Speech Analysis and Synthesis Based on Dynamic Modes

Julio Vargas and Stephen McLaughlin, *Fellow, IEEE*

***Abstract*—In this paper, the source–filter model of speech production is adapted to represent the speech signal as the superposition and convolution of a dynamic source and resonant modes. The aim is to increase the resolution of the time-instantaneous-frequency representation of each of the individual contributions of different sections of the human phonatory system. We present a framework based on dynamic mode predictors and filters, which are adapted, using gradient-based techniques, to track the modal dynamics of speech yielding a representation which is free from quasi-stationary assumptions thus allowing flexible manipulation of the speech signal. Several examples are offered including intonation modifications to illustrate the potential of the proposed approach.**

***Index Terms*—Dynamic features, modal dynamics, instantaneous frequency, instantaneous pitch tracking, intonation modification, nonstationary models, speech resonances.**

I. Introduction

SPEECH is generated by a complex, nonlinear, nonstationary and multi-component process which originates from an intricate interaction between the constituents of the human phonatory system. Over the years numerous researchers have proposed models attempting to capture the nature of speech. While these methods have been successful in capturing elements of the speech generation process and in developing speech systems, these models are often constrained by restrictive assumptions.

If the speech signal is assumed to be locally stationary, the contribution of the vocal tract is manifested in the envelope of short-term spectral representations based on either the Fourier transform or on auto-regressive models. This supports the concept of a source-filter model of speech production inspired by the human phonatory system [1].

This paper builds on ideas that consider speech to be composed of modulated components and the concept of an analytic signal discussed in [2] which itself built on earlier work by Dudley [3], Cherry and Philips [4], and Flanagan [5]. These early works suffered from the estimate of the formants being affected by leakage from the neighboring formats. However, as Rao [2] pointed out, all of these or subsequent approaches ignored the detailed variations in amplitude and phase or frequency.

Some researchers have focused on modeling the speech wave as a quasi-periodic and quasi-stationary signal [6]. They rely on the short-term Fourier Transform to extract the evolution of the parameters which model the sinusoidal components yielding a representation with limited effectiveness in the differentiation of information related to glottis or vocal tract behavior. Despite the fact that the instantaneous frequency of each modeling component is estimated with increased accuracy in [7], marking a departure from the implied quasi-periodicity assumption, the instantaneous variability of the vocal tract is not captured due to the use of quasi-stationary spectral envelope estimators based on linear prediction.

In the search for alternative approaches to seeking an effective characterization of the nonlinear and time varying yet resonant nature of speech, researchers have proposed modeling speech resonances using sinusoids with instantaneously variable parameters (frequency and amplitude [8] or damping [9]). These models introduce a level of flexibility in the parametrization of the characteristics of the resonances allowing a dimensionality reduction in comparison to representations based on quasi-periodicity assumptions. Despite these achievements, the perceptual integrity of the speech signal is not entirely retained because of the use of band-pass filtering and windowing [10] which is required to separate and parameterize the modeling components.

In an attempt to separate vocal tract from source related information, considerable efforts have been devoted in extracting the spectral envelope from short term spectral representations. However, the instantaneous variability of the vocal tract shape and the limited separability of information obtainable in the spectral domain have defeated all efforts. Although sophisticated approaches have been explored recently to extract vocal-tract-related information relying on either multi-cycle correlation measures [11] or spline-based spectral smoothing [12], it can be argued that the separability achievable with these techniques is limited by their inability to faithfully represent vocal tract and glottal or source dynamics. A faithful representation would be one enabling the time-instantaneous-frequency resolvability of the acoustic contribution of different sections of the phonatory system which is an essential requirement to guarantee the necessary flexibility to perform prosodic modifications. This fact is evidenced in speech modification efforts focused on preserving continuity in speech dynamics associated with pitch [13] and resonant frequencies [14] driving the

Manuscript received September 30, 2010; revised January 18, 2011 and April 13, 2011; accepted April 15, 2011. Date of publication May 10, 2011; date of current version September 30, 2011. The work of J. Vargas was supported by the EPSRC via a Visiting Fellowship. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Engin Erzin.

J. Vargas is with the Escuela de Ingenieria Eléctrica, Universidad de Los Andes, La Hechicera, Mérida 5101, Venezuela (e-mail: vargjulio@gmail.com).

S. McLaughlin is with the Institute for Digital Communications, School of Engineering and Electronics, University of Edinburgh, Edinburgh EH9 3JL, U.K. (e-mail: steve.mclaughlin@ed.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2011.2151859

development of techniques that strive to extract instantaneous pitch [15] and resonance related features [17].

Recently proposed techniques to perform prosodic modifications, based on synchronously re-scaling pitch cycles of the linear prediction residual [18], [19], depend on windowing and are not focused on preserving continuity in speech dynamics. This is a limitation if these techniques are to be used to control the fine and continuous evolution of perceptually important parameters observed in real speech.

Based on the brief review given above, it is possible to infer that a challenge still remains to find an adaptable and integrity-preserving technique which does not rely on rigid and distorting predefined frames or filters [10] and which seeks to faithfully represent the dynamics of speech. This technique should focus on the extraction of physically and perceptually relevant information with instantaneous-time-frequency resolvability therefore taking into account the underlying instantaneous properties of the human phonatory and auditory mechanisms.

In this paper, we propose a framework for analysis which accounts for the instantaneously variable and multi-component nature of the speech signal. We seek to capture and preserve continuity in perceptually relevant spectral and pitch related parameters by focusing on the development of a technique free from predefined windowing and filtering. In order to achieve this goal, the speech signal is assumed to be composed of modes or exponential signals characterized by their instantaneous complex frequencies, these modes are termed dynamic modes (DMs) hereafter. The proposed approach seeks to reinforce the idea of extracting the dynamics of elementary components of the speech signal by focusing on the instantaneous complex-frequency features to present a flexible framework for speech analysis.

The DMs are convolved and added together to construct the speech signal and can be grouped according to their association with different components of the phonatory system. In order to extract information from real speech signals, the instantaneous parameters of a set of shadow mode predictors are adapted, relying on gradient-based approaches, to minimize instantaneous prediction measures. The idea of shadow modes was first discussed in the context of nonlinear dynamical systems (see Grebogi *et al.* [20]) and has been further utilized in prediction of nonlinear time series by Judd *et al.* [21]. In [20] the authors state, *For a physical system which exhibits chaos, in what sense does a numerical study reflect the true dynamics of an actual system*. They were focused on demonstrating that a true trajectory could be shown to track closely or *shadow* a noisy, (computer generated) trajectory. In [21], shadowing trajectories were used to assess the reliability of predictions of dynamical time series. In this paper, this idea is adopted seeking to develop shadow predictors that allow tracking the trajectory of components within the speech model.

The ideas introduced here represent a continuation and extension of the work presented in [17] where the contribution of the vocal tract was modeled as superposition of nonstationary exponential signals. Given that the dynamic interaction of vocal tract and glottal (or source) systems was not explored, the proposed model was insufficient to be applied to the task of performing prosodic modifications. In this work, these limitations are overcome with the introduction of a model incorporating vocal tract-source interaction and the glottal or source signal is represented as a sum of modes to account for the possible multimodal nature of the different types of phonation or voice sources. This idea is partially supported by the fact that the glottal waveform has been modeled, explicitly or implicitly, as an oscillatory waveform relying on analytical functions [22], physical models [22], nonlinear function models [23], [24] or hybrid approaches [25].

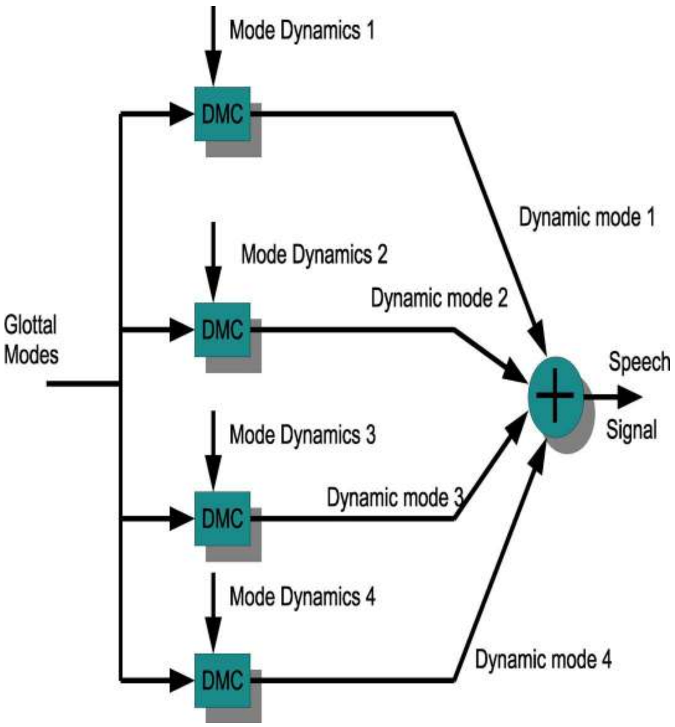

Fig. 1. Illustration of modeling the speech signal as a superposition of four dynamic modes. Each mode is the result of the dynamic mode convolution (DMC) of the glottal modes with the corresponding mode dynamics.

Although more research is required to support the idea of the multimodal nature of the source, in principle, a quasi-periodic glottal signal can be modeled as a superposition of harmonic components (modes). Additionally, the noise-like source for unvoiced speech could be potentially decomposed as a sum of nonharmonic modes [26].

The technique proposed in [17] to adaptively track speech resonances is revised here as a process of parallel de-emphasis and shadow resonant mode adaptation reinforcing a view focused on the intrinsic modal nature of speech. These ideas are extended to track the instantaneous pitch of speech by sharply focusing the "attention" of a shadow dynamic predictor on a fundamental mode of the glottal component of the speech signal. This increased attention is achieved by relying on dynamic Gabor and resonant modes guided by the dynamics of the shadow predictor. As an illustration of the potential of the proposed approach, finely controlled instantaneous pitch (intonation) modification is performed on speech signals by focusing on modifying the instantaneous frequency of the glottal modes to match a target pitch. This is done by relying on instantaneous-pitch-guided time-warping.

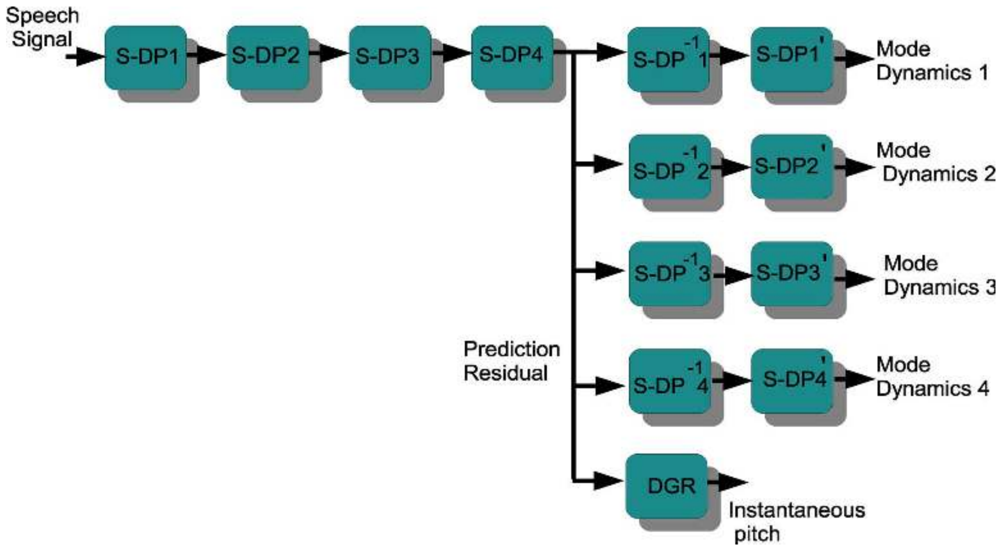

Fig. 2. Example illustrating the implementation of the required cascade for the case of four Shadow dynamic mode predictors (S-DPs) to extract mode dynamics and a dynamic Gabor-resonator (DGR) to track the instantaneous pitch.

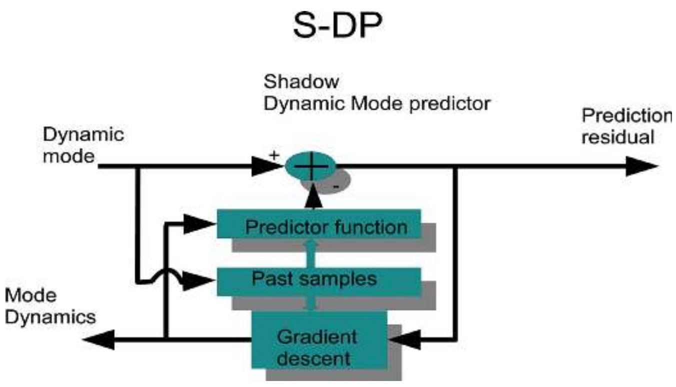

Fig. 3. Illustration of the implementation of a Shadow dynamic mode predictor.

This paper is organized as follows. In Section II, the concept of a dynamic signal is extended to introduce the idea of dynamic modes. After that, a new model of speech production is presented in conjunction with the idea of dynamic mode convolution concluding the section with a detailed description of the idea of dynamic mode predictors, shadow dynamic modes and dynamic Gabor modes. In Section III, the concepts which have been introduced are applied to the analysis of speech signals to extract resonance and pitch related information. The extracted information is then used to allow intonation modification based on instantaneous-pitch-guided time warping of the prediction residual illustrating the proposed approach with several experiments. Finally, the paper draws conclusions based on the analysis and results presented in the paper.

II. Dynamic Modes

The notion of instantaneous frequency of a signal is a concept that has existed for some considerable period of time [27], [28]. Poletti showed [29] how the conditional moments of frequency of a time–frequency distribution can be related to the derivative of the log of the corresponding signal, a complex function whose imaginary part is the instantaneous frequency of the signal. Poletti considered the complex signal

$$s(t) = A(t)e^{j\phi(t)} \tag{1}$$

with derivative

$$\begin{aligned}\frac{\partial s(t)}{\partial t} &= \left[\frac{\frac{\partial A(t)}{\partial t}}{A(t)} + j\frac{\partial \phi}{\partial t}\right] s(t)\\ &= \frac{\partial \log s(t)}{\partial t} s(t) = \zeta(t)s(t) \qquad (2)\end{aligned}$$

where $\zeta(t)$ is termed the dynamic signal. Its real part is the derivative of the log magnitude and its imaginary part is the instantaneous frequency. Consequently, $\zeta(t)$ describes the rate of change of the log magnitude and phase, i.e., the *signal dynamics*.

In this paper, by extending this idea of a dynamic signal to one of dynamic modes, an evolved model of speech production is presented focused on achieving resolution, in an instantaneous-time–frequency sense, on the acoustic contribution of the different components of the phonatory system. A dynamic mode (DM) can be defined as

$$m(t) = e^{\alpha(t)} \cos \phi(t) \tag{3}$$

with dynamics described by

$$\beta(t) = [\sigma(t), \omega(t)] = \left[\frac{\partial \alpha}{\partial t}, \frac{\partial \phi}{\partial t}\right]. \tag{4}$$

In the description above which represents the dynamics of the mode note that $\sigma(t)$ equates to $((\partial A(t))/(\partial t))/A(t)$ in (2) and is also known as the instantaneous bandwidth of the signal [30]. A DM is basically an AM–FM sinusoid and the idea is to allow the introduction of adaptive modes that can be used to analyze and represent the speech signal.

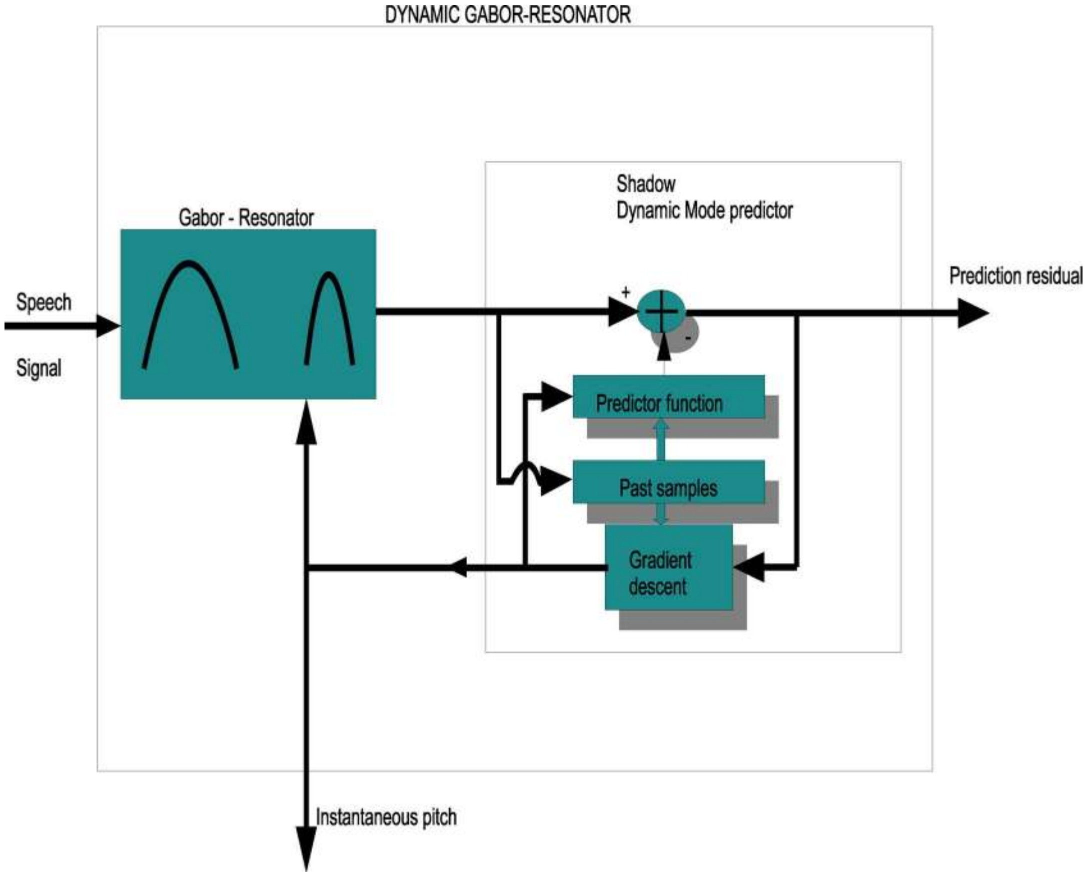

Fig. 4. Dynamic Gabor-Resonator.

A. Dynamic Modes to Model the Speech Signal

In this section, a new model of the speech signal is presented which seeks to characterize, from a time-instantaneous-frequency perspective, the acoustic contribution of different components of the human phonatory system. The model obtained can be seen as an evolution of the quasi-stationary model of speech production reinforcing the nonstationary multimodal and convolutional properties of the system studied. As will be shown in the next section, the dynamics of the proposed model can be extracted by relying on adaptive mode predictors and filters enabling flexible prosodic modifications.

We start by representing the speech signal as a superposition of resonances:

$$s(t) = \sum_{l=1}^{N_R} r_l(t) \tag{5}$$

where each resonance can be modeled as the dynamic convolution of a resonant DM (filter) m_r with a superposition of source modes

$$r_l(t) = m_{r,l}(t) \vec{*} \sum_{k=1}^{N_Q} m_{s,k}(t). \tag{6}$$

The concept of convolution is adapted here to account for the particular properties of dynamic modes. The ***dynamic mode convolution*** (DMC) of a given resonant mode $m(t)$ with any input mode, $m_{in}(t)$, can be defined as a function of the characteristic dynamics of the mode $\beta(t)$ and a vector of its past samples $\bar{m}(t)$

$$m(t) = m(t) \vec{*} m_{in}(t) = \Gamma\{m_{in}(t), \bar{m}(t), \beta(t)\}. \tag{7}$$

For example, for slow resonant dynamics and based on the current input sample and the past two output samples, (7) can be expressed as

$$\begin{aligned} m(t) = {} & m_{in}(t) + 2e^{\hat{\sigma}(t)\tau} \\ & \times \cos[\hat{\omega}(t)\tau] m(t-\tau) - e^{2\hat{\sigma}(t)\tau} m(t-2\tau) \end{aligned} \tag{8}$$

where τ is the sampling period. It should be noted that this dynamic convolution is **not** commutative.

The speech signal can then be finally expressed as

$$s(t) = \sum_{l=1}^{N_R} \left[m_{r,l}(t) \vec{*} \sum_{k=1}^{N_Q} m_{s,k}(t) \right] \tag{9}$$

indicating a concentration of energy on the resonant modes. This model is illustrated in Fig. 1.

By alluding to the physiological origin of the proposed modeling components, the nonstationary resonant modes can be associated with the vocal tract while the source modes can be associated with the glottis in the case of voiced speech.

A related model, assuming linearity and time invariance in discriminated glottal phases, was proposed in [31] to represent

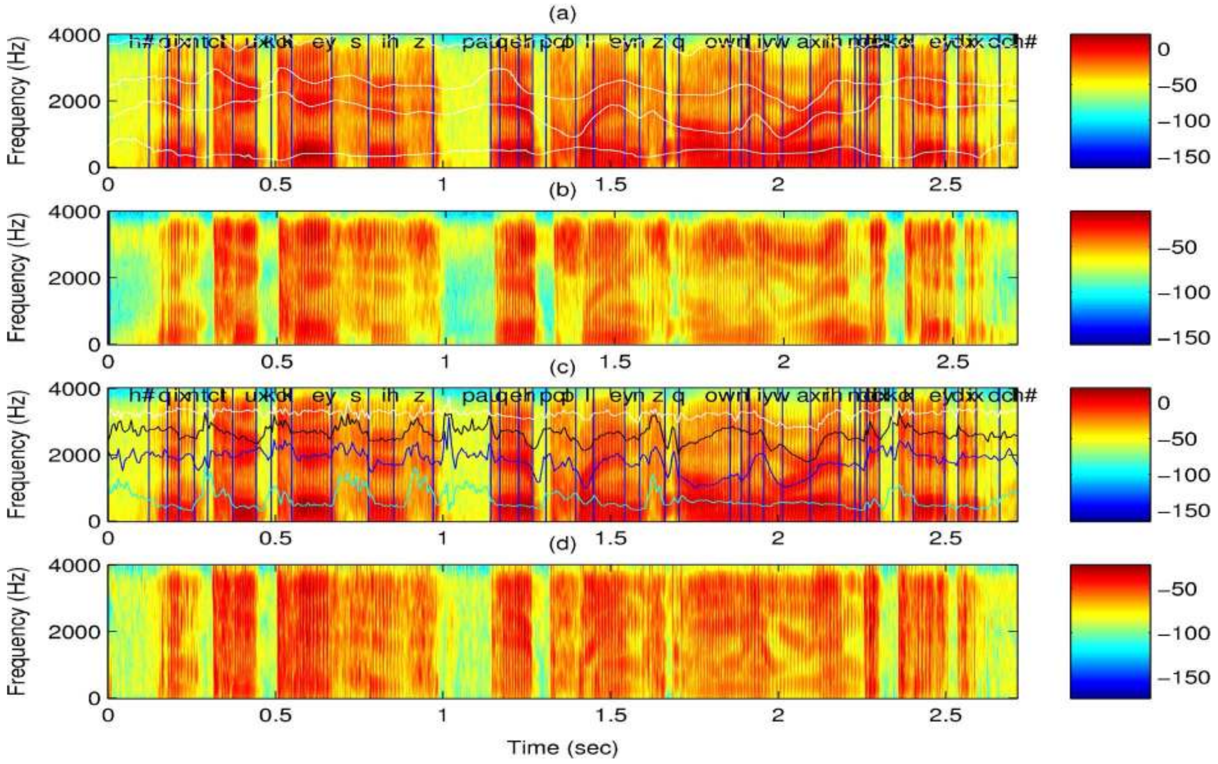

Fig. 5. Instantaneous frequency tracks and prediction residual for sentence S1 ("In two cases airplanes only were indicated"). (a) Hand-labeled "ground truth" data [16]. (b) Prediction residual for the hand-labeled data. (c) Cascade of predictors. (d) Prediction residual for the cascade of predictors.

the speech signal as a sum of exponentials. In the model proposed in this paper, the convolutional representation is retained in an effort to exploit the decomposability associated with source-filter models of speech production.

B. Dynamic Mode Predictors

With each DM we can develop associated dynamic error predictors (DPs) based on the knowledge of the dynamics of the mode

$$\Phi\{\bar{m}(t), \beta(t)\} = 0 \tag{10}$$

where the vector of samples $\bar{m}(t)$ is a vector embedding complementary information, (that is information other than that already included in $\beta(t)$) which describes the state of the signal. In the notation adopted here, Φ represents a functional description of the dynamic predictor which has $\bar{m}(t)$ and $\beta(t)$ as its arguments. Equation (10) is set equal to 0 to illustrate that, in principle, a perfect predictor can be constructed which relies on information extracted from samples of the mode and from its dynamics, $\beta(t)$. If we used all of the derivatives of the dynamics, the information provided by the samples of the mode would be *phase related*. For example, for a DM with constant dynamics $\beta(t) = [\sigma, \omega]$, a DP can be devised using two past samples of the mode as follows:

$$\begin{aligned}\Phi&\{\bar{m}(t), \beta(t)\}\\ &= m(t) - 2e^{\sigma\tau}\cos(\omega\tau)m(t-\tau) + e^{2\sigma\tau}m(t-2\tau).\end{aligned} \tag{11}$$

C. Shadow Dynamic Modes and Predictors

A shadow dynamic mode (S-DM) can be defined as

$$\hat{m}(t) = e^{\hat{\alpha}(t)}\cos\hat{\phi}(t) \tag{12}$$

with characteristic dynamics given by

$$\hat{\beta}(t) = [\hat{\sigma}(t), \hat{\omega}(t)] = \left[\frac{\partial\hat{\alpha}}{\partial t}, \frac{\partial\hat{\phi}}{\partial t}\right]. \tag{13}$$

A shadow predictor (S-DP) (motivated as discussed previously by the work of [20] and [21]), associated with an S-DM, can be used to track and cancel a DM m yielding a prediction error

$$\epsilon_m(t) = \hat{\Phi}\{\bar{m}(t), \hat{\beta}(t)\} \tag{14}$$

which has two components. If we assume that $m(t)$ and $\hat{m}(t)$ have "slow dynamics" (i.e., low values of $\partial\sigma/\partial t, \partial\omega/\partial t, \partial\hat{\sigma}/\partial t$ and $\partial\hat{\omega}/\partial t$), and then apply linear filter theory to obtain the response of shadow predictors for modes (and shadow modes) with null dynamics.[1] In this case, it can be readily shown that the dynamics of the input mode will be unaltered at the output of the predictor. These ideas can be straightforwardly extended to the case of modes with relatively low dynamics to show that a relatively small variation of the output dynamics, in relation to the input dynamics, will be obtained.

[1] As will be made clear later, some of the dynamic modes are associated with the resonances of the vocal tract while others are associated with the source/glottal signal. In the case of the resonant modes, their dynamics will depend on the movement of the articulators and can be considered to be "slow" relative to the sample rates used for speech processing. The same idea can be applied to the dynamics of the glottal signal associated to the evolution of the instantaneous pitch.

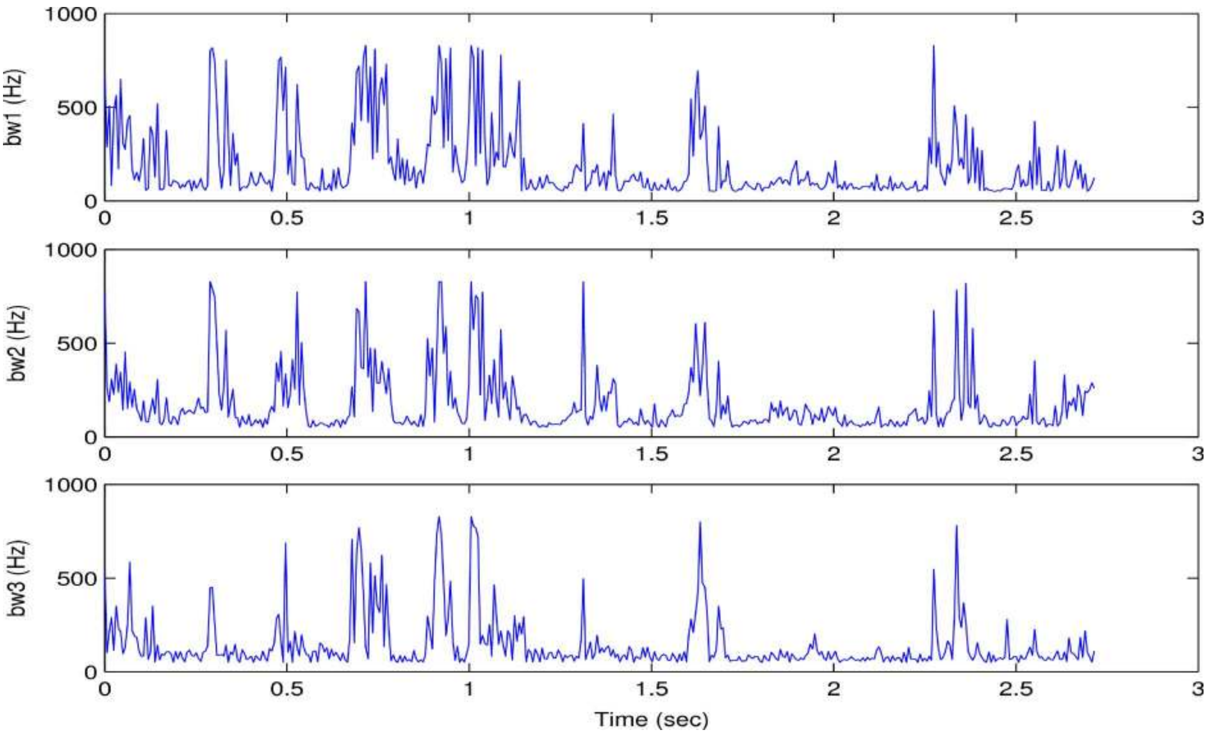

Fig. 6. Instantaneous bandwidth for resonant modes corresponding to the example presented in Fig. 5.

Defining the tracking error as, (and decomposing it into two components)

$$\epsilon_\beta = [\epsilon_\sigma, \epsilon_\omega] = [\hat{\sigma} - \sigma, \hat{\omega} - \omega] \tag{15}$$

and considering a small vicinity of ϵ_β around the origin where ϵ_m^2 is a concave function of ϵ_β; the S-DP can be made adaptive by inducing error dynamics that follow the steepest descent direction (gradient based approach):

$$\begin{aligned} \epsilon_\sigma &= \epsilon_\sigma^- - \Omega_\sigma \\ \epsilon_\omega &= \epsilon_\omega^- - \Omega_\omega \end{aligned} \tag{16}$$

where Ω_σ and Ω_ω are real-valued functions of ϵ_σ and ϵ_ω, respectively, having the same sign as $(\partial\epsilon_m^2)/(\partial\hat{\sigma})$ and $(\partial\epsilon_m^2)/(\partial\hat{\omega})$, respectively. The notation $(\,\cdot\,)^-$ indicates a previous iteration.

Dynamic error induction is effectively implemented adapting the dynamics of the S-PM according to

$$\begin{aligned} \hat{\sigma} &= \hat{\sigma}^- - \Omega_\sigma \\ \hat{\omega} &= \hat{\omega}^- - \Omega_\omega \end{aligned} \tag{17}$$

The S-DP can be finally represented in the following equation:

$$\begin{aligned} \epsilon_m(t) &= \hat{\Phi}\{\bar{m}(t), \hat{\beta}(t)\} \\ \epsilon_\sigma &= \epsilon_\sigma^- - \Omega_\sigma \\ \epsilon_\omega &= \epsilon_\omega^- - \Omega_\omega. \end{aligned} \tag{18}$$

The adaptation functions (Ω_σ and Ω_ω) must be designed focusing on reaching a state where the dynamics of modes are *locked*, where locked means $\omega = \hat{\omega}$.

An S-DM $\hat{m}(t)$ can be dynamically convolved while tracking an input mode $m_{in}(t)$ to obtain a dynamic-tracking resonator (DR)

$$\begin{aligned} m(t) &= \hat{m}(t)\vec{*}m_{in}(t) = \Gamma\{\bar{m}_{in}(t), \bar{m}(t), \hat{\beta}(t)\} \\ \epsilon_m(t) &= \hat{\Phi}(t)\{\bar{m}(t), \hat{\beta}(t)\} \\ \epsilon_\sigma &= \epsilon_\sigma^- - \Omega_\sigma \\ \epsilon_\omega &= \epsilon_\omega^- - \Omega_\omega. \end{aligned} \tag{19}$$

In locked-mode, the DR will reinforce, (i.e., amplify against other superimposed modes with different dynamics) and track the dynamics of the input mode $m_{in}(t)$.

D. Dynamic Gabor Modes and Filters

Gabor filters are well known in signal processing [32] and can serve as excellent band-pass filters for speech signals. A Gabor filter is defined as the product of a Gaussian kernel times a sinusoid. Extending these ideas, a dynamic Gabor mode (G-DM) can be defined as

$$m_g(t) = e^{\delta_0 - (\gamma t)^2} \cos\left(\int_0^t \omega_g(v)dv\right) \tag{20}$$

characterized by dynamics

$$\beta_g(t) = [-2\gamma^2 t, \omega_g(t)]. \tag{21}$$

Also a dynamic Gabor filter, with input $m(t)$ and output $m_o(t)$, can be defined through a non-recursive approximation of the dynamic mode convolution of $m_g(t)$ and $m(t)$:

$$m_o(t) = e^{\delta_0} \int_\tau m(t-\tau) e^{-(\gamma\tau)^2} \cos\left(\int_0^\tau \omega_g(v)dv\right) d\tau. \tag{22}$$

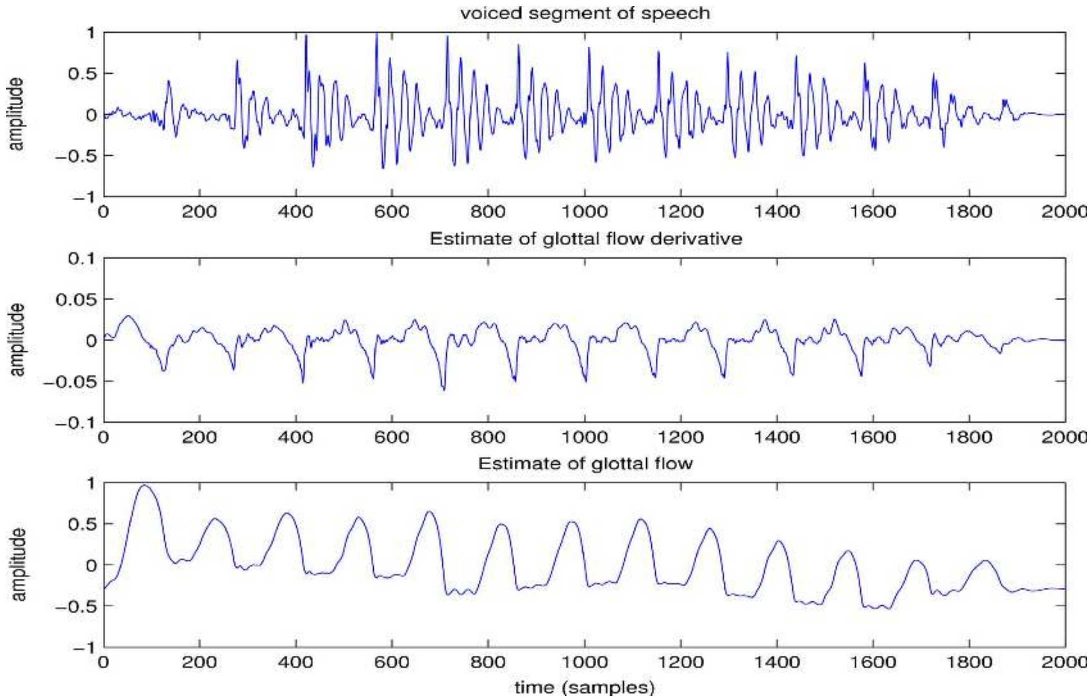

Fig. 7. Estimates of glottal flow derivative and glottal flow for a voiced segment of speech (“(h)ad” from S2).

For the case of $m_g(t)$ and $m(t)$ with “slow dynamics” (low values of $\partial\beta/\partial t = [\partial\sigma/\partial t, \partial\omega/\partial t]$ and $\partial\omega_g/\partial t$), the G-DF will act as a time-varying filter with center frequency ω_g. This fact can be exploited to isolate $m(t)$ if there are additional superimposed modes based on the knowledge of the evolution of $\beta(t)$.

III. Application to Speech Analysis and Synthesis

The application of the above to speech analysis and synthesis involves dynamic (gradient-based adaptation) shadow mode predictors in cascade being used to track the dynamics of the resonant modes of speech without explicitly decomposing the speech signal into its modulated components while still allowing a separation of the vocal tract from source-related information. The prediction residual conveys source-related modal dynamics that can be further extracted via dynamic filters. Additionally, the speech signal can be recovered by inverse filtering (dynamically convolving) the prediction residual with the corresponding dynamic inverse predictors. Fig. 1 illustrates how the speech signal could be modeled as a superposition of four dynamic modes. Each mode is the result of a dynamic mode convolution (DMC) of the glottal modes with the corresponding mode dynamics.

A. Shadow Mode Predictors to Track Resonant Dynamics

Based on the representation presented above, shadow mode predictor adaptation can be used to track the dynamics of each resonant mode. Shadow predictor adaptation can be achieved using a cascade of S-DP to isolate a dynamics-preserving version of each resonant mode while adapting S-DPs

$$\begin{aligned}\tilde{\epsilon}_{m,l}(t) &= \hat{m}_{r,l}(t)\vec{*}\,\amalg_{k=0}^{N_R}\,\hat{\Phi}\{\bar{s}(t),\hat{\beta}(t)\}_k\\ \epsilon_{m,l}(t) &= \hat{\Phi}\{\bar{\tilde{\epsilon}}_{m,l}(t),\hat{\beta}(t)\}_l\\ \epsilon_{\sigma,l} &= \epsilon^{-}_{\sigma,l} - \Omega_{\sigma,l}\\ \epsilon_{\omega,l} &= \epsilon^{-}_{\omega,l} - \Omega_{\omega,l}\\ &\qquad\text{for } l = 0,\ldots,N_R \qquad (23)\end{aligned}$$

where $\amalg[\,\cdot\,]$ denotes the cascade operation, $\hat{m}_{r,l}(t)$ is a shadow resonant mode with dynamics $\hat{\beta}_l(t)$ and $k = 0$ is the “glottal resonance” index. Each of the semi-residual signals $\tilde{\epsilon}_{m,l}$ conveys the corresponding modal dynamics and, in locked mode, the prediction residual $\sum\tilde{m}_s(t)$ will be given by

$$\sum\tilde{m}_s(t) = \epsilon_{m,l}(t) \quad \text{for } l = 0,\ldots,N_R$$

Based on the tracked dynamics of each resonance $(\{\hat{\beta}_l, l = 0,\ldots,N_R\})$, The prediction residual can be expressed in compact form as

$$\sum\tilde{m}_s(t) = \amalg_{l=0}^{N_R}\hat{\Phi}\{\bar{s}(t),\hat{\beta}(t)\}_l \qquad (24)$$

and the speech signal can be recovered through inverse filtering

$$s(t) = \amalg_{l=0}^{N_R}\hat{\Phi}^{-1}\left\{\bar{\sum}\tilde{m}_s(t),\hat{\beta}(t)\right\}_l. \qquad (25)$$

Taking into account that each inverse predictor can be expressed as

$$\hat{\Phi}^{-1}\left\{\bar{\sum}\tilde{m}_s(t),\hat{\beta}(t)\right\}_l = \left[\hat{m}_r(t)\vec{*}\sum\tilde{m}_s(t)\right]_l. \qquad (26)$$

Equation (25) can be rewritten as

$$s(t) = \amalg_{l=0}^{N_R}\left[\hat{m}_r(t)\vec{*}\sum\tilde{m}_s(t)\right]_l. \qquad (27)$$

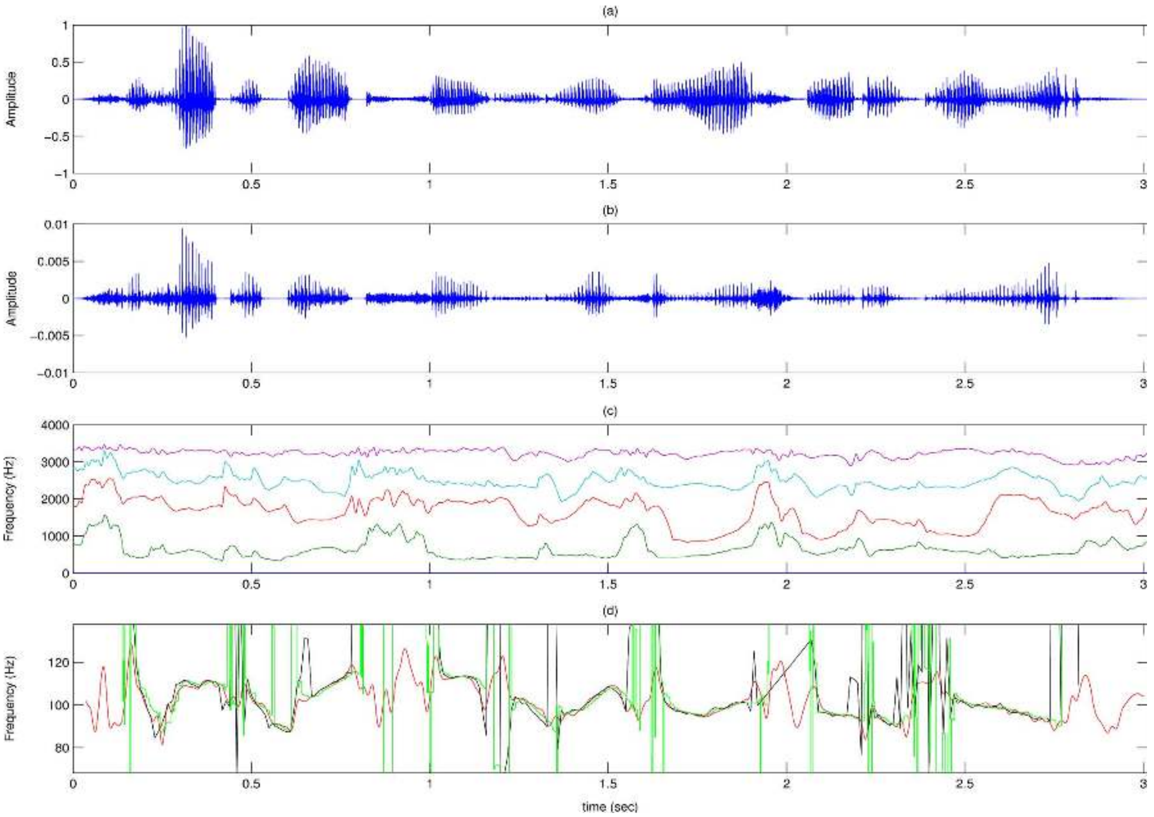

Fig. 8. (a) Segment of speech uttered by a male speaker (S2: "She had your dark suit in greasy wash water all year"). (b) Cascade of predictors residual. (c) Frequency trajectory of the first four resonances. (d) Instantaneous pitch (proposed approach—red line, Yin estimator [39]—green line, RAPT [40]—black line).

It is important to note that the prediction residual is a dynamics-preserving version of the superposition of source modes (6) and (9) and (27) are intrinsically equivalent. An illustration of the process for resonance dynamics tracking (representing the dynamic convolution operation as inverse S-DP) is shown in Fig. 2[2] in conjunction with Fig. 3.

The convergence of the gradient-based algorithm described, for the case of real world speech signals, depends on the appropriateness of the model to describe the intrinsic dynamics. Guaranteeing the necessary concavity depends not only on the proximity of the dynamics of a shadow predictor to that of the tracked modes but also on the efficacy of the model of speech production to faithfully represent the complexity of the speech signal. If the model is successful in providing a search space with reduced dimension, the adaptation functions (Ω_σ and Ω_ω) can be designed to reach the state of locked-modes dynamics, where the dynamics of each constitutive mode is tracked with minimal error.

Since we will not have *a priori* knowledge of the number of resonances N_R in a given frequency range, the exact number of resonances can be determined by monitoring the instantaneous bandwidth of each semi-residual mode $\tilde{\epsilon}_{m,l}$ in (22). A relatively high bandwidth (more than 400 Hz) is a strong indication of a weak or non existent mode.

The number of glottal/source modes N_Q will depend on the type of source excitation and its study is not addressed in this paper.

B. Dynamic Mode Filters to Track the Instantaneous Pitch of Speech

Extending the efforts to capture non-stationary speech features focussing on the instantaneous pitch frequency [33], [34], [12], [35], we rely on a dynamic Gabor-resonator filter, applied to the prediction residual $\sum \tilde{m}_s(t)$, to track the instantaneous frequency of the fundamental source mode

$$\begin{aligned} m_o(t) &= e^{\delta_0} \int_\tau \Big(\sum \tilde{m}_s(t-\tau))e^{-(\gamma\tau)^2} \\ &\qquad \times \cos\left(\int_0^\tau \breve{\omega}_p(v)dv\right)\Big) d\tau \\ m(t) &= \Gamma\{\bar{m}_o(t), \bar{m}(t), \hat{\beta}_p(t)\} \\ \epsilon_m(t) &= \hat{\Phi}\{\bar{m}(t), \hat{\beta}_p(t)\} \\ \epsilon_{\omega_p} &= \epsilon^-_{\omega_p} - \Omega_{\omega_p} \end{aligned} \tag{28}$$

where $\breve{\omega}_p$ is a low-pass version of ω_p.

Finally, the instantaneous pitch $\tilde{\omega}_p$ is obtained as the trend of the estimated instantaneous frequency evolution $\hat{\omega}_p$. The desired trend is extracted using least-squares splines [36] by selecting a temporal separation between knots to obtain smooth estimates of the dynamics of the pitch-related modes.

[2] The apostrophes in Fig. 2 are used to distinguish S-DPs with equal dynamics but different inputs.

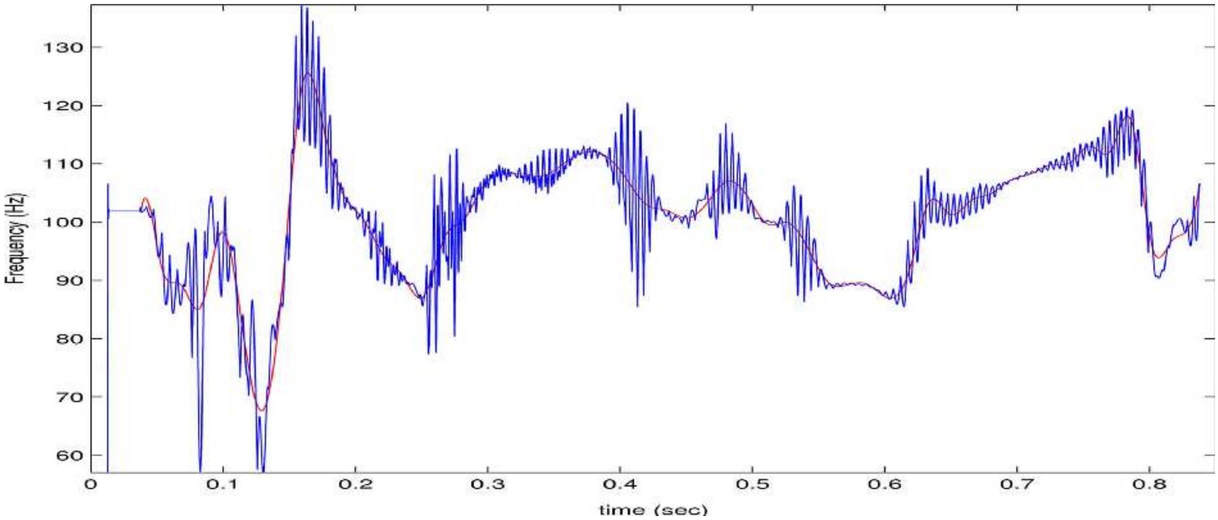

Fig. 9. Close look at the trend extraction task used to obtain the pitch track for the first 0.8 s of the sentence used in 8 (S2).

C. Intonation Modification Based on Instantaneous-Pitch-Guided Time-Warping

A methodology is presented in this section to modify intonation by relying on instantaneous-pitch-guided time-warping. Once the instantaneous pitch is obtained, following the procedure explained previously above, the instantaneous pitch period can be found as

$$\lambda(t) = \frac{1}{2\pi}\int_0^t \tilde{\omega}_p(v)dv. \tag{29}$$

The residual signal $\sum \tilde{m}_s(t)$ can then be time-warped to match a target pitch $\lambda_m(t)$ according to

$$\sum \tilde{m}_{sm}(t) = \sum \tilde{m}_s(\alpha(t)) \tag{30}$$

where $\alpha(t)$ is a nonlinear time-warping function

$$\alpha(t) = \lambda^{-1}(\lambda_m(t)). \tag{31}$$

D. Experimental Results and Observations

In this section, several examples are presented to illustrate the potential of the proposed approach. These examples are based on four sentences selected from the TIMIT database [37]:

- S1: "In two cases, airplanes only were indicated" (timit\ train\ dr1\ mdac0\ si1261);
- S2: "She had your dark suit in greasy wash water all year" (timit\ train\ dr1\ mdac0\ sa1);
- S3: "She had your dark suit in greasy wash water all year" (timit\ train\ dr1\ fcjf0\ sa1);
- S4: "Have they inherited some money or something" (timit\ train\ dr1\ mdpk0\ si1683).

The basic S-DP of (18) is implemented using samples of the signal as

$$\begin{aligned}\epsilon_m(t) &= m(t) - 2\hat{\boldsymbol{\rho}}(t)\cos(\hat{\boldsymbol{\omega}}(t))m(t-\tau) \\ &\quad + \hat{\boldsymbol{\rho}}(t)^2 m(t-2\tau) \\ \hat{\boldsymbol{\rho}}(t) &= \hat{\boldsymbol{\rho}}(t)^- - \mu_{\boldsymbol{\rho}}(t)|\epsilon_m|\text{sig}\{\hat{\boldsymbol{\rho}}(t)m(t-2\tau) \\ &\quad - \cos(\hat{\boldsymbol{\omega}}(t))m(t-\tau)\} \\ \hat{\boldsymbol{\omega}}(t) &= \hat{\boldsymbol{\omega}}(t)^- - \mu_{\boldsymbol{\omega}}|\epsilon_m|\text{sig}\{m(t-\tau)\}\end{aligned} \tag{32}$$

where $\hat{\boldsymbol{\rho}}(t) = e^{\hat{\sigma}(t)\tau}$, $\hat{\boldsymbol{\omega}}(t) = \hat{\omega}(t)\tau$, $\mu_{\boldsymbol{\rho}}$, and $\mu_{\boldsymbol{\omega}}$ are positive, real-valued and constant scalars. In the derivation of these equations, the following adaptation functions were selected according to the properties specified for (16)[3] :

$$\Omega_{\boldsymbol{\rho}} = \mu_{\boldsymbol{\rho}}|\epsilon_m|\text{sig}\left\{\frac{\partial\epsilon_m}{\partial\boldsymbol{\rho}}\right\} \tag{33}$$

$$\Omega_{\boldsymbol{\omega}} = \mu_{\boldsymbol{\omega}}|\epsilon_m|\text{sig}\left\{\frac{\partial\epsilon_m}{\partial\boldsymbol{\omega}}\right\} \tag{34}$$

and in all examples the values selected for the cascade of predictors were, $\mu_{\boldsymbol{\rho}} = 0.001$ and $\mu_{\boldsymbol{\omega}} = 0.007$, with two iterations over segments of 6.3 ms. Values of $\mu_{\boldsymbol{\rho}} = 0$ and $\mu_{\boldsymbol{\omega}} = 0.3$ were selected for the adaptive Gabor-resonator sections.

Resonator sections were implemented as

$$\begin{aligned}m(t) &= \Gamma\{m_{in}(t), \bar{m}(t), \hat{\beta}(t)\} \\ &= m_{in}(t) + 2\hat{\boldsymbol{\rho}}(t)\cos[\hat{\boldsymbol{\omega}}(t)]m(t-\tau) \\ &\quad - \hat{\boldsymbol{\rho}}(t)^2 m(t-2\tau)\end{aligned} \tag{35}$$

where a constant value of $\hat{\boldsymbol{\rho}}(t) = 0.8$ was used in all of the experiments.

Gabor sections were implemented as

$$m_o(t) = e^{\delta_0}\sum_{\tau} m_{in}(t-\tau)e^{-(\gamma\tau)^2}\cos\left(\int_0^{\tau}\breve{\omega}(v)dv\right) \tag{36}$$

with $\delta_0 = 0$, constant γ ($\simeq$60-Hz bandwidth) and $\breve{\omega}$ adjusted iteratively as the average pitch of the segment. The implementation of the proposed algorithm is further illustrated in Figs. 2–4.

All utterances have been low-pass filtered, following a procedure similar to that described in [17], to focus the analysis on the

[3] A more rigorous analysis can be performed, similar to that presented in [38], to further explore the convergence properties of the proposed algorithm. However, this is beyond the scope of the current paper.

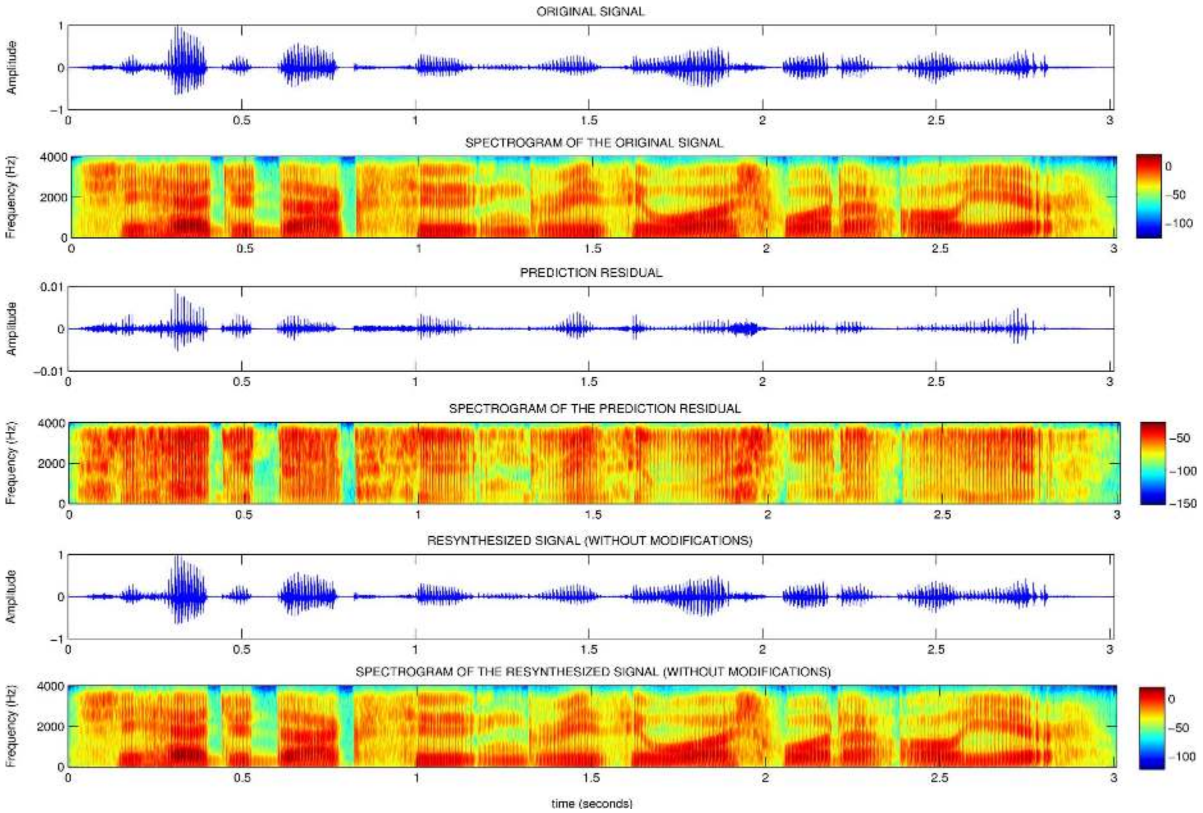

Fig. 10. Cascade of predictors re-synthesis without modifications (S2).

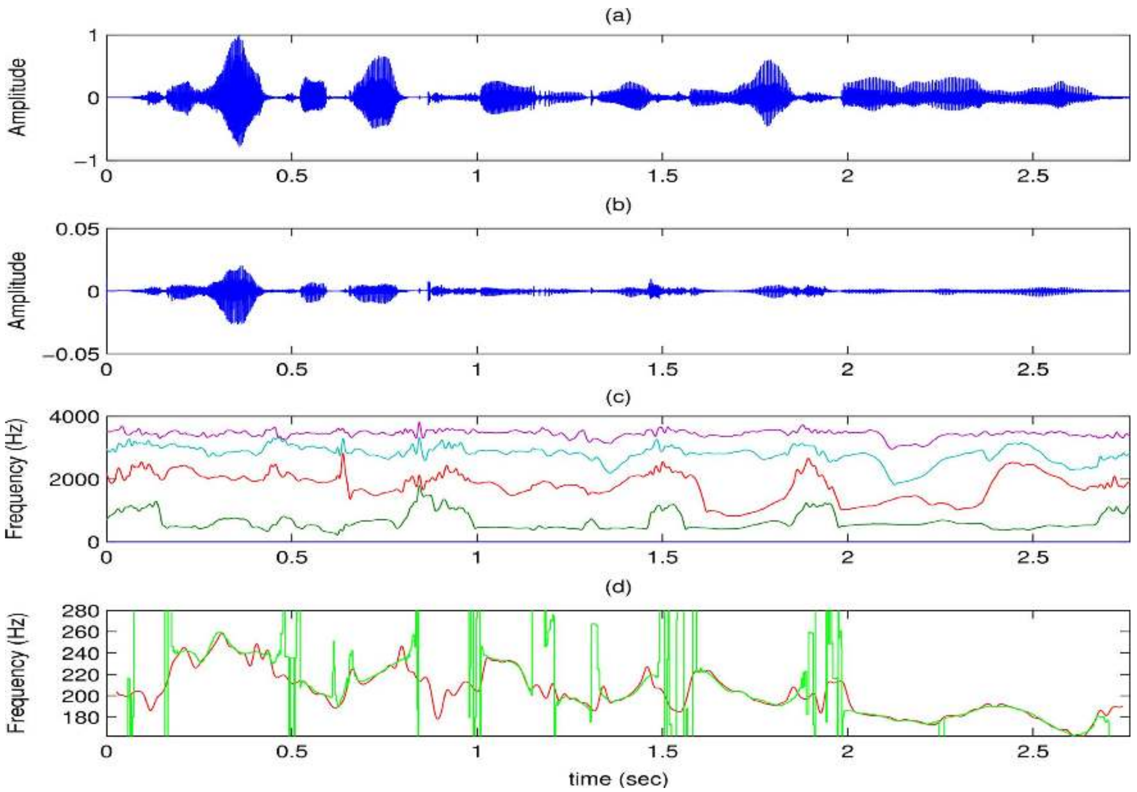

Fig. 11. (a) Segment of speech uttered by a female speaker (S3: “She had your dark suit in greasy wash water all year”). (b) Cascade of predictors residual. (c) Frequency trajectory of the first four resonances. (d) Instantaneous pitch (proposed approach—red line, Yin estimator [39]—green line.

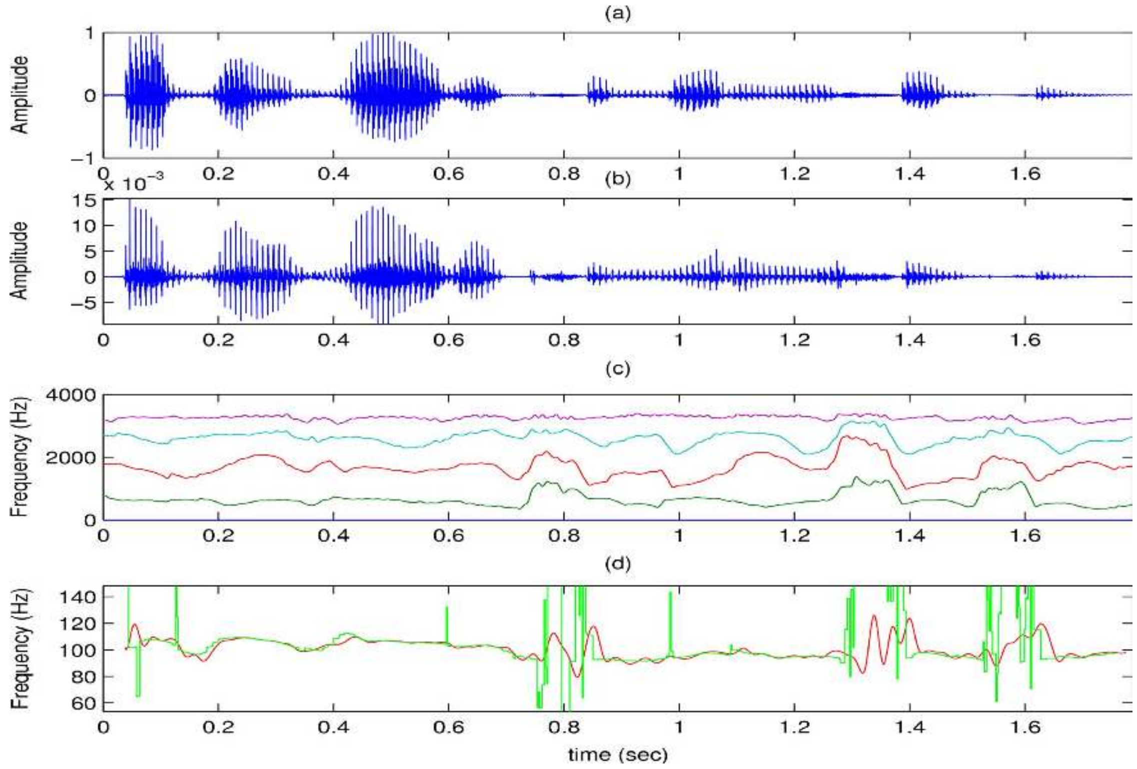

Fig. 12. (a) Segment of speech uttered by a male speaker (S4: "Have they inherited some money or something"). (b) Cascade of predictors residual. (c) Frequency trajectory of the first four resonances. (d) Instantaneous pitch (proposed approach—red line, Yin estimator [39]—green line.

first four resonances. Despite this selected fixed number of resonant modes, the time-varying nature of modal parameters estimated accounts for weakening and disappearing modes. In order to illustrate this fact, Figs. 5 and 6 are provided showing how the time varying bandwidth of the modes are effectively tracked with the proposed approach.[4] The obtained tracks are compared with the corresponding "ground-truth" provided in [16]. More examples and comparisons against the "ground truth" database [16] can be found in [17].

Although the value of each of the adaptation parameters ($\mu_{\boldsymbol{\rho}}$ and $\mu_{\boldsymbol{\omega}}$) was selected empirically, good tracking properties were observed in all experiments with no variance across utterances, although clearly further work is required to demonstrate the robustness of these values. Also the tracking algorithms obtained were not particularly sensitive to the initialization values for the tracked parameters, as long as they were selected appropriately, indicating the possibility to incorporate automatic initialization procedures based on some prior knowledge and coarse estimation algorithms. These facts can be interpreted as an indication of the success of the proposed model in effectively reducing the search space for the modeling parameters as previously observed in [17].

In the following examples, the evolution of the parameters estimated for the resonant modes were interpolated using splines with a separation between knots of 6.25 ms. Similarly, the frequency tracks of the pitch-related modes (instantaneous pitch) were interpolated with 12.5 ms of a separation between the knots. Additionally, down sampling by a factor of 4 has been applied before performing the pitch tracking operation to reduce the computational load of the resulting algorithm.

An example is provided in Fig. 7 to illustrate the accuracy of the proposed approach in the task of recovering an estimate of the glottal flow derivative and the glottal flow for a segment of voiced speech extracted from S2. The estimates were obtained by integrating the prediction residual at the output of the cascade of predictors showing clearly recognizable patterns associated to the glottal flow.

Fig. 8 shows the resonant dynamics and instantaneous pitch extracted from a sentence uttered by a male speaker (S2). The pitch track obtained with the proposed approach is displayed against the YIN [39] and RAPT [40] estimators to highlight its continuous and accurate evolution. A close look at the trend estimation task used to obtain the pitch track, from the instantaneous frequency output of Gabor-resonator sections; is given in Fig. 9 for the first 0.8 s of the sentence.

Fig. 10 illustrates the invertibility of the proposed representation applied to the male utterance S2. It can be seen, in the figure, how the original spectral-temporal features are preserved after re-synthesis using the inverse cascade of predictors. Other examples are illustrated in Figs. 11 and 12 showing resonant dynamics and instantaneous pitch extracted from S3 and S4, respectively.

In addition to the figures provided, several sound examples, obtained from the sentences analyzed before, are available online [42] (including low-pass filtered versions of sentences S2,

[4] The instantaneous frequencies and bandwidths (in Hertz) were computed as

$$f_k = \frac{F_s}{2\pi}\omega_k \text{ Hz}$$

and

$$bw_k = \frac{-F_s}{\pi}\ln(\rho_k) \text{ Hz}.$$

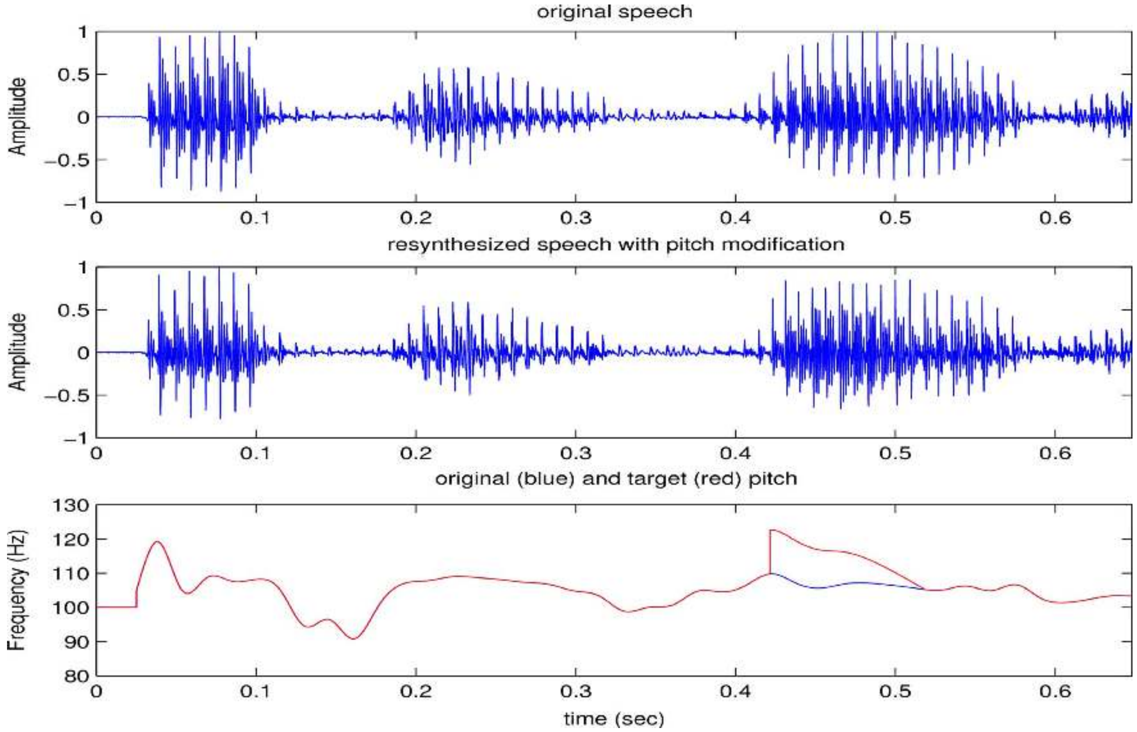

Fig. 13. Pitch modification for a segment of the sentence analyzed in Fig. 12 (S4).

S3, and S4 in conjunction with their modified versions) to illustrate the applicability of the proposed approach in the task of intonation modification. Some of the examples are also intended to illustrate how the instantaneous pitch can be modulated to create a *trembling* voice effect. It is important to remark that, in order to preserve voice source-vocal tract synchronization, the tracks of resonant dynamics can be time warped with function $\alpha(t)$ [obtained from (31)].

Equations (30) and (31) are implemented, in discrete-time, relying on spline-based interpolation to modify the intonation/instantaneous pitch of each analyzed segment according to the intended target pitch evolution. Fig. 13 illustrates the pitch modification process for a segment of the sentence analyzed in Fig. 12.

Finally, a sound-based comparison with the STRAIGHT algorithm [41] is also included, for a segment of S4, to highlight the fact that, although the results show that both techniques allow finely tuned intonation modification; the proposed approach has the potential to flexibly modify all the instantaneous dynamic information (including spectral dynamics) conveyed by the speech signal.

IV. Conclusion

This paper has proposed a new model to represent the speech signal as the convolution and superposition of dynamic modes that can be associated with different sections of the human phonatory system. In order to track the dynamics of the vocal modes the idea of adaptive shadow modes and predictors has been introduced. Similarly, the concept of adaptive Gabor and resonator mode filters have been presented to extract pitch-related dynamics from glottal modes. The potential of the proposed approach is illustrated with several examples including instantaneous pitch tracking and modification. Future work will be focused on performing other prosodic modifications including the control of spectral dynamics.

Acknowledgment

The authors would like to thank the anonymous referees and the Associate Editor for their useful comments that have significantly improved the quality of the paper. In addition, the assistance of B. Mulgrew and Y. Kopsinis in checking our revisions made sense are much appreciated and gratefully acknowledged.

References

[1] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*. New York: Wiley-IEEE, 1999.
[2] A. Rao and R. Kumaresan, "On decomposing speech into modulated components," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 240–254, May 2000.
[3] H. Dudley, "The carrier nature of speech," *Bell Syst. Tech. J.*, vol. 19, pp. 495–515, 1940.
[4] E. C. Cherry and V. J. Phillips, "Some possible uses of single sideband signals in formant-tracking systems," *J. Acoust. Soc. Amer.*, vol. 33, pp. 1067–1077, 1961.
[5] J. L. Flanagan, "Parametric coding of speech spectra," *J. Acoust. Soc. Amer.*, vol. 68, pp. 412–419, 1980.
[6] T. Quatieri and R. McAulay, "Shape invariant time-scale and pitch modification of speech," *IEEE Trans. Signal Process.*, vol. 40, no. 3, pp. 497–510, Mar. 1992.
[7] T. Abe and M. Honda, "Sinusoidal model based on instantaneous frequency attractors," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1292–1300, Jul. 2006.
[8] P. Maragos, J. Kaiser, and T. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Trans. Signal Process.*, vol. 41, no. 10, pp. 3024–3051, Oct. 1993.
[9] H. Ohmura and K. Tanaka, "Speech synthesis using a nonlinear energy damping model for the vocal folds vibration effect," in *Proc. 4th Int. Conf. Spoken Lang. (ICSLP'96)*, 1996, vol. 2, pp. 1241–1244.
[10] F. Gianfelici, G. Biagetti, P. Crippa, and C. Turchetti, "Multicomponent AM & FM representations: An asymptotically exact approach," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 823–837, Mar. 2007.

[11] B. Yegnanarayana and R. Veldhuis, "Extraction of vocal-tract system characteristics from speech signals," *IEEE Speech Audio Process.*, vol. 6, no. 4, pp. 313–327, 1998.

[12] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech Commun.*, vol. 27, no. 3–4, pp. 187–207, 1999.

[13] P. H. Milenkovic, "Voice source model for continuous control of pitch period," *J. Acoust. Soc. Amer.*, vol. 93, no. 2, pp. 1087–1096, 1993.

[14] J. Wouters and M. Macon, "Control of spectral dynamics in concatenative speech synthesis," *IEEE Speech Audio Process.*, vol. 9, no. 1, pp. 30–38, 2001.

[15] B. Resch, M. Nilsson, A. Ekman, and W. Kleijn, "Estimation of the instantaneous pitch of speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 813–822, Mar. 2007.

[16] L. Deng *et al.*, "A database of vocal tract resonance trajectories for research in speech processing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Toulouse, France, May 14-19, 2006, pp. 60–63.

[17] J. Vargas and S. McLaughlin, "Cascade prediction filters with adaptive zeros to track the time-varying resonances of the vocal tract," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 1–7, Jan. 2008.

[18] R. Muralishankar, A. Ramakrishnan, and P. Prathibha, "Modification of pitch using DCT in the source domain," *Speech Commun.*, vol. 42, no. 2, pp. 143–154, 2004.

[19] K. Rao and B. Yegnanarayana, "Prosody modification using instants of significant excitation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 3, pp. 972–980, May 2006.

[20] C. Grebogi *et al.*, "Shadowing of physical trajectories in chaotic dynamics," *Phys. Rev. Lett.*, no. 13, pp. 1527–1530, 1990.

[21] K. Judd *et al.*, "Gradient free descent: Shadowing, and state estimation using limited derivative information," *Physica D*, vol. 90, pp. 153–166, 2004.

[22] D. G. Childers, *Speech Processing and Synthesis Toolboxes*. New York: Wiley, 2000.

[23] E. Rank, "Oscillator-Plus-Noise Modeling of Speech Signals," Ph.D. dissertation, Vienna Univ. of Technol., Vienna, Austria, 2005.

[24] J. Schoentgen, "Shaping function models of the phonatory excitation signal," *J. Acoust. Soc. Amer.*, vol. 114, pp. 2906–2912, 2003.

[25] C. Drioli, "A flow waveform-matched low-dimensional glottal model based on physical knowledge," *J. Acoust. Soc. Amer.*, vol. 117, no. 5, pp. 3184–3195, 2005.

[26] Z. Wu and N. E. Huang, "A study of the characteristics of white noise using the empirical mode decomposition method," in *R. Soc. London Proc., Ser. A*, vol. 460, pp. 1597–1611.

[27] J. Ville, "Theorie et applications de la notion de signal analytique.," *Cables et Transmission*, vol. 2, pp. 61–74, 1948.

[28] H. B. Voelcker, "Toward a unified theory of modulation part i: Phase-envelope relationships," *Proc. IEEE*, vol. 54, no. 3, pp. 340–355, Mar. 1966.

[29] M. Poletti, "Instantaneous frequency and conditional moments in the time–frequency plane," *IEEE Trans. Signal Process.*, vol. 39, no. 3, pp. 755–756, 1991.

[30] L. Cohen and C. Lee, "Instantaneous bandwidth," in *Time-Frequency Signal Analysis-Methods and Applications*, B. Boshash, Ed. London, U.K.: Longman-Cheshire, 1992.

[31] A. Krishnamurthy, *Glottal Source Estimation Using a Sum-of-Exponentials Model*, vol. 40, no. 3, pp. 682–686, 1992.

[32] S. Mallat, *A Wavelet Tour of Signal Processing*. New York: Academic, 1998.

[33] L. Qiu, H. Yang, and S. Koh, "Fundamental frequency determination based on instantaneous frequency estimation," *Signal Process.*, vol. 44, no. 2, pp. 233–241, 1995.

[34] T. Abe, T. Kobayashi, and S. Imai, "Harmonics tracking and pitch extraction based on instantaneous frequency," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP'95)*, May 9–12, 1995, vol. 1, pp. 756–759.

[35] H. Huang and J. Pan, "Speech pitch determination based on Hilbert–Huang transform," *Signal Process.*, vol. 86, no. 4, pp. 792–803, Apr. 2006.

[36] Immoptibox, [Online]. Available: http://www2.imm.dtu.dk/~hbn/immoptibox/

[37] Acoustic-Phonetic Continuous Speech Corpus, NIST Speech Disc 1-1.1, 1990, DARPA-TIMIT.

[38] W. Torres, A. Oppenheim, and R. Rosales, "Generalized frequency modulation," , vol. 48, no. 12, pp. 1405–1412, 2001.

[39] A. Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Amer.* vol. 111, no. 4, pp. 1917–1930, 2002 [Online]. Available: http://link.aip.org/link/?JAS/111/1917/1

[40] D. Talkin, *A Robust Algorithm for Pitch Tracking (RAPT), in Speech Coding and Synthesis*. Amsterdam, The Netherlands: Elseiver Science, 1995.

[41] [Online]. Available: http://www.wakayama-u.ac.jp/~kawahara/STRAIGHTadv/index_e.html

[42] [Online]. Available: http://www.see.ed.ac.uk/~sml/Speech.htm

Julio Vargas was born in Caracas, Venezuela, in 1968. He received the Electrical Engineering degree (*cum laude*) from the Universidad de Los Andes, Merida, Venezuela, in 1992, and the Ph.D. degree in electrical engineering from the University of Edinburgh, Edinburgh, U.K., in 2002.

At the University of Edinburgh, he engaged in research on adaptive techniques to estimate and track the instantaneous frequencies of nonstationary multicomponent signals. Since 1993, he has been a Professor at the Universidad de Los Andes, where he teaches different courses electronics and signal processing at undergraduate and postgraduate levels. In 2008, he was a Visiting Researcher at the Institute for Digital Communications (University of Edinburgh) working on adaptive techniques for speech processing. His current research interests include speech processing and chaos-based information encryption.

Stephen McLaughlin (M'89–SM'04–F'11) was born in Clydebank, U.K., in 1960. He received the B.Sc. degree in electronics and electrical engineering from the University of Glasgow, Glasgow, U.K., in 1981 and the Ph.D. degree from the University of Edinburgh, Edinburgh, U.K., in 1989.

From 1981 to 1984, he was a Development Engineer with Barr & Stroud, Ltd., Glasgow, involved in the design and simulation of integrated thermal imaging and fire control systems. From 1984 to 1986, he worked on the design and development of high-frequency data communication systems with MEL, Ltd. In 1986, he joined the Department of Electrical Engineering, University of Edinburgh, as a Research Associate where he studied the performance of linear adaptive algorithms in high noise and nonstationary environments. In 1988, he joined the teaching staff at Edinburgh, and in 1991 he was awarded a Royal Society University Research Fellowship to study nonlinear signal processing techniques. His research interests lie in the fields of adaptive signal processing and nonlinear dynamical systems theory and their applications to biomedical, communication, and geophysical systems.

Prof. McLaughlin is a Fellow of the Royal Academy of Engineering, the Royal Society of Edinburgh, and the Institute of Engineering Technology, and is a Chartered Engineer.