ROBUST EXPONENTIAL MODELING OF AUDIO SIGNALS

Joost Nieuwenhuijse, Richard Heusdens and Ed F. Deprettere

Delft University of Technology, Dept. of Electrical Engineering,
Mekelweg 4, 2628 CD Delft, The Netherlands

ABSTRACT

In this paper we present a numerically robust method for modeling audio signals which is based on a exponential data model. This model is a generalization of the classical sinusoidal model in the sense that it allows the amplitude of the sinusoids to evolve exponentially. We show that, using this model, so-called attacks can be represented very efficiently and we propose an algorithm for finding the exponentials in a robust way. Moreover, we show that by using a proper segmentation of the input data into variable length segments the signal-to-noise ratio can be drastically improved as compared to a fixed-length analysis.

1. INTRODUCTION

In *audio coding*, the need for lower bit rates ($<$ 64 kbit/s) is growing. The emerging MPEG standardization work items will aim at bit rates well below 64 kbit/s. In the current MPEG-8 overview, bit rates $<$ 32 kbit/s are proposed. Until now, audio coding research has mainly been focusing on *transparent coding* where bit rate constraints are rather mild.

Recently, apart from audio coding schemes delivering a transparent audio quality, other qualifications have gained attention, among which most notably is the *medium quality* coding scheme. As opposed to transparent quality, medium quality audio coding reveals an audible difference between the original and coded signals, yet any coding artifacts must not be perceived as being annoying. Current MPEG-4 standardization negotiations [1] are addressing the full range of possible bit rates and quality assignments.

Traditionally, audio and speech coding have been two completely separated areas of investigation. Speech coders, for example, are exploiting speech specific features almost to the extreme, in particular speech coders which are tuned toward dedicated applications. Examples of such coders include vocoders [2] and sinusoidal coders [3]. The latter, however, are less dependent on data-specific properties and can, therefore, be applied to audio signals as well. In fact, after some modifications of the classical sinusoidal model, this coding technique results in an efficient and robust representation of audio signals.

This paper is organized as follows. In Section 2 we introduce the exponential signal model, an extension to the classical sinusoidal model. In Section 3 we propose an algorithm for finding the exponential components. Since this method is not subject to a stability condition, we investigate in Section 4 the sensitivity of the final reconstruction to perturbations of the model parameters. Next, in Section 5, we investigate the segmentation of the input data in order to represent it with a minimum number of components. Finally, in Section 6, we draw some conclusions.

2. EXPONENTIAL MODELING

Sinusoidal coding aims at modeling a signal x as a sum of, say K, sinusoids, i.e.,

$$\hat{x}(n) = \sum_{k=1}^{K} a_k(n) \cos(n\omega_k(n) + \varphi_k(n)), \tag{1}$$

where $a_k(n)$, $\omega_k(n)$ and $\varphi_k(n)$ are slowly time varying parameters, such that $\|x - \hat{x}\|$ is minimized for some norm and some value of K. Conventional sinusoidal coders divide the signal into segments and assume the parameters a_k, ω_k and φ_k to be constant throughout each segment. In reconstructing the signal, overlap-add or interpolation techniques are used to obtain a smooth transition of the reconstructed signal at the segment boundaries [3].

Audio signals with so-called "attacks" or "transients", like the signal shown in Figure 2a, contain fast variations in amplitude and cannot be modeled efficiently as a sum of constant-amplitude sinusoids. We, therefore, extend the conventional sinusoidal model by allowing the amplitude to evolve exponentially. To do so, we introduce a damping coefficient[1] $\gamma_k \in \mathbb{R}$ and define

$$\begin{aligned}\hat{x}(n) &= \sum_{k=1}^{K} a_k e^{\gamma_k n} \cos(\omega_k n + \varphi_k) \\ &= \sum_{k=1}^{d} r_k \phi_k^n, \quad 0 \le n < N,\end{aligned} \tag{2}$$

where $r_k, \phi_k \in \mathbb{C}$. $N \in \mathbb{N}$ is the segment length. The parameter r_k determines the initial phase and amplitude, while ϕ_k determines the frequency and damping. Note that $d = 2K$. Equation (2) expresses $\hat{x}(n)$ as the sum of d damped (complex) exponentials, in the remainder of this paper referred to as *components*. In order to be able to use this model, we need an analysis method to determine the parameters (r_k, ϕ_k) for d components, that together form a good approximation of a given signal segment.

Signal analysis in conventional sinusoidal coders is based on Fourier transform methods. The performance of these methods is not optimal; most notably, they fail to give an accurate frequency estimation for sinusoids in the low-frequency region [4]. Moreover, the traditional methods take for granted that the sinusoidal components have a constant amplitude and can, therefore, not be used to determine the damping coefficients. In the next section we present a robust analysis method that overcomes the problems mentioned above.

[1] The damping coefficient γ_k can be any real number. Positive values of γ_k, therefore, correspond to expanding amplitudes rather than to truly damped amplitudes.

3. SIGNAL ANALYSIS

3.1. Ideal signals

Let us first suppose the signal x to be "ideal", that is, it really is a sum of d damped complex exponentials,

$$x(n) = \sum_{k=1}^{d} r_k \phi_k^n, \quad 0 \leq n < N, \tag{3}$$

with $r_k \neq 0$ and $\phi_i \neq \phi_j,\ i \neq j$. We can rewrite (3) in matrix notation as

$$x(n) = \mathbf{C}\Phi^n\mathbf{B},$$

where $\mathbf{C} = (1, \ldots, 1) \in \mathbb{C}^{1\times d}, \Phi = \mathrm{diag}(\phi_1, \ldots, \phi_d) \in \mathbb{C}^{d\times d}$ and $\mathbf{B} = (r_1, \ldots, r_d)^t \in \mathbb{C}^{d\times 1}$. The superscript ^{t} denotes matrix transposition.

Let $\mathbf{H} \in \mathbb{C}^{m\times l}$, $m + l - 1 = N$, $m > d$ and $l \geq d$, be a Hankel data matrix built on the signal segment x,

$$\mathbf{H} = \begin{bmatrix} \mathbf{C}\mathbf{B} & \mathbf{C}\Phi\mathbf{B} & \cdots & \mathbf{C}\Phi^{l-1}\mathbf{B} \\ \mathbf{C}\Phi\mathbf{B} & \mathbf{C}\Phi^2\mathbf{B} & \cdots & \mathbf{C}\Phi^{l}\mathbf{B} \\ \vdots & \vdots & & \vdots \\ \mathbf{C}\Phi^{m-1}\mathbf{B} & \mathbf{C}\Phi^{m}\mathbf{B} & \cdots & \mathbf{C}\Phi^{m+l-2}\mathbf{B} \end{bmatrix}.$$

It is well known [5] that there exist matrices $\mathcal{O} \in \mathbb{C}^{m\times d}$ and $\mathcal{C} \in \mathbb{C}^{d\times l}$,

$$\mathcal{O} = \begin{bmatrix} \mathbf{C} \\ \mathbf{C}\Phi \\ \vdots \\ \mathbf{C}\Phi^{m-1} \end{bmatrix}, \quad \mathcal{C} = \begin{bmatrix} \mathbf{B} & \Phi\mathbf{B} & \cdots & \Phi^{l-1}\mathbf{B} \end{bmatrix},$$

such that $\mathbf{H} = \mathcal{O}\mathcal{C}$. This decomposition is unique up to a similarity transformation. $\mathcal{O}$ and $\mathcal{C}$ are of full rank d since they are Vandermonde matrices. Hence, $\mathrm{rank}(\mathbf{H}) = d$. It is also well known [5] that Φ can be computed from $\mathcal{O}$ by exploiting its shift-invariance structure. Thus, let $\mathcal{O}^{\uparrow}$ be $\mathcal{O}$ without the top row and $\mathcal{O}^{\downarrow}$ be $\mathcal{O}$ without the bottom row. Then

$$\mathcal{O}^{\downarrow}\Phi = \mathcal{O}^{\uparrow}, \tag{4}$$

which can be solved for the unknown Φ.

One way for finding $\mathcal{O}$ and $\mathcal{C}$ is through singular value decomposition (SVD). Let $\mathbf{H} = \mathbf{U}\Sigma\mathbf{V}^*$ be the SVD of $\mathbf{H}$, where $\mathbf{U} \in \mathbb{C}^{m\times m}$ and $\mathbf{V} \in \mathbb{C}^{l\times l}$ are unitary matrices and $\Sigma \in \mathbf{R}^{m\times l}$ is a diagonal matrix with diagonal entries $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_d > 0$ and $\sigma_k = 0$ for $k = d+1, \ldots, \min(m, l)$. The superscript * denotes matrix transposition and complex conjugation. A suitable rank-d decomposition $\mathbf{H} = \mathcal{O}'\mathcal{C}'$ can then be obtained by setting, for example, $\mathcal{O}' = \mathbf{U}\Sigma$ and $\mathcal{C}' = \mathbf{V}^*$. As stated above, a similarity transformation will transform them into $\mathcal{O}$ and $\mathcal{C}$, from which the component parameters (r_k, ϕ_k) can be computed.

3.2. Audio signals

The analysis method given in the previous section assumes that the signal segment is exactly of the form (3). Audio signals of reasonable length almost always obey (3) with $d \leq \frac{N}{2}$. From a coding point of view this number is usually too large and we, therefore, are interested in an approximation of the signal segment with a lower number of components $\hat{d} < d$.

As we stated above, the number of components d in a signal segment equals the rank of the corresponding Hankel data matrix. An approximation of the signal segment with a lower number of components can thus be obtained from a lower rank approximation of $\mathbf{H}$. It is well known [6] that the best rank-$\hat{d}$ approximation, in a least squares sense, of a matrix can be obtained by setting the smallest singular values equal to zero and leave the $\hat{d}$ largest singular values unaffected. However, the resulting rank-$\hat{d}$ approximation of $\mathbf{H}$, say $\hat{\mathbf{H}}$, is not Hankel anymore. As a result, $\mathcal{O}$ is not shift-invariant and no Φ satisfying (4) does exist. We, therefore, determine Φ as the *least square* solution to (4). Once we have found Φ we have to determine $\mathbf{B}$. For the same reason, $\mathbf{B}$ cannot be taken directly from $\mathcal{C}$. Instead, given Φ, we determine $\mathbf{B}$ by solving the least squares problem

$$\min_{\mathbf{B}} \|x - \hat{x}\|_2.$$

By inspection of (3), we conclude that this is equivalent to solving

$$\min_{\mathbf{B}} \left\| \begin{bmatrix} x(0) \\ x(1) \\ \vdots \\ x(N-1) \end{bmatrix} - \begin{bmatrix} 1 & \cdots & 1 \\ \phi_1 & \cdots & \phi_{\hat{d}} \\ \vdots & & \vdots \\ \phi_1^{N-1} & \cdots & \phi_{\hat{d}}^{N-1} \end{bmatrix} \mathbf{B} \right\|_2. \tag{5}$$

Other methods to obtain a reduced rank approximation of a signal segment include the Cadzow algorithm [7], which can be used to determine a rank-$\hat{d}$ *Hankel* approximation of $\mathbf{H}$, and the structured total least norm algorithm [8], which can be used to determine a rank-$\hat{d}$ approximation of the signal segment itself, both in a least square sense. With both methods, Φ and $\mathbf{B}$ can be computed as described in Section 3.1 since the rank-reduced Hankel matrix in that case does have the shift-invariant structure.

4. SENSITIVITY

Using the procedure described above, we approximate a given signal segment with a sum of components of the form $r_k\phi_k^n$. If $|\phi_k| < 1$, the component has a decaying envelope and we will call such a component a stable component. The analysis described above is not subject to a stability condition and can thus also output components for which $|\phi_k| > 1$. Such components have an expanding envelope and we refer to them as unstable components. The names stem from the fact that in the case the segment length is unbounded, components for which $|\phi_k| < 1$ will converge to zero while components for which $|\phi_k| > 1$ will grow unlimited.

In practice components will not in general decay to zero or expand to infinity, as they are confined to a finite length segment. However, even with finite length segments, the unstable components will be very sensitive to perturbations of ϕ_k. An example that illustrates this sensitivity phenomenon is shown in Figure 1. In Figure 1a, a segment of a music signal of 160 samples is depicted. This signal is approximated with 60 exponentials with the method described in the previous section. The reconstruction of the signal using these components is shown in Figure 1b (solid line) as well as the reconstruction error (dotted line). It is clear that the reconstruction is almost perfect.

Next we take one of the unstable components and increase the modulus of the corresponding parameter ϕ_k. In our experiment, we increased one parameter for which $|\phi_k| = 1.06$ to $|\phi_k| =$

1.09. The resulting reconstruction, using the perturbed parameter, is shown in Figure 1c. It is clear that the reconstruction "explodes" at the right-hand side boundary which has to be prevented.

The sensitivity to perturbations of unstable components can be greatly reduced. Before discussing how this can be done, we first take a closer look to the influence of perturbed parameters on the final reconstruction.

Let $\phi_k = e^{\gamma_k} e^{j\omega_k}$, where $\gamma_k, \omega_k \in \mathbb{R}$. The parameters γ_k and ω_k are the damping coefficient and frequency, respectively, of the kth component, which we will denote by x_k. If we perturb γ_k, say $\gamma_k' = \gamma_k + \varepsilon$, resulting in a perturbed component x_k', we have that $x_k'(n) = e^{\varepsilon n} r_k \phi_k^n$ so that

$$\begin{aligned} \|\Delta_k(n)\|_\infty &= \|x_k'(n) - x_k(n)\|_\infty \\ &= \|(e^{\varepsilon n} - 1) r_k \phi_k^n\|_\infty, \quad 0 \le n < N. \end{aligned} \tag{6}$$

Equation (6) shows that, if $|\phi_k| > 1$, the error is zero at $n = 0$ and increases exponentially as n increases, which was illustrated in Figure 1c. On the other hand, if $|\phi_k| < 1$, such problems do not occur since in that case the error decreases exponentially for sufficiently large n, assuming that $|\varepsilon| < |\gamma_k|$.

We can greatly improve the numerical stability of the signal reconstruction by using an alternative representation for the components x_k for which $|\phi_k| > 1$. This can be seen as follows.

Let $p_k = r_k \phi_k^{N-1}$, the value of x_k at $n = N - 1$. We then can rewrite x_k as

$$\begin{aligned} x_k(n) &= r_k \phi_k^n \qquad (7) \\ &= r_k \phi_k^{N-1} \phi_k^{n-N+1} \\ &= p_k \phi_k^{n-N+1}, \quad 0 \le n < N. \qquad (8) \end{aligned}$$

In fact, (8) is a backward description of x_k whereas (7) is a forward description of x_k. With (8), the error $\|\Delta_k(n)\|_\infty$ becomes

$$\begin{aligned} \|\Delta_k(n)\|_\infty &= \|x_k'(n) - x_k(n)\|_\infty \\ &= \|(e^{\varepsilon(n-N+1)} - 1) p_k \phi_k^{n-N+1}\|_\infty \\ &= \|(e^{\varepsilon(n-N+1)} - 1) r_k \phi_k^n\|_\infty, \quad 0 \le n < N. \end{aligned}$$

In this case, if $|\phi_k| > 1$, the error does *not* increase exponentially as n increases but becomes zero at $n = N - 1$. In fact, by time-reversing the data we turn unstable components into stable ones and vice versa. This means that we can significantly improve the numerical stability of the overall reconstruction by choosing different representations for different components; if $|\phi_k| \le 1$ we use the forward description (7) whereas if $|\phi_k| > 1$ we use the backward description (8). This procedure guarantees us that *all* components can be regarded as being stable, which is of great importance when N is large, as is the case with high-quality audio signals (typically $N > 500$ for 44.1 kHz sampled audio).

Back to our experiment, Figure 1d shows the result of reconstructing the audio signal of Figure 1a using the two different component representations (solid line) together with the corresponding reconstruction error (dotted line). This signal is reconstructed with the same parameters as the ones used to reconstruct the signal of Figure 1c, that is, including the perturbed parameter ϕ_k. It is obvious that the reconstruction error has been reduced significantly.

The analysis procedure does not change by allowing the two representations. To see this, assume that $|\phi_1|, \ldots, |\phi_j| \le 1$ and $|\phi_{j+1}|, \ldots, |\phi_{\hat{d}}| > 1$. If we let $\mathbf{D} = \mathrm{diag}(1, \ldots, 1, \phi_{j+1}^{1-N}, \ldots, \phi_{\hat{d}}^{1-N}) \in \mathbb{C}^{\hat{d} \times \hat{d}}$ and define $\mathbf{V}_\phi$ as

$$\mathbf{V}_\phi = \begin{bmatrix} 1 & \cdots & 1 \\ \phi_1 & \cdots & \phi_{\hat{d}} \\ \vdots & & \vdots \\ \phi_1^{N-1} & \cdots & \phi_{\hat{d}}^{N-1} \end{bmatrix},$$

the minimization problem (5) can be formulated as

$$\min_{\mathbf{B}} \|\mathbf{x} - \mathbf{V}_\phi \mathbf{B}\|_2 = \min_{\mathbf{B}'} \|\mathbf{x} - (\mathbf{V}_\phi \mathbf{D})\, \mathbf{B}'\|_2,$$

where $\mathbf{B}' = \mathbf{D}^{-1}\mathbf{B} = (r_1, \ldots, r_j, p_{j+1}, \ldots, p_{\hat{d}})^t$ and $\mathbf{x} = (x(0), \ldots, x(N-1))^t$. Hence, rather than finding the parameters $r_1, \ldots, r_{\hat{d}}$, we now find the parameter r_k if $|\phi_k| \le 1$ and p_k if $|\phi_k| > 1$. Note that, in coding or transmission applications, we do not have to code or transmit additional data for discriminating between both representations at the decoder c.q. receiver. If $|\phi_k| \le 1$ we use the forward representation, otherwise we use the backward representation.

5. SEGMENTATION

One of the main problems in audio coders is how to handle so-called "attacks" or "transients". With the exponential modeling, these attacks can be represented very efficiently. The reason for this is that attacks can almost perfectly be described as the impulse response of a linear time-invariant system, which is of the form (3) with $|\phi_k| \le 1$ for all k.

In order to model these attacks with a minimum number of components, it is important that the attack starts at $n = 0$. If this condition is not satisfied, we have to model the signal with considerable more components, the additional components needed to compensate for the samples preceding the attack. This is illustrated by Figure 2. Figure 2a shows a recording of the attack of a castanet. Figure 2c shows the same signal, shifted in time. Figures 2b and 2d show the reconstruction of these two fragments using $\hat{d} = 32$ components. The SNR of the reconstruction of Figure 2b is 13.8 dB, while Figure 2d has a SNR of 26.0 dB. We, therefore, conclude that it is important to split up the input data into segments which can be modeled with low-order systems. As a consequence, the start points of the analysis windows, as well as the length, must be variable.

One way of finding a (possible) split point within a data segment is to divide the segment in two parts, and model each part with an equal number of components $\hat{d}/2$. Since both parts are modeled with a fixed number of components, regardless of their respective lengths, the SNR tends to have an optimum for a split point near the middle of the segment. However, if the segment contains an attack, the optimal split point will be located at the beginning of the attack. The optimal split point found this way is then taken as the boundary of the next segment, and so forth. The procedure is illustrated in Figure 3 for one single segment. Figure 3a shows a segment of 320 samples of a recording of a castanet, sampled at a rate of 8 kHz. The signal is split in two parts where both parts are modeled with 32 components ($\hat{d} = 64$). Figure 3b shows the SNR in the reconstruction of this signal versus the location of the split point. As we see, the optimal split point aligns well with the attack.

It is impractical to determine the optimal split point by trying each split point in between the start and the end of the segment,

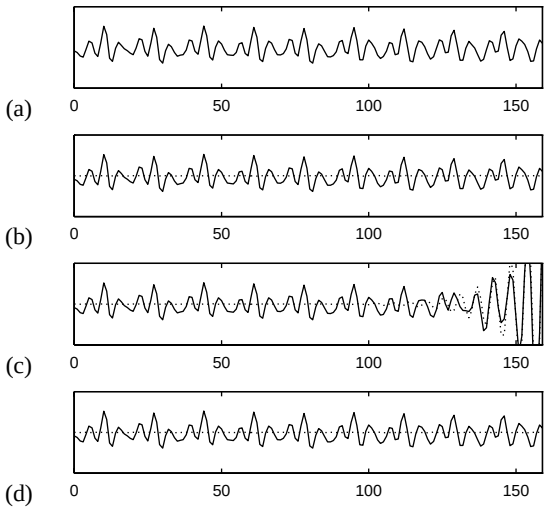

Figure 1: *a) Fragment of a music signal, b) reconstruction using 60 components (solid line) and reconstruction error (dotted line), c) reconstruction with perturbed parameter ϕ_k, and d) corresponding reconstruction using forward and backward representation.*

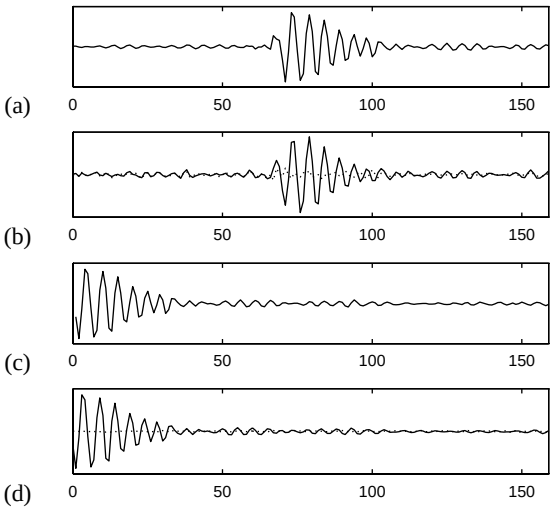

Figure 2: *a) Recording of a castanet signal, b) reconstruction with 32 components (solid line) and reconstruction error (dotted line), c) recording of the (time shifted) castanet signal, and d) reconstruction of c) with 32 components (solid line) and reconstruction error (dotted line).*

since this would require $2N$ Hankel decompositions for each segment. Instead we use a two-step approach. In the first step, the SNR for a few split points is computed in order to roughly locate the optimal split point. In the next step, some split points around the optimal one are computed to more accurately determine the optimal position.

We have used this procedure for the segmentation of a few seconds recording of the castanet, containing several attacks. We used an analysis frame of length $N = 320$ and $\hat{d} = 64$. The resulting mean segment length is 168 and the resulting SNR, measured over the entire signal, is 20.5 dB. If we do the same experiment for a fixed length window of 160 samples (using 32 components per frame), the resulting SNR is only 12.9 dB. We, therefore, conclude that proper segmentation results in a potentially large improvement of SNR for signals containing attacks.

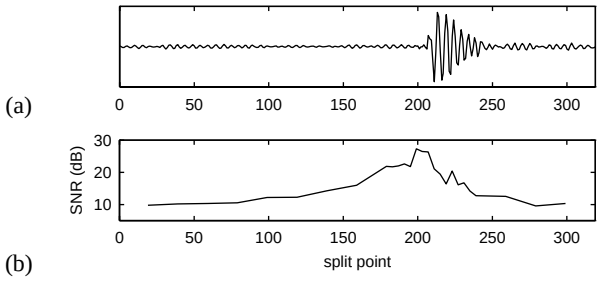

Figure 3: *a) Recording of a castanet signal and b) the signal-to-noise ratio for different split points.*

6. CONCLUSIONS

We have investigated the modeling of audio signals with complex exponentials. We showed that such modeling can very efficiently represent attacks in the audio signal, one of the main bottlenecks in state of the art audio coders. We presented a numerically robust algorithm for determining the exponential components and showed that, by using a proper segmentation of the input data, the total number of components needed for the reconstruction can be significantly reduced as compared to a fixed-length analysis.

7. REFERENCES

[1] *ISO/IEC JTC1/SC29/WG11 N1730, Overview of the MPEG-4 Standard.* Stockholm, July 1997. http://drogo.cselt.stet.it/mpeg/standards/mpeg-4.htm.

[2] W.B. Kleijn and K.K. Paliwal, editors. *Speech coding and synthesis*. Elsevier Science Publishers, Amsterdam, 1995.

[3] R.J. McAulay and T.F. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Trans. on ASSP*, 34(4):744–754, August 1986.

[4] S.M. Kay and S.L. Marple, jr. Spectrum analysis – a modern perspective. *Proceedings of the IEEE*, 69(11):1380–1419, November 1981.

[5] S.Y. Kung, K.S. Arun, and D.V. Bhaskar Rao. State-space and singular-value decomposition-based approximation methods for the harmonic retrieval problem. *J. Opt. Soc. Am.*, 73(12):1799–1811, December 1983.

[6] G.H. Golub and C.F. Van Loan. *Matrix Computations*. North Oxford Academic, Oxford, second edition, 1983.

[7] J.A. Cadzow. Signal enhancement – A composite property mapping algorithm. *IEEE Trans. on ASSP*, 36(1):49–62, January 1988.

[8] S. van Huffel, H. Park, and J.B. Rosen. Formulation and solution of structured total least norm problems for parameter estimation. *IEEE Trans. on Signal Processing*, 44(10):2464–2474, October 1996.