Autoregressive Modeling of Temporal Envelopes

Marios Athineos, *Student Member, IEEE*, and Daniel P. W. Ellis, *Senior Member, IEEE*

***Abstract*—Autoregressive (AR) models are commonly obtained from the linear autocorrelation of a discrete-time signal to obtain an all-pole estimate of the signal's power spectrum. We are concerned with the dual, frequency-domain problem. We derive the relationship between the discrete-*frequency* linear autocorrelation of a spectrum and the temporal envelope of a signal. In particular, we focus on the real spectrum obtained by a type-I odd-length discrete cosine transform (DCT-Io) which leads to the all-pole envelope of the corresponding symmetric squared Hilbert temporal envelope. A compact linear algebra notation for the familiar concepts of AR modeling clearly reveals the dual symmetries between modeling in time and frequency domains. By using AR models in both domains in cascade, we can jointly estimate the temporal and spectral envelopes of a signal. We model the temporal envelope of the residual of regular AR modeling to efficiently capture signal structure in the most appropriate domain.**

***Index Terms*—Autoregressive (AR) modeling, frequency-domain linear prediction (FDLP), Hilbert envelope, linear prediction in spectral domain (LPSD), temporal noise shaping (TNS).**

I. Introduction

Autoregressive (AR) modeling identifies and exploits a particularly simple form of redundancy in signal sequences by finding the optimal linear combination of a fixed-length history to predict the next sample—hence the alternative name of linear predictive (LP) modeling. By extracting the linear dependence (correlation) in a signal, AR models find many applications in signal compression and communications, but the form of the model itself—which describes the original sequence as the result of passing a temporally uncorrelated (white) excitation sequence through a fixed all-pole digital filter—leads to some interesting and important applications in itself. The filter comprises a low-dimensional parametric approximation of the signal, or specifically its broad spectral structure, since the magnitude of the Fourier transform of the signal is the product of the white excitation's spectrum (expected to be flat) and coarse spectral variation provided by the poles of the AR filter.

One prominent application domain relates to human voice, since voiced speech is well modeled as a broadband, pseudo-periodic glottal pulse train filtered by resonances (poles) in the vocal tract, identified with formants. Formants carry much of the linguistic information in speech, and many formant tracking applications operate by fitting AR models to short-time windows of the speech signal, factoring to identify the individual poles, then constructing formant trajectories from the succession of center frequencies of these poles [1]. While explicit formant tracks turn out to be a brittle basis for speech recognition, the properties of AR modeling to suppress fine detail while preserving the broad structure of formants has led to its widespread use in speech recognition preprocessing, for instance in "perceptual linear prediction" (PLP) features [2]. In this application, the all-pole filter defined by the optimal difference equation coefficients is taken as the description of a smoothed spectral envelope—the magnitude of the z-transform of that filter evaluated on the unit circle—which is then described by its cepstral coefficients for further statistical modeling in a speech recognizer.

This paper is concerned with using AR models to form smoothed, parametric models of *temporal* rather than spectral envelopes. The duality between the time and frequency domains means that AR modeling can be applied equally well to sequences which are discrete *spectra* instead of time-domain sequences. In this case the magnitude evaluated on the unit circle in the z-plane describes the *time-domain* envelope—specifically the squared Hilbert envelope. Just as conventional AR models are used most effectively on signals with sharp spectral peaks that can be well modeled with individual complex pole pairs, so AR models of the temporal envelope are most appropriate for "peaky" temporal envelopes, and individual poles in the resulting polynomial may be directly associated with specific energy maxima in the time waveform.

In the same way that a parametric description of the spectral envelope leads to numerous applications, a temporal envelope model can be useful. For signals that are expected to consist of a number of distinct transients, (be they the isolated vocal pitch pulses of low-pitch male speech, or the irregularly spaced crackling of burning log) fitting an AR model can constrain the modeled envelope to be a sequence of maxima, and the AR fitting procedure can remove finer-scale detail. This suppression of detail is particularly useful in classification applications, where the goal is to extract the general form of the signal regardless of minor variations. Because the envelopes modeled by AR can include sharp, narrow maxima even for low-order models, this approach can be preferable to the implicit low-pass filtering of a low-order Fourier approximation.

This idea was first applied in audio coding by Herre and Johnston [3] who dubbed it temporal noise shaping (TNS). This frequency-domain version of D*PCM [4] was used to eliminate pre-echo artifacts associated with transients in perceptual audio coders such as MPEG2 AAC by factoring-out the parameterized time envelope prior to quantization, then reintroducing it during reconstruction. Traditional transform coding introduced

Manuscript received March 24, 2006; revised December 27, 2006. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Kostas Berberidis.This work was supported in part by the Defense Advanced Research Projects Agency (DARPA) under the EARS Novel Approaches Grant MDA972-02-1-0024, in part by the National Science Foundation (NSF) Project IIS-0629345 "Separating Speech from Speech Noise to Improve Intelligibility," and in part by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811) through the International Computer Science Institute, Berkeley, CA.

The authors are with the Laboratory for the Recognition and Organization of Speech and Audio (LabROSA), Electrical Engineering Department, Columbia University, New York, NY 10027 (e-mail: marios@ee.columbia.edu; dpwe@ee.columbia.edu).

Digital Object Identifier 10.1109/TSP.2007.898783

"reverberation" or temporal-smearing pre-echo artifacts to signals that were "peaky" in time, and TNS eliminated these artifacts. In their original and subsequent papers [3], [5]–[7] Herre and Johnston motivated TNS by citing the duality between the squared Hilbert envelope and the power spectrum for continuous signals, but no exact derivation for finite-length discrete-time signals was given.

Kumaresan *et al.* [8]–[13] have also addressed the problem of AR modeling of the temporal envelope of a signal. Specifically, in [8] Kumaresan formulated the so-called linear prediction in the spectral domain (LPSD) equations. Working with a continuous, periodic time-domain signal and its corresponding infinite-length Fourier series, he used an all-pole model in order to fit the Hilbert envelope without calculating the corresponding analytic signal. By considering the discrete, periodic signal obtained by sampling the continuous signal, the AR model of the infinite series was taken as an approximate model for the finite-length discrete Fourier transform of a finite discrete signal.

We solve the same problem of finding an AR model of a discrete spectrum and relating it to the envelope of the finite, discrete time-domain signal, but our solution is expressed entirely in the discrete, finite domain through matrix operations. One particular issue we examine concerns continuity at boundaries: because the envelope calculated on the z-plane is intrinsically periodic, discontinuities between the first and last values of a finite-length sequence, which is treated as periodic by the analysis, will lead to Gibbs-type ringing. To avoid this, we symmetrize the time signal prior to analysis and model the envelope of the resulting, double-length signal. As a consequence, we end up using a discrete cosine transform (DCT) to shift between temporal and spectral domains, and our envelopes, while constrained to have zero slope at the edges, do not suffer from discontinuities in value. This leads to our central result, that to obtain an all-pole model describing the squared Hilbert envelope of an odd-length symmetrized discrete sequence, one should apply AR modeling to the real spectrum computed by the DCT type I odd (DCT-Io).

We present a linear-algebra derivation of AR models that is completely symmetric for spectral and temporal envelopes. By using orthogonal versions of all transforms, energy preservation and perfect reconstruction properties are guaranteed. This leads to a joint fit of a cascade of temporal and spectral AR models that obtains a lower average minimum total squared error when compared to independent modeling in each domain.

Section II presents the mathematical background needed to relate autocorrelation and envelopes. In Section III we formulate the time-domain and the frequency-domain AR models in a dual fashion and combine them to form the cascade and joint time-frequency models. Section IV contains two examples, modeling the temporal envelope of voiced speech, and modeling the speech residual. We draw conclusions and discuss future work in Section V.

II. Relating Autocorrelation and Envelopes

The aspect of AR modeling that interests us is its ability to approximate the envelope of the transform of a signal by starting from the autocorrelation in the nontransformed domain. Conventionally, this is estimating the spectral envelope from the autocorrelation of the time signal. In the dual domain, this becomes estimating the temporal envelope from the autocorrelation of the spectrum—in our case the DCT. We show that by concentrating on odd-length, symmetric signals this relationship is particularly tidy.

A. Linear Autocorrelation

Let $\mathbf{x}$ be an N-dimensional real column vector that represents a finite-duration discrete-time real signal $x(n)$

$$\mathbf{x} = [x(0)\, x(1)\, \ldots\, x(N-1)]^T \tag{1}$$

where $\{\cdot\}^T$ denotes transposition. Let $\tilde{\mathbf{r}}_\mathbf{x}$ be a column vector that represents the biased linear (aperiodic) autocorrelation of $x(n)$, where $|m| \leq N-1$. In column vector form we have

$$\tilde{\mathbf{r}}_\mathbf{x} = [\tilde{r}_x(-N+1)\ldots\tilde{r}_x(-1)\ldots\tilde{r}_x(0)\tilde{r}_x(1)\ldots\tilde{r}_x(N-1)]^T \tag{2}$$

where

$$\tilde{r}_x(m) = \frac{1}{N}\sum_{n=0}^{N-|m|-1} x(n)x\left(n+|m|\right) \tag{3}$$

$\tilde{\mathbf{r}}_\mathbf{x}$ is always odd-length with $M = 2N-1$, and is even-symmetric over the lag $m = 0$ i.e., $\tilde{r}_x(m) = \tilde{r}_x(-m)$.

We define $\mathbf{Z}_{r,l}$ as the $(l+n+r) \times n$ zero-padding matrix

$$\mathbf{Z}_{r,l} = [\mathbf{0}_{n\times l}\, \mathbf{I}_n\, \mathbf{0}_{n\times r}]^T \tag{4}$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix. Left-multiplying $\mathbf{x}$ by the matrix $\mathbf{Z}_{r,l}$ pads an input sequence $\mathbf{x}$ to the right and left by r and l zeros respectively, where the dimension n can be inferred by the vector that $\mathbf{Z}_{r,l}$ is applied to. For the most part of this paper l will be zero (meaning no left padding) in which case we will drop the parameter l and denote the matrix as $\mathbf{Z}_r$. The special case $\mathbf{Z}_{N-1}$ (right padding by $N-1$ zeros) is simply denoted as $\mathbf{Z}$. Note that the transposed zero-padding matrix $\mathbf{Z}^T$ applied to an M-dimensional vector (where $M = 2N-1$ as above) simply selects the first N elements, and the product $\mathbf{Z}\mathbf{Z}^T$ applied to an M-dimensional vector zeroes out the last $N-1$ elements.

Let $\mathbf{F}$ be the $M \times M$ unitary discrete Fourier transform (DFT) matrix defined as

$$\mathbf{F} = \frac{1}{\sqrt{M}}\exp\left(-\jmath\frac{2\pi mn}{M}\right) \tag{5}$$

where the row and column indexes are $m, n = 0, 1, \ldots, M-1$. For the unitary DFT we have $\mathbf{F}^{-1} = \mathbf{F}^H$ and $\mathbf{F}^H = \mathbf{F}^*$.

Defining $\hat{\mathbf{x}} = \mathbf{F}\mathbf{Z}\mathbf{x}$ as the forward DFT of the zero-padded input signal $\mathbf{x}$, the autocorrelation of (3) is

$$\mathbf{r}_\mathbf{x} = \frac{\sqrt{M}}{N}\mathbf{F}^H(\hat{\mathbf{x}} \odot \hat{\mathbf{x}}^*) \tag{6}$$

where $\{\cdot\}^*$ denotes complex conjugation and $\mathbf{A} \odot \mathbf{B}$ denotes the Hadamard (element wise) product, giving the familiar relationship between autocorrelation and the magnitude of the transform-domain signal. Note the distinction between $\mathbf{r}_\mathbf{x}$ of (6) and $\tilde{\mathbf{r}}_\mathbf{x}$ of (2), which is rotated by $N-1$ elements to start with $x(0)$.

B. WSHS Symmetry of Autocorrelation

In the terminology of Martucci [14], autocorrelation is left whole-sample symmetric—right half-sample symmetric (WSHS). Formally, an M-point sequence $\bar{x}(n)$ is WSHS symmetric if

$$\bar{x}(n) = \begin{cases} x(n) & n = 0, 1, \ldots, N-1 \\ x(M-n) & n = N, \ldots, M-1 \end{cases} \tag{7}$$

and $\bar{x}(n)$ is always an odd-length sequence. An infinite periodic extension of $\bar{x}(n)$ is symmetric over the sample at index $n = 0$ (the whole-sample part) and also symmetric over the "half-sample" $n = N - (1/2)$ i.e., $\bar{x}(N-1) = \bar{x}(N)$.

The $M \times N$-dimensional WSHS symmetric extension operator (SEO) matrix $\mathbf{S}$ is defined [14] as

$$\mathbf{S} = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \\ & & & 1 \\ & & ⋰ & \\ & 1 & & \end{bmatrix} = \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{N-1} \\ \mathbf{0} & \mathbf{J}_{N-1} \end{bmatrix} \tag{8}$$

where $\mathbf{J}_{N-1}$ is the $N-1 \times N-1$ reverse identity matrix (1's on the antidiagonal). We can right-symmetrize an N-dimensional signal $\mathbf{x}$ by left multiplying it by S i.e., $\bar{\mathbf{x}} = \mathbf{S}\mathbf{x}$. $\mathbf{S}^T$ is essentially an aliasing operator: it "folds" and adds the signal onto itself, to reduce M points to N.

Note that the DFT and inverse DFT (IDFT) of a WSHS sequence are also WSHS. In our case this means that since the autocorrelation $\mathbf{r_x}$ is WSHS, then the corresponding sampled power spectrum $\hat{\mathbf{x}} \odot \hat{\mathbf{x}}^*$ in (6) will also be WSHS.

C. Discrete Cosine Transform Type I Odd (DCT-Io)

Out of the 16 discrete trigonometric transforms (DTT) first tabulated by Wang [15], we are interested in the DCT-Io which is the only one related to the DFT through the WSHS SEO operator [14], meaning that the WSHS symmetry property of the autocorrelation can be preserved through the DCT-Io. The orthogonal $N \times N$ DCT-Io matrix $\mathbf{C}$ is defined as

$$\mathbf{C} = \frac{2}{\sqrt{M}} k_m k_n \cos\left(\frac{2\pi mn}{M}\right) \tag{9}$$

where $m, n = 0, 1, \ldots, N-1$ and the coefficients k_j are

$$k_j = \begin{cases} 1/\sqrt{2}, & j = 0 \\ 1, & j \neq 0. \end{cases} \tag{10}$$

We place the weights k_j on the main diagonal of an $N \times N$ matrix $\mathbf{W}$ that we define as

$$\mathbf{W} = \begin{bmatrix} 1/\sqrt{2} & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix} = \begin{bmatrix} 1/\sqrt{2} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{N-1} \end{bmatrix} \tag{11}$$

and we derive the following orthogonal DCT-Io factorization (see Appendix I):

$$\mathbf{C} = \mathbf{W}(\mathbf{Z}^T\mathbf{F}\mathbf{S})\mathbf{W}^{-1}. \tag{12}$$

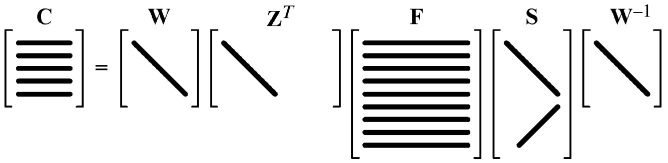

Fig. 1. DCT-Io factorization. A pictorial representation of (12). Diagonal matrices are represented by diagonal lines and basis vectors are represented by horizontal lines.

Notice that the DFT matrix $\mathbf{F}$ is $M \times M$ and complex whereas the DCT-Io matrix $\mathbf{C}$ is $N \times N$ and real; moreover the inverse of $\mathbf{W}$ is trivial since it is diagonal and non singular. One way to interpret this equation is that the DCT-Io can be seen simply as the first N elements $(\mathbf{Z}^T)$ of the DFT $(\mathbf{F})$ of the WSHS-symmetrized $(\mathbf{S})$ input vector. In order to make the columns and rows orthogonal, we left and right multiply by $\mathbf{W}$ and $\mathbf{W}^{-1}$ respectively. This interpretation is depicted pictorially in Fig. 1.

Although $\mathbf{C}$ is orthogonal and self-inverse (involutary), we wish to derive a *nonorthogonal* forward $(\mathbf{C}_F)$ and corresponding inverse $(\mathbf{C}_I)$ pair of $N \times N$ DCT-Io transforms. We have

$$\begin{aligned} \mathbf{I}_N &= \mathbf{C}\mathbf{C}^T \\ &= \mathbf{W}(\mathbf{Z}^T\mathbf{F}\mathbf{S})\mathbf{W}^{-1}\mathbf{W}^{-1}(\mathbf{S}^T\mathbf{F}^H\mathbf{Z})\mathbf{W} \\ &= \mathbf{W}^2(\mathbf{Z}^T\mathbf{F}\mathbf{S})\mathbf{W}^{-2}(\mathbf{S}^T\mathbf{F}^H\mathbf{Z}). \end{aligned} \tag{13}$$

(We can move the $\mathbf{W}$ like this because of the identity matrix on the left-hand side i.e., if $\mathbf{I} = \mathbf{A}\mathbf{X}$ and $\mathbf{A}$ is nonsingular, then $\mathbf{X} = \mathbf{A}^{-1}$ and thus $\mathbf{A}\mathbf{X} = \mathbf{X}\mathbf{A}$.) Associating the scalar weights 2 and 1/2 with the forward and inverse nonorthogonal DCT-Io we can divide into a pair of new transforms

$$\mathbf{C}_F = 2\mathbf{W}^2(\mathbf{Z}^T\mathbf{F}\mathbf{S}) \tag{14}$$

$$\mathbf{C}_I = 1/2\mathbf{W}^{-2}(\mathbf{S}^T\mathbf{F}^H\mathbf{Z}) \tag{15}$$

and thus $\mathbf{C}_F\mathbf{C}_I = \mathbf{I}_N$. Note that, like the orthogonal DCT-Io $\mathbf{C}$, $\mathbf{C}_F$ includes the $\mathbf{Z}^T\mathbf{F}\mathbf{S}$ term i.e., a truncated Fourier transform of a WSHS-symmetrized sequence. The main difference is that in $\mathbf{C}_F$ the vector being processed is exactly the input sequence, whereas in $\mathbf{C}$ the intervening $\mathbf{W}^{-1}$ factor modifies the input vector prior to the transform. Note that $\mathbf{W}$ affects only the first element of the time and frequency domain signals.

D. Discrete-Time "Analytic" Signal

The analytic signal was introduced by Gabor [16]. Its fundamental property is that its spectrum vanishes for negative frequencies or, put another way, it is "causal" in the frequency domain. By this definition, a discrete-time signal cannot be analytic because its spectrum is periodic and thus not causal. One way to define a discrete-time "analytic" signal is by forcing the spectrum to be "periodically causal" [17] meaning that the second half of each periodic repetition of the spectrum is forced to be zero. Marple [18] used this definition in order to derive a discrete-time analytic signal using the DFT.

In the time domain, the analytic signal is complex with its real part being the original signal and its imaginary part being the Hilbert transform of the original signal [19]. The squared magnitude of this time-domain signal is the temporal envelope we will approximate through AR modeling, by showing its dual relationship to the sampled power spectrum.

Conversion to an analytic signal can be expressed as a matrix multiply. The $M \times M$ analytic transformation matrix $\mathbf{A}$ is

$$\mathbf{A} = \mathbf{F}^H(2\mathbf{W}^2)(\mathbf{Z}\mathbf{Z}^T)\mathbf{F}. \tag{16}$$

We can interpret $\mathbf{A}$ as follows. After taking the DFT $(\mathbf{F})$ of the input signal we first zero-out the negative frequencies by multiplying with $\mathbf{Z}\mathbf{Z}^T$ thus forcing the spectrum to be periodically causal. Then we scale with the appropriate weights $(2\mathbf{W}^2)$ as defined in [18] to ensure orthogonality of the real and imaginary parts. Finally we take the IDFT $(\mathbf{F}^H)$ to return to the time domain.

E. Autocorrelation of the DCT-Io

Eqn. (6) showed that the autocorrelation of the time-domain signal is the IDFT of the power spectrum. The dual of this is that the autocorrelation of the DCT-Io transform of a signal $\mathbf{x}$ is the DFT of the WSHS-symmetric squared Hilbert envelope. Starting from the nonorthogonal DCT-Io of $\mathbf{x}$, $\mathbf{y} = \mathbf{C}_F\mathbf{x}$, its autocorrelation is given by analogy to (6) as

$$\mathbf{r}_\mathbf{y} = \frac{\sqrt{M}}{N}\mathbf{F}(\hat{\mathbf{y}} \odot \hat{\mathbf{y}}^*) \tag{17}$$

where $\hat{\mathbf{y}} = \mathbf{F}^H\mathbf{Z}\mathbf{y}$ analogously to $\hat{\mathbf{x}}$ of (6). (Because $\mathbf{r}_\mathbf{y}$ is WSHS-symmetric and has a pure-real transform, $\mathbf{F}$ and $\mathbf{F}^H$ are interchangeable here.)

Using the forward DCT-Io of (14) and the analytic transformation matrix $\mathbf{A}$ of (16) we can express $\hat{\mathbf{y}}$ as follows:

$$\begin{aligned}\hat{\mathbf{y}} &= \mathbf{F}^H\mathbf{Z}\mathbf{y}\\ &= \mathbf{F}^H\mathbf{Z}\mathbf{C}_F\mathbf{x}\\ &= \mathbf{F}^H\mathbf{Z}\left(2\mathbf{W}_N^2\mathbf{Z}^T\mathbf{F}\mathbf{S}\right)\mathbf{x}\\ &= \mathbf{F}^H\left(2\mathbf{W}_M^2\right)(\mathbf{Z}\mathbf{Z}^T)\mathbf{F}\mathbf{S}\mathbf{x} \Rightarrow\\ \hat{\mathbf{y}} &= \mathbf{A}\mathbf{S}\mathbf{x}. \end{aligned} \tag{18}$$

(The last equation holds since $\mathbf{Z}\mathbf{W}_N^2 = \mathbf{W}_M^2\mathbf{Z}$ with an appropriate change of dimensionality of $\mathbf{W}^2$ from $N \times N$ to $M \times M$.) The important interpretation of this formula is that the inverse DFT of the zero-padded DCT-Io of $\mathbf{x}$ is equal to the analytic WSHS-symmetrized signal. This means that the Hadamard product $\hat{\mathbf{y}} \odot \hat{\mathbf{y}}^*$ in (17) is just the squared Hilbert envelope of the WSHS-symmetrized signal. This equality allows us to model the temporal envelope of $\mathbf{x}$ by fitting an AR model to the DCT-Io.

This derivation depended on using the nonorthogonal $\mathbf{C}_F$. However, the material differences between $\mathbf{C}_F$ and the orthogonal DCT-Io $\mathbf{C}$ lie only in the zero-index terms in both time and frequency. If the signal being analyzed has zero values for these terms, we could equally use $\mathbf{C}$. We can construct such a signal by 1) subtracting its mean in the time domain, and 2) left-padding with a zero, since this first time-domain value is, by duality, the mean value in the frequency domain.

III. Autoregressive Modeling

Since the autocorrelation of the DCT-Io relates to the WSHS-symmetrized temporal envelope just as the time-domain autocorrelation determines the power spectrum, we can use autocorrelation-based AR techniques to approximate the temporal envelope. We present three variants. In Section III-A we review the standard time-domain AR model and its frequency-domain counterpart. Section III-B uses time and frequency domain AR models in cascade to model both spectral and temporal envelopes. Lastly, Section III-C optimizes the cascade, jointly minimizing a total quadratic error by iterating on the partial derivatives with respect to the time- and frequency-domain AR models.

A. Time and Frequency Duality

Consider convolution expressed in a matrix form. If $\mathbf{x}$ and $\mathbf{y}$ are N- and L-dimensional vectors respectively, we must zero-pad each one to length $N+L-1$ to avoid circular aliasing when convolving them, i.e.,

$$\mathbf{x} \circledast \mathbf{y} = \mathbf{X}\mathbf{Z}\mathbf{y} \tag{19}$$

where $\mathbf{X}$ is a right-circulant matrix [20], [21] with $\mathbf{Z}_{L-1}\mathbf{x}$ as its first column (i.e., generated by $\mathbf{Z}_{L-1}\mathbf{x}$). Convolution is commutative, so (19) is also equal to $\mathbf{y} \circledast \mathbf{x} = \mathbf{Y}\mathbf{Z}_{L-1}\mathbf{x}$ where $\mathbf{Y}$ is the right-circulant matrix generated by $\mathbf{Z}\mathbf{y}$.

AR modeling is equivalent to finding the FIR filter, with 1 as its first coefficient, that minimizes the energy of the output when applied to the sequence being modeled i.e.,

$$\mathbf{X}\mathbf{Z}\mathbf{a} = \mathbf{d} \tag{20}$$

where $\mathbf{X}$ is the right circulant matrix generated by the zero-padded signal $\mathbf{Z}_{L-1}\mathbf{x}$, $\mathbf{a} = [1 \; a(1) \; \ldots \; a(L-1)]^T$ are the AR model coefficients to be found, and the residual $\mathbf{d}$ is the $N+L-1 \times 1$ column vector residual whose norm is to be minimized. Note that the $\mathbf{Z}$ can be viewed as zero-padding $\mathbf{a}$ to facilitate convolution, or as right-multiplying $\mathbf{X}$ to truncate the last L columns, making the system of equations over-determined.

The solution of (20) is given by minimizing the quadratic $D = \mathbf{d}^T\mathbf{d}$ by setting its first derivative with respect to the free elements in $\mathbf{a}$ to zero. This leads to the well-known Yule-Walker equations, which give the classic result that the AR solution depends only on the autocorrelation $\mathbf{r}_\mathbf{x}$ of the signal $\mathbf{x}$ being modeled. In our notation, this becomes

$$\left(\mathbf{Z}_N^T\mathbf{R}_\mathbf{x}\mathbf{Z}_N\right)\mathbf{Z}_{0,1}^T\mathbf{a} = -\mathbf{Z}_{N-1,1}^T\mathbf{r}_\mathbf{x} \tag{21}$$

where $\mathbf{R}_\mathbf{x}$ is the circulant matrix generated by $\mathbf{r}_\mathbf{x}$, and the $\mathbf{Z}$ matrices are simply trimming the other elements appropriately.

The residual from passing the original signal through the FIR filter defined by the AR coefficients is simply the first N elements of $\mathbf{d}$ from (20) i.e., $\mathbf{d}_N = \mathbf{Z}_{L-1}^T\mathbf{d}$. The average minimum total squared error as defined in [22] is given by

$$G^2 = \frac{1}{N}\mathbf{d}_N^T\mathbf{d}_N \tag{22}$$

and it will be used as a goodness-of-fit measure in Section IV-B.

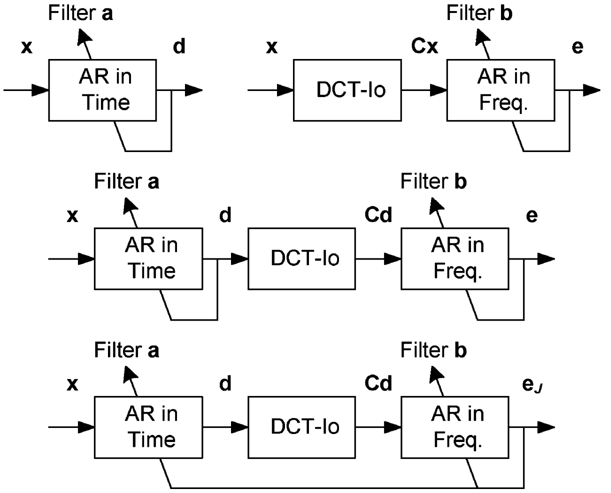

Fig. 2. Block diagrams of AR models. On the left of the first row we plot the regular time-domain AR model and on the right the frequency domain AR model. On the middle row we plot the cascade time-frequency AR model and on the bottom we plot the joint time-frequency AR model.

In the dual domain, (20) becomes

$$\mathbf{YZb} = \mathbf{e} \tag{23}$$

where $\mathbf{Y}$ is the circulant matrix generated by the zero-padded DCT-Io transformed signal $\mathbf{Z}_{K-1}\mathbf{y}$, $\mathbf{b} = [1\ b(1)\ \ldots\ b(K-1)]^T$ are the coefficients of the filter we are estimating—this time in the DCT-Io domain—and $\mathbf{e}$ is again a residual to be minimized. We call this model frequency-domain linear prediction or FDLP [23].

The solution of this equation involves the autocorrelation of $\mathbf{y}$, i.e., the circulant matrix $\mathbf{R_y}$ and the vector $\mathbf{r_y}$. From (17) and (18), we can see that the magnitude of the AR filter defined by $\mathbf{b}$ evaluated on the unit circle in the z-plane—the analog of the conventional AR spectral magnitude approximation—is in this case an approximation of the WSHS-symmetric squared Hilbert envelope of the original sequence. One implication of this is that sharp transients in the temporal signal result in significant, short-lag correlation among values in the DCT-Io which can be effectively captured by AR models.

The block diagrams of temporal and frequency domain AR models are plotted on the top row of Fig. 2.

B. Cascade Time-Frequency AR Modeling

The time-domain residual signal $\mathbf{d}_N$ is spectrally flat since its spectral peaks are balanced by the zeros of the estimated FIR filter. However, any remaining temporal structure is modified but not in general eliminated. For instance, in voiced speech the residual still carries clearly audible pitch pulses. We propose a second AR model operating on the DCT-Io-transformed residual in order to generate a *temporally* flat frequency-domain residual.

This second model is obtained by calculating the DCT-Io transform of the residual and estimating a new filter $\mathbf{b}$ that operates on the spectrum of the residual; we have

$$\mathbf{DZb} = \mathbf{e} \tag{24}$$

where $\mathbf{D}$ is the circulant matrix generated by $\mathbf{Z}_{K-1}\mathbf{Cd}_N$ (the zero-padded DCT-Io transformed residual). The filters $\mathbf{a}$ and $\mathbf{b}$ can be calculated by solving (20) and (24) independently, one after the other. This process is depicted on the block diagram plotted on the second row of Fig. 2. Note that the ordering of filters (time domain first) was suggested by our chosen application of modeling the speech residual (Section IV-B). From a theoretical point of view, one as well could estimate the frequency domain filter first.

C. Joint Time-Frequency AR Modeling

The separate, sequential optimization of the two filters in the cascade model above ultimately aims to minimize a quadratic error in the frequency domain by eliminating temporal structure. But the first filter optimized a time-domain error; a fully optimal solution will solve for both filters jointly. Our objective is to minimize the final frequency domain residual and we can try to do that by estimating the two filters jointly. The error measure we minimize is

$$\mathbf{DZb} = \mathbf{e}_J \tag{25}$$

but this time both filters $\mathbf{a}$ (which affects $\mathbf{D}$) and $\mathbf{b}$ are variable. In order to minimize the quadratic $E_J = \mathbf{e}_J^T\mathbf{e}_J$ we need to calculate the two partial derivatives with respect to the vectors $\mathbf{a}$ and $\mathbf{b}$ and set them to zero. The partial with respect to $\mathbf{b}$ is the same as solving the frequency domain (24) but this time the solution for $\mathbf{b}$ is a function of $\mathbf{a}$.

In order to calculate the partial with respect to $\mathbf{a}$ we use the commutativity property of convolution to write (25) as

$$\mathbf{BZ}_{K-1}\mathbf{Cd}_N = \mathbf{e}_J \tag{26}$$

where $\mathbf{B}$ is the right-circulant matrix generated by $\mathbf{Zb}$. Finally, the error becomes

$$\mathbf{BZ}_{K-1}\mathbf{CZ}_{L-1}^T\mathbf{XZa} = \mathbf{e}_J \tag{27}$$

where we have substituted $\mathbf{d}$ using (20). Now the partial with respect to $\mathbf{a}$ is easy to calculate which means that we can express the optimal $\mathbf{a}$ as a function of $\mathbf{b}$. We do not have an exact solution of this system of equations, but repeatedly solving for $\mathbf{a}$ and $\mathbf{b}$ in turn (and using each new value in the solution for the other) settles to a minimal value for E_J. Both of (25) and (27) are equivalent to "normal equation" forms, solving for the free elements in the $\mathbf{b}$ and $\mathbf{a}$ vectors, respectively, that minimize the length of the residual $\mathbf{e}_J$—which will therefore always be orthogonal to the subspace defined by the relevant columns of the left-multiplying matrix. In each iteration, the particular normal problem being solved depends on the values assigned to the other AR model, which appear in the left-multiplying matrix and of course change in each iteration. However, the particular value of $\mathbf{e}_J$ that was optimal in the solution of (25) exists as a possible solution of (27) (i.e., if $\mathbf{a}$ were unchanged), but the normal solution allows us to find the global minimum for that particular fixed value of $\mathbf{b}$. Thus, E_J is guaranteed to get smaller at each iteration, and the iterative procedure will always converge to a local minimum. In our experience, 5–10 iterations are sufficient to converge to a value that improves by less than

1% in successive steps. The question remains whether this is the unique, global optimal solution: we believe it is, given the simple convexity of each subproblem, but we do not yet have a proof. The results, however, show that the joint solution does represent a real improvement over the cascade in any case. A block diagram representation of the joint model is plotted on the last row of Fig. 2.

IV. Examples and Applications

Here we present two applications of the AR models we have introduced. Section IV-A illustrates the use of the time- and frequency-domain linear prediction filters to model the formant structure and pitch pulses, respectively, of a segment of speech. This simple example demonstrates the flexibility and power of the parametric model. FDLP poles represent temporal peaks, with the pole's distance to the unit circle relating to peak sharpness and the pole's angle reflecting the peak's position in time. This parametric description provides novel features for applications such as automatic speech recognition [23].

In Section IV-B we show how the second-stage frequency domain model can be used to parameterize the residual of a standard time-domain model, for instance for speech coding applications. We show that the joint model improves on the cascade in terms of the average minimum total squared error.

We have also used the cascade model for audio synthesis, specifically in the modeling of audio textures. Sounds such as rain or footsteps are rich in temporal micro-transients which are well represented by the FDLP model [24]. A similar application in coding was investigated by Schuijers [25]. He used FDLP to model the temporal envelope of a noise-excited segment but substituted the spectral AR-moving average (ARMA) model with a Laguerre filter. Schuijers' model later became part of the MPEG-4 SSC coding standard [26].

A. Temporal Envelope Modeling

Fig. 3 shows the spectral and temporal envelopes of 50 ms of the /aa/ (arpabet) sound from the word "c_o_ttage," extracted from the TIMIT corpus and sampled at 8 kHz. On the upper left pane we plot the original signal and on the upper right the output of the DCT-Io transform. On the middle right pane we plot the first half of the sampled power spectrum, the other half being WSHS-symmetric. Specifically, we plot $\mathbf{Z}^T(\hat{\mathbf{x}} \odot \hat{\mathbf{x}}^*)$ as in (6). On the middle left pane we plot the corresponding first half of the WSHS-symmetric squared Hilbert envelope; similarly this is given by $\mathbf{Z}^T(\hat{\mathbf{y}} \odot \hat{\mathbf{y}}^*)$ from (17). On the bottom two panes we plot AR envelopes modeled by two 24th order filters. Notice the classic all-pole behavior of fitting the peaks of the signal—meaning the pitch pulses in the time domain and the formants in the frequency domain—which are well represented by the poles of the corresponding filters. The valleys of the signal in time and frequency are smoothed since they contribute little to the quadratic error. Note that since the Hilbert envelope being modeled has been WSHS-symmetrized, the temporal envelope from filter $\mathbf{b}$ has zero slope at the boundaries (as does the spectral model).

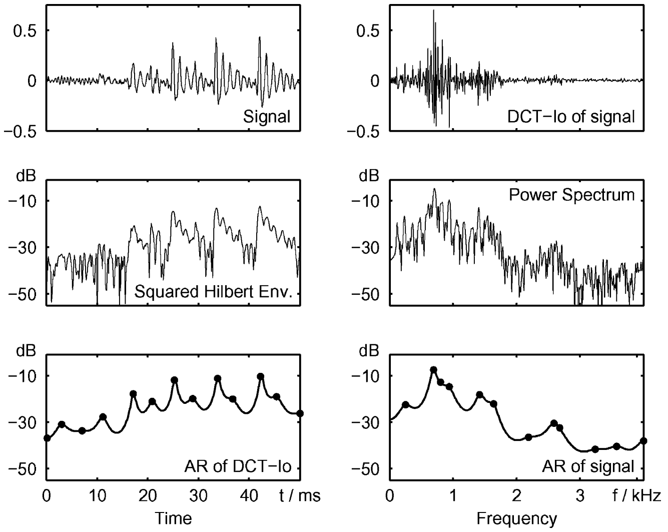

Fig. 3. Dual forms of AR modeling. On the left column (time) we plot 50 ms of a speech signal, the corresponding squared Hilbert envelope and the all-pole fit. On the right column (frequency) we display the DCT-Io of the same signal, the corresponding power spectrum and the conventional time-domain all-pole fit. Both envelope models are 24th order and the dots show pole angles.

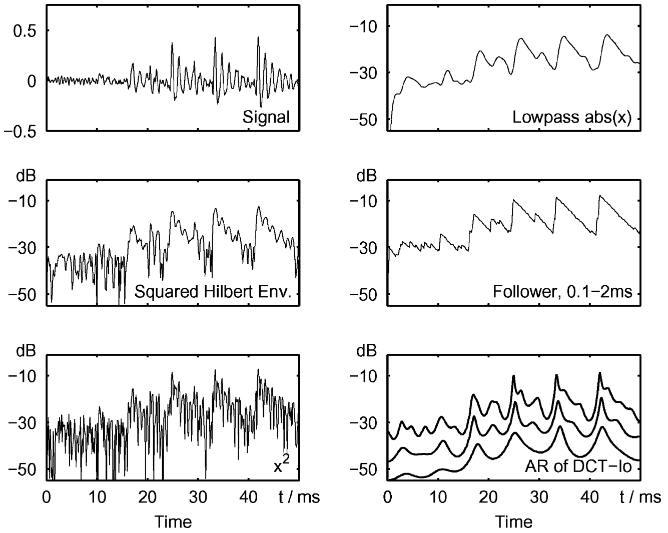

Fig. 4. Comparison of envelope representations. On the left column we plot the voiced speech signal, the squared Hilbert envelope, and on the bottom the signal squared. On the upper right is the low-pass filtered full-rectified signal and on the middle right is the output of an envelope follower. On the bottom right we plot three frequency domain AR models of order 16, 24, and 48 shifted by -20, -10, and 0 dB for readability.

Fig. 4 compares a range of methods for extracting temporal envelopes. Notice the noisiness of the envelope obtained by squaring the signal (bottom left); the squared Hilbert envelope (middle left) is much less noisy by comparison. A low-pass filtered version on the top right smooths out both peaks and valleys of the signal as all sharp edges are eliminated. On the middle right a simple envelope follower with exponential attack of 0.1 and exponential decay of 2 ms follows the signal well but it is approximately piecewise-linear on a log scale. For comparison on the bottom right we present the estimated envelope using model orders of 16, 24, and 48, where the detail increases with the model order. The angles of poles of the FDLP model indicate the timing of the peaks of the signal and are particularly accurate if the poles are sharply "tuned" (close to the unit circle in the z-plane). Applied to subbands this can provide useful features for speech recognition [23].

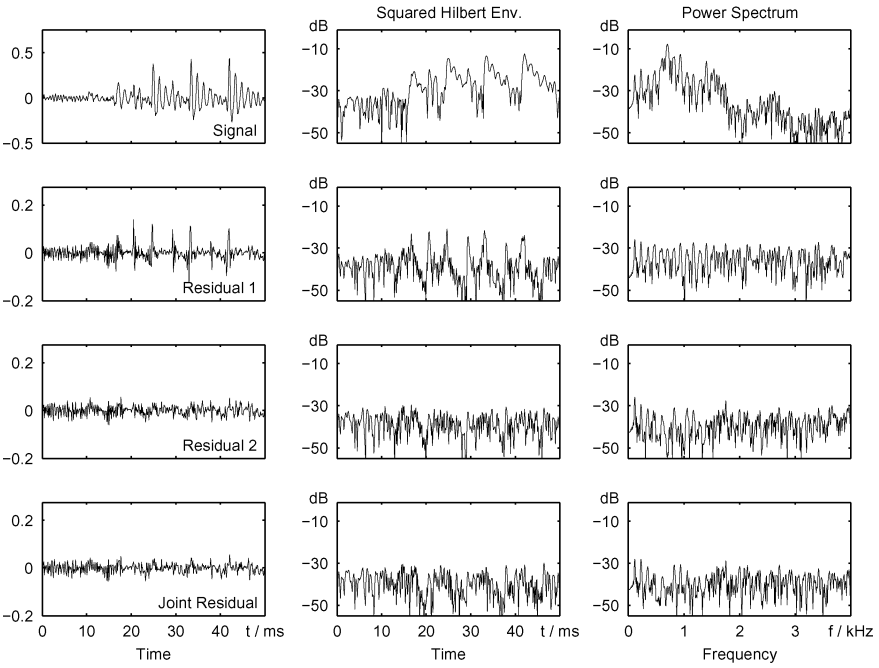

Fig. 5. Cascade time-frequency modeling. On the upper row, we present the time-domain speech signal, its corresponding squared Hilbert envelope and its sampled power spectrum. On the second row we plot the residual in time after whitening the spectral envelope through a regular time-domain AR modeling. On the third row we present the residual after the second (cascade) AR model, this time filtering the DCT-Io. On the last row we plot the residual after joint modeling. All models are 24th order.

B. Modeling of the Speech Residual

The second application we present is modeling the residual of a regular time-domain AR model. A common way to model the pitch pulses typically remaining in the residual is through a second long-term predictor (LTP) [27]. Our cascade and joint AR models can parameterize each pitch pulse in the second, frequency-domain AR model, and thereby flatten the temporal envelope of the overall residual.

In Fig. 5, we plot the effect of modeling speech using the cascade model. On the upper row we plot the same speech segment used in Fig. 3 along with the corresponding squared Hilbert envelope and power spectrum. After filtering the signal with the first-stage time-domain AR filter we get the residual 1 plotted on the second row along with its two envelopes. Notice how the pitch pulses persist in the time domain whereas the spectral envelope is broadly flat since the formant peaks have been captured by the first stage time-domain AR filter. Plotting the Hilbert envelope in the log domain reveals sharp details in the temporal envelope that are difficult to capture with traditional LTP. In fact even though the original time domain signal appears to contain four pronounced pitch cycles, the first stage residual contains six dominant temporal peaks which could have perceptual importance, since they identify moments within the pitch cycle where the waveform does not correspond simply to decaying resonances—the kind of textural detail that may contribute to perceived voice quality.

These temporal peaks are modeled by the second stage AR model operating on the DCT-Io of the residual 1. This operation is illustrated on the third row where we plot the residual 2 in the time domain after filtering the DCT-Io by the second filter. Notice that both the temporal and the spectral envelopes are flattened. Also notice that due to the nature of all-pole modeling the dips (or zeros) of the signal are barely affected and persist in both temporal and spectral envelopes. The residuals 1 and 2 correspond to $\mathbf{d}$ and $\mathbf{e}$ respectively in the cascade block diagram of Fig. 2.

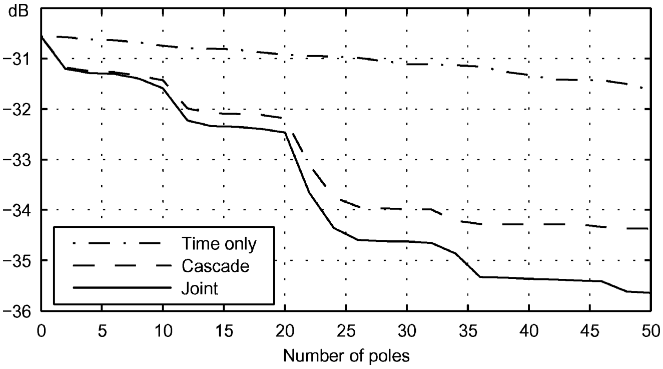

Fig. 6. Average minimum total squared error comparison. All three models start with 24th order time-domain models. On the time-only plot we keep adding poles from 0 to 50 for a total of 24 up to 74 poles. On the cascade and joint models we add 0 to 50 poles on the frequency domain model keeping the time domain fixed at 24.

Fig. 5 shows the results of the simple cascade model with independent, noniterative filter optimization; the joint model should improve on this. Specifically, we note that the power spectrum of the final residual has been distorted by the application of the frequency-domain AR filter, which might be avoided by a joint optimization.

Fig. 6 shows the average minimum total squared error of (22) as a function of the model order for the three models. We plot the average minimum total squared error of the regular time-only

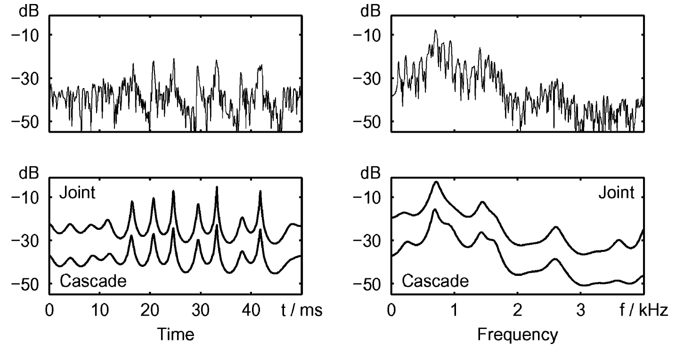

Fig. 7. Filter comparison. On the left column we plot the Hilbert envelope of the residual and its two models (joint and cascade) and on the right we plot the power spectrum and its two models. The joint and cascade models have been shifted by $+10$ dB and -5 dB respectively for readability purposes. All filters are 24th order.

model for filter orders between 24 and 74. For comparison, in the cascade and joint models we fix the temporal filter order at 24 then add 0 to 50 extra coefficients to the second stage filter, capturing the temporal peaks of the voiced speech signal. Even the purely time-domain model improves performance as model order increases, but only by about 1 dB for 50 extra poles. Adding the same number of coefficients to the second model gains 4 dB extra at $+24$ poles onwards. The joint model is able to squeeze an extra 1 dB over the cascade when we add 35 poles or more.

Notice the jumps in the joint and cascade model errors at 12 and 24 coefficients. This is because, as observed, this excerpt has 6 main temporal peaks in its residual, so 12 poles allow each to be modeled by a separate peak (pole-pair), and 24 poles allow a pair of peaks for each temporal feature. This highlights that, in general, the order of FDLP models will depend on the anticipated density of temporal transients within an analysis window.

To compare the results of the cascade and joint approaches, Fig. 7 displays their models. On the upper right, we plot the power spectrum of the original signal and on the left the squared Hilbert envelope of the first stage residual. Below each are the AR model envelopes (magnitudes from the z-domain) resulting from joint (above) and cascade (below) optimization. All models are 24th order, and the joint optimization employed ten iterations. Notice that the time-domain filter peaks for the joint model are sharper, and that the jointly estimated spectral envelope is smoother in the low frequencies while amplifying some high frequencies consistent with the sharper peaks in time.

V. Conclusion

As the dual of traditional time-domain AR modeling in which the resulting filter approximates the spectrum, we have presented frequency-domain LP modeling where the z polynomial provides an approximation of the temporal (Hilbert) envelope, incorporating all the desirable sharpness-preserving properties of all-pole modeling.

This technique has potential applications wherever temporal envelopes are of interest. In comparison to traditional rectify-and-smooth envelope extraction, the AR model has a well-defined optimal relation to the original signal. All-pole models also have a number of useful properties: Varying the length of the window over which envelopes are estimated and/or the order of the models used controls the tradeoff between local detail and average representation size, since AR models may distribute poles nonuniformly to minimize error. Unlike subsampling, a low-order AR model does not necessarily remove temporal detail, since each pole pair can result in an arbitrarily sharp peak in the envelope. Finally, there is a vast literature on representation and manipulation of AR models which can be brought to bear in this alternative domain.

The cascade model combines AR modeling of both spectral and temporal structure, and we have shown how to jointly optimize this structure to balance the signal modeling across domains. Insofar as traditional linear-predictive modeling has proven to be a versatile and popular model for spectral structure, we foresee many applications in signal analysis, manipulation, and compression using joint and cascade structures of the kind we have proposed. We also note that the approach extends very simply to modeling distinct subranges of the spectrum (subbands), to estimate separate temporal envelopes for these bands [23]. This is simply the dual of AR modeling for short-time spectral analysis of successive time frames.

Future work includes using ARMA models in the frequency domain. We believe that zeros on the Hilbert envelope are important, for instance to model regions in speech such as stop consonants, and we expect that the pure-real spectrum generated by the DCT-Io will be an effective domain in which to pursue such representations.

Appendix I
Proof of the DCT-Io Factorization

Let $\mathbf{T}$ be the matrix of cosines $\mathbf{T} = \cos(2\pi mn/M)$ for m, $n = 0, 1, \ldots, N-1$ as it appears in (9). Using (11) we write the $N \times N$ orthogonal DCT-Io matrix $\mathbf{C}$ as

$$\mathbf{C} = \frac{2}{\sqrt{M}}\mathbf{WTW}. \tag{28}$$

Substituting $\mathbf{T} = (\sqrt{M}/2)\mathbf{Z}^T(\mathbf{F} + \mathbf{F}^H)\mathbf{Z}$ we finally have

$$\begin{aligned}\mathbf{C} &= \mathbf{WZ}^T(\mathbf{F} + \mathbf{F}^H)\mathbf{ZW}\\ &= \mathbf{WZ}^T\mathbf{FSW}^{-1}\end{aligned} \tag{29}$$

since $\mathbf{FS}$ equals $(\mathbf{F} + \mathbf{F}^H)\mathbf{Z}$ except on the first column. The right multiplication by $\mathbf{W}^{-1}$ establishes the final equality.

Acknowledgment

The authors would like to thank Prof. H. Hermansky of IDIAP Research Institute for his valuable insights and hospitality during part of this work. They would also like to thank the anonymous reviewers for their great help in improving the paper.

References

[1] D. Talkin, "Speech formant trajectory estimation using dynamic programming with modulated transition costs," AT&T Bell Laboratories, Tech. Rep. 11222-870720-07TM, 1987.

[2] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, no. 4, pp. 85–111, Apr. 1990.

[3] J. Herre and J. D. Johnston, "Enhancing the performance of perceptual audio coders by using temporal noise shaping (TNS)," presented at the 101st AES Conv., Nov. 1996.

[4] N. S. Jayant and P. Noll, *Digital Coding of Waveforms: Principles and Applications to Speech and Video*. Englewood Cliffs, NJ: Prentice-Hall, 1984.

[5] J. Herre and J. D. Johnston, "Continuously signal-adaptive filterbank for high-quality perceptual audio coding," presented at the IEEE WASPAA, Oct. 1997.

[6] J. Herre and J. D. Johnston, "Exploiting both time and frequency structure in a system that uses an analysis/synthesis filterbank with high frequency resolution," presented at the 103rd AES Conv., Sep. 1997.

[7] J. Herre, "Temporal noise shaping, quantization and coding methods in perceptual audio coding: A tutorial introduction," presented at the AES 17th Conf. High Quality Audio Coding, Sep. 1999.

[8] R. Kumaresan, "An inverse signal approach to computing the envelope of a real valued signal," *IEEE Signal Process. Lett.*, vol. 5, no. 10, pp. 256–259, Oct. 1998.

[9] R. Kumaresan and A. Rao, "Unique positive FM-AM decomposition of signals," *Multidimen. Syst. Signal Process.*, vol. 9, no. 4, pp. 411–418, Oct. 1998.

[10] R. Kumaresan and A. Rao, "Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications," *J. Acoust. Soc. Amer.*, vol. 105, no. 3, pp. 1912–1924, Mar. 1999.

[11] R. Kumaresan, "On minimum/maximum/all-pass decompositions in time and frequency domains," *IEEE Trans. Signal Process.*, vol. 48, no. 10, pp. 2973–2976, Oct. 2000.

[12] R. Kumaresan and Y. Wang, "On the relationship between line-spectral frequencies and zero-crossings of signals," *IEEE Trans. Speech Audio Processing*, vol. 9, no. 4, pp. 458–461, May 2001.

[13] R. Kumaresan and Y. Wang, "On representing signals using only timing information," *J. Acoust. Soc. Am.*, vol. 110, no. 5, pp. 2421–2439, Nov. 2002.

[14] S. A. Martucci, "Symmetric convolution and the discrete sine and cosine transforms," *IEEE Trans. Signal Process.*, vol. 42, no. 5, pp. 1038–1051, May 1994.

[15] Z. Wang and B. R. Hunt, "The discrete W transform," *Appl. Math. Comput.*, vol. 16, no. 1, pp. 19–48, Jan. 1985.

[16] D. Gabor, "Theory of communication," *J. IEE*, vol. 93, pp. 429–457, 1946.

[17] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1999.

[18] L. S. Marple, "Computing the discrete-time 'analytic' signal via FFT," *IEEE Trans. Signal Process.*, vol. 47, no. 9, pp. 2600–2603, Sep. 1999.

[19] L. Cohen, *Time-Frequency Analysis*. Englewood Cliffs, NJ: Prentice-Hall, 1995.

[20] P. J. Davis, *Circulant Matrices*. New York: Wiley, 1979.

[21] H. A. Stone, "Convolution theorems for linear transforms," *IEEE Trans. Signal Process.*, vol. 46, no. 10, pp. 2819–2821, Oct. 1998.

[22] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, Apr. 1975.

[23] M. Athineos and D. Ellis, "Frequency-domain linear prediction for temporal features," in *Proc. IEEE ASRU Workshop*, Dec. 2003, pp. 261–266.

[24] M. Athineos and D. Ellis, "Sound texture modelling with linear prediction in both time and frequency domains," in *Proc. IEEE ICASSP*, Apr. 2003, pp. 648–651.

[25] E. Schuijers, W. Oomen, B. den Brinker, and J. Breebaart, "Advances in parametric coding for high-quality audio," presented at the 114st AES Conv., Mar. 2003.

[26] *Coding of Audio-Visual Objects. Part3: Audio, AMENDMENT 2: Parametric Coding of High Quality Audio*, ISO/IEC Int. Std. 14496-3:2001/Amd2:2004, ISO/IEC, Jul. 2004, ISO/IEC.

[27] M. Schroeder and B. Atal, "Code-excited linear prediction (CELP): High-quality speech at very low bit rates," in *Proc. IEEE ICASSP*, Apr. 1985, vol. 10, pp. 937–940.

Marios Athineos (S'03) received the B.S. degree in physics from the University of Patras, Patras, Greece, and the M.S. degree in electrical engineering from Columbia University, New York, where he is currently working toward the Ph.D. degree, conducting research on novel feature extraction methods for automatic recognition of speech and audio.

He is a member of the Laboratory for the Recognition and Organization of Speech and Audio (LabROSA), Electrical Engineering Department, Columbia University, New York.

Daniel P. W. Ellis (S'92–M'96–SM'04) received the Ph.D. degree in electrical engineering from The Massachussets Institute of Technology (MIT), Cambridge.

He was a Research Assistant at the Media Lab, MIT. He is currently an Associate Professor in the Electrical Engineering Department, Columbia University, New York, and an External Fellow of the International Computer Science Institute, Berkeley, CA. His Laboratory for Recognition and Organization of Speech and Audio (LabROSA) is concerned with all aspects of extracting high-level information from audio, including speech recognition, music description, and environmental sound processing. He also runs the AUDITORY e-mail list of 1700 worldwide researchers in perception and cognition of sound.