

SPEECH WATERMARKING WITH OBJECTIVE FIDELITY AND ROBUSTNESS CRITERIA*

Aparna R. Gurijala and J.R. Deller, Jr.

Michigan State University

Department of Electrical & Computer Engineering / 2120 EB

East Lansing, MI 48824 USA

ABSTRACT

Speech watermarking strategies inevitably alter original signal content. Fidelity is adversely affected by increased perturbation while watermark robustness to attack is generally improved. Parameter-embedded watermarking is effected through slight perturbations of parametric models of some deeply integrated dynamics of the speech. Within this framework, a specific algorithm is presented in which the fidelity-robustness tradeoff can be objectively assessed and quantifiably adjusted according to specific measures. An overview of the general parameter-embedding strategy is followed by presentation of the featured algorithm, analysis of its properties, and experiments with speech data to assess fidelity, robustness, and other performance properties.

1. INTRODUCTION

The motivating application for this work is the creation of the *National Gallery of the Spoken Word* (NGSW), a Digital Libraries Initiative II project whose goal is the development of a on-line repository of spoken word collections. An introduction to the NGSW project is found in [1], and further information is available at www.ngsw.org. The research reported here is concerned with attempts to develop digital watermarking strategies to protect the copyrighted materials in the NGSW collection.

Digital watermarking has been the subject of intense research activity in recent years. Speech, the focus of the present research, is not typical of music and other forms of audio for which watermarking strategies might be sought. For all its well-known variability, speech has relatively well-understood properties that have been intricately modeled and quantified through decades of research. Further, the present application

enjoys abundant computational and storage resources arising from the digital library application, and the ability to do off-line watermarking. These factors have made it possible to develop relatively complex watermarking strategies for the NGSW that require nontrivial computational effort, but which are exceptionally flexible, secure, and robust as a consequence.

Digital watermarking is the process of embedding data (the *watermark*) imperceptibly into a host signal (the *coversignal*) to create a *stegosignal* [2]. The design of a watermarking strategy involves the balancing of two principal criteria. First, watermarks must be imperceptible to the listener (or viewer). Second, watermarks must be robust. That is, they must be able to survive *attacks* – those deliberately designed to destroy or remove them, as well as distortions inadvertently imposed upon the watermarks by specific technical processes (e.g., compression) or by systemic processes like channel noise or computational roundoff errors. The fidelity and robustness criteria are generally competing in the sense that greater robustness requires more watermark energy, more manipulation of the coversignal, etc., which, in turn, ultimately lead to noticeable distortion of the original content.

Although theoretical models (e.g., channel models) and simplifying assumptions (e.g., Gaussian distributions) have made it possible to quantify figures of merit for certain watermarking approaches (e.g., [3, 4]), the ability to quantify and specify the fidelity-robustness tradeoff for watermarking algorithms has remained elusive. The present research has produced a fairly general signal-processing formulation for signal watermarking within which this tradeoff can be optimized. The first component of this strategy is the class of techniques called *parameter-embedded watermarking*. Parameter-embedded watermarking is effected through slight perturbations of parametric models of some deeply integrated dynamics of the speech. A second component of the strategy is the deployment of innovative new methods for estimation of parametric models known as *set-membership filtering* (SMF). SMF provides the

*Work supported in part by the National Science Foundation of the United States under Cooperative Agreement No. IIS-9817485, and also by the Michigan State University Cybersecurity Initiative. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

means to identify parametric watermarks that result in a rigorously quantified fidelity criterion. The third component of the strategy is also a consequence of the SMF approach. SMF results in a set-solution of watermark candidates, each element of which adheres to the fidelity requirement. The properties of this set solution make it possible to study the robustness potential of a given solution, or to develop multi-watermark strategies designed to maximize robustness against attacks.

2. BACKGROUND

2.1. Parametric Speech Watermarking

A general formulation of parameter-embedded watermarking is given in previous papers [5, 6]. In the present work, a simple parametric signal model is used in this context to good effect, but this approach can almost surely be researched in many directions to produce interesting new results. Experiments have shown [5, 6] that perturbations made to the autocorrelation sequence of the coversignal (effecting a watermark) are relatively robust to many of the common watermark attacks. In turn, sets of changes to the autocorrelation sequence are in one-to-one correspondence with sets of alterations to the linear-predictive (LP) coefficients of speech [6]. Thus, changes to the LP parametric model can be embedded indirectly by altering the autocorrelation values of the coversignal. Accordingly, the coversignal $\{y_n\}$ is assumed to follow the LP model,

$$y_n = \sum_{i=1}^M a_i y_{n-i} + \xi_n \stackrel{\text{def}}{=} \mathbf{a}^T \mathbf{y}_n + \xi_n \quad (1)$$

with coefficients $\{\bar{a}_i\}_{i=1}^M$ and prediction residual $\{\xi_n\}$. In an algorithm reported at ICSLP 2002 [6], the autocorrelation values of the cover-signal were modified by the addition of an independent watermark vector. The stegosignal was constructed using the correspondingly perturbed LP coefficients and the exact prediction residual, $\{\xi_n\}$, using the FIR filter

$$\bar{y}_n = \sum_{i=1}^M \bar{a}_i y_{n-i} + \xi_n \stackrel{\text{def}}{=} \bar{\mathbf{a}}^T \mathbf{y}_n + \xi_n \quad (2)$$

Parametric watermarking was found to be fairly robust against a wide variety of attacks such as addition of noise, MP3 compression, and jitter [6]. A main reason for good robustness is that the watermark signal is concentrated into a parametric representation during watermark embedding and recovery, while it is spread across the entire work otherwise. The following aspects of the algorithm contribute to its security: speech frames to be watermarked can be selected randomly, and the LP model order can be different for each watermarked frame of the coversignal (model order also de-

pends on the fidelity constraint). In addition, a copy of the coversignal is required for watermark recovery. Because the prediction residual associated with the coversignal is used for reconstructing the stegosignal, the autocorrelation values of the stegosignal are different from the modified autocorrelation values obtained derived from the perturbed LP coefficients and the prediction residual $\{\xi_n\}$. Hence watermark recovery is impossible without a copy of the coversignal.

2.2. SMF-Based Fidelity Criterion

As a step toward quantifying the relationships between the two competing performance measures and the watermarking strategy, the authors have posed a general problem framework in which parametric watermarks are sought, subject to an ℓ_∞ fidelity constraint [5]. In the present work, we generalize the fidelity constraint to allow for more “local” fidelity considerations in time as the signal properties change. For the present paper, a simple instance of this general framework is adopted which is well-suited to the application of emerging set-theoretic techniques.

The *set-membership filtering* (SMF) concept, first published by Gollamudi *et al.* [7, 8], can be viewed as a reformulation of the broadly-researched class of algorithms concerned with *set-membership identification* (e.g. [9]). The SMF problem is used to design systems that are *affine-in-parameters* (but not necessarily in the data), subject to a bound on the absolute error between a desired sequence and a linearly-filtered version of another sequence. The two sequences may be directly observed or they may be nonlinear combinations of other sequences considered to be the system inputs and outputs. Formally, the SMF problem is stated as follows:

Given a sequence $\{\mathbf{x}_m \in \mathbb{R}^M\}_{m=1}^n$ of observations, a “desired” sequence $\{z_m \in \mathbb{R}\}_{m=1}^n$, and a sequence of error “tolerances” $\{\gamma_m\}_{m=1}^n$, find the the exact feasibility set, $\mathcal{P}_n \subseteq \mathbb{R}^M$, of filters $\theta \in \mathbb{R}^M$, at time n ,

$$\mathcal{P}_n = \left\{ \theta \mid |z_m - \theta^T \mathbf{x}_m| < \gamma_m, \quad m \in [1, n] \right\}. \quad (3)$$

The SMF problem is solved using a series of recursions which ultimately return an hyperellipsoidal *membership set*, say $\mathcal{E}_n \supset \mathcal{P}_n$, and the ellipsoid’s center, say θ_n . The recursions execute an optimization strategy designed to tightly bound \mathcal{P}_n by \mathcal{E}_n in some sense. Accordingly, the broad class of algorithms employed in the SMF problem are often called the *optimal bounding ellipsoid* (OBE) algorithms. Results on the theory and application of OBE algorithms abound, and the reader is referred, for example, to the tutorial paper [9].

The construction of a watermark set guaranteed to satisfy a fidelity criterion is readily solved as an SMF

problem. Subtracting y_n from each side of (2), then rearranging, yields

$$\bar{y}_n - y_n = \sum_{i=1}^M \bar{a}_i y_{n-i} + \xi_n - y_n = \bar{\mathbf{a}}^T \mathbf{y}_n - (y_n - \xi_n). \quad (4)$$

A fidelity criterion is prescribed in the form of a sequence of pointwise absolute bounds, $\{\gamma_n\}_{n=1}^N$, on the coversignal perturbation: $|y_n - \bar{y}_n| < \gamma_n$ for each $n \in [1, N]$. Upon defining the sequence $z_n = y_n - \xi_n$, $n = 1, 2, \dots, N$, (recall that $\{\xi_n\}$ is known), the search for the constrained watermark parameters is reduced to a SMF problem as in (3). The result of applying the SMF estimation is the hyperellipsoidal set of watermark (perturbed model parameter) candidates, \mathcal{E}_N , guaranteed to tightly bound, the exact set¹

$$\mathcal{P}_N = \left\{ \bar{\mathbf{a}} \in \mathbb{R}^M \mid |z_n - \bar{\mathbf{a}}^T \mathbf{y}_n| < \gamma_n, \quad n \in [1, N] \right\} \quad (5)$$

3. ROBUSTNESS STUDIES

3.1. Introduction

At EuroSpeech 2003 [5], the authors proposed the use of SMF processing as a means of adding a well-quantified fidelity criterion to parameter-embedded watermarking. It is suggested that the set solution resulting from the SMF design can be used to select robust watermarks, but the point is not pursued there. A principal purpose of the present paper is to report results relating to the robustness issue.

In this section we analyze the effects of two principal modes of attack – additive noise and linear filtering – on watermarks inserted by the reported algorithm. In each case, watermark characteristics that lead to the most robust watermark for the particular attack are identified. The balancing of two principal watermarking requirements — fidelity and robustness – and the role played by the hyperellipsoidal set in this trade-off, are illustrated with examples.

The robustness property is dependent on the fidelity constraint, appropriate watermark selection from the ellipsoidal set, and watermark detection. In general,

¹In previous work [5], we referred to an “ ℓ_∞ ” fidelity constraint on the difference between the cover- and stegosignals. In that work γ_n was taken to be a constant, say γ , wherein the set in (5) may be written $\mathcal{P}_N = \left\{ \bar{\mathbf{a}} \mid \|\mathbf{y}_1^N - \bar{\mathbf{y}}_1^N(\bar{\mathbf{a}})\|_\infty < \gamma \right\}$ in which \mathbf{y}_1^N and $\bar{\mathbf{y}}_1^N$ are the N -vectors of cover- and stegosignal samples $\mathbf{y}_1^N = [y_1 \ \dots \ y_N]^T$ and $\bar{\mathbf{y}}_1^N = [\bar{y}_1 \ \dots \ \bar{y}_N]^T$. In this case, therefore, γ is identical to (or at least a tight upper bound on) the ℓ_∞ norm of $\mathbf{y}_1^N - \bar{\mathbf{y}}_1^N(\bar{\mathbf{a}})$. In the present formulation in which γ_n is allowed to vary with time, it is still true that $\max_{n \in [1, N]} \gamma_n$ is the ℓ_∞ norm on the signal differences, but to refer to the constraint in these terms tends to obscure the “local” (in time) control over the fidelity that is afforded by this generalization.

greater robustness can be obtained by embedding more energetic watermarks, in turn affecting stegosignal fidelity. By default, the center of the hyperellipsoid can be used as the, generally suboptimal, watermark solution. In the present work, a conservative definition of watermark detection is employed. A watermark is considered detected only if every element of the watermark vector is above a threshold.

3.2. Watermark Recovery from the Stegosignal

In the present application, the LP model is used as a device to parameterize long intervals of stationary or nonstationary speech without the intention of properly parameterizing any particular dynamics in the waveform. In either case – stationary or nonstationary – the LP parameters are derived according to the usual optimization criterion – to minimize the total energy in the residual. However, to understand the robustness aspects of the watermarks, it is necessary to consider stationary segments of the coversignal and the stegosignal. That is, segments of y_n , w_n , and, hence, \bar{y}_n , are assumed to be partial realizations of wide-sense stationary (WSS) and ergodic random processes.

Watermark recovery is effected through least-square-error (LSE) estimation of the perturbed parameters, $\{\bar{a}_i\}_{i=1}^M$ in the following manner. Let us rewrite the stegosignal generation equation (2) as

$$d_n = \sum_{i=1}^M \bar{a}_i y_{n-i} = \bar{\mathbf{a}}^T \mathbf{y}_n \quad \text{with} \quad d_n \stackrel{\text{def}}{=} \bar{y}_n - \xi_n. \quad (6)$$

In principle, the system of equations consisting of (6) taken over $n = 1, 2, \dots, N$ is noise free and can be solved for $\bar{\mathbf{a}}$ using any subset of M equations. For generality, to smooth roundoff and other errors and to support further developments, we pose the problem as an attempt to compute the LSE linear estimator of the “desired” signal, d_n , given observations \mathbf{y}_n . The following normal equations are solved,

$$\mathbf{C}_y \bar{\mathbf{a}} = \mathbf{c}_{yd}. \quad (7)$$

in which $\mathbf{C}_y = \sum_{n=1}^N \mathbf{y}_n \mathbf{y}_n^T = \mathbf{Y}_N \mathbf{Y}_N^T$ and $\mathbf{c}_{yd} = \sum_{n=1}^N \mathbf{y}_n d_n = \mathbf{Y}_N \mathbf{d}_1^N$, where

$$\mathbf{Y}_N = \begin{bmatrix} \mathbf{y}_N & \mathbf{y}_{N-1} & \dots & \mathbf{y}_1 \end{bmatrix} \in \mathbb{R}^{M \times N} \quad (8)$$

$$\mathbf{d}_1^N = \begin{bmatrix} d_N & d_{N-1} & \dots & d_1 \end{bmatrix}^T \in \mathbb{R}^{N \times 1} \quad (9)$$

3.3. Robustness to White Noise Attack

The watermarking strategy is extremely robust to white-noise corruption of the stegosignal. Let $\{\eta_n\}_{n=1}^N$ be a partial realization of a zero mean, uncorrelated noise process which is added to the stegosignal samples $\{\bar{y}_n\}_{n=1}^N$.

Let the corrupted stegosignal be denoted $\{\bar{y}_n^\eta\}_{n=1}^N$. In this case, the “desired” signal used in the LSE estimation problem (system of equations of form (6) for $n \in [1, N]$) will be likewise corrupted. That is the clean signal d_n is replaced by, say,

$$d_n^\eta = \bar{y}_n^\eta - \xi_n = d_n + \eta_n, \quad n = 1, 2, \dots, N. \quad (10)$$

Accordingly, the cross-correlation vector [i.e., right side of normal equations (7)], but, *only* this vector, is affected by the attack. It is readily shown, however, that the “corrupted” cross-correlation, say $\mathbf{c}_{y d^\eta}$, asymptotically becomes $\mathbf{c}_{y d}$. For large N , therefore, the watermark is immune to the white noise attack.

3.4. Robustness to Correlated Noise Attack

Suppose that the stegosignal is distorted by the addition of a correlated noise process, $\{\rho_n\}_{n=1}^N$. The distorted stegosignal is denoted $\{\bar{y}_n^\rho\}_{n=1}^N$. As in the white noise case, the “desired” signal in the watermark recovery process is likewise corrupted. Instead of d_n , we have access to

$$d_n^\rho = \bar{y}_n^\rho - \xi_n = d_n + \rho_n, \quad n = 1, 2, \dots, N. \quad (11)$$

Consequently, the cross-correlation vector in the normal equations is altered by the attack. Because of the correlation in the noise, however, $\mathbf{c}_{y d^\rho}$ no longer approaches $\mathbf{c}_{y d}$ asymptotically. Depending on the relative magnitudes of the cross-correlation elements in $\mathbf{c}_{y d^\rho}$, the LSE estimation of the perturbed coefficients, and hence the watermark signal, may be affected. A solution to this problem is to whiten the noise process, in which case, the effect of noise will be similar to the white noise attack. Whitening entails the knowledge of noise correlation properties and these properties can be easily determined as this technique involves informed watermark detection. The estimation of the perturbed LP coefficients is the solution to (7) with \mathbf{C}_y replaced by $\mathbf{C}_y^\rho = (\mathbf{Y}_N \mathbf{C}_\rho^{-1} \mathbf{Y}_N)$ and $\mathbf{c}_{y d}$ replaced by $\mathbf{c}_{y d}^\rho = \mathbf{Y}_N \mathbf{C}_\rho^{-1} \mathbf{d}_1^N$, in which \mathbf{C}_ρ is the covariance matrix associated with the correlated noise process.

3.5. Robustness to Filtering Attacks

Let $\{\bar{y}_n^f\}_{n=1}^N$ be the result of filtering the stegosignal. At time n ,

$$\bar{y}_n^f = \bar{y}_n * h_n = y_n * h_n + w_n * h_n \quad (12)$$

in which $\{h_n\}$ is the impulse response of the filter, $*$ denotes linear convolution, and where we have defined the *watermark signal* $w_n = \bar{y}_n - y_n$. In the first analysis, it seems very reasonable to state that an ideal attack will result in $\bar{y}_n^f \approx y_n$. This indicates that the ideal attack filter will maximize (in some sense) the contribution of the first term in the sum, and minimize the second –

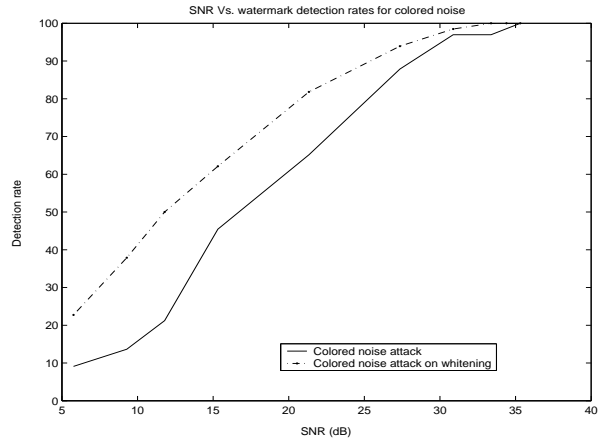


Fig. 1. Detection rates for correlated noise attack

similarly to any optimal filter design to remove noise.² On the other hand, (12) reveals that good watermark design requires that the watermark signal be as spectrally similar to the coversignal as possible, so that any attack on the watermark will also degrade the coversignal component thereby degrading fidelity.

However, the analysis above can be confounded by less-easily quantified factors. For example, the watermark designer must be aware of perceptual properties of speech that may allow some distortion of y_n without loss of perceptual quality. Since any distortion of the watermark signal can potentially render it undetectable, this means that spectrally matching w_n and y_n is not sufficient. Rather, the watermark spectrum should match that of any obtainable \bar{y}_n^f that is perceptually indistinguishable from y_n . Further, watermarks from diverse frequency bands can be embedded across speech segments to increase the recovery probability against filtering attacks.

4. EXPERIMENTS AND DISCUSSION

Correlated noise attack. The coversignal consisted of 3 seconds of Thomas Edison’s speech from the Vincent Voice Library (VVL) [1], sampled at 44.1 kHz. Each coversignal frame consisted of 3969 samples and a watermark was embedded in each of the 33 frames. A 4th order LP model was used for watermarking and the fidelity constraint was set so that $|\bar{y}_n - y_n| < 0.25|y_n|$ for each n . Correlated noise of various SNRs was added to the stegosignals. Figure 1 shows the watermark detection rates vs. SNR, with and without the use of a prewhitening filter. Improved performance is observed when prewhitening is employed.

²Since the attacker does not have access to the watermark signal $\{w_n\}$, truly optimal design is not possible.

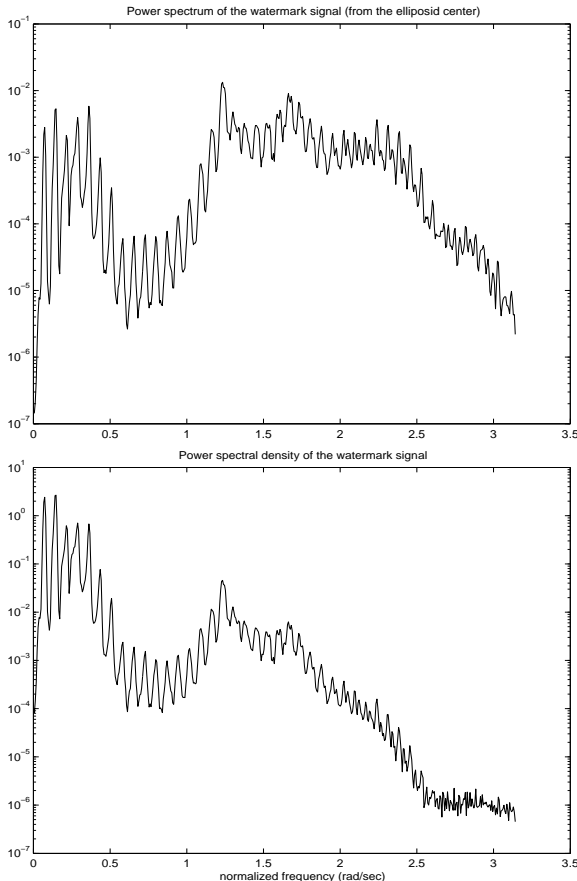


Fig. 2. (a) PSD of watermark signal from the ellipsoid center and (b) PSD for LP coefficients from ellipsoid exterior

Filtering attack. The coversignal consisted of 3000 samples of the vowel /a/, sampled at 10 kHz. The watermark was embedded using a 3rd order LP model for easy visualization. Practically a higher order model would be preferred to facilitate embedding a longer vector. The fidelity constraint was such that $|\bar{y}_n - y_n| < 0.25|y_n|$ for each n .

The experiment involved the selection of two different sets of LP parameters from the ellipsoid. The power spectral density (PSD) spectra of the two watermark signals are shown in Fig. 2. The first watermark was embedded using the LP coefficients at the ellipsoid center. For the second watermark the LP vector from the ellipsoid boundary at the intersection of the longest axis was used. In both cases the attack involves filtering using a 4th Butterworth high pass filter, with cut-off frequency 100 Hz. The watermark was not detectable when the center of ellipsoid was used for watermarking, while it was detected in the latter case.

Since the highpass filter cut-off frequency was at

100 kHz, stegosignal fidelity was not significantly affected, even though the watermark from ellipsoid center was destroyed. The PSD of watermark 2 has power spectrum similar to the PSD of the coversignal. Watermark 2 has more energetic low frequency components, with not much energy within the 0 to 100 Hz frequency range. Hence, watermark robustness can be improved by obtaining appropriate LP coefficients from the set.

5. CONCLUSIONS

A parametric speech algorithm with capability to objectively adjust the fidelity-robustness trade-off is presented. The fidelity criterion, which specifies the sample-wise distortion on the stegosignal, fits well into the SMF paradigm. The resulting watermarks exhibit good robustness to additive white noise, and detection in colored noise can be improved by prewhitening. From the SMF solution set, an appropriate watermark for possible filter attacks can be embedded.

6. REFERENCES

- [1] J.H.L. HANSEN *et al.*, "Audio stream phrase recognition ...," *Proc. Int'l. Conf. Spoken Lang. Proc.*, Beijing, 1089-92, Oct. 2000.
- [2] I.J. COX *et al.*, *Digital Watermarking*, New York: Academic Press, 2002.
- [3] J.A. O'SULLIVAN *et al.*, "Info-theoretic analysis of steganography," *Proc. IEEE Int'l. Symp. Info. Theory*, Cambridge MA, Aug. 1998.
- [4] J.K. SU *et al.*, "Analysis of digital watermarks subjected to optimum linear filtering ...," *IEEE Trans. Signal Proc.*, vol. 81, June 2001.
- [5] A. GURIJALA, J.R. DELLER, JR. "Speech watermarking by parametric embedding ...," *Proc. EUROSPEECH*, Geneva, Sep. 2003, CD-ROM.
- [6] A. GURIJALA *et al.* "Speech watermarking through parametric modelling," *Proc. Int'l. Conf. Spoken Language Proc.*, Denver, Sep. 2002, CD-ROM.
- [7] S. GOLLAMUDI *et al.*, "SMART: A toolbox for SMF," *Proc. European Conf. Circuit Theory and Design*, Budapest, 1997.
- [8] S. NAGARAJ *et al.*, "BEACON: An adaptive set-membership filtering technique ...," *IEEE Trans. Signal Proc.*, **47**: 2928-41, 1999.
- [9] J.R. DELLER, JR., Y.F. HUANG, "Set-membership identification and filtering ...," *Circuits, Systems, and Signal Processing*, Feb. 2002.