Kalman-based autoregressive moving average modeling and inference for formant and antiformant tracking[a)]

Daryush D. Mehta[b)] and Daniel Rudoy
School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138

Patrick J. Wolfe[c)]
Department of Statistical Science, University College London, London WC1E 6BT, United Kingdom

(Received 30 June 2011; revised 26 June 2012; accepted 27 June 2012)

Vocal tract resonance characteristics in acoustic speech signals are classically tracked using frame-by-frame point estimates of formant frequencies followed by candidate selection and smoothing using dynamic programming methods that minimize *ad hoc* cost functions. The goal of the current work is to provide both point estimates and associated uncertainties of center frequencies and bandwidths in a statistically principled state-space framework. Extended Kalman (K) algorithms take advantage of a linearized mapping to infer formant and antiformant parameters from frame-based estimates of autoregressive moving average (ARMA) cepstral coefficients. Error analysis of KARMA, WAVESURFER, and PRAAT is accomplished in the all-pole case using a manually marked formant database and synthesized speech waveforms. KARMA formant tracks exhibit lower overall root-mean-square error relative to the two benchmark algorithms with the ability to modify parameters in a controlled manner to trade off bias and variance. Antiformant tracking performance of KARMA is illustrated using synthesized and spoken nasal phonemes. The simultaneous tracking of uncertainty levels enables practitioners to recognize time-varying confidence in parameters of interest and adjust algorithmic settings accordingly.
 [http://dx.doi.org/10.1121/1.4739462]

PACS number(s): 43.72.Ar, 43.70.Bk, 43.60.Cg, 43.60.Uv [CYE] Pages: 1732–1746

I. INTRODUCTION

Speech formant tracking has received continued attention during the past 60 years to better characterize formant motion during vowels as well as vowel-consonant boundaries. The *de facto* approach to resonance estimation involves waveform segmentation and the assumption of an all-pole model characterized by second-order digital resonators (Schafer and Rabiner, 1970). The center frequency and bandwidth of each resonator are then estimated through picking peaks in the all-pole spectrum or finding roots of the prediction polynomial. Tracking these estimates across frames is typically accomplished via dynamic programming methods that minimize cost functions to produce smoothly varying trajectories.

This general formant tracking approach is implemented in WAVESURFER (Sjölander and Beskow, 2005) and PRAAT (Boersma and Weenink, 2009), speech analysis tools that enjoy widespread use in the speech recognition, clinical, and linguistic communities. There are, however, numerous shortcomings to this classical approach. For example, formant track smoothing (e.g., correcting for large frequency jumps) are performed in an *ad hoc* manner that precludes the ability to apply statistical analysis to obtain confidence intervals around the estimated tracks.

Initial development of the formant tracking approach described here has been reported by Rudoy *et al.* (2007) using a manually marked formant database for error analysis (Deng *et al.*, 2006b). The current work continues this line of research and offers two main contributions. The first provides improvements to the Kalman-based autoregressive approach of Deng *et al.* (2007) and extensions to enable antiformant frequency and bandwidth tracking in a Kalman-based autoregressive moving average (KARMA) framework. The second empirically determines the performance of the KARMA approach through visual and quantitative error analysis and compares this performance with that of WAVESURFER and PRAAT.

A. Classical formant tracking algorithms

Linear predictive coding (LPC) models have been shown to efficiently encode source/filter characteristics of the acoustic speech signal (Atal and Hanauer, 1971). To extract frame-by-frame formant parameters, the poles of the LPC spectrum can be computed as the roots of the prediction polynomial, peaks in the LPC spectrum, or peaks in the second derivative of the frequency spectrum (Christensen *et al.*, 1976). The first complete formant *tracker* over multiple continuous speech frames employed spectral peak-picking, selection of formants from the candidate peaks using continuity constraints, and voicing detection to handle silent and unvoiced speech segments (McCandless, 1974). Extensions to LPC analysis

[a)]Portions of this work were presented at the INTERSPEECH conference in Antwerp, Belgium, in August 2007 (Rudoy *et al.*, 2007).
[b)]Author to whom correspondence should be addressed. Also with the Center for Laryngeal Surgery and Voice Rehabilitation, Department of Otolaryngology, The University of Melbourne, Parkville, Victoria 3010, Australia. Electronic mail: daryush.mehta@alum.mit.edu
[c)]Also with the Department of Computer Science, University College London, London WC1E 6BT, United Kingdom and the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801.

incorporated autoregressive moving average (ARMA) models that added estimates of candidate zeros associated with antiresonances during consonantal and nasal speech sounds (Steiglitz, 1977; Atal and Schroeder, 1978).

Figure 1(A) illustrates the classical tracking process for the all-pole case. Following pre-processing steps, LPC spectral coefficients yield intra-frame point estimates of candidate frequency and bandwidth parameters via root finding, peak-picking, or a number of other methods (Atal and Hanauer, 1971; Atal and Schroeder, 1978; Broad and Clermont, 1989; Yegnanarayana, 1978). Inter-frame parameter assignment and smoothing is performed by minimizing various cost functions in a dynamic programming environment (Sjölander and Beskow, 2005; Boersma and Weenink, 2009). Note that the required root-finding (or peak-picking) procedure cannot be written in closed form. Consequently, statistical analysis (distributions, bias, variance) of the resultant formant and bandwidth estimates is challenging. Alternative spectrographic representations primarily apply to sustained vowels and require significant manual interaction (Fulop, 2010).

B. Statistical formant tracking algorithms

Probabilistic and statistical models for tracking formants have gained widespread use in the past 15 years with motivation from automatic speech recognition applications. The first such probabilistic model, introduced by Kopec (1986), featured a hidden Markov model that was used to constrain the evolution of vector-quantized sets of formant frequencies and bandwidths. Similarly, a state-space dynamical model can appropriately constrain the evolution of formant parameters where observations of the acoustic speech waveform are linked through nonlinear relationships to desired "hidden" states (formant parameters) that evolve over time. Inference of the state values can be performed by variants of the Kalman filter (Kalman, 1960).

In these algorithms, *ad hoc* assignment of poles and zeros to appropriate formant indices is precluded by the inherent association of spectral/cepstral coefficients to formant and antiformant frequencies and bandwidths. The first reported state-space approach to formant tracking inferred formant frequencies and bandwidths directly from LPC spectral coefficients (Rigoll, 1986). An extension to this LPC approach was made by Toyoshima *et al.* (1991) to build a tracker that inferred frequencies and bandwidths of both formants and antiformants from time-varying ARMA spectral coefficients (Miyanaga *et al.*, 1986). More recent state-space models define the observations as coefficients in the LPC *cepstral* domain (Zheng and Hasegawa-Johnson, 2004; Deng *et al.*, 2007), providing statistical methods to support or refute empirical relations obtained between low-order cepstral coefficients and formant frequencies (Broad and Clermont, 1989).

The proposed KARMA approach explores the performance of such a state-space model with ARMA cepstral coefficients as observations to track formant and antiformant parameters. Taking advantage of a linearized mapping between frequency and bandwidth values and cepstral coefficients, KARMA applies Kalman inference to yield point estimates and uncertainties of the output trajectories.

II. METHODS

Figure 1(B) illustrates the proposed statistical modeling approach to formant and antiformant tracking. This approach affords several advantages over classical approaches: (1) both formant and antiformant trajectories are tracked, (2) both frequency and bandwidth estimates are propagated as distributions instead of point estimates to provide for uncertainty quantification, and (3) pole/zero assignment to formants/antiformants is made through a linearized cepstral mapping instead of candidate selection using *ad hoc* cost functions.

A. Step 1: pre-processing

The acoustic speech waveform is first resampled to rate f_s, depending on the number of trajectories the user expects to track within the speech bandwidth $f_s/2$. The resulting signal $s[m]$ is then segmented into short-time frames $s_t[m] = s[m]w_t[m]$ using overlapping windows $w_t[m]$, with frame index t and sample index m. Each frame $s_t[m]$ is then filtered via

$$s_t[m] = s_t[m] - \gamma s_t[m-1], \tag{1}$$

where γ is the pre-emphasis coefficient defining the inherent high-pass filter characteristic that is typically applied to equalize the energy across the speech spectrum and improve spectral model fitting.

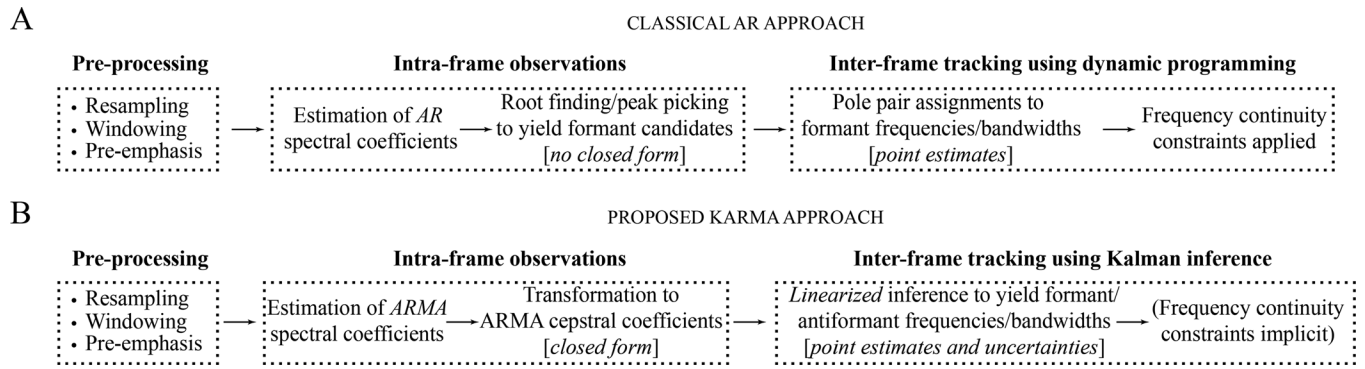

FIG. 1. Illustration of the (A) classical and (B) proposed approaches to formant tracking. Key advantages to the proposed KARMA approach include intra-frame observation of autoregressive moving average parameters for both formant and antiformant tracking, inter-frame tracking using linearized Kalman inference, and the availability of both point estimates and uncertainties for each trajectory.

B. Step 2: intra-frame observation generation

1. ARMA model of speech

Following windowing and pre-emphasis, the acoustic waveform $s_t[m]$ is modeled as a stochastic ARMA(p, q) process:

$$s_t[m] = \sum_{i=1}^{p} a_i s_t[m-i] + \sum_{j=1}^{q} b_j u[m-j] + u[m], \tag{2}$$

where a_i are the p AR coefficients, b_j are the q MA coefficients, and $u[m]$ is the stochastic excitation waveform. The z-domain transfer function associated with Eq. (2) is

$$T(z) \triangleq \frac{1 - \sum_{j=1}^{q} b_j z^{-j}}{1 - \sum_{i=1}^{p} a_i z^{-i}}. \tag{3}$$

A number of standard spectral estimation techniques can be employed to fit data to the ARMA(p, q) model (see Marelli and Balazs, 2010, for a recent review of ARMA estimation methods). In the current study, ARMA estimation was performed using the "armax" function in MATLAB's System Identification toolbox (The MathWorks, Natick, MA), which implements an iterative method to minimize a quadratic error prediction criterion (Ljung, 1999).

2. Generation of observations: ARMA cepstral coefficients

In the proposed approach, the ARMA spectral coefficients in Eq. (3) are transformed to the complex cepstrum before inferring formant characteristics. This mapping from ARMA spectral coefficients to ARMA cepstral coefficients has been derived in the all-pole case (e.g., Deng *et al.*, 2006a) and can be extended to account for the presence of zeros in the spectrum. Letting C_n denote the nth cepstral coefficient,

$$C_n = c_n - c'_n, \tag{4}$$

where C_n depends on separate contributions from the denominator and numerator of the ARMA model through the following recursive relationships:

$$c_n = \begin{cases} a_n & \text{if} \quad n = 1 \\ a_n + \sum_{i=1}^{n-1} \left(\frac{i}{n}\right) a_{n-i} c_i & \text{if} \quad 1 < n \le p \\ \sum_{i=n-p}^{n-1} \left(\frac{i}{n}\right) a_{n-i} c_i & \text{if} \quad p < n, \end{cases} \tag{5a}$$

$$c'_n = \begin{cases} b_n & \text{if} \quad n = 1 \\ b_n + \sum_{j=1}^{n-1} \left(\frac{j}{n}\right) b_{n-j} c'_j & \text{if} \quad 1 < n \le q \\ \sum_{j=n-q}^{n-1} \left(\frac{j}{n}\right) b_{n-j} c'_j & \text{if} \quad q < n. \end{cases} \tag{5b}$$

Derivation of Eqs. (5a) and (5b) is given in the Appendix. The proof is derived under the minimum-phase assumption that constrains the poles and zeros of the ARMA transfer function to lie within the unit circle.

C. Step 3: inter-frame tracking

The proposed KARMA algorithm tracks point estimates and uncertainties for I formants and J antiformants from frame to frame. To accommodate the temporal dimension, the states at frame t are placed in column vector $\boldsymbol{x}_t$:

$$\boldsymbol{x}_t \triangleq (f_1 \cdots f_I \;\; b_1 \cdots b_I \;\; f'_1 \cdots f'_J \;\; b'_1 \cdots b'_J)^T, \tag{6}$$

where (f_i, b_i) is the frequency/bandwidth pair of the ith formant and (f'_j, b'_j) is the frequency/bandwidth pair for the jth antiformant.

1. Observation model

Inference of the output parameters is facilitated by the closed-form mapping from the state vector $\boldsymbol{x}_t$ to the observed cepstral coefficients C_n in Eq. (4). Extending the speech production model of Schafer and Rabiner (1970) to capture zeros, we assume that the transfer function $T(z)$ of the ARMA model can be written as a cascade of I second-order digital resonators (formants) and J second-order digital antiresonators (antiformants):

$$T(z) = \frac{\prod_{j=1}^{J} (1 - \beta_j z^{-1})(1 - \bar{\beta}_j z^{-1})}{\prod_{i=1}^{I} (1 - \alpha_i z^{-1})(1 - \bar{\alpha}_i z^{-1})}, \tag{7}$$

where $(\alpha_i, \bar{\alpha}_i)$ and $(\beta_j, \bar{\beta}_j)$ denote complex-conjugate pole and zero pairs, respectively. Each pole and zero are parameterized by a center frequency and 3-dB bandwidth (both in units of hertz) using the following relations:

$$(\alpha_i, \bar{\alpha}_i) = \exp\left(\frac{-\pi b_i \pm 2\pi\sqrt{-1} f_i}{f_s}\right), \tag{8a}$$

$$(\beta_j, \bar{\beta}_j) = \exp\left(\frac{-\pi b'_j \pm 2\pi\sqrt{-1} f'_j}{f_s}\right), \tag{8b}$$

where f_s is the sampling rate (in hertz).

Performing a Taylor-series expansion of $\log T(z)$ yields

$$\log T(z) = \sum_{i=1}^{I} \sum_{n=1}^{\infty} \frac{\alpha_i^n + \bar{\alpha}_i^n}{n} z^{-n} - \sum_{j=1}^{J} \sum_{n=1}^{\infty} \frac{\beta_j^n + \bar{\beta}_j^n}{n} z^{-n}. \tag{9}$$

Recalling that C_n is the nth cepstral coefficient, $\log T(z) = C_0 + \Sigma_{n=1}^{\infty} C_n z^{-n}$. Thus equating the coefficients of powers of z^{-1} leads to

$$C_n = \frac{1}{n} \sum_{i=1}^{I} (\alpha_i^n + \bar{\alpha}_i^n) - \frac{1}{n} \sum_{j=1}^{J} (\beta_j^n + \bar{\beta}_j^n). \tag{10}$$

Finally, inserting α_i and β_j from Eqs. (8a) and (8b) into Eq. (10) yields the following observation model $h(\boldsymbol{x}_t)$ that maps elements of $\boldsymbol{x}_t$ to C_n:

$$h(\boldsymbol{x}_t) \triangleq C_n = \frac{2}{n}\sum_{i=1}^{I} \exp\left(-\frac{\pi n}{f_s} b_i\right) \cos\left(\frac{2\pi n}{f_s} f_i\right) - \frac{2}{n}\sum_{j=1}^{J} \exp\left(-\frac{\pi n}{f_s} b'_j\right) \cos\left(\frac{2\pi n}{f_s} f'_j\right). \quad (11)$$

2. State-space model

We adopt a state-space framework similar to that by Deng *et al.* (2007) to model the evolution of the state vector in Eq. (6) from frame t to frame $t+1$:

$$\boldsymbol{x}_{t+1} = \boldsymbol{F}\boldsymbol{x}_t + \boldsymbol{w}_t, \quad (12a)$$

$$\boldsymbol{y}_t = h(\boldsymbol{x}_t) + \boldsymbol{v}_t, \quad (12b)$$

where $\boldsymbol{F}$ is the state transition matrix, and $\boldsymbol{w}_t$ and $\boldsymbol{v}_t$ are uncorrelated white Gaussian sequences with covariance matrices $\boldsymbol{Q}$ and $\boldsymbol{R}$, respectively. The function $h(\boldsymbol{x}_t)$ is the nonlinear mapping of Eq. (11), and vector $\boldsymbol{y}_t$ consists of *estimates* of the first N cepstral coefficients of C_n (not including the zeroth coefficient). The initial state $\boldsymbol{x}_0$ follows a normal distribution with mean $\boldsymbol{\mu}_0$ and covariance $\boldsymbol{\Sigma}_0$. The state-space model of Eqs. (12a) and (12b) is thus parameterized by the set $\boldsymbol{\theta}$:

$$\boldsymbol{\theta} \triangleq (\boldsymbol{F}, \boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0). \quad (13)$$

3. Linearization via Taylor approximation

The cepstral mapping in Eq. (11) must be linearized to enable *approximate* minimum-mean-square-error (MMSE) estimation of the tracked states via the extended Kalman filter. Linearization of the mapping $h(\boldsymbol{x}_t)$ arises from the first-order terms of the Taylor-series expansion of C_n in Eq. (11):

$$\frac{\partial C_n}{\partial f_i} = -\frac{4\pi}{f_s} \exp\left(-\frac{\pi n}{f_s} b_i\right) \sin\left(\frac{2\pi n}{f_s} f_i\right),$$
$$\frac{\partial C_n}{\partial b_i} = -\frac{2\pi}{f_s} \exp\left(-\frac{\pi n}{f_s} b_i\right) \cos\left(\frac{2\pi n}{f_s} f_i\right),$$
$$\frac{\partial C_n}{\partial f'_j} = \frac{4\pi}{f_s} \exp\left(-\frac{\pi n}{f_s} b'_j\right) \sin\left(\frac{2\pi n}{f_s} f'_j\right),$$
$$\frac{\partial C_n}{\partial b'_j} = \frac{2\pi}{f_s} \exp\left(-\frac{\pi n}{f_s} b'_j\right) \cos\left(\frac{2\pi n}{f_s} f'_j\right).$$

The Jacobian matrix $\boldsymbol{H}_t$ thus consists of four sub-matrices for each frame t:

$$\boldsymbol{H}_t \triangleq \left(\boldsymbol{H}(f_i)\, \boldsymbol{H}(b_i)\, \boldsymbol{H}(f'_j)\, \boldsymbol{H}(b'_j)\right), \quad (14)$$

where $\boldsymbol{H}(f_i)$ and $\boldsymbol{H}(b_i)$ each consists of N rows and $p/2$ columns:

$$\boldsymbol{H}(f_i) \triangleq \begin{pmatrix} \frac{\partial C_1}{\partial f_1} & \frac{\partial C_1}{\partial f_2} & \cdots & \frac{\partial C_1}{\partial f_{p/2}} \\ \frac{\partial C_2}{\partial f_1} & \frac{\partial C_2}{\partial f_2} & \cdots & \frac{\partial C_2}{\partial f_{p/2}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial C_N}{\partial f_1} & \frac{\partial C_N}{\partial f_2} & \cdots & \frac{\partial C_N}{\partial f_{p/2}} \end{pmatrix}, \quad (15a)$$

$$\boldsymbol{H}(b_i) \triangleq \begin{pmatrix} \frac{\partial C_1}{\partial b_1} & \frac{\partial C_1}{\partial b_2} & \cdots & \frac{\partial C_1}{\partial b_{p/2}} \\ \frac{\partial C_2}{\partial b_1} & \frac{\partial C_2}{\partial b_2} & \cdots & \frac{\partial C_2}{\partial b_{p/2}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial C_N}{\partial b_1} & \frac{\partial C_N}{\partial b_2} & \cdots & \frac{\partial C_N}{\partial b_{p/2}} \end{pmatrix}, \quad (15b)$$

and $\boldsymbol{H}(f'_j)$ and $\boldsymbol{H}(b'_j)$ are each defined analogously with N rows and $q/2$ columns. This frame-local linearization of the observation model differs from the complex piecewise-linear approach of Deng *et al.* (2007) in which a trainable residual term is included to absorb any errors due to the approximation.

4. Kalman-based inference

Given observations $\boldsymbol{y}_t$ for frame indices 1 to T, the extended Kalman smoother (EKS) can be used to compute the mean $\boldsymbol{m}_{t|T}$ (point estimates) and covariance $\boldsymbol{P}_{t|T}$ (estimate uncertainties) of each parameter in $\boldsymbol{x}_t$ (Kalman, 1960). Table I displays the steps of the EKS, which employs a two-pass filtering (forward) and smoothing (backward) procedure. For real-time processing, the forward filtering stage may be applied without the backward pass; naturally, this will lead to larger uncertainties in the corresponding parameter estimates.

Care must be taken when linearly approximating the observation model of Eq. (11) to avoid suboptimal performance or algorithm divergence in the case of the Kalman filter (Julier and Uhlmann, 1997). Although the EKS loses the statistical optimality of the linear Kalman filter, extended Kalman inference is widely used in engineering practice with satisfactory results due to careful selection of parameters (Julier and Uhlmann, 2004). The application of the EKS to formant and antiformant tracking is no exception, and thus this study explores performance effects due to certain parameter selections and algorithmic decisions.

To verify the appropriateness of the linearization in Sec. II C 3 in the current setting, a comparison is made to the more computationally intensive method of particle filtering.

TABLE I. The extended Kalman algorithms for yielding point estimates and associated uncertainties of tracked parameters. See text for definition of variables.

1. Initialization: Set $\boldsymbol{m}_{0|0} = \boldsymbol{\mu}_0$ and $\boldsymbol{P}_{0|0} = \boldsymbol{\Sigma}_0$
2. Filtering: Repeat for $t = 1,\ldots, T$

$$\begin{aligned} \boldsymbol{m}_{t|t-1} &= \boldsymbol{F}\boldsymbol{m}_{t-1|t-1} \\ \boldsymbol{P}_{t|t-1} &= \boldsymbol{F}\boldsymbol{P}_{t-1|t-1}\boldsymbol{F}^T + \boldsymbol{Q} \\ \boldsymbol{K}_t &= \boldsymbol{P}_{t|t-1}\boldsymbol{H}_t^T(\boldsymbol{H}_t\boldsymbol{P}_{t|t-1}\boldsymbol{H}_t^T + \boldsymbol{R})^{-1} \\ \boldsymbol{m}_{t|t} &= \boldsymbol{m}_{t|t-1} + \boldsymbol{K}_t\left(\boldsymbol{y}_t - h(\boldsymbol{m}_{t|t-1})\right) \\ \boldsymbol{P}_{t|t} &= \boldsymbol{P}_{t|t-1} - \boldsymbol{K}_t\boldsymbol{H}_t\boldsymbol{P}_{t|t-1} \end{aligned} \quad (16)$$

3. Smoothing: Repeat for $t = T,\ldots, 1$

$$\begin{aligned} \boldsymbol{S}_t &= \boldsymbol{P}_{t-1|t-1}\boldsymbol{F}^T\boldsymbol{P}_{t|t-1}^{-1} \\ \boldsymbol{m}_{t-1|T} &= \boldsymbol{m}_{t-1|t-1} + \boldsymbol{S}_t(\boldsymbol{m}_{t|T} - \boldsymbol{F}\boldsymbol{m}_{t-1|t-1}) \\ \boldsymbol{P}_{t-1|T} &= \boldsymbol{P}_{t-1|t-1} + \boldsymbol{S}_t(\boldsymbol{P}_{t|T} - \boldsymbol{P}_{t-1|t-1})\boldsymbol{S}_t^T \end{aligned}$$

An approach based on particle filtering involves approximating the posterior density though a mixture of discrete samples ("particles") that are propagated directly through the nonlinear system dynamics at each time step (Gordon *et al.*, 1993; Zheng and Hasegawa-Johnson, 2004).

The state-space framework of Eqs. (12a) and (12b) is used here as a generative model to simulate the dynamics of $I = 4$ formant tracks. The model generates 25 Monte Carlo simulations, each 100 samples in length. The distributions of the initial state vector $\boldsymbol{x}_0 = (500\ 1500\ 2500\ 3500)$ Hz propagate independently ($\boldsymbol{F}$ is the identity matrix) with a 30 Hz standard deviation at each time step. Corresponding formant bandwidths are (80 120 160 200) Hz, which propagate with one-fifth of the center frequency's standard deviation. The state vector at each time step is transformed to $N = 15$ cepstral coefficients via the observation model of Eq. (11) with $f_s = 10$ kHz. Simulated noise is added to the cepstral coefficients to yield a signal-to-noise ratio of 15 dB.

Figure 2 compares the output of the extended Kalman filter of Table I and that of a particle filter in terms of root-mean-square error (RMSE) averaged over the three formant frequency tracks as a function of number of particles. Both algorithms are provided with oracle values of formant bandwidth, $\boldsymbol{x}_0$, $\boldsymbol{F}$, and $\boldsymbol{R}$. The diagonal noise covariance $\boldsymbol{Q}$ is estimated as the variance of the first difference function of each formant track. The performance of the extended Kalman filter compares favorably to that of the particle filter, even when a large number of particles is used. Similar results hold over a broad range of parameter values.

5. Observability of states

The model of Eqs. (12a) and (12b) does not explicitly take into account the existence of speech and non-speech states. To continue to track or *coast* parameters during non-speech frames, the state vector $\boldsymbol{x}_t$ can be augmented with a binary indicator variable to specify the presence of speech in the frame. The approximate MMSE state estimate is then obtained by modifying the Kalman gain in Eq. (16) (Table I):

$$\boldsymbol{K}_t = \boldsymbol{M}_t \boldsymbol{P}_{t|t-1} \boldsymbol{H}_t^T (\boldsymbol{H}_t \boldsymbol{P}_{t|t-1} \boldsymbol{H}_t^T + \boldsymbol{R})^{-1},$$

where $\boldsymbol{M}_t$ is a diagonal matrix with entries equal to 1 or 0 depending on the presence or absence, respectively, of speech energy in frame t.

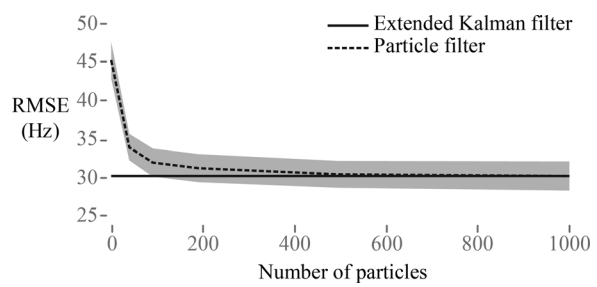

FIG. 2. Comparison of extended Kalman filter (solid) and particle filter (dashed) tracking performance in terms of root-mean-square error (RMSE) averaged over 25 Monte Carlo trials and reported with 95% confidence intervals (gray).

In addition, to handle the presence or absence of particular tracks, the state vector $\boldsymbol{x}_t$ can be dynamically modified to include or omit corresponding frequency/bandwidth states in Eq. (6). The approximate MMSE state estimate is then obtained via the Kalman recursions in Table I with the modified state vector. If an absent state reappears in a given frame, that state can be re-initialized with corresponding entries from $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$.

D. Summary of KARMA algorithm

Table II outlines the three steps of the proposed KARMA algorithm, consisting of a pre-processing stage, intra-frame cepstral coefficient estimation, and inter-frame tracking of formant and antiformant frequencies using Kalman inference. Table III lists the modifiable parameters of KARMA in each step along with baseline values found to empirically follow most temporal variations of resonances due to articulatory motion.

As is commonly done, the orders p and q of the ARMA model are selected to capture as much information as possible on the peaks and valleys in the resonance spectrum while avoiding overfitting and errantly capturing source-related information. The cepstral order N is chosen to be at least $\max(p, q)$ so that all pole/zero information is incorporated per Eq. (5a) and (5b). Choosing appropriate values for I and J in Eq. (11) depends on the expected number of formants and antiformants, respectively, present within the speech bandwidth $[0, f_s/2]$ Hz.

The state transition covariance matrix $\boldsymbol{Q}$ sets the expected frame-to-frame frequency variation of the center frequencies and bandwidths and can be assigned manually or using *a priori* knowledge of formant transition rates. The baseline matrix for $\boldsymbol{Q}$ describes the evolution of each formant as a normal random walk with standard deviation of 224 Hz.

Formants do not evolve independently of one another, and their temporal trajectories are not independent in

TABLE II. Proposed KARMA algorithm for formant and antiformant tracking.

Repeat for frames $t = 1,\ldots, T$ (Online or batch mode)
1. Pre-processing of input speech waveform $s[m]$
   (a) Resample waveform to f_s
   (b) Segment and window: $s_t[m] = s[m]w_t[m]$
   (c) Pre-emphasize $s_t[m]$ via Eq. (1)
   (d) Identify frames containing speech
2. Intra-frame observation of N cepstral coefficients
   (a) Estimate ARMA(p, q) spectral coefficients $\hat{a}_i$ and $\hat{b}_j$ in Eq. (3)
   (b) Convert $\hat{a}_i$ and $\hat{b}_j$ to N cepstral coefficients using Eq. (4)
3. Inter-frame tracking of I formants and J antiformants
   (a) Apply Kalman filter from Table I
   (b) $\boldsymbol{m}_{t|t}$ are point estimates, diagonal elements of $\boldsymbol{P}_{t|t}$ are variances of the point estimates

Repeat for frames $t = T,\ldots, 1$ (Batch mode only)
4. Inter-frame tracking of I formants and J antiformants
   (a) Apply Kalman smoothing step in Table I
   (b) $\boldsymbol{m}_{t|T}$ are point estimates, diagonal elements of $\boldsymbol{P}_{t|T}$ are variances of the point estimates

TABLE III. Modifiable parameters and their baseline values for the three steps in the proposed KARMA approach.

Step	Parameter	Symbol	Units	Baseline
1. Pre-processing				
	Resampling rate	f_s	Hz	7000
	Window type/duration/overlap	$w_t[m]$	—	Hamming/20 ms/50%
	Pre-emphasis coefficient	γ	—	0.7
2. Intra-frame observation				
	AR order	p	—	12
	MA order	q	—	0
	Cepstral order	N	—	15
3. Inter-frame tracking				
	Number of formants	I	—	3[a]
	Number of antiformants	J	—	0
	State transition matrix	$\boldsymbol{F}$	—	Lag-one correlation[b]
	State transition covariance	$\boldsymbol{Q}$	Hz^2	$(50000\ 50000\ 50000)^T$
	Observation noise covariance	$\boldsymbol{R}$	—	$(1\ 1/2\ 1/3)^T$
	Initial state	$\boldsymbol{\mu}_0$	Hz	$(500\ 1500\ 2500)^T$
	Initial state covariance	$\boldsymbol{\Sigma}_0$	Hz^2	$(10000\ 10000\ 10000)^T$

[a]Due to general issues with bandwidth estimation in real speech spectra, formant bandwidths can be either tracked or fixed, e.g., to 80, 120, and 160 Hz.
[b]$\boldsymbol{F}$ is estimated *a priori* based on first-pass formant information for a particular utterance using a linear least-squares estimator (Hamilton, 1994).

frequency. For example, in the synthesis of front vowels, it is common practice to employ a linear regression of f_3 onto f_1 and f_2 (e.g., Nearey, 1989). Empirically, we found the formant cross-correlation function to decay slowly (Rudoy *et al.*, 2007), implying that a set of formant values at frame t might be helpful in predicting values of all formants at frame $t+1$ (*lag-one correlation*). Thus instead of setting the state transition matrix $\boldsymbol{F}$ to the identity matrix (Deng *et al.*, 2006a; Deng *et al.*, 2007), $\boldsymbol{F}$ can be estimated *a priori*, e.g., based on first-pass formant information using a linear least-squares estimator (Hamilton, 1994).

The covariance matrix $\boldsymbol{R}$ models the signal-to-noise ratio of the cepstral coefficients and is set to a diagonal matrix with elements $\boldsymbol{R}_{nn} = 1/n \text{ for } n \in \{1, 2, \ldots, N\}$. Empirically, this setting for $\boldsymbol{R}$ is observed to be in reasonable agreement with the variance of the residual vector of the cepstral coefficients derived from speech waveforms. The initialized values of the center frequencies and bandwidths in $\boldsymbol{\mu}_0$ can be fixed according to neutral vowel formant frequencies (the baseline vector here) or assigned depending on the experimental setup.

E. Benchmark algorithms

Performance of KARMA is compared with that of WAVESURFER (Sjölander and Beskow, 2005) and PRAAT (Boersma and Weenink, 2009), two software packages that see wide use among voice and speech researchers. WAVESURFER and PRAAT both follow the classical formant tracking approach in which frame-by-frame formant frequency candidates are obtained from an all-pole spectrum model and smoothed across the entire speech utterance to remove outliers and constrain the trajectories to physiologically plausible values. Smoothing is accomplished through dynamic programming to minimize the sum of the following three cost functions: (1) the deviation between the frequency for each formant from baseline values of each frequency, (2) a measure of the quality factor f_i/b_i of a formant where higher quality factors are favored, and (3) a transition cost that penalizes large frequency jumps. The user can set weights to these cost functions to tune the algorithm's performance. Antiformant tracking is not included as an option in either software package.

III. RESULTS

Evaluation of the KARMA approach is accomplished in the all-pole case using a vocal tract resonance (VTR) database with manually corrected formant trajectories (Deng *et al.*, 2006b). Because the VTR database itself exhibits observable labeling errors and thus yields approximations to ground truth, two synthesized ground-truth databases (VTRsynth and VTRsynthf0) are created using overlap-add synthesis of short-time AR sequences using the four formant frequency/bandwidth trajectory pairs in the VTR database. Finally, the performance of antiformant tracking using KARMA is illustrated with synthesized and spoken nasal phonemes.

A. Error analysis using a hand-corrected formant database

The VTR database contains a representative subset of the TIMIT speech corpus (Garofolo *et al.*, 1993) that consists of 516 diverse, phonetically balanced utterances collated across gender, individual speakers, dialects, and phonetic contexts (Deng *et al.*, 2006b). The database contains state information for the first four formant trajectory pairs (center frequency and bandwidth), where the first *three* center frequency trajectories were inspected after an initial automated pass (Deng *et al.*, 2004). Manual corrections were made using knowledge-based intervention based on the speech waveform, its wideband spectrogram, word- and phoneme-level transcriptions, and phonemic boundaries. Ground truth values from the VTR database were computed using formant modeling only with antiformant and other spectral information effectively absorbed by the bandwidth parameters of the modeled formants.

To provide a fair comparison with the VTR database, antiformants are not modeled and tracked in the KARMA algorithm. Analysis parameters of KARMA are set to the following values: $f_s = 7$ kHz, 20 ms Hamming windows with 50% overlap, $\gamma = 0.7$, $p = 12$ ($q = 0$), and $I = 3$ ($J = 0$). Thus each frame is fit with an ARMA(12, 0) model using the autocorrelation method of linear prediction and, subsequently, transformed to $N = 15$ cepstral coefficients via Eq. (5a). This baseline parameter set is listed in Table III. The extended Kalman smoother is applied for inter-frame tracking.

All formant tracks are coasted during frames labeled as non-speech. Although speech activity detection may be accomplished using a variety of approaches (e.g., simple energy thresholding in Rudoy *et al.*, 2007), errors can arise due to inadvertent assignment of inhalation/exhalation frames as speech. Thus for the VTR-based error analysis, the speech/non-speech decision is derived from TIMIT phoneme

labels so errors due to automatic speech activity detection would not confound performance metrics.

Available TIMIT phone transcriptions provide a speech sound label for each waveform sample. A frame is considered non-speech if all of its samples are labeled as a pause (*pau, epi, h#*), closure interval (*bcl, dcl, gcl, pcl, tcl, kcl*), or glottal stop (*q*). Speech-labeled frames are classified into one of six phonetic categories to reveal any phone-dependent error patterns. The six TIMIT phonetic categories are vowel, semivowel/glide, nasal, fricative, affricate, and stop. Each speech frame is assigned the phonetic category that labels a majority of the samples within the frame where the given ordering indicates priority in the case of ties.

WAVESURFER estimates are computed using the "formant" command of the SNACK SOUND TOOLKIT (Sjölander, 2005). PRAAT estimates are obtained using the "To Formant (burg)…" command (Boersma and Weenink, 2009). Preprocessing and intra-frame analysis parameters are matched to that of KARMA, and default smoothing settings are set within WAVESURFER and PRAAT.

Table IV summarizes the performances of KARMA, WAVESURFER, and PRAAT on the VTR database. The cepstral-based KARMA approach results in lower overall error compared to the classical algorithms with particular improvement for tracking through frames containing obstruents (fricative, affricate, or stop) relative to WAVESURFER and PRAAT performance. The lower error potentially stems from the smoothly varying nature of estimated formant tracks by the KARMA algorithm relative to those of the benchmark approaches.

Figure 3 illustrates the formant tracks output from the three algorithms for VTR utterance 19 spoken by an adult female. Compared to the WAVESURFER and PRAAT tracks, KARMA trajectories are smoother and better behaved, reflecting the slow-moving nature of the speech articulators. The classical algorithms exhibit errant tracking of f_3 during the word "lion" at 1.5 s that is handled by KARMA. KARMA formant estimates "coast" linearly through non-speech frames during which the estimate uncertainties increase. Uncertainty shading in the plots are ovoid due to the backward pass of the Kalman smoother.

TABLE IV. Formant tracking performance of KARMA, WAVESURFER, and PRAAT in terms of root-mean-square error (RMSE) taken per formant across all 516 utterances in the VTR database (Deng *et al.*, 2006b). Reported RMSE (in Hz) is computed over speech-labeled frames and further categorized by 6 phonetic classes.

	KARMA			WAVESURFER			PRAAT		
Phonetic class	f_1	f_2	f_3	f_1	f_2	f_3	f_1	f_2	f_3
Vowel	82	258	336	112	254	262	134	269	341
Semivowel/glide	104	336	437	128	229	320	139	295	474
Nasal	112	292	297	176	478	357	216	405	327
Fricative	177	231	316	231	359	423	306	291	322
Affricate	201	239	276	266	418	374	331	310	299
Stop	172	274	328	220	328	370	247	269	343
Overall per formant	123	267	341	163	308	326	202	291	353
Overall		260			276			289	

Table V reports the influence of gender on tracking performance. The VTR database consists of 322 utterances by male speakers and 194 by female speakers. KARMA error is lower than that of the benchmark algorithms on female speakers, whose speech waveforms have been traditionally difficult to analyze spectrally due to high fundamental frequencies. Note that f_3 tracking performance using KARMA seems to be highly dependent on gender. Differences in performance on males and females can also be ascribed to variations in vocal tract length. Cepstral coefficients, derived from the log-magnitude spectrum, are dependent on spectral slope effects that are directly tied to vocal tract length. As discussed in Olive (1971), the log-magnitude spectrum allows the use of fixed formant bandwidths in the KARMA model because formant peaks become less important in the spectral fit, whereas the breakpoints in spectrum slope relating to the formant frequencies become more important.

Exploring these issues on a case-by-case basis reveals that KARMA may track f_3 through spectral regions with fourth formant energy. Because vocal tract resonances of male speakers are lower, on average, than those of female speakers, the 3500 Hz speech bandwidth may include fourth-formant energy bands. In these cases, KARMA tracking can be improved by increasing the number of tracked formants I or by reducing the resampling rate f_s and thus the effective speech bandwidth. Some analyses might benefit from utterance-dependent estimation for $\boldsymbol{x}_0$ instead of the application of fixed baseline values. These results show the robustness of Kalman tracking across utterances but also acknowledge the appropriate selection of parameters in certain cases.

Figure 4 illustrates the effects of bandwidth tracking and $\boldsymbol{Q}$ on KARMA performance. In particular, Fig. 4(A) shows that the f_2 track with baseline parameters underestimates the second formant of the utterance during obstruents (at 0.75 and 1.5 s) and frames that exhibit rapid formant transitions around high front vowels (most notably at 0.25 s). Both WAVESURFER and PRAAT exhibit similar error. Fixing bandwidths in the KARMA approach usually yields accurate overall tracking performance (Table IV), which is expected due to the general difficulties related to bandwidth estimation in speech (Iseli *et al.*, 2007). When bandwidths are allowed to vary, the RMSE across all formant trajectories in the VTR database is 320 Hz compared with an error of 260 Hz when bandwidths are fixed.

Figure 4(B) displays the KARMA output with bandwidth tracking enabled. For this utterance, f_2 tracking improves when bandwidths b_i are included in the state vector $\boldsymbol{x}_t$, especially during the obstruent segments and rapid second formant transitions at 1.75 and 2.25 s. Two other regions with large second formant slopes must be addressed further: /ɹi/ in "reading" at 0.25 s and /ɑi/ in "light" at 1.25 s.

Increasing the diagonal elements of $\boldsymbol{Q}$ to $(949\,\text{Hz})^2$ permits tracks like f_2 to respond faster to formant transitions that change by 1000 Hz from frame to frame. Figure 4(C) shows output tracks with this larger variance parameter. KARMA performance improves in the regions mentioned but with the concomitant characteristics of less smoothly varying trajectories and increased estimator uncertainty. Knowledge of such a bias-variance trading relation due to

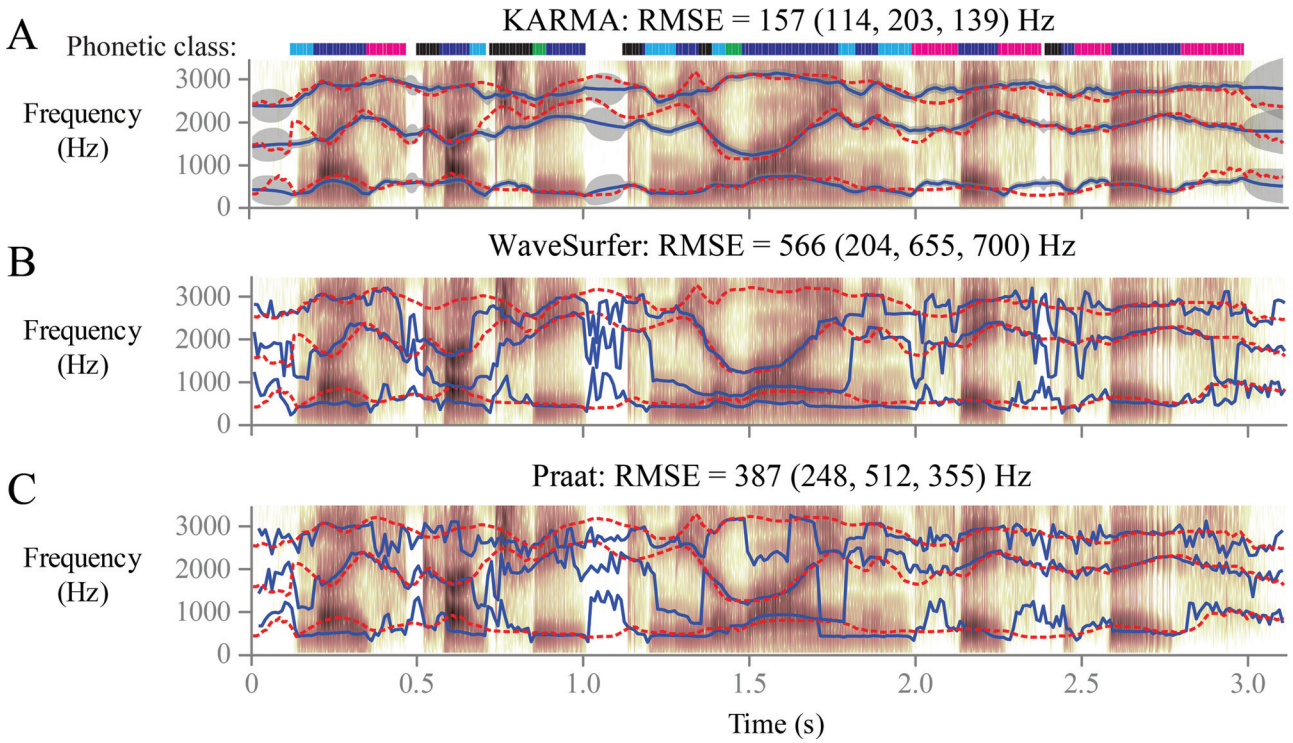

FIG. 3. Estimated formant tracks on spectrogram of VTR utterance 19 by an adult female from New England: "Nice country to meet a lion in face to face." Reference trajectories from the VTR database are shown (red, dashed) along with the formant frequency tracks (blue, solid) from (A) KARMA, (B) WAVESURFER, and (C) PRAAT. Overall root-mean-square error (RMSE) is reported across all formants and frames labeled as speech, in addition to separate RMSE values for f_1, f_2, and f_3. The KARMA output additionally displays uncertainty (gray shading, ± 1 standard deviation) for each formant trajectory. Frames are categorized using TIMIT labels of phonetic class: vowel (blue), semivowel/glide (green), nasal (cyan), fricative (magenta), affricate (red), stop (black).

parameter settings allows the user to judiciously employ the proposed algorithm to handle certain phonetic contexts.

B. Error analysis using synthesized databases

The VTRsynth database consists of synthesized speech waveforms through overlap-add of frames that each follow the ARMA model of Eq. (2). ARMA spectral coefficients in each frame are derived from the four formant frequency/bandwidth pairs in the corresponding frame of the VTR database utterance using the impulse-invariant transformation of a digital resonator (Klatt, 1980). The Klatt-like synthesizer has a parameter similar to Klatt's TL parameter (Klatt, 1980) that modifies overall spectral tilt in addition to the spectral tilt inherently created by the resonance peaks and tails. In the synthesized databases of the current study, the TL parameter is set to 0 dB. The source excitation is white Gaussian noise, and no distinction is made between speech and non-speech frames. Synthesis parameters are set to the following values: $f_s = 16$ kHz, 20 ms Hanning windows with 50% overlap, and $p = 8$ ($q = 0$).

TABLE V. Formant tracking performance of KARMA, WAVESURFER, and PRAAT in terms of root-mean-square error (RMSE) taken per formant across all 516 utterances in the VTR database (Deng *et al.*, 2006b). RMSE (in Hz) is reported over speech-labeled frames and further categorized by speaker gender (male, female).

Formant	KARMA	WAVESURFER	PRAAT
f_1	123 (124, 121)	163 (152, 180)	202 (197, 209)
f_2	267 (239, 307)	308 (246, 388)	291 (251, 345)
f_3	341 (373, 280)	326 (320, 335)	353 (363, 337)
Overall	260 (266, 250)	276 (249, 314)	289 (279, 304)

The second synthesized database (VTRsynthf0) introduces a model mismatch between synthesis and analysis by applying the Rosenberg C derivative source waveform (Rosenberg, 1971) instead of white noise for voiced frames. Synthesis parameters are set as in the VTRsynth database. The fundamental frequency of each voiced frame is taken from WAVESURFER estimates of the original VTR utterance. The VTRsynthf0 database thus includes voiced, unvoiced, and silence frames. Synthesized formant trajectories act as truer ground truth contours than those in the VTR database to test the performance of the cepstral-based KARMA inference.

Tables VI and VII display algorithmic performance on the VTRsynth and VTRsynthf0 databases, respectively. Parameters of the three tested algorithms are set as described in the previous section except bandwidths are tracked in the KARMA approach. KARMA performance compares well with that of WAVESURFER and PRAAT with error significantly higher for the synthesized waveforms with higher fundamental frequencies (derived from utterances by female speakers). The similar overall error of KARMA and WAVESURFER on both synthesized databases provides validation for the use of LPC cepstral coefficients as observations as an alternative to LPC spectral coefficients. Some of the slightly higher error is due to the initial startup time during which the KARMA trajectories lock onto their respective tracks.

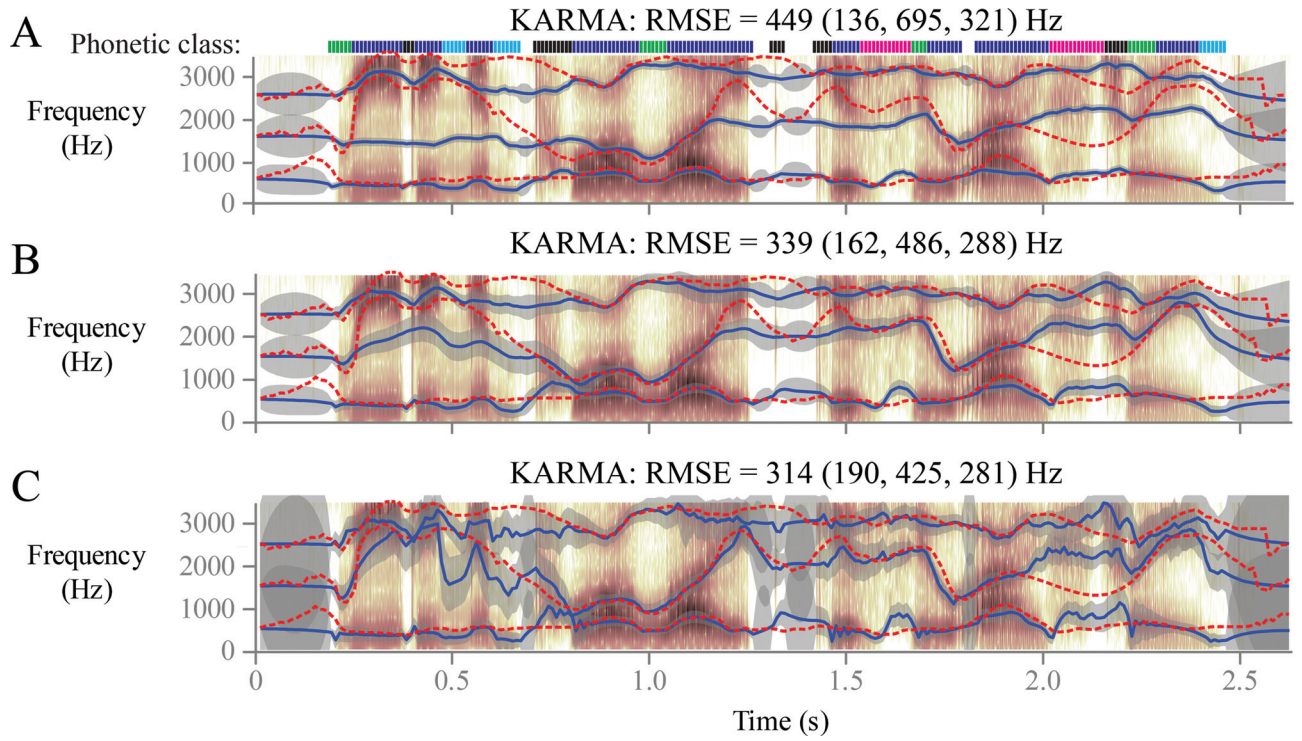

FIG. 4. Effect of bandwidth tracking and state covariance matrix $\boldsymbol{Q}$ on KARMA formant tracking. VTR utterance 10 by an adult female from New England: "Reading in poor light gives you eyestrain." (A) Bandwidths fixed to baseline values in Table III with diagonal elements of $\boldsymbol{Q}$ equal to $(224\ \text{Hz})^2$, (B) bandwidth values tracked with $\boldsymbol{Q}$, as in (A), and (C) bandwidth values tracked with diagonal elements of $\boldsymbol{Q}$ increased to $(949\ \text{Hz})^2$. Overall root-mean-square error (RMSE) is reported across all formants and frames labeled as speech in addition to separate RMSE values for f_1, f_2, and f_3. Color coding as in Fig. 3.

C. Antiformant tracking

The KARMA approach to formant and antiformant tracking is illustrated in this section. Synthesized and real speech examples are presented to determine the ability of the ARMA-derived cepstral coefficients to capture pole and zero information.

1. Synthesized waveform

In the synthesized case, a speech-like waveform /nɑn/ is generated with varying frame-by-frame formant and antiformant characteristics and a periodic source excitation ($f_0 = 100$) as was implemented for the VTRsynthf0 database. The /nɑn/waveform is synthesized at $f_s = 10\,\text{kHz}$ using 100 ms frames with 50% overlap. Formant frequencies (bandwidths) of the /n/ phonemes are set to 257 Hz (32 Hz) and 1891 Hz (100 Hz). One antiformant is placed at 1223 Hz (bandwidth of 52 Hz) to mimic the location of an alveolar nasal antiformant. Formant frequencies (bandwidths) of /ɑ/ are are set to 850 Hz (80 Hz) and 1500 Hz (120 Hz). A random term with zero mean and standard deviation of 10 Hz is added to each trajectory to simulate realistic variation.

Figure 5 shows the results of formant and antiformant tracking using KARMA on the synthesized phoneme string /nɑn/. Two different visualizations are displayed. Figure 5(A) plots point estimates and uncertainties of the center frequency *and* bandwidth trajectories for each frame. Figure 5(B) displays the wideband spectrogram with overlaid center frequency tracks whose widths reflect the corresponding 3-dB bandwidth values. Note that the length of the state vector in the state-space model is modified depending on the presence or absence of antiformant energy. Estimated trajectories fit the ground truth values well once initialized values reach a steady state.

2. Spoken nasal consonants

During real speech, a vocal tract configuration consisting of multiple acoustic paths results in the possible existence of

TABLE VI. RMSE (in Hz) of KARMA, WAVESURFER, and PRAAT formant tracking of the first three formant trajectories in the VTRsynth database that resynthesizes utterances using a stochastic source.

Formant	KARMA	WAVESURFER	PRAAT
f_1	32	47	74
f_2	69	73	163
f_3	79	67	176
Overall	63	63	145

TABLE VII. RMSE of KARMA, WAVESURFER, and PRAAT formant tracking of the first three formant trajectories in the VTRsynthf0 database that resynthesizes VTR database utterances using stochastic and periodic sources. RMSE (in Hz) is reported over speech-labeled frames and further categorized by original speaker gender (male, female) to reveal any fundamental frequency effects.

Formant	KARMA	WAVESURFER	PRAAT
f_1	51 (50, 52)	73 (55, 96)	76 (74, 80)
f_2	84 (64, 110)	57 (50, 68)	197 (107, 289)
f_3	96 (65, 131)	57 (44, 74)	186 (101, 272)
Overall	79 (60, 103)	63 (50, 80)	163 (95, 234)

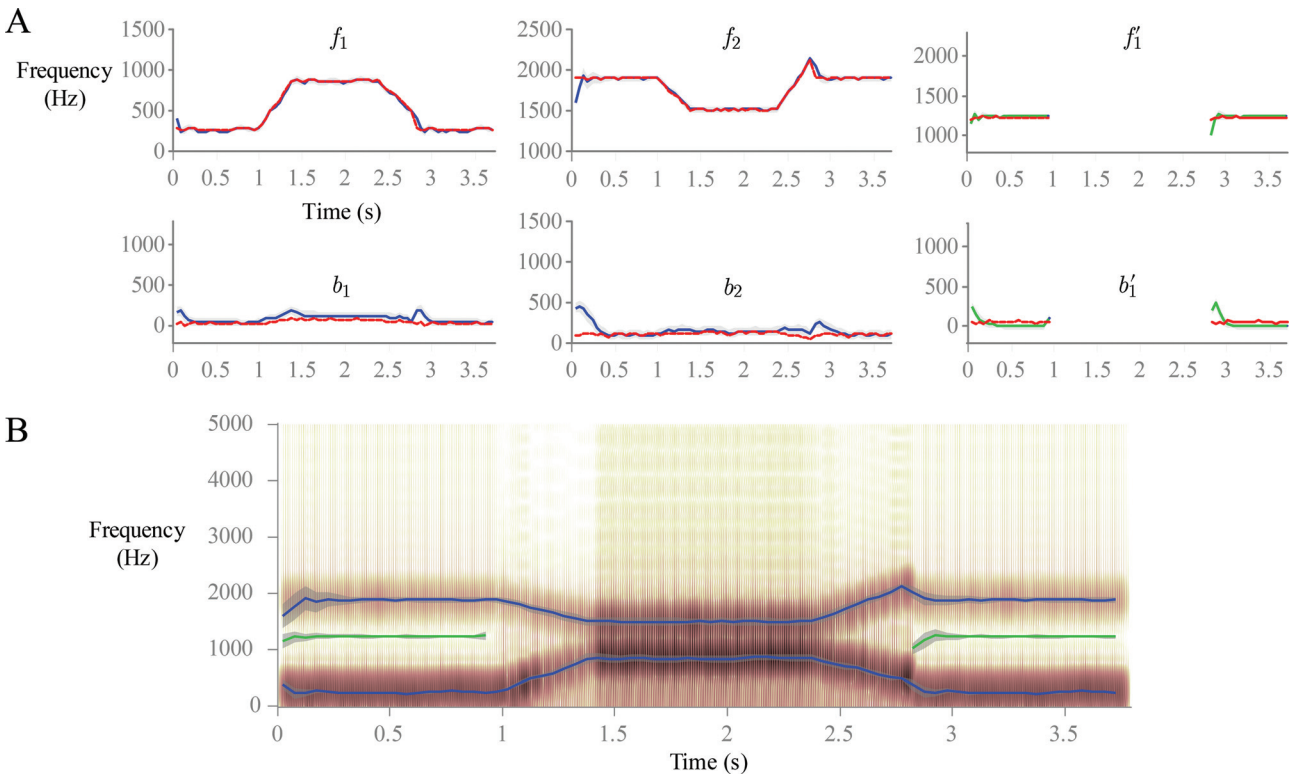

FIG. 5. Illustration of the output from KARMA for the synthesized utterance /nɑn/. (A) True trajectories (red, dashed) are shown with the mean estimates (solid blue for formants, solid green for antiformants) and uncertainties (gray shading) for each frequency and bandwidth. (B) plots an alternative display is shown with a wideband spectrogram along with estimated frequency and bandwidth tracks of formants (blue) and antiformants (green). The 3-dB bandwidths dictate the width of the corresponding frequency tracks.

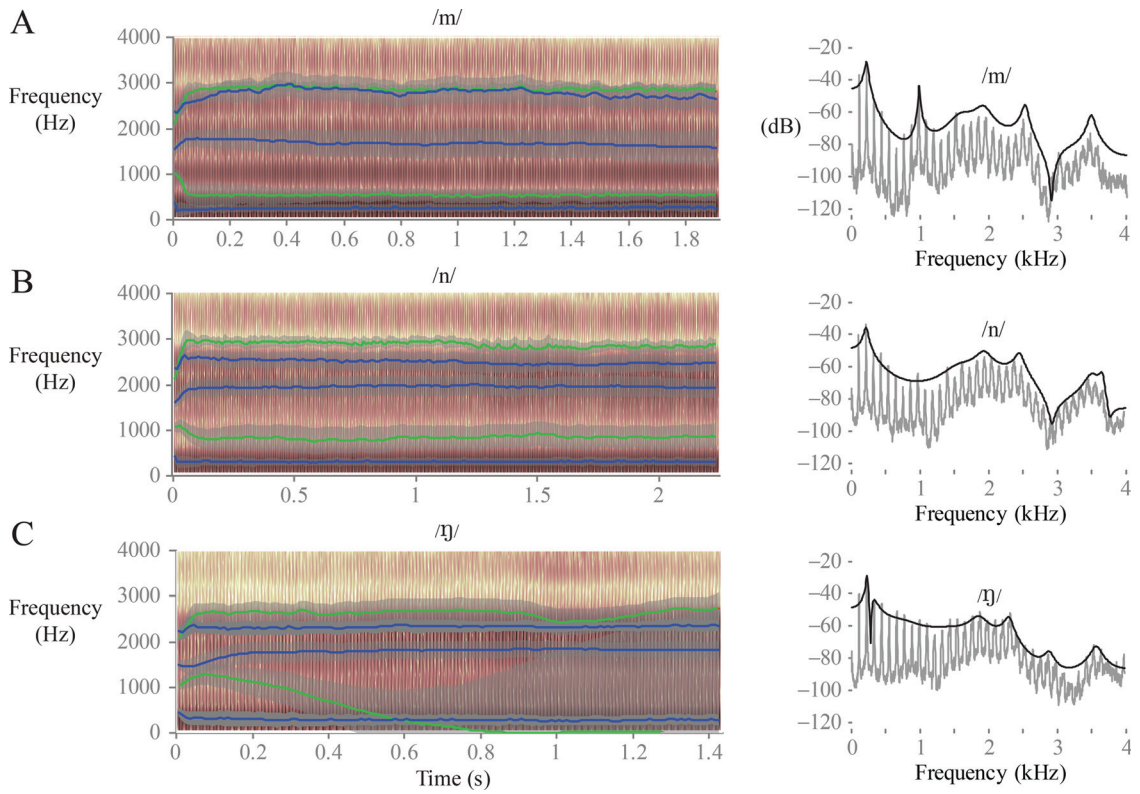

FIG. 6. KARMA output for three spoken nasal consonants: (A) /m/, (B) /n/, and (C) /ŋ/. On the left, spectrograms overlay the mean estimates (blue for formants, green for antiformants) and uncertainties (gray shading) for each frequency and bandwidth. Plots to the right display the corresponding periodogram (gray) and spectral ARMA model fit (black).

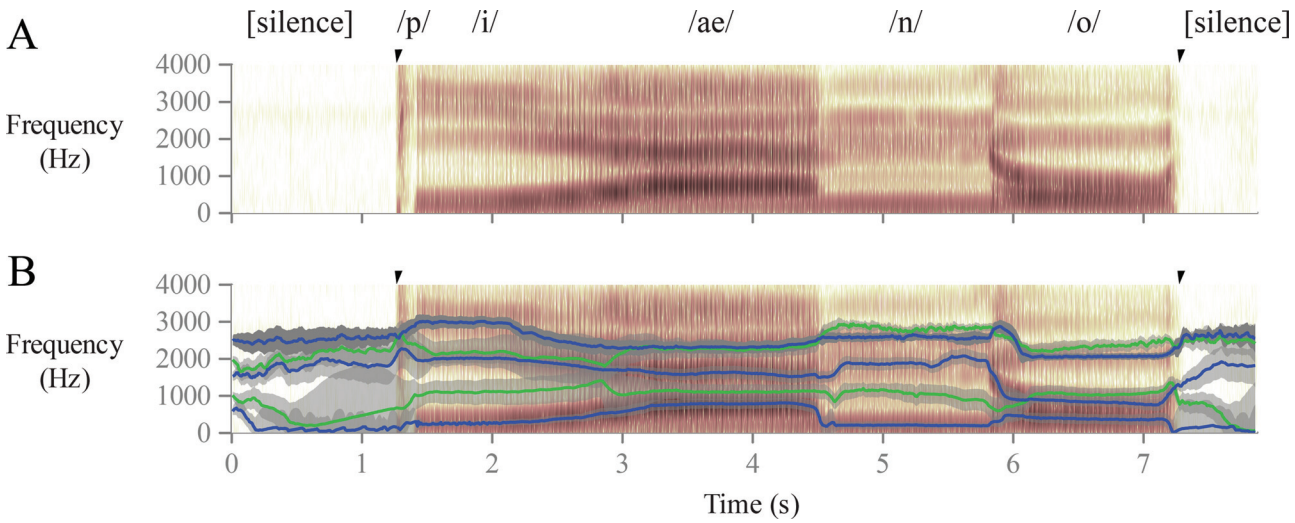

FIG. 7. KARMA formant and antiformant tracks of utterance by adult male: "piano." Displayed are the (A) wideband spectrogram of the speech waveform and (B) the spectrogram overlaid with formant frequency estimates (blue), antiformant frequency estimates (green), and uncertainties (±1 standard deviation) for each track (gray). Arrows indicate beginning and ending of utterance. Note that the increase in uncertainty during silence regions.

both poles and zeros in the transfer function $T(z)$ [Eq. (7)]. For example, in a nasal consonant configuration, acoustic energy from the glottal source travels through the pharynx toward both the oral cavity and the nasal cavities. Complete closure at the end of the oral cavity introduces antiformants in the overall transfer function.

Ignoring the effects of the sinus cavities (Pruthi *et al.*, 2007), the frequency of the lowest antiformant depends on the length of the oral cavity, which in turn depends on tongue position. For the labial nasal consonant /m/, the frequency of this antiformant is approximately 1100 Hz (Stevens, 1998, p. 495). As the point of closure moves toward the back of the oral cavity—such as for the alveolar and velar nasal consonants—the length of the resonating cavity decreases and the frequency of the antiformant increases. The frequency of a second antiformant is approximately three times the frequency of this antiformant due to the quarter-wavelength oral cavity configuration.

KARMA performance is evaluated visually on spoken nasal consonants produced with closure at the labial (/m/), alveolar (/n/), and velar (/ŋ/) positions. The extended Kalman smoother is applied using an ARMA(16, 4) model, $f_s = 8$ kHz, 20 ms Hamming windows with 50% overlap, $N = 20$ cepstral coefficients, and $\gamma = 0.7$. The frequencies (bandwidths) of the formants are initialized as in Table III. The frequencies (bandwidths) of the antiformants are initialized to 1000 Hz (80 Hz) and 2000 Hz (80 Hz).

Figure 6 displays KARMA outputs (point estimate and uncertainty of center frequency tracks) and averaged spectra for the three sustained consonants. The KARMA algorithm takes a few frames to settle to its steady-state estimates. As expected, the frequency of the antiformant increases as the position of closure moves toward the back of the oral cavity. The uncertainty of the first antiformant of /ŋ/ increases significantly and indicates that this antiformant is not well observed in the waveform, agreeing with the

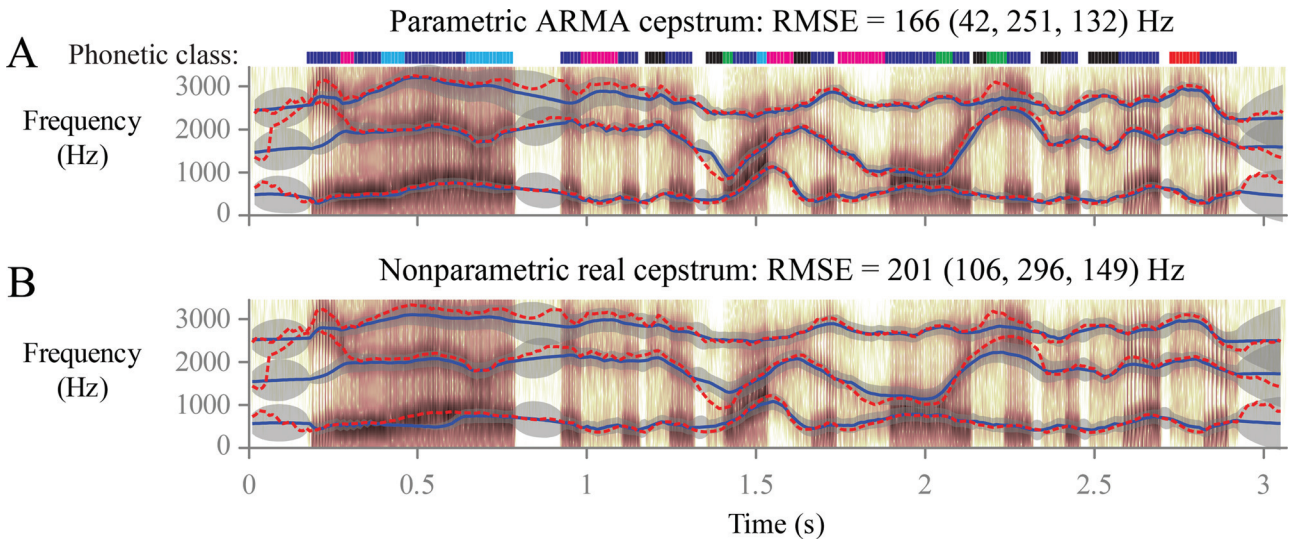

FIG. 8. Kalman-based formant tracks using the (A) parametric ARMA cepstrum and (B) nonparametric real cepstrum as observations. VTRsynth f0 waveform is a synthesized version of VTR utterance 1: "Even then, if she took one step forward, he could catch her." Color coding as in Fig. 3.

observation that velar articulation shortens the oral cavity length to such an extent that antiformants may not be generated (Stevens, 1998, p. 507). Note that the inclusion of zeros improves the ability of the ARMA model to fit the underlying waveform spectra.

Finally, the KARMA tracker is applied to the spoken word "piano" to determine if the antiformant tracks would capture any zeros during the nasal phoneme. Figure 7 displays the KARMA formant and antiformant tracks with their associated uncertainties. During the non-nasalized regions, the uncertainty around the point estimates of the antiformant track is large, reflecting the lack of antiformant information. During the /n/ segment, the uncertainty of the antiformant tracks decreases to reveal underlying antiformant information.

IV. DISCUSSION

In this article, the task of tracking frequencies and bandwidths of formants and antiformants is approached from a statistical point of view. The evolution of parameters is cast in a state-space model to provide access to point estimates and uncertainties of each track. The key relationship is a linearized mapping between cepstral coefficients and formant and antiformant parameters that allows for the use of the extended family of Kalman inference algorithms.

The VTR database provides an initial benchmark of ground truth for the first three formant frequency values to which multiple algorithm outputs can be compared. The values in the VTR database, however, should be interpreted with caution because baseline tracks were initially obtained via a first-pass automatic algorithm (Deng *et al.*, 2004). It is unclear how much manual intervention was required and what types of errors were corrected. In particular, VTR tracks are observed to not always track high-energy spectral regions. Despite the presence of various labeling errors in the VTR database, it is still useful to obtain initial performance metrics of formant tracking algorithms on real speech.

In the current framework, cepstral coefficients are derived from the spectral coefficients of the fitted stochastic ARMA model (Sec. II B). Source information related to phonation is thus separated from vocal tract resonances by assuming that the source is a white Gaussian noise process. This is not the case in reality, especially for voiced speech, where the source excitation component has its own characteristics in frequency (e.g., spectral slope) and time (e.g., periodicity). This model mismatch has been explored here via VTRsynthf0, although we note that it is also possible to incorporate more sophisticated source modeling through the use flexible basis functions such as wavelets (Mehta *et al.*, 2011).

An alternative approach to ARMA modeling is to compute the nonparametric (real) cepstrum directly from the speech samples. Based on the convolutional model of speech, low-frequency cepstral coefficients are largely linked to vocal tract information up to about 5 ms (Childers *et al.*, 1977), depending on the fundamental frequency. Skipping the 0th cepstral coefficient that quantifies the overall spectral level, this would translate to including up to the first 35 cepstral coefficients in the observation vector $\boldsymbol{y}_t$ in the state-space framework [Eqs. (12a) and (12b)].

Although coefficients from the real cepstrum do not strictly adhere to the observation model derived in Eq. (11), the approximate separation of source and filter in the nonparametric cepstral domain makes this approach viable. Some insight into the reasons why the real cepstrum provides a meaningful acoustic observation space comes from the ability of the logarithm function to yield a linear superposition of contributions from different fixed formants. Further analysis of "noise" in the estimated ARMA cepstrum can also be expected to improve overall robustness in the presence of various sources of uncertainty (Tourneret and Lacaze, 1995).

Figure 8 illustrates the output of the algorithm using the first 35 coefficients of the real cepstrum as observations. Interestingly the performance of the nonparametric cepstrum is comparable to that of the parametric ARMA cepstrum, especially for the first formant frequency. Most of the error stems from underestimating the second and third formant frequencies. Advantages to using the nonparametric cepstrum include computational efficiency and freedom from ARMA model constraints.

The capability of automated methods to track the third formant strongly depends on the resampling frequency, which controls the amount of energy in the spectrum at higher frequencies. For example, if the signal was resampled to 10 kHz, a given algorithm might erroneously track the third formant frequency through spectral regions typically ascribed to the fourth formant. Traditional formant tracking algorithms have access to multiple candidate frequencies, which are constantly re-sorted so that $f_4 > f_3 > f_2 > f_1$. In the proposed statistical approach, the ordering of formant indices is inherent in the mapping of formants to cepstral coefficients [Eq. (11)], and further empirical study of the formants-to-cepstrum mapping can be expected to lead to improved methods when there are additional resonances present in the speech bandwidth.

Overall, the proposed KARMA approach compares favorably with WAVESURFER and PRAAT in terms of RMSE. WAVESURFER exhibits fairly accurate performance during voiced regions, while less smooth trajectories are output for obstruent phonemes. RMSE, however, is only one selected error metric, which must be validated by observing how well raw trajectories behave. The proposed KARMA tracker yields smoother outputs and offers parameters that allow the user to tune the performance of the algorithm in a statistically principled manner. Such well behaved trajectories may be particularly desirable for the resynthesis of perceptually natural speech and automatic speech recognition systems.

It has been observed in the speech analysis literature that it is difficult to obtain accurate formant bandwidth estimates (Hanson and Chuang, 1999); this has sometimes led to the use of *a priori* fixed bandwidth values in lieu of their estimates (Olive, 1971; Iseli *et al.*, 2007). Fixing bandwidths was found to improve the performance of KARMA on the VTR database when error was taken across all formant tracks. These results are not unexpected because the Cramér–Rao bounds for AR parameter estimation indicate

that the lower bounds on estimator variances are different for each AR coefficient (Friedlander, 1984; Friedlander and Porat, 1989). These differences are potentially further amplified by the nonlinear transformation of the AR coefficients into frequency/bandwidth pairs. We have observed, empirically, that the variance of bandwidth estimators may be an order of magnitude larger than the variance of the formant frequency estimators.

Although considerably more complex and more sensitive to model assumptions, a time-varying autoregressive moving average (TV-ARMA) model has been previously proposed for formant and antiformant tracking (Toyoshima *et al.*, 1991) with little follow-up investigation. In their study, Toyoshima *et al.* used an extended Kalman filter to solve for ARMA spectral coefficients at each speech sample. One real zero and one real pole were included to model changes in gross spectral shape over time. While a frame-based approach (as taken in KARMA) appears to yield salient parameters at a lower computational cost, future work could consider alternative time-varying models (Rudoy *et al.*, 2011) and spectral-shaping parameters to help explain nonlinear source-filter coupling effects (Titze, 2008).

Antiformant tracking remains a challenging task in speech analysis. Antiformants are typically less strong than their resonant counterparts during nasalized phonation, and the estimation of subglottal resonances continues to rely on empirical relationships rather than direct acoustic observation (Arsikere *et al.*, 2011). Phone recognizers and spontaneous speech recognition systems that only incorporate formant information in their cepstral mapping could potentially benefit from antiformant trajectory modeling (Ma and Deng, 2000, 2004; Deng *et al.*, 2006d; Deng *et al.*, 2006c; Deng and Ma, 2000). Previously any antiformant information that is not explained by the formants-to-cepstrum mapping was absorbed into the bandwidth parameters, residual vectors, or observation noise (Deng *et al.*, 2007). The proposed KARMA approach allows the user the option of tracking antiformants during speech regions of interest. Potential improvements here include the use of formal statistical tests for detecting the presence of zeros within a frame prior to tracking them.

V. CONCLUSIONS

This article has presented KARMA, a Kalman-based autoregressive moving average modeling approach to formant and antiformant tracking. The contributions of this work are twofold. The first is methodological with improvements to the Kalman-based AR approach of Deng *et al.* (2007) and extensions to enable antiformant frequency and bandwidth tracking in an ARMA framework. The second is empirical with visual and quantitative error analysis of the KARMA algorithm to demonstrate improvements over two standard speech processing tools, WAVESURFER (Sjölander and Beskow, 2005) and PRAAT (Boersma and Weenink, 2009). The full KARMA algorithm provides for several features, such as zero modeling and formant cross-correlation, that can be activated or deactivated depending on the specific speech segment (e.g., nasalized phonemes).

It is expected that additional improvements will come with better understanding of precisely how formant information is captured through this class of nonlinear ARMA (or nonparametric) cepstral coefficient models. As noted, antiformant tracking remains challenging, although it has been shown here that appropriate results can be obtained for selected cases exhibiting antiformants. The demonstrated effectiveness of this approach, coupled with its ability to capture uncertainty in the frequency and bandwidth estimates, yields a statistically principled tool appropriate for use in clinical and other applications where it is desired, for example, to quantitatively assess acoustic features such as nasality, subglottal resonances, and coarticulation.

ACKNOWLEDGMENTS

This work was supported in part by the Army Research Office under PECASE Award W911NF-09-1-0555 and by the Office of Naval Research under MURI Award 58153-MA-MUR. The work of D.R. was supported by a National Defense Science and Engineering Fellowship.

APPENDIX: DERIVATION OF CEPSTRAL COEFFICIENTS FROM THE ARMA SPECTRUM

Assume an ARMA process with the minimum-phase rational transfer function

$$T(z) \triangleq \frac{B(z)}{A(z)} = \frac{1 - \sum_{j=1}^{q} b_j z^{-j}}{1 - \sum_{i=1}^{p} a_i z^{-i}}, \tag{A1}$$

which in turn implies a right-sided complex cepstrum. For the moment, assume $b_j = 0$ for $1 \le j \le q$ to initially derive the all-pole LPC cepstrum, the $\mathcal{Z}$-transform of which is denoted by $C(z)$:

$$C(z) \triangleq \log T(z) = \sum_{n=0}^{\infty} c_n z^{-n}, \tag{A2}$$

where

$$c_n = \frac{1}{2\pi} \oint_{z=e^{i\omega}} (\log T(z)) z^{n-1} dz$$

is the nth coefficient of the LPC cepstrum. Using the chain rule, $dT(z)/dz^{-1}$ can be obtained independently from Eq. (A1) or Eq. (A2), yielding the relation

$$\frac{dC(z)}{dz^{-1}} = \frac{1}{T(z)} \frac{dT(z)}{dz^{-1}} = \frac{\sum_{i=1}^{p} i a_i z^{-i+1}}{1 - \sum_{i=1}^{p} a_i z^{-i}},$$

which implies that

$$\frac{\sum_{i=1}^{p} i a_i z^{-i+1}}{1 - \sum_{i=1}^{p} a_i z^{-i}} = \sum_{n=0}^{\infty} c_n \frac{d}{dz^{-1}} (z^{-n}) = \sum_{n=0}^{\infty} n c_n z^{-n+1}.$$

Rearranging terms, we obtain

$$\sum_{n=0}^{\infty} nc_n z^{-n+1} = \sum_{i=1}^{p} ia_i z^{-i+1} + \sum_{i=1}^{p} a_i z^{-i} \sum_{n=0}^{\infty} nc_n z^{-n+1}. \tag{A3}$$

Using Eq. (A3), we can match the coefficients of terms on both sides with equal exponents. In the constant-coefficient case (associated to z^0), we have $c_1 = a_1$. For $1 < n \leq p$, we obtain

$$c_n = a_n + \sum_{i=1}^{n-1} \frac{n-i}{n} a_i c_{n-i} = a_n + \sum_{i=1}^{n-1} \left(1 - \frac{i}{n}\right) a_i c_{n-i}.$$

On the other hand, if $n > p$, then Eq. (A3) implies that

$$c_n = \sum_{i=n-p}^{n-1} \frac{n-i}{n} a_i c_{n-i} = \sum_{i=1}^{n-1} \left(1 - \frac{i}{n}\right) a_i c_{n-i}.$$

In summary, we have obtained the following relationship between the prediction polynomial coefficients and the complex cepstrum:

$$c_n = \begin{cases} a_1 & \text{if} \quad n = 1 \\ a_n + \sum_{i=1}^{n-1} \left(\frac{n-i}{n}\right) a_i c_{n-i} & \text{if} \quad 1 < n \leq p \\ \sum_{i=n-p}^{n-1} \left(\frac{n-i}{n}\right) a_i c_{n-i} & \text{if} \quad p < n. \end{cases}$$

Reversing the roles of i and $(n-i)$ yields the all-pole version in Eq. (5a).

To allow for nonzero b_j coefficients in Eq. (A1), we separate contributions from the numerator and denominator of Eq. (A1) to obtain the ARMA cepstral coefficients C_n:

$$\begin{aligned} C_n &= \mathcal{Z}^{-1} \log T(z) \\ &= \mathcal{Z}^{-1} \log \frac{1}{A(z)} - \mathcal{Z}^{-1} \log \frac{1}{B(z)} \\ &= c_n - c'_n, \end{aligned}$$

yielding the respective pole and zero recursions of Eqs. (5a) and (5b).

Arsikere, H., Lulich, S. M., and Alwan, A. (**2011**). "Automatic estimation of the first subglottal resonance," J. Acoust. Soc. Am. **129**, EL197–EL203.

Atal, B. S., and Hanauer, S. L. (**1971**). "Speech analysis and synthesis by linear prediction of the speech wave," J. Acoust. Soc. Am. **50**, 637–655.

Atal, B. S., and Schroeder, M. R. (**1978**). "Linear prediction analysis of speech based on a pole-zero representation," J. Acoust. Soc. Am. **64**, 1310–1318.

Boersma, P., and Weenink, D. (**2009**). "PRAAT: Doing phonetics by computer (version 5.1.40) [computer program]," http://www.fon.hum.uva.nl/praat/ (Last viewed July 13, 2009).

Broad, D. J., and Clermont, F. (**1989**). "Formant estimation by linear transformation of the LPC cepstrum," J. Acoust. Soc. Am. **86**, 2013–2017.

Childers, D. G., Skinner, D. P., and Kemerait, R. C. (**1977**). "The cepstrum: A guide to processing," Proc. IEEE **65**, 1428–1443.

Christensen, R., Strong, W., and Palmer, E. (**1976**). "A comparison of three methods of extracting resonance information from predictor-coefficient coded speech," IEEE Trans. Acoust. **24**, 8–14.

Deng, L., Acero, A., and Bazzi, I. (**2006a**). "Tracking vocal tract resonances using a quantized nonlinear function embedded in a temporal constraint," IEEE Trans. Audio Speech Lang. Process. **14**, 425–434.

Deng, L., Cui, X., Pruvenok, R., Huang, J., Momen, S., Chen, Y., and Alwan, A. (**2006b**). "A database of vocal tract resonance trajectories for research in speech processing," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* pp. I-369–372.

Deng, L., Lee, L. J., Attias, H., and Acero, A. (**2004**). "A structured speech model with continuous hidden dynamics and prediction-residual training for tracking vocal tract resonances," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* pp. I-557–560.

Deng, L., Lee, L. J., Attias, H., and Acero, A. (**2007**). "Adaptive Kalman filtering and smoothing for tracking vocal tract resonances using a continuous-valued hidden dynamic model," IEEE Trans. Audio Speech Lang. Process. **15**, 13–23.

Deng, L., and Ma, J. (**2000**). "Spontaneous speech recognition using a statistical coarticulatory model for the vocal-tract-resonance dynamics," J. Acoust. Soc. Am. **108**, 3036–3048.

Deng, L., Yu, D., and Acero, A. (**2006c**). "A bidirectional target-filtering model of speech coarticulation and reduction: Two-stage implementation for phonetic recognition," IEEE Trans. Audio Speech Lang. Process. **14**, 256–265.

Deng, L., Yu, D., and Acero, A. (**2006d**). "Structured speech modeling," IEEE Trans. Audio Speech Lang. Process. **14**, 1492–1504.

Friedlander, B. (**1984**). "On the computation of the Cramer-Rao bound for ARMA parameter estimation," IEEE Trans. Acoust. **32**, 721–727.

Friedlander, B., and Porat, B. (**1989**). "The exact Cramer-Rao bound for Gaussian autoregressive processes," IEEE Trans. Aerosp. Electron. Syst. **25**, 3–7.

Fulop, S. A. (**2010**). "Accuracy of formant measurement for synthesized vowels using the reassigned spectrogram and comparison with linear prediction," J. Acoust. Soc. Am. **127**, 2114–2117.

Garofolo, J. S., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren, N., and Zue, V. (**1993**). *TIMIT Acoustic-Phonetic Continuous Speech Corpus* (Linguistic Data Consortium, Philadelphia, PA).

Gordon, N. J., Salmond, D. J., and Smith, A. F. M. (**1993**). "Novel approach to nonlinear/non-Gaussian Bayesian state estimation," IEE Proc. F, Radar Signal Process. **140**, 107–113.

Hamilton, J. D. (**1994**). *Time Series Analysis* (Princeton University Press, Princeton, NJ), Sec. (11.1).

Hanson, H. M., and Chuang, E. S. (**1999**). "Glottal characteristics of male speakers: Acoustic correlates and comparison with female data," J. Acoust. Soc. Am. **106**, 1064–1077.

Iseli, M., Shue, Y.-L., and Alwan, A. (**2007**). "Age, sex, and vowel dependencies of acoustic measures related to the voice source," J. Acoust. Soc. Am. **121**, 2283–2295.

Julier, S., and Uhlmann, J. (**1997**). "A new extension of the Kalman filter to nonlinear systems," in *Proc. AeroSense: The 11th Int. Symp. on Aerospace/Defense Sensing, Simulation, and Controls*, pp. 182–193.

Julier, S., and Uhlmann, J. (**2004**). "Unscented filtering and nonlinear estimation," Proc. IEEE **92**, 401–422.

Kalman, R. E. (**1960**). "A new approach to linear filtering and prediction problems," Trans. ASME J. Basic Eng. **82**, 35–45.

Klatt, D. H. (**1980**). "Software for a cascade/parallel formant synthesizer," J. Acoust. Soc. Am. **67**, 971–995.

Kopec, G. (**1986**). "Formant tracking using hidden Markov models and vector quantization," IEEE Trans. Acoust. **34**, 709–729.

Ljung, L. (**1999**). *System Identification: Theory for the User* (Prentice Hall PTR, Upper Saddle River, NJ), Sec. (10.2).

Ma, J. Z., and Deng, L. (**2000**). "A path-stack algorithm for optimizing dynamic regimes in a statistical hidden dynamic model of speech," Comput. Speech Lang. **14**, 101–114.

Ma, J. Z., and Deng, L. (**2004**). "Target-directed mixture dynamic models for spontaneous speech recognition," IEEE Trans. Speech Audio Process. **12**, 47–58.

Marelli, D., and Balazs, P. (**2010**). "On pole-zero model estimation methods minimizing a logarithmic criterion for speech analysis," IEEE Trans. Audio Speech Lang. Process. **18**, 237–248.

McCandless, S. (**1974**). "An algorithm for automatic formant extraction using linear prediction spectra," IEEE Trans. Acoust. **22**, 134–141.

Mehta, D. D., Rudoy, D., and Wolfe, P. J. (**2011**). "Joint source-filter modeling using flexible basis functions," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 5888–5891.

Miyanaga, Y., Miki, N., and Nagai, N. (**1986**). "Adaptive identification of a time-varying ARMA speech model," IEEE Trans. Acoust. **34**, 423–433.
Nearey, T. M. (**1989**). "Static, dynamic, and relational properties in vowel perception," J. Acoust. Soc. Am. **85**, 2088–2113.
Olive, J. P. (**1971**). "Automatic formant tracking by a Newton-Raphson technique," J. Acoust. Soc. Am. **50**, 661–670.
Pruthi, T., Espy-Wilson, C. Y., and Story, B. H. (**2007**). "Simulation and analysis of nasalized vowels based on magnetic resonance imaging data," J. Acoust. Soc. Am. **121**, 3858–3873.
Rigoll, G. (**1986**). "A new algorithm for estimation of formant trajectories directly from the speech signal based on an extended Kalman-filter," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* pp. 1229–1232.
Rosenberg, A. E. (**1971**). "Effect of glottal pulse shape on the quality of natural vowels," J. Acoust. Soc. Am. **49**, 583–590.
Rudoy, D., Quatieri, T. F., and Wolfe, P. J. (**2011**). "Time-varying autoregressions in speech: Detection theory and applications," IEEE Trans. Audio Speech Lang. Process. **19**, 977–989.
Rudoy, D., Spendley, D. N., and Wolfe, P. J. (**2007**). "Conditionally linear Gaussian models for estimating vocal tract resonances," in *Proc. INTERSPEECH*, pp. 526–529.
Schafer, R. W., and Rabiner, L. R. (**1970**). "System for automatic formant analysis of voiced speech," J. Acoust. Soc. Am. **47**, 634–648.
Sjölander, K. (**2005**). "SNACK SOUND TOOLKIT (version 2.2.10) [computer program]," http://www.speech.kth.se/snack/ (Last viewed November 19, 2011).
Sjölander, K., and Beskow, J. (**2005**). "WAVESURFER FOR WINDOWS (version 1.8.5) [computer program], http://www.speech.kth.se/wavesurfer/ (Last viewed November 19, 2011).
Steiglitz, K. (**1977**). "On the simultaneous estimation of poles and zeros in speech analysis," IEEE Trans. Acoust. **25**, 229–234.
Stevens, K. N. (**1998**). *Acoustic Phonetics* (MIT Press, Cambridge, MA), Chap. 9.
Titze, I. R. (**2008**). "Nonlinear source-filter coupling in phonation: Theory," J. Acoust. Soc. Am. **123**, 2733–2749.
Tourneret, J.-Y., and Lacaze, B. (**1995**). "On the statistics of estimated reflection and cepstrum coefficients of an autoregressive process," Signal Process. **43**, 253–267.
Toyoshima, T., Miki, N., and Nagai, N. (**1991**). "Adapative formant estimation with compensation for gross spectral shape," Electron. Commun. Jpn. Part III: Fundam. Electron. Sci. **74**, 58–68.
Yegnanarayana, B. (**1978**). "Formant extraction from linear-prediction phase spectra," J. Acoust. Soc. Am. **63**, 1638–1640.
Zheng, Y., and Hasegawa-Johnson, M. (**2004**). "Formant tracking by mixture state particle filter," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* pp. I-565–568.