Tracking Vocal Tract Resonances Using a Quantized Nonlinear Function Embedded in a Temporal Constraint

Li Deng, *Fellow, IEEE*, Alex Acero, *Fellow, IEEE*, and Issam Bazzi

***Abstract*—This paper presents a new technique for high-accuracy tracking of vocal-tract resonances (which coincide with formants for nonnasalized vowels) in natural speech. The technique is based on a discretized nonlinear prediction function, which is embedded in a temporal constraint on the quantized input values over adjacent time frames as the prior knowledge for their temporal behavior. The nonlinear prediction is constructed, based on its analytical form derived in detail in this paper, as a parameter-free, discrete mapping function that approximates the "forward" relationship from the resonance frequencies and bandwidths to the Linear Predictive Coding (LPC) cepstra of real speech. Discretization of the function permits the "inversion" of the function via a search operation. We further introduce the nonlinear-prediction residual, characterized by a multivariate Gaussian vector with trainable mean vectors and covariance matrices, to account for the errors due to the functional approximation. We develop and describe an expectation–maximization (EM)-based algorithm for training the parameters of the residual, and a dynamic programming-based algorithm for resonance tracking. Details of the algorithm implementation for computation speedup are provided. Experimental results are presented which demonstrate the effectiveness of our new paradigm for tracking vocal-tract resonances. In particular, we show the effectiveness of training the prediction-residual parameters in obtaining high-accuracy resonance estimates, especially during consonantal closure.**

***Index Terms*—Continuity constraint, dynamic programming, expectation–maximization (EM) optimization, formant, greedy search, linear predictive coding (LPC) cepstrum, nonlinear prediction, prediction residual, quantization, vocal-tract resonance (VTR).**

I. Introduction

RESONANCE frequencies of the human vocal tract are of fundamental importance in speech production and perception [12], [24]. They are the natural frequencies, or eigenfrequencies, of the air path in the vocal tract from glottis to lips, and the air path is shaped principally by the tongue, jaw, and other articulators. Since such vocal tract resonances (VTRs) are defined as characteristics of a physical system, they are required to exist at some frequency values at all times, even if the mouth is closed (and during any other consonantal constriction or closure) with weak or no measurable emitting acoustic signals. VTR frequencies are constrained to change continuously over time since they are a smooth function of the vocal tract's airway shape, which is determined by the continuously moving articulators. As such, VTRs do not disappear, split, or merge during any portion of a speech utterance, regardless of the acoustic evidence which may suggest otherwise.

The VTRs defined above largely in articulatory terms are sometimes called formants in the speech synthesis and analysis literature [2], [15], [19], [25]. However, the predominant use of the term "formant" in the literature is based on the acoustic definition: Formants are associated with peaks or prominences in the smoothed power spectrum of the acoustic signal of speech. With this acoustic definition, formants would "disappear" during complete consonantal closure, and may "split" or "merge" under other conditions when the peaks in the acoustic spectrum become ambiguous [20], [21]. One extreme example of equating formants with acoustic spectral peaks in the literature is the use of a mixture of Gaussians to fit the spectrum in order to identify the formant frequencies as the Gaussian means [30]. Most of the auditory modeling studies on formant extraction are also based on the similar concept of formants as correlates of spectral prominences of speech acoustics instead of the underlying vocal tract resonances [5], [9]. For the vocalic sounds with no narrow local constrictions in the vocal tract, the articulatorily defined "formants" (i.e., VTRs) and the acoustically defined formants coincide with each other; indeed, the natural frequencies of the vocal tract are evidenced by the energy concentration in the acoustic power spectrum for such sounds. However, during typical consonant production where narrow constriction or full closure in the vocal tract is made, VTRs often deviate from the spectral peaks. In this case, the VTRs are frequently unobservable due to pole-zero cancellation, where sometimes simply no acoustic signal is emitted to the air during the closure.

This paper addresses the issue of automatic tracking of continuous VTRs in natural speech utterances. In particular, we note that estimation of VTRs during consonantal closure has not been adequately dealt with in the existing literature on formant tracking, and this will hence be the major focus of the research presented in this paper. Our approach is to use a temporal smoothness constraint, together with the local quantized VTR matching function, to obtain an optimal track of the VTR sequence for all segments of speech utterances including all types of consonantal constriction and closure. The local VTR matching function is established via a novel analytical nonlinear predictor characterizing the relationship between the VTR variables (resonance frequencies and bandwidths) and the linear

Manuscript received March 19, 2004; revised September 11, 2004. A preliminary version of this work was presented at the Eurospeech Conference in September 2003, Geneva, Switzerland. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Kuldip K. Paliwal.

The authors are with the Microsoft Research, Redmond, WA 98052 USA (e-mail: deng@microsoft.com; alexac@microsoft.com).

Digital Object Identifier 10.1109/TSA.2005.855841

predictive coding (LPC) cepstral coefficients. To implement optimal VTR tracking algorithms, we quantize the VTR space, and search for all possible (quantized) VTR candidates in order to locally match the observed acoustic observation represented in the same LPC cepstral space.

The estimation technique for the VTR variables via the quantized nonlinear function presented in this paper differs from previous published work on formant estimation in several ways. Most formant estimation methods rely on some type of LPC spectral analysis, followed by peak picking or root finding [1], [21], [27], which may then be further combined with temporal continuity constraints. More recent work exploits dynamic programming along the frequency axis to optimize the locations of these formant candidates [28]. In contrast, the new technique presented in this paper explores all possible candidates, via the full-space search of all frequency values (subject only to the quantization error), for fitting the measured acoustic data to both the nonlinear prediction function and the temporal constraint simultaneously. No explicit spectral analysis and error-prone peak picking or root finding are needed.

The aspect of quantizing the VTR space in our technique is similar to the hidden Markov model (HMM) state-space construction in an earlier HMM approach to formant tracking [20]. That HMM approach uses state-dependent discrete distributions on the vector-quantized LPC spectra to map from the formants to the LPC spectra, and the distributions require extensive training with labeled data. In contrast, our new technique is based on direct, analytical mapping from the VTR variables to LPC cepstra, and no training data or process is required. (In our technique, the use of LPC cepstra is advantageous over LPC spectra because the former provides a much simpler form of the analytical function.) Further, the earlier HMM approach requires a special "null" state to represent "missing" formants during consonantal closure. Our new approach, however, is able to track continuous time-varying VTR variables including those during the closure, eliminating the "missing formant" problem entirely. The ability to produce VTRs for each time frame without any missing values is important for applications to speech recognition based on standard statistical pattern-matching approaches that require consistent front-end features over time and over the feature dimension. The concept of providing continuity constraints across difficult speech regions to avoid the "missing formant" problem has been explored in other related work (e.g., [17] and [29]). The new approach presented in this paper, however, is the only one that integrates this concept with the use of a quantized analytical function that enables an efficient search over the full VTR space. A number of successful signal processing techniques, including parameter quantization, expectation–maximization (EM)-based parameter training, constrained search, and dynamic programming, are embedded within a unified discretized dynamic-system framework for estimating VTR sequences without explicit hand-labeling of training data. This integrated approach gives our method its novelty.

The organization of this paper is as follows. In Section II, we derive and present an approximate nonlinear mapping or prediction function from the VTR variables to the speech acoustics represented in terms of LPC cepstra. Graphical illustrations of the components of the function are provided. We also describe details of a quantization scheme on the VTR variables as input to the function, permitting direct inversion of this function to give a crude VTR estimate. In Section III, we introduce the nonlinear-prediction residual, characterized by a multivariate Gaussian vector with the trainable mean vector and covariance matrix, to account for errors due to the functional approximation. We further introduce a simplified constraint on the quantized VTR values over adjacent time frames as the prior knowledge for the VTRs temporal behavior. A combination of the constraint and the mapping function with trainable residuals constitutes a dynamic system model of speech. Then, in Sections IV and V, respectively, we present algorithms for EM training of the residual's parameters and for dynamic programming based VTR tracking. In Section VI, some implementation details for computation speedup and experimental results are presented. The experimental results demonstrate the effectiveness of our new paradigm for VTR training and of the related algorithms with their efficient implementation.

II. Nonlinear Function From VTR to LPC Cepstrum and Its Quantization

A. Derivation of the Nonlinear Function

Denote the VTR vector, which consists of a set of K resonant frequencies $\mathbf{f}$ and bandwidths $\mathbf{b}$, as

$$\mathbf{x} = \begin{pmatrix} \mathbf{f} \\ \mathbf{b} \end{pmatrix}$$

where

$$\mathbf{f} = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_K \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_K \end{pmatrix}.$$

As a basis for local, frame-level VTR estimates based on the match between the observed and predicted speech acoustics, we in this section present an approximate, vector-valued, nonlinear mapping function

$$\mathbf{o} \approx \mathbf{C}(\mathbf{x})$$

from the VTR vector, $\mathbf{x}$, to the observed speech acoustics, $\mathbf{o}$. Depending on the type of the acoustic measurements as the output in the mapping function, closed-form computation for $\mathbf{C}(\mathbf{x})$ may be impossible, or its in-line computation may be too expensive. To overcome these difficulties, we quantize each dimension of $\mathbf{x}$ over a range of frequencies or bandwidths, and then compute $\mathbf{C}(\mathbf{x})$ for every quantized value of $\mathbf{x}$. In the earlier work reported in [3], we described a procedure for constructing a function $\mathbf{C}(\mathbf{x})$ when the output acoustic measurements are the Mel-frequency cepstral coefficients (MFCCs). This procedure incurred very high costs of computation and memory due to the interaction of several resonances in determining the output MFCC values and also due to a lack of closed-form expressions of the nonlinear function. In this section of the paper, we present a new scheme where the closed form of the function can be easily derived when the output of the nonlinear function becomes LPC cepstra instead of MFCCs. The use of the new, analytical function from VTRs to LPC cepstra offers a significant

advantage of computation efficiency due to the decomposition property which we derive and describe below.

Consider an all-pole model of speech, with each of its poles represented as a frequency-bandwidth pair (f_k, b_k). Then the corresponding complex root is given by [1]

$$z_k = e^{-\pi \frac{b_k}{f_s} + j2\pi \frac{f_k}{f_s}} \quad z_k^* = e^{-\pi \frac{b_k}{f_s} - j2\pi \frac{f_k}{f_s}} \tag{1}$$

where f_s is the sampling frequency. The transfer function with K poles and a gain of G is

$$H(z) = G \prod_{k=1}^{K} \frac{1}{(1 - z_k z^{-1})(1 - z_k^* z^{-1})}. \tag{2}$$

Taking logarithm on both sides of (2), we obtain

$$\log H(z) = \log G - \sum_{k=1}^{K} \log(1 - z_k z^{-1}) - \sum_{k=1}^{K} \log(1 - z_k^* z^{-1}). \tag{3}$$

Now using the well-known infinite series expansion formula

$$\log(1 - v) = -\sum_{n=1}^{\infty} \frac{v^n}{n}, \qquad |v| \le 1$$

and with $v = z_k z^{-1}$, we obtain

$$\begin{aligned} \log H(z) &= \log G + \sum_{k=1}^{K} \sum_{n=1}^{\infty} \frac{z_k^n z^{-n}}{n} + \sum_{k=1}^{K} \sum_{n=1}^{\infty} \frac{z_k^{*n} z^{-n}}{n} \\ &= \log G + \sum_{n=1}^{\infty} \left[\sum_{k=1}^{K} \frac{z_k^n + z_k^{*n}}{n} \right] z^{-n}. \end{aligned} \tag{4}$$

Comparing (4) with the definition of the one-sided z-transform

$$C(z) = \sum_{n=0}^{\infty} c_n z^{-n} = c_0 + \sum_{n=1}^{\infty} c_n z^{-n}$$

we immediately see that the inverse z-transform of $\log H(z)$ in (4), which by definition is the LPC cepstrum, is

$$c_n = \sum_{k=1}^{K} \frac{z_k^n + z_k^{*n}}{n}, \qquad n > 0 \tag{5}$$

and $c_0 = \log G$.

Using (1) to expand and simplify (5), we obtain the final form of the nonlinear function (for $n > 0$)

$$\begin{aligned} c_n &= \frac{1}{n} \sum_{k=1}^{K} \left[e^{-\pi n \frac{b_k}{f_s} + j2\pi n \frac{f_k}{f_s}} + e^{-\pi n \frac{b_k}{f_s} - j2\pi n \frac{f_k}{f_s}} \right] \\ &= \frac{1}{n} \sum_{k=1}^{K} e^{-\pi n \frac{b_k}{f_s}} \left[e^{j2\pi n \frac{f_k}{f_s}} + e^{-j2\pi n \frac{f_k}{f_s}} \right] \\ &= \frac{1}{n} \sum_{k=1}^{K} e^{-\pi n \frac{b_k}{f_s}} \left[\cos\left(2\pi n \frac{f_k}{f_s}\right) + j \sin\left(2\pi n \frac{f_k}{f_s}\right) \right. \\ &\quad \left. + \cos\left(2\pi n \frac{f_k}{f_s}\right) - j \sin\left(2\pi n \frac{f_k}{f_s}\right) \right] \\ &= \frac{2}{n} \sum_{k=1}^{K} e^{-\pi n \frac{b_k}{f_s}} \cos\left(2\pi n \frac{f_k}{f_s}\right). \end{aligned} \tag{6}$$

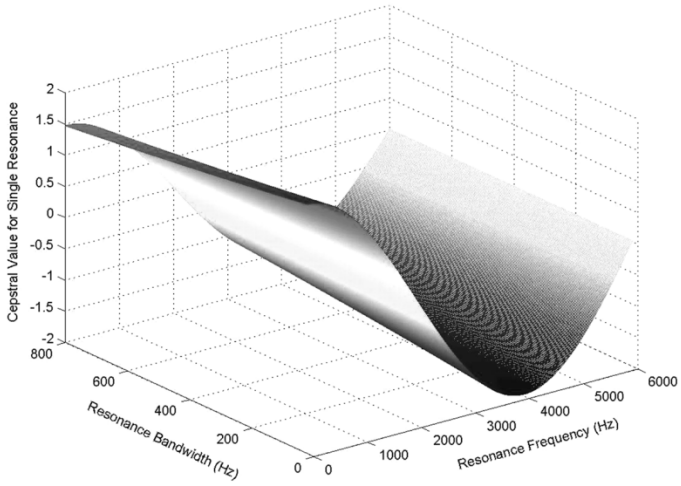

Fig. 1. First-order cepstral value of a one-pole (single-resonance) filter as a function of the resonance frequency and bandwidth. This plots the value of one term in (6) versus f_k and b_k with fixed $n = 1$ and $f_s = 8000$ Hz.

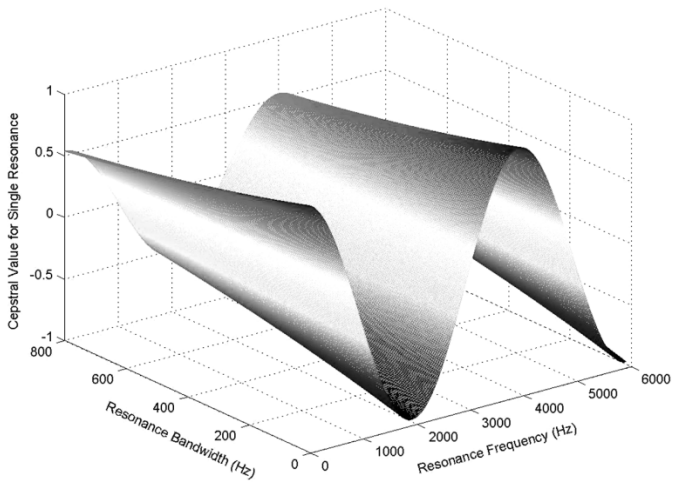

Fig. 2. Second-order cepstral value of a one-pole (single-resonance) filter as a function of the resonance frequency and bandwidth ($n = 1$ and $f_s = 8000$ Hz).

B. Discussion and Illustration

Equation (6) gives the decomposition property of the LPC cepstrum—each of the LPC cepstral coefficients is a sum of the contributions from separate resonances without interacting with each other. This contrasts the use of the MFCC as the output of the nonlinear function studied in [3], which is a function of all interacting VTRs, instead of in a simple additive form as in (6). (Also, the nonlinear function in [3] can only be determined in a tabulated form using a numerical procedure rather than in an analytical form developed here.) The key advantage of the decomposition property is that it makes the optimization procedure highly efficient for inverting the nonlinear function from the acoustic measurement to the VTR which we will detail in Section VI-A.

As an illustration, in Figs. 1–3, we plot the value of one term, $e^{-\pi n(b/f_s)} \cos(2\pi n(f/f_s))$, in (6) as a function of the resonance frequency f and bandwidth b, for the first-order $(n = 1)$, second-order $(n = 2)$, and the fifth-order $(n = 5)$ cepstrum, respectively. (The sampling frequency $f_s = 8000$ Hz is used in all the plots.)

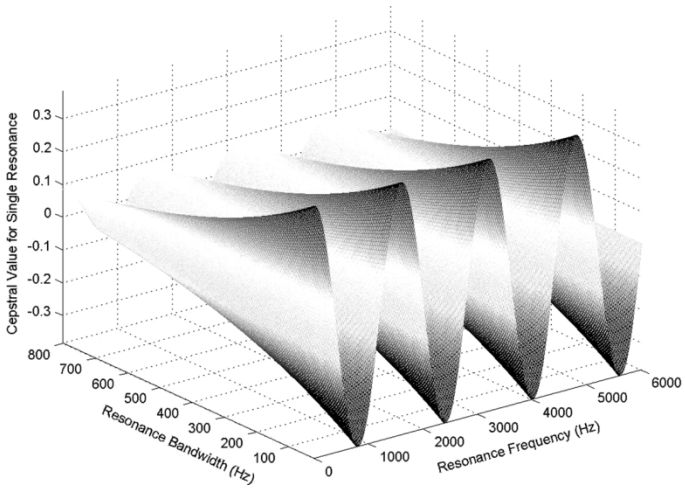

Fig. 3. Fifth-order cepstral value of a one-pole (single-resonance) filter as a function of the resonance frequency and bandwidth $n = 5$ and $f_s = 8000$ Hz.

These are the cepstra corresponding to the transfer function of a single-resonance (i.e., one pole with no zeros) linear system. Due to the decomposition property of the LPC cepstrum, for multiple-resonance systems, the corresponding cepstrum is simply a sum of those for the single-resonance systems.

Examining Figs. 1–3, we easily observe some key properties of the (single-resonance) cepstrum. First, the mapping function from the VTR frequency and bandwidth variables to the cepstrum, while nonlinear, is well behaved. That is, the relationship is smooth, and there is no sharp discontinuity. Second, for a fixed resonance bandwidth, the frequency of the sinusoidal relation between the cepstrum and the resonance frequency increases as the cepstral order increases. The implication is that when piecewise linear functions are to be used to approximate the nonlinear function of (6), more "pieces" will be needed for the higher order than for the lower order cepstra. Third, for a fixed resonance frequency, the dependence of the low-order cepstral values on the resonance bandwidth is relatively weak. The cause of this weak dependence is the low ratio of the bandwidth (up to 800 Hz) to the sampling frequency (e.g., 16 000 Hz) in the exponent of the cepstral expression in (6). For example, as shown in Fig. 1 for the first-order cepstrum, the extreme values of bandwidths from 20–800 Hz reduce the peak cepstral values only from 1.9844 to 1.4608 (computed by $2\exp(-20\pi/8000)$ and $2\exp(-800\pi/8000)$, respectively). The corresponding reduction for the second-order cepstrum is from 0.9844 to 0.5335 (computed by $\exp(-2\times 20\pi/8000)$ and $\exp(-2\times 800\pi/8000)$, respectively). In general, the exponential decay of the cepstral value, as the resonance bandwidth increases, becomes only slightly more rapid for the higher order than for the lower order cepstra (see Fig. 3). This weak dependence is desirable since the VTR bandwidths are known to be highly variable with respect to the acoustic environment [18], and to be less correlated with the phonetic content of speech and with human speech perception than the VTR frequencies [12], [24].

In the earlier literature, a special case of (6), with the expression for only a single resonance (i.e., $K = 1$), was shown in [4] (with no derivation). However, the tentative conclusion made in [4] that the resonance frequency can be approximated by a weighted sum of LPC cepstra seems questionable. Based on (6) and its components' illustration in Figs. 1–3, the inversion from LPC cepstra to resonance frequencies is clearly a nonlinear process. The empirical plots [4, Fig. 1] which show a gross linear trend between linearly combined cepstra and measured formants may be somewhat misleading in two ways. First, the plots were made over a wide frequency range, e.g., F2 or f_2 is ranged from 700–2600 Hz. If a smaller frequency range were to be examined, the deviation from linear regression, which can be as large as 600 Hz for F2 (the middle plot of Fig. 1 in [4]), would give a rather poor linear fit. For example, Fig. 1(b) of [4] used the entire frequency range from 700–2600 Hz for the F2 variation, which gave an artificially high correlation coefficient of 0.971. If the range of the F2 variation is limited to only that for each speech sound in context (e.g., 1800–2400 Hz for sound/i/), then the correlation coefficient would be much lower, likely to be of a similar value to what the (mildly) nonlinear relationship of (6) can predict. Second, the weights of combining LPC cepstra reported in [4] were trained using multiple linear regression with the same data as given in the plots. It is not clear whether the gross linear trend over a wide frequency range can still hold for different data sets.

C. Quantization of VTR Variables as Input to the Nonlinear Function

Due to the definitive nonlinearity, analytically derived as (6) and graphically shown in Figs. 1–3, accurate inversion from LPC cepstra to VTR variables (resonance frequencies and bandwidths) requires nonlinear techniques. In our current work, we adopt a novel quantization and search technique to accomplish the purpose of nonlinear function inversion.

In our implementation and experimental work, we choose to use four poles in the LPC model of speech [i.e., using $K = 4$ in (6)], since these lowest VTRs carry the most important phonetic information of the speech signal. That is, an eight-dimensional vector $\mathbf{x} = (f_1, f_2, f_3, f_4, b_1, b_2, b_3, b_4)$ is used as the input to the nonlinear function $\mathbf{C}(\mathbf{x})$. For the output of the nonlinear function, up to 15 orders of LPC cepstra are used. The zeroth order cepstrum, c_0, is excluded from the output vector, making the nonlinear mapping from VTRs to cepstra independent of the energy level in the speech signal. This corresponds to setting the gain $G = 1$ in the all-pole model of (2). (In the earlier work of [3], a six-dimensional input vector of $(f_1, f_2, f_3, b_1, b_2, b_3)$ and a 12-dimensional output vector of MFCCs were used. We found that using the additional VTR improves the overall VTR tracking accuracy, and using the LPC cepstra instead of MFCCs improves computational efficiency.)

For each of the eight dimensions in the VTR vector, we use scalar quantization in our current VTR tracker implementation. Since $\mathbf{C}(\mathbf{x})$ is relevant to all possible phones in speech, we select the appropriate range for each VTR frequency and its corresponding bandwidth to cover all phones according to the considerations discussed in [2] and [7]. Table I lists the range, from minimal to maximal frequencies in hertz, for each of the four VTR frequencies and bandwidths. It also lists the corresponding number of quantization levels used. Bandwidths are quantized uniformly with five levels while frequencies are

TABLE I
QUANTIZATION SCHEME FOR THE VTR VARIABLES, INCLUDING THE RANGES OF THE FOUR VTR FREQUENCIES AND BANDWIDTHS AND THE CORRESPONDING NUMBERS OF QUANTIZATION LEVELS

	Minimum (Hz)	Maximum (Hz)	No. of Quantization
f_1	200	900	20
f_2	600	2800	20
f_3	1400	3800	20
f_4	1700	5000	20
b_1	40	300	5
b_2	60	300	5
b_3	60	500	5
b_4	100	700	5

mapped to the Mel-frequency scale and then uniformly quantized with 20 levels. The total number of quantization levels shown in Table I yields a total of 100 million $(20^4 \times 5^4)$ entries for $\mathbf{C}(\mathbf{x})$, but because of the constraint $f_1 < f_2 < f_3 < f_4$, the resulting number is reduced by about 25%. This is still a very large space to do exhaustive search in order to find the optimal fit to the observed acoustics and to accomplish the optimal function inversion. To overcome this difficulty, in Section VI-A, we will discuss a greedy search technique we have implemented exploiting the decomposition property of the nonlinear function.

Quantization of the VTR variables discussed above gives rise to the discrete nature in the "codebook" construction for the nonlinear function $\mathbf{C}(\mathbf{x})$ expressed in (6). In this work, we use a precomputed, tabulated form to represent this nonlinear function. As a notation, we denote the VTR vector value of $\mathbf{x}$ at the ith level of quantization as $\mathbf{x}[i]$. When the VTR vector occurs at time frame t, we denote the value of $\mathbf{x}_t$ at the ith level of quantization as $\mathbf{x}_t[i]$, or simply $i_t = i$. That is, the index to the codebook, i, is used interchangeably with the value stored at that index, $\mathbf{x}_t[i]$. When the index is used alone, it is intended to represent the value stored at that index.

III. NONLINEAR-PREDICTION RESIDUAL AND TEMPORAL CONSTRAINT

In practical implementation of the VTR tracker, computation of the LPC cepstrum from VTRs according to (6) can necessarily include only a finite number of poles. As mentioned above, in our implementation and experiments, we chose to use $K = 4$. The remaining (higher order) poles and possible zeros (as well as their possible interactions with poles) in actual speech are known to affect acoustics and to create prediction errors using the nonlinear mapping function of (6) based on the all-pole filter model of speech with low orders. One way to improve the mapping function is to introduce the trainable prediction residuals in order to (blindly) compensate for all sources of errors.

Let us denote the residual from the nonlinear prediction function of (6) by $\mathbf{v}$, which we assume is a Gaussian random vector with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{D}$. That is

$$\mathbf{v} \sim \mathcal{N}(\mathbf{v}; \boldsymbol{\mu}, \mathbf{D}).$$

After accounting for the approximation error by the IID residual $\mathbf{v}_t$ for each time frame t, the exact relationship between the VTR vector $\mathbf{x}_t$ and the LPC cepstral vector $\mathbf{o}_t$ now becomes

$$\mathbf{o}_t = \mathbf{C}(\mathbf{x}_t[i]) + \mathbf{v}_t$$

or

$$\mathbf{o}_t = \mathbf{C}(\mathbf{x}_t[i]) + \boldsymbol{\mu} + \mathbf{v}'_t \tag{7}$$

where $\mathbf{v}'_t$ is a zero-mean Gaussian random vector: $\mathbf{v}' \sim \mathcal{N}(\mathbf{v}'; \mathbf{0}, \mathbf{D})$. This forms the *observation equation* of a dynamic system model with the state-space formulation [7].

We can further improve the prediction from the VTR *sequence* to the LPC cepstrum *sequence* by exploiting the prior knowledge about the VTRs temporal behavior. This prior knowledge is expressed as the temporal smoothness constraint using the following discretized *state equation* in the dynamic system model

$$\mathbf{x}_t[i] = \mathbf{x}_{t-1}[j] + \mathbf{w}_t \tag{8}$$

where the state noise at frame t is assumed to be an IID, zero-mean Gaussian random variable

$$\mathbf{w}_t \sim \mathcal{N}(\mathbf{w}_t; \mathbf{0}, \mathbf{B})$$

with the covariance matrix of $\mathbf{B}$.

IV. ALGORITHM FOR ESTIMATING RESIDUAL PARAMETERS

A. Reestimation Formulae

Following the general spirit of the EM algorithm [6], we have derived reestimation formulae (M-step) for the parameters in the state-space model consisting of (7) and (8). In particular, the mean vector in the nonlinear prediction residual of (7) is reestimated in each EM iteration by

$$\hat{\boldsymbol{\mu}} = \frac{\sum_{t=1}^{N} \sum_{i=1}^{I} \gamma_t(i) \{\mathbf{o}_t - \mathbf{C}(\mathbf{x}_t[i])\}}{N} \tag{9}$$

where N is the total number of frames in the training data, I is the total number of quantization levels for the VTRs, and the posterior

$$\gamma_t(i) \equiv p\left(\mathbf{x}_t[i] \,\middle|\, \mathbf{o}_1^N\right)$$

is computed efficiently using a novel forward-backward recursion (E-step) described below.

The reestimation formula for the covariance matrix $\mathbf{D}$ in the prediction residual is

$$\hat{\mathbf{D}} = \frac{\sum_{t=1}^{N} \sum_{i=1}^{I} \gamma_t(i) [\mathbf{o}_t - \mathbf{C}(\mathbf{x}_t[i]) - \hat{\boldsymbol{\mu}}][\mathbf{o}_t - \mathbf{C}(\mathbf{x}_t[i]) - \hat{\boldsymbol{\mu}}]^*}{N}. \tag{10}$$

B. Forward–Backward Recursion for Computing Posterior Probability

We now derive and describe a novel efficient algorithm we have developed, based on the quantized state-space model of (7)

and (8), for computing the posterior probability, $\gamma_t(i)$, required in the reestimation formulae in (9) and (10).

First, define the forward probability of

$$\alpha(i_t) \equiv p\left(\mathbf{o}_1^t, i_t\right).$$

The forward recursive formula is established using the following derivation:

$$\begin{aligned}\alpha(i_{t+1}) &\equiv p\left(\mathbf{o}_1^{t+1}, i_{t+1}\right) = \sum_{i_t} p\left(\mathbf{o}_1^t, \mathbf{o}_{t+1}, i_{t+1}, i_t\right) \\ &= \sum_{i_t} p\left(\mathbf{o}_{t+1}, i_{t+1} | \mathbf{o}_1^t, i_t\right) p\left(\mathbf{o}_1^t, i_t\right) \\ &= \sum_{i_t} p(\mathbf{o}_{t+1}, i_{t+1} \,|\, i_t)\alpha(i_t) \\ &= \sum_{i_t} p(\mathbf{o}_{t+1} | i_{t+1}, i_t) p(i_{t+1} \,|\, i_t)\alpha(i_t) \\ &= \sum_{i_t} p(\mathbf{o}_{t+1} \,|\, i_{t+1}) p(i_{t+1} \,|\, i_t)\alpha(i_t). \end{aligned} \tag{11}$$

In (11), $p(\mathbf{o}_{t+1} \,|\, i_{t+1})$ is determined by the observation equation according to

$$p(\mathbf{o}_{t+1} \,|\, i_{t+1} = i) = \mathcal{N}(\mathbf{o}_{t+1}; \mathbf{C}(\mathbf{x}_{t+1}[i]) + \hat{\boldsymbol{\mu}}_0, \hat{\mathbf{D}}_0)$$

where $\hat{\boldsymbol{\mu}}_0$ and $\hat{\mathbf{D}}_0$ were estimated from the previous EM iteration. And $p(i_{t+1} \,|\, i_t)$ is determined by the state equation according to

$$p(i_{t+1} = i \,|\, i_t = j) = \mathcal{N}(\mathbf{x}_{t+1}[i]; \mathbf{x}_t[j], \hat{\mathbf{B}}_0). \tag{12}$$

Next, we establish the following backward recursion to efficiently compute the posterior probability:

$$\begin{aligned}\gamma(i_t) &\equiv p\left(i_t \,\middle|\, \mathbf{o}_1^N\right) = \sum_{i_{t+1}} p\left(i_t, i_{t+1} \,|\, \mathbf{o}_1^N\right) \\ &= \sum_{i_{t+1}} p\left(i_t \,|\, i_{t+1}, \mathbf{o}_1^N\right) p\left(i_{t+1} \,\middle|\, \mathbf{o}_1^N\right) \\ &= \sum_{i_{t+1}} \frac{p\left(i_t, i_{t+1}, \mathbf{o}_1^t\right)}{p\left(i_{t+1}, \mathbf{o}_1^t\right)} \gamma(i_{t+1}) \\ &= \sum_{i_{t+1}} \frac{p\left(i_t, i_{t+1}, \mathbf{o}_1^t\right)}{\sum_{i_t} p\left(i_t, i_{t+1}, \mathbf{o}_1^t\right)} \gamma(i_{t+1}) \\ &= \sum_{i_{t+1}} \frac{p\left(i_t, \mathbf{o}_1^t\right) p\left(i_{t+1} \,|\, i_t, \mathbf{o}_1^t\right)}{\sum_{i_t} p\left(i_t, \mathbf{o}_1^t\right) p\left(i_{t+1} \,|\, i_t, \mathbf{o}_1^t\right)} \gamma(i_{t+1}) \\ &= \sum_{i_{t+1}} \frac{\alpha(i_t) p(i_{t+1} \,|\, i_t)}{\sum_{i_t} \alpha(i_t) p(i_{t+1} \,|\, i_t)} \gamma(i_{t+1}) \end{aligned} \tag{13}$$

where $\alpha(i_t)$ and $p(i_{t+1} \,|\, i_t)$ on the right-hand side of the above have been computed already in the forward recursion. Initialization for the above γ recursion is $\gamma(i_N) = \alpha(i_N)$.

V. Algorithm for VTR Tracking

After the parameters of the quantized state-space model are trained using a speech utterance as just described, the model can be used for optimal VTR sequence estimation (i.e., tracking) for the same speech utterance. The dynamic programming algorithm described in this section is aimed to find the best single quantized VTR sequence $i_1^N = (i_1, i_2, \ldots, i_N)$ for a given acoustic observation sequence $\mathbf{o}_1^N$.

Let us define the optimal partial score of

$$\begin{aligned}\delta_t(i) &= \max_{i_1^{t-1}} p\left(\mathbf{o}_1^t, i_1^{t-1}, x_t[i]\right) = \max_{i_1^{t-1}} p\left(\mathbf{o}_1^t, i_1^{t-1}, i_t = i\right) \\ &= \max_{i_1^{t-1}} \left[p\left(\mathbf{o}_1^t \,|\, i_1^{t-1}, i_t = i\right) p(i_1^{t-1}, i_t = i)\right] \\ &= \max_{i_1^{t-1}} \left\{ \left[\prod_{\tau=1}^{t-1} p(\mathbf{o}_\tau \,|\, i_\tau) p(\mathbf{o}_t \,|\, i_t = i)\right] \right. \\ &\qquad \left. \times \left[p(i_1) \prod_{\tau=2}^{t-1} p(i_\tau \,|\, i_{\tau-1}) p(i_t = i \,|\, i_{t-1})\right] \right\}. \end{aligned} \tag{14}$$

Each $\delta_t(i)$ defined in (14) is associated with a node in the trellis diagram. Each increment of time corresponds to reaching a new stage in dynamic programming. At the final stage $t = N$, we have the objective function of $\delta_N(i)$, which is accomplished via all the previous stages of computation for $t \leq N - 1$. Based on the optimality principle, the optimal partial likelihood at the processing stage of $t + 1$ can be computed using the following dynamic programming (Viterbi) recursion:

$$\delta_{t+1}(i) = \max_j \delta_t(j) p(i_{t+1} = i \,|\, i_t = j) p(\mathbf{o}_{t+1} \,|\, i_{t+1} = i). \tag{15}$$

In the above, the "transition probability" is computed by

$$p(i_{t+1} = i \,|\, i_t = j) = \mathcal{N}(\mathbf{x}_{t+1}[i]; \mathbf{x}_t[j], \mathbf{B}) \tag{16}$$

and the "observation probability" is computed by

$$p(\mathbf{o}_{t+1} \,|\, i_{t+1} = i) = \mathcal{N}(\mathbf{o}_{t+1}; \mathbf{C}(\mathbf{x}_{t+1}[i]) + \boldsymbol{\mu}, \mathbf{D}). \tag{17}$$

Back tracing of the optimal VTR quantization index j in (15) gives the optimally estimated VTR sequence in terms of the quantized values at each time frame.

VI. VTR Tracking Experiments

In this section, we first present some implementation detail of the algorithms for residual parameter training and for VTR tracking presented so far, aimed to overcome computational difficulties associated with the large quantization space on the VTR variables. We then present experimental results demonstrating the effectiveness of the algorithms in accurate tracking of VTRs for conversational speech. In particular, we demonstrate the crucial role of the use of trainable residual parameters in enhancing the VTR tracking accuracy.

A. Computation Speedup

One principal implementation difficulty for the training and tracking algorithms presented in the preceding sections is the high computational cost in summing and in searching over the huge space in the quantized VTR variables. The sum is required in the reestimation formulae of (9)–(11) and (13), and the search

is required in dynamic programming recursion of (15). To overcome this difficulty, we have successfully developed and implemented a suboptimal, greedy technique capitalizing on the decomposition property of the nonlinear mapping function presented in Section II. We describe this technique now.

Let us consider an objective function F, to be optimized with respect to M noninteracting variables that determine the function's value. An example is the following decomposable function consisting of M terms $F_m, m = 1, 2, \ldots, M$, each of which contains independent variables (α_m) to be searched for

$$F = \sum_{m=1}^{M} F_m(\alpha_m).$$

The greedy optimization technique proceeds as follows. First, initialize $\alpha_m, m = 1, 2, \ldots, M$ to reasonable values. Then, fix all $\alpha_m's$ except one, say α_n, and optimize α_n with respect to the new objective function of

$$F - \sum_{m=1}^{n-1} F_m(\alpha_m) - \sum_{m=n+1}^{M} F_m(\alpha_m).$$

Next, after the low-dimensional, inexpensive search problem for $\hat{\alpha_n}$ is solved, fix it and optimize a new $\alpha_m, m \neq n$. Repeat this for all $\alpha_m's$. Finally, iterate the above process until all optimized $\alpha_m's$ become stabilized.

In our implementation of this technique for VTR tracking and parameter estimation, each of the K resonances is treated as a separate, noninteractive variable to optimize. We found that only two to three overall iterations above are already sufficient to stabilize the parameter estimates.[1] Further, surprisingly, in our experiments we found that initialization of all VTR variables to zero gives virtually the same estimates as those that are obtained by more carefully thought-out initialization schemes.

With the use of the above greedy, suboptimal technique instead of full optimal search, the computation cost of VTR tracking is reduced by over 4000-fold compared with the brute-force implementation of the algorithms presented in the preceding sections. As a result, the VTR tracker as currently implemented in Matlab runs close to real time on a P-III machine.

B. Experimental Results and Analysis on VTR Tracking

The above greedy technique has been incorporated into the VTR tracking algorithm and into the EM training algorithm for the nonlinear-prediction residual parameters. We have compared the results obtained by the rigorous algorithms with those obtained by the computationally intensive, brute-force search. No qualitative differences in the results are found, and hence all the results described in this section have been obtained with the use of the greedy technique for optimization.

We first demonstrate the effectiveness of the EM training. Note that the training does not require any data labeling and is fully unsupervised. Fig. 4 shows the VTR tracking (f_1, f_2, f_3, f_4) results, superimposed on the spectrogram of a telephone speech utterance (excised from Switchboard database [14]) of "*the way you dress*" by a male speaker, when the residual mean vector $\boldsymbol{\mu}$ in (7) is set to zero and the covariance matrix $\mathbf{D}$ is set to be diagonal with empirically determined diagonal values.[2] Setting $\boldsymbol{\mu}$ to zero corresponds to the assumption that the nonlinear function of (6) is an unbiased predictor of the real speech data in the form of LPC cepstra. Under this assumption we observe from Fig. 4 that while f_1 and f_2 are accurately tracked through the entire utterance, f_3 and f_4 are incorrectly tracked during the later half of the utterance.[3] One iteration of the EM training on the residual mean vector and covariance matrix does not correct the errors (see Fig. 5), but two iterations are able to correct the errors in the utterance for

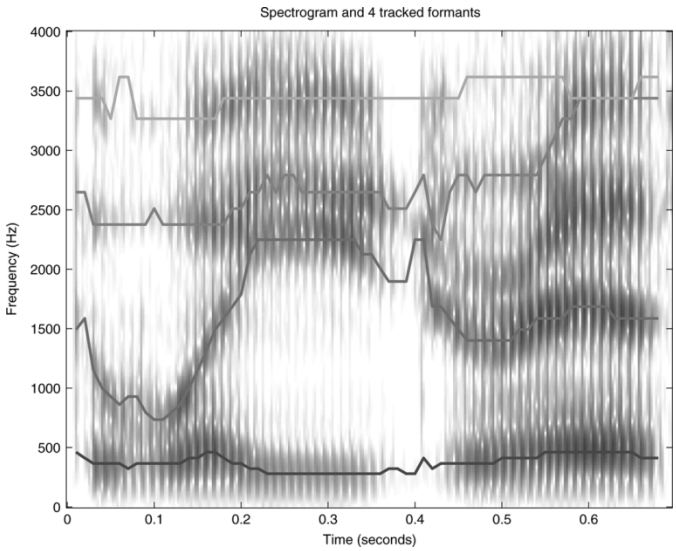

Fig. 4. VTR tracking by setting the residual mean vector to zero.

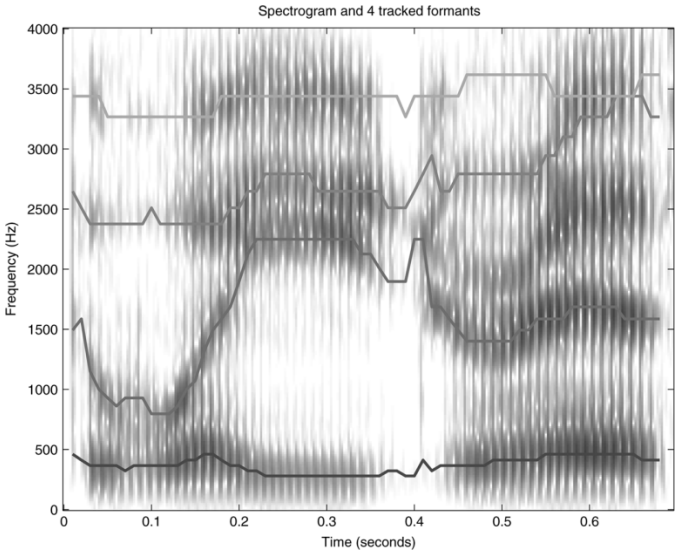

Fig. 5. VTR tracking with one iteration of residual training.

[1]During the training of the residual parameters, these (inner) iterations are embedded in each of the (outer) EM iterations.

[2]The initialized variances are those computed from the codebook entries that are constructed from quantizing the nonlinear function discussed in Section II-C.

[3]Note that the many step jumps in the VTR estimates are due to the quantization of the VTR frequencies discussed in Section II-C.

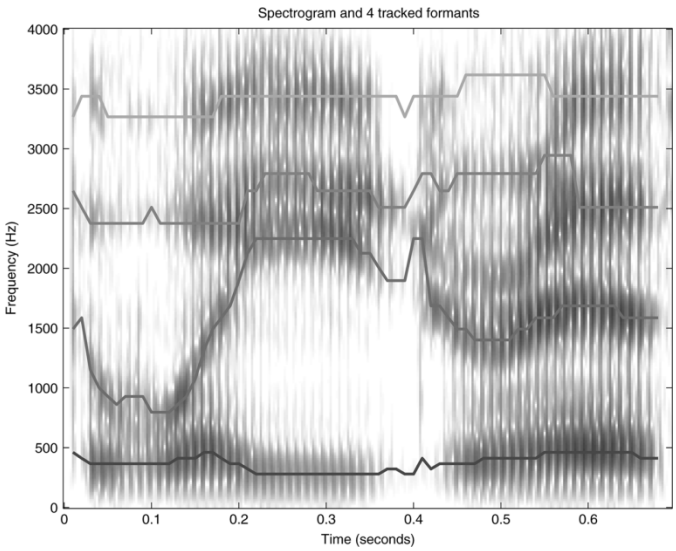

Fig. 6. VTR tracking with two iterations of residual training.

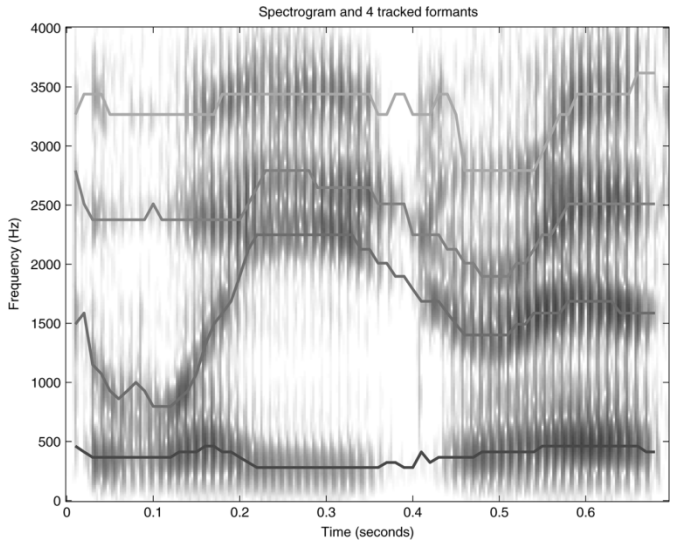

Fig. 7. VTR tracking with three iterations of residual training.

about 20 frames (after time mark of 0.6 s in Fig. 6). One further iteration is able to correct almost all errors as shown in Fig. 7.

To examine the quantitative behavior of the residual parameter training, we list the log-likelihood score as a function of the EM iteration number in Table II. Three iterations of the training appear to have reached the EM convergence. When we examine the VTR tracking results after five and 20 iterations, they are found to be identical to Fig. 7, consistent with the near constant converging log-likelihood score reached after three iterations of training. Note that the regions in the utterance where the speech energy is relatively low are where consonantal constriction or closure is formed, e.g., near time mark of 0.1 s for/w/constriction and near time mark of 0.4 s for/d/closure). The VTR tracker gives almost as accurate estimates for the resonance frequencies in these regions as for the vowel regions.

In Figs. 8 and 9, we show two typical long conversational speech utterances randomly selected from the Switchboard database. As is typical for all the utterances that we have examined, the VTR tracker is able to correctly identify all formants

TABLE II
LOG-LIKELIHOOD SCORE AS A FUNCTION OF THE EM ITERATION NUMBER IN TRAINING THE NONLINEAR-PREDICTION RESIDUAL PARAMETERS

EM Iteration No.	Log-Likelihood Score
0	1.7680
1	2.0813
2	2.0918
3	2.1209
5	2.1220
20	2.1222

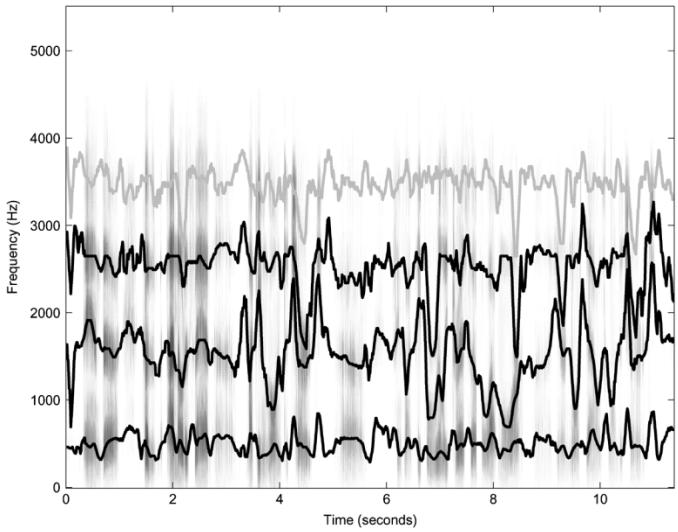

Fig. 8. VTR tracking results for a typical Switchboard utterance: And the, cause I notice that. . .uh. . . like even going to church huh things like people really dress up a lot more. . .huh. . . is going to church.

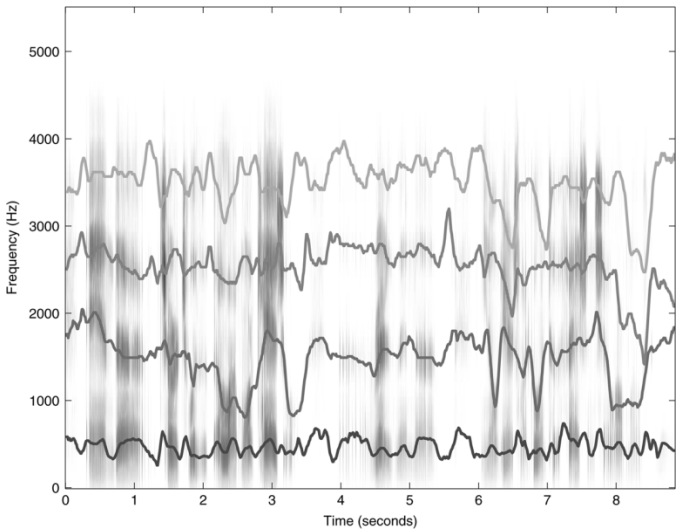

Fig. 9. VTR tracking for another typical Switchboard sentence: And the church was a lot more casual huh rather than. . .uh. . . you know here it is like going to a fashion show almost.

for vowels and glides, which coincide with VTR frequencies. No obvious errors are found where a lower formant or VTR is missed to cause all higher formants to be misidentified, which is a common problem for many conventional formant tracking

methods. For the regions corresponding to consonantal constriction and closure where relatively weak signal energies are present, we observe the general desirable behavior of the VTR tracker—the estimated VTR frequencies tend to be smoothing through these weak-energy regions, based on the adjacent vowel formant estimates, while maintaining correct VTR tracks in the vowel-consonant or consonant-vowel transitions. No clear cases are found where the weak or null energies around a VTR frequency region (e.g., in the f_1 region for/k/closure, around the 1-s time in Fig. 8) cause missing of VTR identification and shifts of higher VTR frequencies. For example, during the above/k/closure, although virtually no signal is present, the VTR tracker provides appropriate values for all f_1, f_2, f_3, and f_4 throughout the closure. This contrasts with other methods where the estimated "formant" values during such closure are either treated as "missing" [20], [21], or they are shifted up inappropriately to high-frequency values [28], [30], deviating substantially from the resonances that are physically present even if no acoustic evidence is available when full closure in the vocal tract is made.

VII. Discussion and Summary

In recent years, there has been a growing interest in developing accurate, efficient, and compact representations, as well as related statistical models, of speech dynamics. Such representations include articulatory variables [22], [26], vocal tract shapes [11], and formants or vocal tract resonances [3], [10], [13], [17], [23], [28]. In this paper, we present a novel technique of tracking vocal tract resonances (VTRs) as a compact representation for time-varying characteristics of speech. VTRs share some common, desirable temporal properties with articulatory variables and yet have a lower dimensionality and more intuitive acoustic interpretation. The VTRs defined as natural frequencies of the vocal tract from glottis to lips (rather than spectral prominences in acoustics) are related to but are also different from formants. The VTRs exist at all times, even when the mouth is closed, just as articulatory variables exist at all times. Since VTRs correspond to natural frequencies of the physical system, they cannot "disappear" even if the acoustic signal does not directly reveal them. Importantly, VTRs are a smooth function of the articulatory variables, whose movement uniquely determines the time-varying vocal tract area function shaping the dynamics of the acoustic resonances. This "noninterrupted" sequence of dynamic, physically meaningful VTR variables is expected to benefit speech recognition, and its automatic extraction from acoustic signals of speech forms the main subject of this study.

While VTRs may not correspond to spectral prominences where zeros in the vocal tract transfer function exist in fricatives, stops, and nasals, they coincide with formants for nonnasalized vowels where no vocal tract side branches and no supra-glottal excitation sources are involved in speech production. In contrast to all the existing formant tracking techniques (e.g., [20], [21], [27], [28], and [30]) that rely, directly or indirectly, on the spectral prominence information from speech acoustics only, the new technique presented in this paper exploits additional dynamic prior information, which we call hidden dynamics expressed in terms of the temporal constraint. This prior captures general smoothness properties of VTR trajectories even if supraglottal excitation may eliminate acoustic spectral prominences (e.g., during consonantal closure). The joint use of the VTR temporal constraint and the speech acoustics forms a discretized state-space model that enables accurate tracking of VTR trajectories at all times and for all manner and voicing classes of speech.

In this paper, we first present the construction and use of a quantized VTR space in an analytical nonlinear prediction function from the VTR variables to the LPC cepstra. We note that this function has been recently used in the work of [29] for formant tracking, where a very different approximation technique (i.e., particle filtering) was explored in dealing with nonlinearity from the quantization scheme that we presented in this paper. The tracking results we present in this paper are significantly more accurate than those in the very limited examples shown in [29]. Also, as discussed in Section II-B, we note that the work of [4] was inspired by the relationship between the VTR variables and the LPC cepstra which we develop and report in this paper. The more recent study described in [16] extends the ideas proposed in [4] and uses a piece-wise linear model where the entire frequency range of each formant is divided into four bands. The success of this extension adds support to the nonlinearity in the analytical relationship between the formant frequency and cepstrum. Instead of using piece-wise linearization in [16], we in the current study deal with the nonlinearity in a very different way—by quantization and by systematically exploiting the entire quantized input-VTR space.

Another innovation in the work presented in this paper is the introduction of the prediction-residual parameters, which are optimally trained by a novel EM algorithm, to effectively compensate for the prediction error. Further, the quantized prediction function with the trained residuals is embedded into a temporal constraint to establish a discretized nonlinear dynamic system model, enabling high-accuracy VTR tracking using a novel dynamic-programming based algorithm. We outline in this paper the development of a greedy technique that exploits the decomposition property of the nonlinear function to drastically reduce the training and tracking algorithms' computational costs. The experimental results on VTR tracking presented in this paper provide evidence that the discretized nonlinear dynamic system approach is effective in modeling the hidden dynamics of speech, as represented by VTR trajectories, and its causal relationship to measurable speech acoustics, as represented by LPC cepstra. Our new work is aimed to expand the current implementation of the dynamic system model so as to include discrete phonological states and its phonetic correlates of speaker-normalized VTR targets for the purpose of speech recognition.

Acknowledgment

The authors thank Dr. H. Hermansky for introducing [4] and [15] to us and for insightful discussions. They also thank an anonymous reviewer who introduced the work of [16] as related to the technique presented in this paper.

REFERENCES

[1] A. Acero, "Formant analysis and synthesis using hidden Markov models," in *Proc. Eurospeech*, Budapest, Hungary, Sep. 1999.

[2] J. Allen, M. S. Hunnicutt, and D. Klatt, *From Text to Speech: The MITalk System*. Cambridge, U.K.: Cambridge Univ. Press, 1987.

[3] I. Bazzi, A. Acero, and L. Deng, "An expectation-maximization approach for formant tracking using a parameter-free nonlinear predictor," in *Proc. ICASSP*, Hong Kong, Apr. 2003.

[4] D. Broad and F. Clermont, "Formant estimation by linear transformation of the LPC cepstrum," *J. Acoust. Soc. Amer.*, vol. 86, pp. 2013–2017, 1989.

[5] I. Bruce, N. Karkhanis, E. Young, and M. Sachs, "Robust formant tracking in noise," in *Proc. ICASSP*, Orlando, FL, 2002, pp. 281–284.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc.*, vol. 39, no. 1, pp. 1–38, 1977.

[7] L. Deng and D. O'Shaughnessy, *Speech Processing—A Dynamic and Optimization-Oriented Approach*. New York: Marcel Dekker, 2003.

[8] L. Deng, I. Bazzi, and A. Acero, "Tracking vocal tract resonances using an analytical nonlinear predictor and a target-guided temporal constraint," in *Proc. Eurospeech*, vol. I, 2003, pp. 73–76.

[9] L. Deng and D. Geisler, "A composite auditory model for processing speech sounds," *J. Acoust. Soc. Amer.*, vol. 82, pp. 2001–2012, Dec. 1987.

[10] L. Deng and J. Ma, "Spontaneous speech recognition using a statistical coarticulatory model for vocal-tract-resonance dynamics," *J. Acoust. Soc. Amer.*, vol. 108, pp. 3036–3048, 2000.

[11] S. Dusan and L. Deng, "Recovering vocal tract shapes from MFCC parameters," in *Proc. ICSLP*, 1998, pp. 3087–3090.

[12] G. Fant, *Acoustic Theory of Speech Production*. The Hague, The Netherlands: Mouton, 1960.

[13] Y. Gao, R. Bakis, J. Huang, and B. Zhang, "Multistage coarticulation model combining articulatory, formant, and cepstral features," in *Proc. ICSLP*, vol. 1, 2000, pp. 25–28.

[14] J. Godfrey, E. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Proc. ICASSP*, 1992, pp. 517–520.

[15] H. Hermansky and D. Broad, "The effective second formant F2 and the vocal tract front-cavity," in *Proc. ICASSP*, vol. 1, 1989, pp. 480–483.

[16] J. Hogberg, "Prediction of formant frequencies from linear combinations of filterbank and cepstral coefficients," Royal Inst. Technol., Stockholm, Sweden, KTH-STL Quarterly Progress Rep., 1997.

[17] J. Holmes, W. Holmes, and P. Garner, "Using formant frequencies in speech recognition," in *Proc. Eurospeech*, Rhodes, Greece, Sep. 1997, pp. 2083–2086.

[18] C. S. Huang and H. C. Wang, "Bandwidth-adjusted LPC analysis for robust speech recognition," *Pattern Recognit. Lett.*, vol. 24, pp. 1583–1587, 2003.

[19] D. Klatt, "Software for a cascade/parallel formant synthesizer," *J. Acoust. Soc. Amer.*, vol. 67, pp. 971–995, 1980.

[20] G. Kopec, "Formant tracking using HMM's and vector quantization," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 709–729, 1986.

[21] S. McCandless, "An algorithm for automatic formant extraction using linear prediction spectra," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 135–141, 1974.

[22] K. Richmond, S. King, and P. Taylor, "Modeling uncertainty in recovering articulation from acoustics," *Comput. Speech Lang.*, vol. 17, pp. 153–172, 2003.

[23] F. Seide, J. Zhou, and L. Deng, "Coarticulation modeling by embedding a target-directed hidden trajectory model into HMM—MAP decoding and evaluation," in *Proc. ICASSP*, 2003, pp. 748–751.

[24] K. Stevens, *Acoustic Phonetics*. Cambridge, MA: MIT Press, 1998.

[25] K. Stevens and C. Bickley, "Constraints among parameters simplify control of Klatt formant synthesizer," *J. Phonetics*, vol. 19, pp. 161–174, 1991.

[26] J. Sun, L. Deng, and X. Jing, "Data-driven model construction for continuous speech recognition using overlapping articulatory features," in *Proc. ICSLP*, vol. 1, 2000, pp. 437–440.

[27] D. Talkin, "Speech formant trajectory estimation using dynamic programming with modulated transition costs," *J. Acoust. Soc. Amer.*, vol. S1, p. S55, 1987.

[28] L. Welling and H. Ney, "Formant tracking for speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 36–48, 1998.

[29] Y. Zheng and M. Hasegawa-Johnson, "Formant tracking by mixture state particle filter," in *Proc. ICASSP*, vol. 1, 2004, pp. 565–568.

[30] P. Zolfaghari and T. Robinson, "Formant analysis using mixtures of Gaussians," in *Proc. Eurospeech*, Rhodes, Greece, 1997, pp. 2539–2542.

Li Deng (M'86–SM'91–F'04) received the B.S. degree in biophysics from University of Science and Technology of China, Hefei, in 1982, the M.S. degree in electrical engineering from University of Wisconsin-Madison in 1984, and the Ph.D. degree from University of Wisconsin-Madison in 1986.

He worked on large vocabulary automatic speech recognition at INRS-Telecommunications in Montreal, QC, Canada, from 1986 to 1989. In 1989, he joined the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada, as Assistant Professor, where he became a tenured Full Professor in 1996. From 1992 to 1993, he conducted sabbatical research at the Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge, and from 1997 to 1998, at ATR Interpreting Telecommunications Research Laboratories, Kyoto, Japan. In 1999, he joined Microsoft Research, Redmond, WA, as Senior Researcher, and as Affiliate Professor of Electrical Engineering at the University of Washington, Seattle. His research interests include acoustic-phonetic modeling of speech, speech and speaker recognition, speech synthesis and enhancement, speech production and perception, auditory speech processing, noise robust speech processing, statistical methods and machine learning, nonlinear signal processing, spoken language systems, multimedia signal processing, and multimodal human–computer interaction. In these areas, he has published over 200 technical papers and book chapters, and is inventor and co-inventor of numerous patents. He is principal author of the book *Speech Processing—A Dynamic and Optimization-Oriented Approach* (New York: Marcel Dekker, 2003).

Dr. Deng served on the Education Committee and Speech Processing Technical Committee of the IEEE Signal Processing Society during 1996–2000. He served as Associate Editor for IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING during 2002–2005 and is currently serving as a member of the IEEE Signal Processing Society's Technical Directions Committee and Multimedia Signal Processing Technical Committee. He was a Technical Chair of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP04). He is Fellow of the Acoustical Society of America.

Alex Acero (S'83–M'90–SM'00–F'03) received the engineering degree from the Polytechnic University of Madrid, Madrid, Spain, in 1985, the M.S. degree from Rice University, Houston, TX, in 1987, and the Ph.D. degree from Carnegie Mellon University, Pittsburgh, PA, in 1990, all in electrical engineering.

He was a Senior Voice Engineer at Apple Computer (1990–1991) and Manager of the Speech Technology Group at Telefonica Investigacion y Desarrollo (1991–1993). He joined Microsoft Research, Redmond, WA, in 1994, where he is currently Manager of the Speech Group. He is also an Affiliate Professor at the University of Washington, Seattle. He is author of the books *Spoken Language Processing* (Englewood Cliffs, NJ: Prentice-Hall, 2000) and *Acoustical and Environmental Robustness in Automatic Speech Recognition* (Norwell, MA: Kluwer, 1993). He also has written chapters in three edited books, has nine granted U.S. patents, and written over 90 journal and conference papers. His research interests include noise robustness, signal processing, acoustic modeling, statistical language modeling, spoken language processing, speech-centric multimodal interfaces, and machine learning. He is Associate Editor of *Computer Speech and Language*.

Dr. Acero has had several positions within the IEEE Signal Processing Society, including Member-at-Large of the Board of Governors, Associate Editor of IEEE SIGNAL PROCESSING LETTERS, and as Member (1996–2000) and Chair (2000–2002) of the Speech Technical Committee. He was General Co-Chair of the 2001 IEEE Workshop on Automatic Speech Recognition and Understanding, Sponsorship Chair of the 1999 IEEE Workshop on Automatic Speech Recognition and Understanding, and Publications Chair of ICASSP'98.

Issam Bazzi, photograph and biography not available at the time of publication.