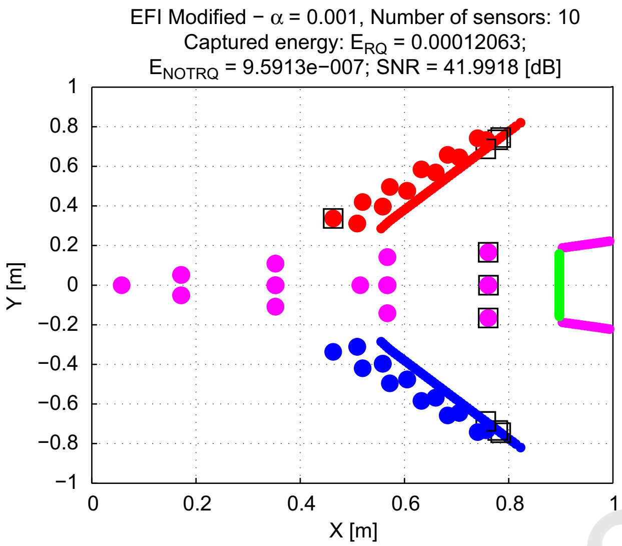


Fig. 18. Optimal sensor positions.

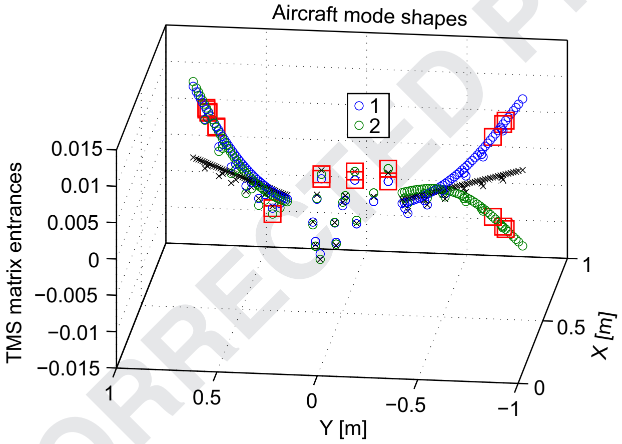


Fig. 19. Optimal sensor positions (red squares) plotted in required mode shapes (first mode shape—blue o and second mode shape—green o) and sensors reference positions with zero deflection (black x). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

results in positions near the most outboard nodes of the not-required modes.

The case of two highest modeled modes to be rejected is considered next. Last two modes in this case can be considered as a "high frequency noise" with defined spatial distribution to be filtered out by our OSP method. The 29th and 30th target mode shapes are plotted in Fig. 17.

Optimal sensors placement for the case of first symmetrical and anti-symmetrical modes versus last two symmetrical and anti-symmetrical modes is plotted in Fig. 18. Similarly as in the previous case, the most outboard sensors are involved due to nodes of not-required modes, but now also sensors in rear fuselage are selected. This behavior can be explained by comparison of target mode shapes plotted in Figs. 10 and 17. One can see a diving aircraft tail in case of first symmetrical wing bending mode and the fuselage rotation along longitudinal axis in the case of first anti-symmetrical wing bending mode (Fig. 10). On the other hand no deflection of fuselage occurs in 29th and 30th symmetrical and anti-symmetrical wing bending modes. This can also be seen from comparison of selected sensor sets superimposed into target mode shapes of required modes (Fig. 19) and undesirable modes (Fig. 20).

6. Conclusions

Novel approach to optimal sensors placement which takes into account the spillover issues has been presented in this paper. The information based approach was adapted, and a related effective algorithm was developed from the standard

Please cite this article as: T. Hanis, M. Hromcik, Optimal sensors placement and spillover suppression, Mech. Syst. Signal Process. (2011), doi:10.1016/j.ymssp.2011.12.007

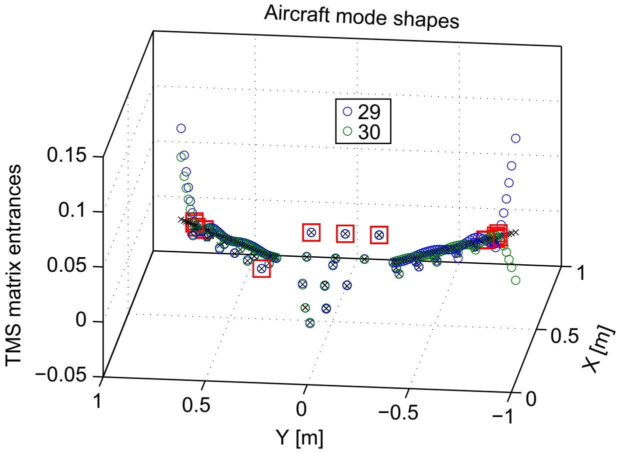


Fig. 20. Optimal sensors positions (red squares) plotted in undesirable modes shapes (29th mode shape—blue o and 30th mode shape—green o) and sensors reference position with zero deflection (black x). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

EFI procedure. Performance of the algorithm was assessed by means of a simple example and a Blended-Wing-Body aircraft case study.

Acknowledgments

The work of M. Hromcik was supported by the Ministry of Education of the Czech Republic (MSMT) under contract No. 1M0567. The work of T. Hanis was supported by the EC project ACFA 2020—Active Control for Flexible 2020 Aircraft under No. 213321.

References

[1] 〈http://www.acfa2020.eu/〉.
[2] D.C. Kammer, M.L. Tinker, Optimal placement of triaxial accelerometers for modal vibration tests, Mech. Syst. Signal Process. 18 (2004) 29–41.
[3] D.C. Kammer, Sensor set expansion for modal vibration testing, Mech. Syst. Signal Process. 19 (2005) 700–713.
[4] D.C. Kammer, Sensor placement for on-orbit modal identification and correlation of large space structures, J. Guidance Control Dyn. 14 (1991) 251–259.
[5] D.C. Kammer, Optimal sensor placement for modal identification using system-realization methods, J. Guidance Control Dyn. 19 (3) (1995).
[6] D.C. Kammer, Effect of model error on sensor placement for on-orbit modal identification of large space structures, J. Guidance Control Dyn. 15 (2) (1992).
[7] W.L. Poston, R.H. Tolson, Maximizing the determinant of the information matrix with the effective independence method, J. Guidance Control Dyn. 15 (6) (1992).
[8] D.S. Li, H.N. Li, C.P. Fritzen, The connection between effective independence and modal kinetic energy methods for sensor placement, J. Sound Vib. 305 (2007) 945–955.
[9] M. Meo, G. Zumpano, On the optimal sensor placement techniques for a bridge structure, Eng. Struct. 27 (2005) 1488–1497.
[10] W. Liu, Z. Hou, M.A. Demetriou, A computational scheme for the optimal sensor/actuator placement of flexible structures using spatial H2 measures, Mech. Syst. Signal Process. Syst. 20 (4) (2006) 881–895.
[11] J.W. Choi, U.S. Park, Spillover suppression via eigenstructure assignment in large flexible structures, J. Guidance Control Dyn. 25 (3) (2001).
[12] Y. Chait, C.J. Radcliffet, Control of flexible structures with spillover using an augmented observer, J. Guidance Control Dyn. 12 (2) (1989).
[13] J.W. Choi, U.S. Park, Spillover stabilization of large space structures, J. Guidance Control Dyn. 13 (6) (1989).
[14] Ch. Benatzky, M. Kozek, A. Schirrer, A. Stribersky, Vibration damping of a flexible car body structure using Piezo-Stack actuators, in: 17th IFAC World Congress, Seoul, Korea, 2008.
[15] L. Yao, W.A. Sethares, D.C. Kammer, Sensor placement for on-orbit modal identification of large spacestructure via a genetic algorithm, in: IEEE International Conference on Systems Engineering, 1992, Kobe, Japan, 1992, pp. 332–335.
[16] W.K. Gawronski, Advanced Structural Dynamics and Active Control of Structures, Springer-Verlag, New York, 2004.
[17] M.-H. Kim, D.J. Inman, Reduction of observation spillover in vibration suppression using a sliding mode observer, J. Vib. Control 7 (2001) 1087–1105.