A Novel Method of Glottal Inverse Filtering

Subhasmita Sahoo, *Student Member, IEEE*, and Aurobinda Routray, *Member, IEEE*

***Abstract*—This paper presents a new technique for glottal inverse filtering using a distributed model of the vocal tract. A discrete state space model has been constructed for the speech production system by combining the concatenated tube model of the vocal tract and Liljencrants–Fant (LF) model of the glottal flow derivative waveform. An adaptive system identification technique, based on extended Kalman filtering, has been used for estimation of the states and model parameters from continuous speech. The glottal signal, represented by the LF model, is piecewise differentiable in one glottal cycle. Hence, the hybrid system has been characterized by separate models during two different modes. Multiple model estimation has been performed by switching between the two models at the mode jumps. The open phase of the glottal cycle has been considered as Mode 1; whereas, the return phase and closed phase combined has been taken as Mode 2. The starting point of Mode 1, also known as glottal opening instant, was estimated by observing formant modulation, which remains negligible during closed phase, and starts to increase at the onset of opening. The starting point of Mode 2, also known as the glottal closing instant, was computed by peak-picking from linear prediction (LP) residual signal. The proposed method estimates the glottal waveform as well as changes in flow occurring at different sections of the vocal tract during speech production. This technique has been found to be accurate and robust to variations in pitch as compared to other LP-based methods in the literature. The method also estimates the air pressure distribution at different sections of the vocal tract.**

***Index Terms*—Concatenated tube model, extended Kalman filter, glottal inverse filtering, LF model, multiple model estimation.**

I. Introduction

GLOTTAL inverse filtering (GIF) helps in dissociating complex speech production system—which is inaccessible by noninvasive techniques—into source and filter form. Use of inverse estimation methods on the observable speech signal facilitates interpretation of the associated system in a simplified way. Starting from the 1950s [1], [2], GIF has been attracting significant research attention due to its broad applicability, such as: Speech synthesis [3], speaker identification [4], speech recognition [5], emotion and stress analysis [6], [7], medical applications [1], [2], etc.

Numerous work on inverse filtering of the speech signal has been reported in the literature [1] [8]. Most of them are based on linear prediction (LP) technique. The LP-based methods are established on the all-pole model of the vocal tract, in which, future values of the output signal are reproduced from its current and past samples except at the places of input pulses [9], [10]. However, due to gradual opening and closing of the glottis, the input to the vocal tract can be approximated to be triangular shaped instead of pure impulses. In [11], the impulse response of the system is estimated from the speech signal during closed glottis. Subsequently, the impulse response is used to find the glottal flow throughout the pitch period. However, the technique may not work during high pitch periods of speech due to the fast closing of the glottis, rendering a very small number of samples for approximation. One of the most popular techniques based on similar LP analysis is iterative adaptive inverse filtering (IAIF) as in [12]. In this method, the glottal flow and vocal tract contributions are eliminated in an iterative manner to get the estimated glottal signal. The quasi closed phase (QCP) method of inverse filtering [13] uses a technique similar to multipulse method as in [14] to obtain a better estimation of the linear model of the vocal tract. In multipulse method, the effect of the source is eliminated from the response by repeated computation and subtraction of LP residual from the signal. However, in QCP method, the speech signal is first down weighted at the places of glottal excitation. Subsequently, weighted LP [15] is used to estimate the impulse response of the system.

Another glottal source estimation method that uses zeros of Z-transform representation is presented in [16] and [17]. It uses the LP polynomial to separate causal and anti-causal roots contributing to source and filter spectrum respectively. In [18], homomorphic filtering is used for deconvolution of the speech signal to retrieve the glottal source and vocal tract filter. In [4], closed phase covariance technique, a method similar to [11], is used to obtain an initial estimate of the glottal signal. Subsequently, the predicted signal is refined by fitting with the Liljencrants–Fant (LF) model [19]. Another method that uses similar LF model fitting has been given in [20]. A lattice structured cascaded digital filter for inverse filtering of speech is presented in [20]. The filter provides a solution for the LP coefficients. It is implemented using reflection coefficients from the Levinson–Durbin [9] recursion. The speech signal is given as input to the first filter, and the LP residual is produced at the other end.

Previous studies demonstrate enough accuracy in the inverse computation of glottal and vocal tract responses from the recorded speech signal. However, in lumped model of the vocal tract, the shape is assumed to be constant in an interval. Hence, it is not possible to know the spatial distribution of air pressure or flow within the tract. Moreover, in LP-based methods, performance degrades with the increase in fundamental frequency (pitch) of the speech signal due to short closed phase duration. In this paper, a new technique has been proposed for inverse estimation of the glottal signal along with air pressure distribution in the vocal tract. The method is based on distributed model of the vocal tract. The state space model of the speech production system has been formulated using concatenated tube model of vocal tract [9] and LF [19] model of glottal flow. The system

Manuscript received October 25, 2015; revised March 29, 2016; accepted April 3, 2016. Date of publication April 7, 2016; date of current version May 9, 2016. The associate editor coordinating the reviewof this manuscript and approving it for publication was Dr. Sin-Horng Chen.

The authors are with the Department of Electrical Engineering, Indian Institute of Technology Kharagpur, Kharagpur 721302, India (e-mail: subhasmita@ee.iitkgp.ernet.in; aurobinda.routray@gmail.com).

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/TASLP.2016.2551864

becomes hybrid due to the piecewise differentiable nature of the glottal signal in LF model. Therefore, separate state space models have been used to characterize the speech production system at different time intervals. The proposed formulation is different from that in [21]. Instead of presenting the vocal tract as a cascade of systems, it has been represented as a single system with state variables distributed across the length of the tube. The glottal source acts as the input, and the speech signal acts as the output of the whole system. However, the system becomes nonlinear because of the unknown glottal source signal and the model parameters.

There are several state estimation techniques for nonlinear systems. The EKF [22] linearizes the nonlinear system equations and uses Jacobian matrices for linear transformations in the Kalman filter algorithm. The unscented Kalman filter (UKF) [23] is used to overcome the issues in EKF like difficulties in implementation, tuning, and restricted applicability to nearly linear systems. The EKF and UKF are applied to nonlinear systems having known model parameters. However, for nonlinear systems with unknown model parameters, both state and parameters need to be estimated simultaneously. An extension of EKF for adaptive system identification [24] estimates the model parameters along with the states. As the formulated model of the speech production system is nonlinear, hybrid, and contains unknown parameters, EKF based adaptive system identification has been used with multiple model estimation (MME) method [25], [26]. The EKF estimates the glottal source signal by switching between the two set of models.

The major contributions of this paper are: (1) Formulation of a new state space model of the speech production system using the existing models of the vocal tract and glottis, (2) estimation of the glottal source signal as well as the air pressure distribution in the vocal tract by applying EKF on the formulated model. The proposed method of GIF has been compared with other existing methods of glottal signal estimation. The method has been found to be more accurate, robust to the increase in pitch, and gives provision to visualize the temporal variation of air pressure at different sections of the vocal tract. However, the proposed method has been found to be computationally complex as compared to other techniques with respect to computational time.

Rest of this paper has been organized as follows. Section II describes the state space formulation of the speech production system. The state equations, derived from the vocal tract model and glottal source signal model, have been described separately. The non-linear state space model used with the EKF algorithm has also been explained in this section. The estimation procedure has been described in Section III. Section IV presents the results of inverse filtering. Possible applications of the method have also been described in this section. Section V presents the conclusion and future work. The EKF estimation algorithm has been described in Appendix A.

II. State Space Formulation of the Speech Production System

The glottal source signal is produced by the continuous vibration of vocal folds. It subsequently stimulates the vocal tract to produce the speech signal. The state space model of the speech production system has been formulated by combining the state equations of the glottis and the vocal tract. The concatenated tube model of vocal tract [9] and the LF model [19] of glottal flow derivative has been used in the formulation.

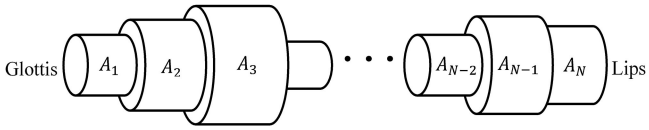

Fig. 1. The concatenated tube model.

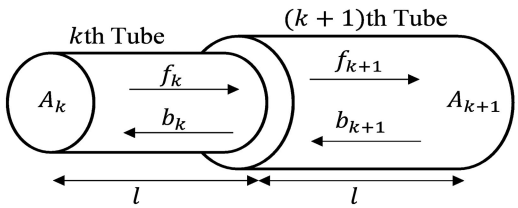

Fig. 2. The junction of kth and $(k+1)$th tube.

A. State Equations From Vocal Tract Model

The concatenated tube model represents the vocal tract as the concatenation of several discrete cylindrical tubes of uniform length [27], [28]. The sound propagation through the vocal tract is assumed as a one-dimensional (1-D) propagation of air-volume through a set of confined tubes (see Fig. 1). Let the number of discrete tubes be N. The 1-D waveguide equations, written in terms of the forward and backward traveling air volume velocities through different tube sections, are as follows (see [9] for derivation):

1) For the 1st tube adjoining glottis
$f_1(t) = \frac{1}{2}(1+r_g)\,u_g(t) + r_g b_1(t)$
$b_1(t+\tau) = -r_1 f_1(t-\tau) + (1-r_1)\,b_2(t).$
2) For $k = 2$ to $N-1$, i.e., for all the intermediate tubes
$f_k(t) = (1+r_{k-1})\,f_{k-1}(t-\tau) + r_{k-1} b_k(t)$
$b_k(t+\tau) = -r_k f_k(t-\tau) + (1-r_k)\,b_{k+1}(t).$
3) For the Nth tube adjoining lips
$f_N(t) = (1+r_{N-1})\,f_{N-1}(t-\tau) + r_{N-1} b_N(t)$
$b_N(t+\tau) = -r_N f_N(t-\tau)$
4) Output volume velocity at the lips
$u_N(t) = (1+r_N)\,f_N(t-\tau)$

where, r_g, r_k, and r_N are the reflection coefficients represented as, $r_g = \left(\frac{Z_g - \frac{\rho c}{A_1}}{Z_g + \frac{\rho c}{A_1}}\right)$, $r_k = \left(\frac{A_{k+1} - A_k}{A_k + A_{k+1}}\right)$, $r_N = \left(\frac{\frac{\rho c}{A_N} - Z_r}{\frac{\rho c}{A_N} - Z_r}\right)$. A_k and A_{k+1} are cross-sectional areas of kth and $(k+1)$th tube. $f_k(t)$ and $b_k(t)$ represent the forward and backward traveling air volume velocities through the kth tube (see Fig. 2). ρ is the density of air, c is the velocity of sound in air, Z_g is the glottal impedance, Z_r is the radiation impedance at lips, and $u_g(t)$ is the input to the vocal tract from glottis. τ represents the time consumed by air volume to pass through one tube section. $\tau = \frac{l}{c}$, where l is the length of each tube section.

B. The Linear State Space Model

A discrete state space model of the speech production system has been formulated from the 1-D waveguide equations. The air-volume velocities at different tube junctions have been considered as the state variables. However, to retain the stability of finite difference equations, the spatial sampling interval of vocal tract Δx and the temporal sampling interval of speech signal ΔT should be selected so as to satisfy the condition $c\Delta T \leq \Delta x$. Let the speech signal be sampled with sampling period exactly equal to the time consumed by the air-volume to pass through a tube section, i.e., $\Delta T = \tau$. Replacing t with n, and τ with one time unit, the 1-D waveguide equations can be arranged in a state space form:

$$\begin{aligned} X[n+1] &= A_n X[n] + B_n u_g[n] \\ u_N[n+1] &= C_n X[n]. \end{aligned} \tag{1}$$

The state vector $X[n]$, state matrices A_n, B_n, and C_n, are illustrated in Eqs. (2), (3), (4) and (5), respectively. Matrix elements relating forward and backward traveling volume velocities have been demarcated by solid lines

$$X[n] = \left[\begin{array}{c} X_1[n] \\ \hline X_2[n] \end{array}\right] \tag{2}$$

where,

$$X_1[n] = \begin{bmatrix} f_1[n-1] \\ f_2[n-1] \\ \vdots \\ f_N[n-1] \end{bmatrix}, X_2[n] = \begin{bmatrix} b_1[n] \\ b_2[n] \\ \vdots \\ b_N[n] \end{bmatrix}.$$

The state variables in $X_1[n]$ and $X_2[n]$ are the air-volume velocity of forward and backward traveling waves respectively, through tube 1 to N,

$$A = \left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right] \tag{3}$$

where,

$$A_{11} = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ 1+r_1 & 0 & \cdots & 0 & 0 \\ 0 & 1+r_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1+r_{N-1} & 0 \end{bmatrix}$$

$$A_{12} = \begin{bmatrix} r_g & 0 & \cdots & 0 & 0 \\ 0 & r_1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & \cdots & r_{N-2} & 0 \\ 0 & 0 & \cdots & 0 & r_{N-1} \end{bmatrix}$$

$$A_{21} = \begin{bmatrix} -r_1 & 0 & \cdots & 0 & 0 \\ 0 & -r_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & -r_{N-1} & 0 \\ 0 & 0 & \cdots & 0 & -r_N \end{bmatrix}$$

$$A_{22} = \begin{bmatrix} 0 & 1-r_1 & 0 & \cdots & 0 \\ 0 & 0 & 1-r_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1-r_{N-1} \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}.$$

A_{11} relates the air-volume velocity of forward traveling waves at time n and at time $n-1$.
A_{12} relates the air-volume velocity of forward traveling waves at time n and air-volume velocity of backward traveling waves at time n.
A_{21} relates the air-volume velocity of backward traveling waves at time $n+1$ and forward traveling waves at time $n-1$.
A_{22} relates the air-volume velocity of backward traveling waves at time $n+1$ and at time n.

$$B = \left[\begin{array}{c} B1 \\ \hline B2 \end{array}\right] \tag{4}$$

where,

$$B_1 = \begin{bmatrix} \frac{1}{2}(1+r_g) \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, B_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

B_1 relates the volume velocity of forward traveling waves with the glottal input signal $u_g[n]$. B_2 relates the backward traveling waves with the glottal input signal

$$C = \left[\begin{array}{c|c} C_1 & C_2 \end{array}\right], \tag{5}$$

where,

$$\begin{aligned} C_1 &= \left[0\ 0 \cdots 0\ 1+r_N\right], \\ C_2 &= \left[0\ 0\ 0 \cdots 0\right]. \end{aligned}$$

C_1 and C_2 relate the output speech signal to the forward traveling waves and the backward traveling waves, respectively.

State space representation of waveguide equations helps simultaneous update of all the state variables at the advent of every incoming sample of $u_g[n]$. This process resembles the natural speech production mechanism, where, the pressure change throughout the vocal tract takes place in a time synchronous manner. An important point to be noted is that all state variables in the state space model are in volume velocity form. Whereas, the measured speech signal is a pressure signal. The air volume velocity is converted to pressure through lip radiation, which is

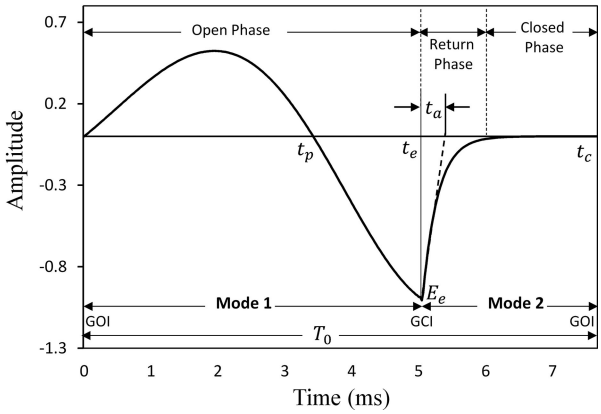

Fig. 3. The glottal flow derivative.

modeled by a simple derivative [9]. As the system is linear, the lip radiation effect can be inserted after the glottis. Therefore, instead of taking glottal flow as input, its derivative has to be taken.

In the state space formulation of (1), the only measurable entity is the speech signal. The input to the system, i.e., the glottal flow derivative $u_g[n]$, is unknown. Therefore, a mathematical model of the glottal flow derivative, given by Liljencrants–Fant in [19], has been used to obtain the state equations. Later, the state equations of the glottis and the vocal tract have been combined in one model to estimate the glottal signal as well as the pressure distribution in the vocal tract.

C. State Equations From Glottal Flow Model

Let one cycle of the glottal flow derivative waveform (shown in Fig. 3) be denoted by $E(t)$. The mathematical expression of glottal flow derivative signal in the LF model is

$$\begin{aligned} E(t) &= E_0 e^{\alpha t}\sin(\omega_g t)\,; t < t_e \\ &= -\frac{E_e}{\varepsilon t_a}\left[e^{-\varepsilon(t-t_e)} - e^{-\varepsilon(t_c-t_e)}\right]; t_e < t < t_c \end{aligned} \quad (6)$$

$$\text{with constraints,} \begin{matrix} \int_0^{T_0} E(t)\,dt = 0, \text{where,} T_0 = t_c = \frac{1}{F_0}, \\ \varepsilon t_a = 1 - e^{-\varepsilon(t_c - t_e)} \end{matrix} \quad (7)$$

where t_e is the glottal closing instant at which the negative peak E_e of the glottal flow derivative occurs. t_p is the time instant at which positive peak occurs in the glottal flow signal, and the glottal flow derivative becomes zero. t_a is related to the return phase duration. t_c is the time instant at which the glottis starts to open after a closed phase. The parameters are shown in Fig. 3. T_0 is the fundamental period, and F_0 is the fundamental frequency (pitch) of the speech signal. E_0, ω_g, and α describe the shape of the signal during open phase. E_e and ϵ describe the shape of the return phase. ϵ, α, and E_0 need to be determined from the constraints of Eq. (7). These constraints are imposed to achieve area balance and to maintain continuity at t_e. The requirement of area balance is to make sure that the net air-flow in one pitch cycle is zero.

As the glottal flow derivative signal is piecewise differentiable, one pitch period has been divided into two separate time intervals: Mode 1 ($t < t_e$) and Mode 2 ($t_e < t < t_c$). The discrete state equations of the glottal flow model in both the intervals have been obtained as follows:

1) For Mode 1, i.e., $t < t_e$:

Let $u_1(t) = E_0 e^{\alpha t}\sin(\omega_g t)$ and $u_2(t) = \dot{u_1}(t)$.

$$\dot{u_1}(t) = \alpha u_1(t) + \omega_g E_0 e^{\alpha t}\cos(\omega_g t), \quad (8)$$

$$\dot{u_2}(t) = 2\alpha u_2(t) - \left({\omega_g}^2 + \alpha^2\right) u_1(t). \quad (9)$$

Now, using the forward difference formula for derivatives $\left(\text{i.e., } \dot{u}(t) = \frac{u(t+\tau)-u(t)}{\tau}\right)$ in Eq. (8) and Eq. (9), the expression of $u_1(t+\tau)$ and $u_2(t+\tau)$ becomes

$$u_1(t+\tau) = u_1(t) + \tau u_2(t), \quad (10)$$

$$u_2(t+\tau) = -\left({\omega_g}^2 + \alpha^2\right)\tau u_1(t) + (1 + 2\alpha\tau)\,u_2(t). \quad (11)$$

2) For Mode 2, i.e., $t_e < t < t_c$:

Let $u_1(t) = -\frac{E_e}{\varepsilon t_a}\left[e^{-\varepsilon(t-t_e)} - e^{-\varepsilon(t_c-t_e)}\right]$ and $u_2(t) = \dot{u_1}(t)$.

$$\dot{u_1}(t) = -\varepsilon u_1(t) + \frac{E_e}{t_a} e^{-\varepsilon(t_c - t_e)}, \quad (12)$$

$$\dot{u_2}(t) = -\varepsilon u_1(t). \quad (13)$$

Proceeding in a similar manner as in the Mode 1, the expression of $u_1(t+\tau)$ and $u_2(t+\tau)$ can be obtained as

$$u_1(t+\tau) = u_1(t) + \tau u_2(t), \quad (14)$$

$$u_2(t+\tau) = (1 - \tau\epsilon)u_2(t). \quad (15)$$

D. The Non-Linear State Space Model

Inserting the state equations of the glottal signal in the linear state space model of (1), the resulting state space model can be written in the non-linear form:

$$\begin{aligned} \underline{x}_{n+1} &= A_n(\underline{\theta}_n)\underline{x}_n + \underline{\xi}_n, \\ u_n &= C_n(\underline{\theta}_n)\underline{x}_n + \eta_n, \end{aligned} \quad (16)$$

where $\underline{x}_n$ is the state vector, $\underline{\theta}_n$ is the parameter vector, and u_n is the measured speech signal. $\underline{\xi}$ and η represent the Gaussian white noise sequences associated with the system and observations, respectively. $A_n(\underline{\theta}_n)x_n$ and $\underline{\theta}_n$ will be different for the two modes of the glottal cycle due to distinct state equations of the glottal signal in Mode 1 and Mode 2.

1) For $t < t_e$, i.e., Mode 1:

$$\underline{x}_n = \begin{bmatrix} f_1[n-1] \\ f_2[n-1] \\ \vdots \\ f_N[n-1] \\ \hline b_1[n] \\ b_2[n] \\ \vdots \\ b_N[n] \\ \hline u_1[n] \\ u_2[n] \end{bmatrix}, \underline{\theta}_n = \begin{bmatrix} r_g[n] \\ r_1[n] \\ r_2[n] \\ \vdots \\ r_{N-1}[n] \\ r_N[n] \\ \omega_g[n] \\ \alpha[n] \end{bmatrix} \tag{17}$$

$$A_n(\underline{\theta}_n)x_n = \begin{bmatrix} r_g b_1[n] + \frac{1}{2}(1+r_g)u_1[n] \\ (1+r_1)f_1[n-1] + r_1 b_2[n] \\ \vdots \\ (1+r_{N-1})f_{N-1}[n-1] + r_{N-1}b_N[n] \\ \hline -r_1 f_1[n-1] + (1-r_1)b_2[n] \\ -r_2 f_2[n-1] + (1-r_2)b_3[n] \\ \vdots \\ -r_{N-1}f_{N-1}[n] + (1-r_{N-1})b_N[n] \\ -r_N f_N[n-1] \\ \hline u_1[n] + \tau u_2[n] \\ -\left(\omega_g^2 + \alpha^2\right)\tau u_1[n] + (1+2\alpha\tau)\,u_2[n] \end{bmatrix} \tag{18}$$

$$C(\underline{\theta}[n]) = [\,0 \cdots 0\ 1 + r_N \,|\, 0\ 0 \cdots 0\ 0 \,|\, 0\ 0\,]. \tag{19}$$

2) For $t_e < t < t_c$, i.e., Mode 2: All the state matrix elements, except those corresponding to the expression of $u_g[n]$, remain same. The last two elements of $A_n(\underline{\theta}_n)x_n$ are replaced by the discrete forms of Eqs. (14) and (15), respectively. The last two elements of $\underline{\theta}_n$ (i.e, ω_g and α) are replaced by ϵ.

The non-linear state space model of (16) contains unknown parameters. Because the reflection coefficients of the vocal tract model and the parameters of the glottal flow signal are unknown. EKF for adaptive system identification [24] has to be applied to estimate the unknown parameters as well as the states. The steps of estimation algorithm have been illustrated in Appendix A. However, due to the distinct state space models of the two modes in one glottal cycle, the EKF algorithm has to be applied to Mode 1 and Mode 2 by switching between the models given in Section II-D1 and Section II-D2.

III. The Estimation Method

A. Estimation of Glottal Closing Instants (GCI) and Glottal Opening Instants (GOI)

The first requirement in the proposed analysis is the estimation of GOI and GCI. In Fig. 3, GCI and GOI have been marked by t_e and t_c, respectively. GCI is the time instant at which the Mode 1 ends and the Mode 2 starts; whereas, GOI is the time instant at which the Mode 2 ends and the Mode 1 starts. The algorithm described by Plumpe *et al.* in [4] has been used to determine GCIs and GOIs. In this algorithm, GCI are detected by looking at the peak locations in the LP residual signal. Pitch synchronous tracking of peaks is carried out to achieve better accuracy. The peaks are searched only in those regions, where it is expected to occur. For estimation of GOI, closed phase regions are detected between each pair of successive GCIs. After every closed phase, when glottis starts to open, formant change occurs due to the interaction between glottis and vocal tract. LP covariance based method is employed for formant tracking due to better accuracy as compared to correlation based method. The end point of the extracted closed phase region is considered as the opening instant.

B. Initialization of EKF for Proper Convergence

Next major step is the suitable initialization of EKF for quick convergence (see Appendix A for EKF algorithm). The state vector $\underline{x}$, parameter vector $\underline{\theta}$, estimation error covariance matrix P need to be properly initialized. System noise covariance matrix Q, parameter noise covariance matrix S, and observation noise variance R should be assigned proper values for convergence. $\underline{x}$ has been initialized to zeros assuming that there is no air flow initially. The reflection coefficients (r_1 to r_{N-1}) of $\underline{\theta}$ have been initialized to values calculated from experimentally obtained vocal tract area given in [29]. The chosen areas should correspond to the vowel associated with the speech. In a continuous speech, the vowels can be determined from the formants. r_g and r_N have been chosen as positive values close to, but less than one. ω_g, α, and ϵ have been assigned random positive values less than one. The P matrix has been initialized to an identity matrix. Diagonal matrices of appropriate dimensions, with a small value (0.001 for Q and 0.01 for S) at all the diagonal locations, have been assigned to Q and S. These matrices have been considered as diagonal because of the assumption that the system noises and parameter noises, associated with the states and parameters respectively at different tube sections, are independent. R has been chosen to be 0.01.

The values assigned to the system noise covariance matrix Q, parameter noise covariance matrix S, and the initial values of the unknown parameters play an important role in the convergence of the algorithm. For unknown vocal tract parameters, the initial values computed from experimentally obtained vocal tract areas (corresponding to the vowel in speech) are sufficient for the algorithm to converge. For the unknown glottal parameters, any positive value less than one works well. Convergence is not much affected by the initial values of the glottal parameters, because the number of unknown glottal parameters is very less as compared to the total number of unknown parameters in the system. The values assigned to the parameter noise covariance S do not much affect convergence due to less deviation in unknown parameters with variation in speaker and speech type. However, a suitable value for Q is very much crucial to ensure convergence. As the matrix Q represents system noise covariance (i.e., the inability of the model in describing the process),

its value for convergence largely depends on the amplification level of the speech signal. If a large value is assigned, the estimated states may contain large error. If the assigned value is too small as compared to the speech amplitude, the Kalman gain matrix elements may become very high, and sometimes may go beyond the computational limits. As a rule of thumb, if the algorithm does not converge for the given values of Q, the speech signal should be scaled down (by multiplying with a suitably chosen positive value less than one) to keep the elements of Kalman gain matrix within computable limits.

The sampling time of speech should be chosen carefully in order to satisfy the stability condition $c\Delta T \leq \Delta X$. For example, in the case of the vowel /a/ (as in *"bart"*), there are 44 sections of the vocal tract [29]. Considering the length of the vocal tract to be 17.46 cm, the spatial sampling ΔX (same as the length of each small tube) becomes 0.397 cm. In order to satisfy the stability condition, the sampling frequency $1/\Delta T$ should be at least 84, 170 Hz. The velocity of air is assumed to be 33, 400 cm/s.

C. The Multiple Model Estimation

The MME technique is used in hybrid systems which are represented by different model equations at different time intervals [25], [26]. MME has been employed for inverse filtering of the continuous speech signal due to the piecewise differentiable nature of the glottal flow derivative. In this approach, at every GCI and GOI, the EKF estimation algorithm (given in Appendix A) switches from one model to the other. While switching, the values of common state variables such as $\underline{x}$ and the impedance terms in $\underline{\theta}$ are retained. However, as the parameters of the glottal waveform change at every switch, these parameters and their corresponding terms in P assume values from the preceding cycle. For instance, during Mode 1, $\underline{\theta}$ is of order 47×1, $\underline{x}$ is of order 90×90, and P is of order 137×137. In this mode, the 136th and 137th row and column of P correspond to the last two elements of $\underline{\theta}$, i.e., ω_g and α. Similarly, during Mode 2, $\underline{\theta}$ is of order 46×1, P is of order 136×136, and order of $\underline{x}$ remains the same. In this period, the 136th row and column of P correspond to the last element of $\underline{\theta}$, i.e., ϵ. So, while switching from the Mode 1 to the Mode 2 (at GCI), the last two elements of $\underline{\theta}$, which correspond to ω_g and α, are removed and saved for use in the next switching, i.e., from Mode 2 to Mode 1. The value of ϵ from the previous glottal cycle is appended at the 46th place of $\underline{\theta}$. In a similar manner, the 136th and 137th row and column of P are replaced by the 136th row and column of P from the previous cycle. The values of the replaced rows and columns are saved for use in the next cycle. A similar procedure is followed while switching from Mode 2 to Mode 1 (at GOI). However, $\underline{x}$ retains its values at every switch as all states are identical for both the systems.

IV. Results of Inverse Filtering

A. Test Results With Synthesized Test Data

The proposed inverse filtering technique was first tested on synthesized speech with known glottal flow derivative waveform. Vowels having different fundamental frequency were synthesized using the linear state space model given in (1). The glottal flow derivative signals were generated using LF model of (6). A comparison of the original glottal flow derivative waveform used for synthesizing the vowel /a/, and the glottal signal estimated by the proposed method, has been shown in Fig. 4(a) and (b), respectively. The figure shows waveforms for three different fundamental frequencies. For comparison, glottal waveforms obtained by QCP inverse filtering method [13] and IAIF method [12], [30]—when applied to the same speech signal—have also been shown in Fig. 4(c) and (d), respectively. It is apparent from the figure that, the glottal signal estimated by the proposed technique has maximum similarity with the original, as compared to QCP and IAIF method. Besides, with an increase in pitch, degradation in the performance of the proposed method is less as compared to others.

Several parameters that describe the glottal signal have been used for quantitative evaluation of inverse filtering performance of the proposed technique and other existing methods (QCP and IAIF). These parameters are normalized amplitude quotient (NAQ) [31], quasi-open quotient (QOQ) [32], H1H2 [33], and mean square error (MSE) between the original and the estimated glottal signal [13]. NAQ and QOQ represent the relative closed phase and open phase duration of the glottal cycle respectively (see Fig. 3). H1H2, a frequency domain parameter that represents voice quality, is computed as the difference between the first and second harmonic. MSE represents error between amplitude of the original and estimated time domain glottal signal. The mean of the absolute relative errors between the NAQ and QOQ parameters of the original glottal signal and that of the estimated one has been computed as

$$ENAQ = \frac{1}{P}\sum_{p=1}^{P}\frac{|NAQ_{\mathrm{ref}}[p] - NAQ_{\mathrm{est}}[p]|}{NAQ_{\mathrm{ref}}[p]} \tag{20}$$

$$EQOQ = \frac{1}{P}\sum_{p=1}^{P}\frac{|QOQ_{\mathrm{ref}}[p] - QOQ_{\mathrm{est}}[p]|}{QOQ_{\mathrm{ref}}[p]} \tag{21}$$

where $NAQ_{\mathrm{ref}}[p]$ and $NAQ_{\mathrm{est}}[p]$ represent the NAQ values of the reference and the estimated signal for the pth glottal cycle. $QOQ_{\mathrm{ref}}[p]$ and $QOQ_{\mathrm{est}}[p]$ represent the QOQ values of the reference and estimated signal for the pth cycle. P denotes the total number of glottal periods present in the vowel speech. For H1H2, mean of the absolute error has been considered:

$$EH1H2 = \frac{1}{P}\sum_{p=1}^{P}|H1H2_{\mathrm{ref}}[p] - H1H2_{\mathrm{est}}[p]| \tag{22}$$

where $H1H2_{\mathrm{ref}}[p]$ and $H1H2_{\mathrm{est}}[p]$ represent the H1H2 values of the reference and estimated signal, respectively, for the pth glottal period. The MSE has been calculated as the mean of the squared errors as in [13]:

$$EMSE = \frac{1}{N}\sum_{n=1}^{N}\left(g_{\mathrm{ref}}[n] - g_{\mathrm{est}}[n]\right)^2 \tag{23}$$

where N represents the number of samples in the inverse filtered speech, g_{ref} represents the original glottal waveform, and g_{est}

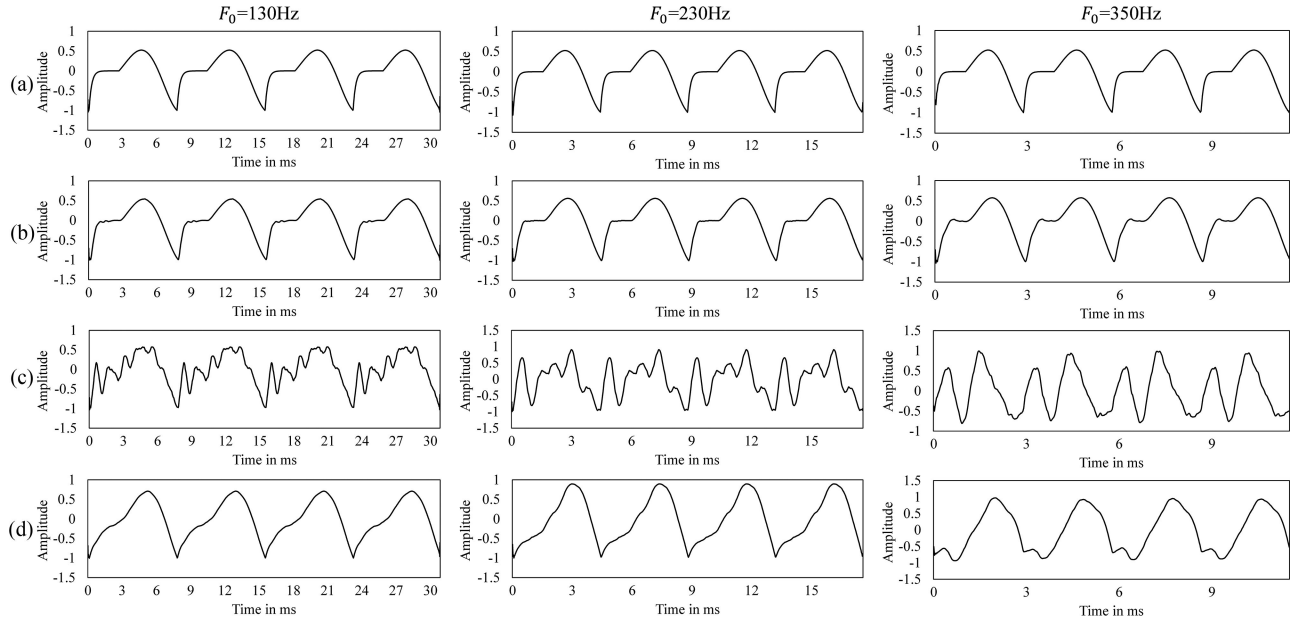

Fig. 4. (a) The original glottal flow derivative waveform. (b) Signal estimated by proposed method. (c) Signal estimated by QCP method. (d) Signal estimated by IAIF method.

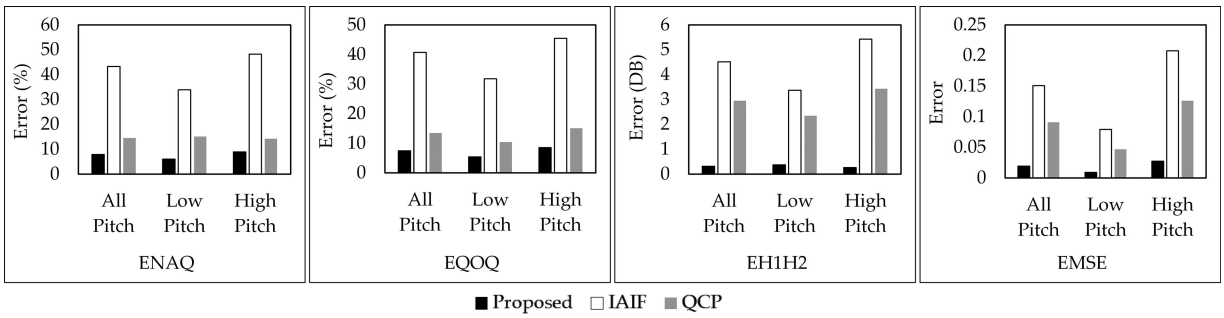

Fig. 5. Comparison of errors in various parameters of the glottal signal estimated by the proposed method, IAIF method, and QCP method.

represents the estimated glottal signal. The parameters (NAQ, QOQ, H1H2, and MSE) were evaluated for speech signals with different ranges of fundamental frequency. The pitch values as within $80 \leq F_0 \leq 200$ were considered as the low pitch range, and $200 \leq F_0 \leq 350$ as the high pitch range. The fundamental frequency of the synthesized vowels was varied in steps of 10 Hz starting from the lower limit, and mean of the errors was computed over the entire range. Fig. 5 compares the errors in parameters of the glottal signal estimated by all the three methods for low, high, and all frequency range. As can be seen from the figure, the errors are less for the proposed method as compared to QCP and IAIF method. In most of the parameters for all the three methods, error for high-pitched speech signal can be observed to be more as compared to the low pitched speech. However, for all the pitch ranges, the performance of the proposed technique is better than that of other methods.

B. Test Results With Reduced Number of Vocal Tract Sections

In order to maintain stability, the temporal sampling time should be at most same as the time consumed by the sound wave to cross one section of the vocal tract. If the number of sections in the vocal tract would be very high—thereby making the length of each section very small—to maintain the stability condition, the required sampling period will be very short. As given in the example of Section III-B, the required sampling period for a 44 section vocal tract is at least 84.17 kHz. The high sampling frequency requirement limits the practical applicability of the technique.

Therefore, the system performance was also evaluated while considering less number of sections in the vocal tract. The number of sections was reduced to $1/3$rd of the original numbers. The initial values for these reduced number of vocal tract sections were taken by down sampling the experimentally obtained vocal tract area given in [29]. With the reduction in number of sections, the required sampling frequency for stability was also reduced. For instance, for the vowel /a/ (as in "*bart*"), 44 sections in the original data was reduced to only 15 sections. Therefore, the required sampling frequency of the speech signal became 28.7 kHz, instead of 84.17 kHz. As the maximum number of vocal tract area samples

for any particular vowel in the data of [29] is 46, the minimum required frequency remains within 28.7 kHz.

The performance of the proposed technique, with original and reduced number of vocal tract sections, was compared to that of IAIF and QCP method. All the three methods were applied to the same original synthesized speech signal as well as that with reduced sampling frequency. However, in the proposed algorithm, less number of vocal tract sections was considered for the signal with reduced sampling frequency. Comparison of the errors in glottal parameters for all the three methods, applied to speech signals with original and reduced sampling frequency has been shown in Fig. 6. The errors have been averaged over the entire pitch range (80–350 Hz). As can be seen in the figures, with a reduction in sampling frequency the performance of all the three methods is degrading. However, the performance of the proposed method can be observed to be still better than the other two.

C. Sensitivity to Error in GCI and GOI Estimation

As the proposed method involves detection of the glottal opening (GOIs) and closing instants (GCIs), the robustness of the technique to GCI estimation error has also been evaluated. This evaluation has been performed on the synthesized speech data with reduced sampling frequency. Similar to the procedures followed in the preceding section, the number of vocal tract sections was reduced to $1/3$rd of the original numbers given in [29]. Therefore, the sampling frequency was also reduced by the same factor. Uniform deviations of fixed number of samples (16, 32, and 64), representing errors of about 0.5, 1.1, and 2.2 ms, were added to the GCIs and GOIs. Then, the speech signal was inverse filtered taking these erroneous GCIs and GOIs. The errors in NAQ, QOQ, H1H2, and the MSEs at different sample deviations (averaged over the entire pitch range 80–350 Hz) are shown in Fig. 7. As evident from the figures, added deviation has a very negligible effect on the performance. The reason might be the fact that there is a very small difference in the state equations of both the systems. Therefore, even if the place of switching between the two systems is altered by changing the position of GCI and GOI, the algorithm is still able to estimate the glottal signal with almost negligible effect on performance.

D. Test Results With Natural Speech Data

The GIF performance of the proposed method, when applied to different vowels of natural speech signal, has been compared with those of the IAIF and QCP method in Fig. 8. It is difficult to evaluate the GIF performance for natural speech due to the absence of any standard reference glottal signal. In [20], a goodness of fit comparison has been used. In this technique, the MSE between a reconstructed glottal signal and the estimated signal is compared. However, as the LF parameters for reconstruction of the reference glottal signal are also estimated using tracking methods, the efficiency of the technique is questionable. The GIF results can also be compared visually, and inferences can be drawn [8]. The estimated glottal signals shown in Fig. 8 are not exactly similar in shape to the synthesized one; because the vowel samples have been taken from natural conversations. The glottal signal estimated by the IAIF method is smoother as compared to the other two. However, the estimated waveform by the proposed and the QCP method is close to the shape represented in the LF model. To some extent, the proposed method is able to reduce the ripples present in the closed phase. It can also be observed that the glottal signal in the open phase is less abrupt for soft voice as compared to the pressed voice.

E. Estimation of Air Pressure Variation in the Vocal Tract

The proposed GIF technique also keeps track of the flow variation within the vocal tract during speech production. The air-volume velocity at any section of the vocal tract is the difference between the forward and backward traveling volume velocities, one time instant before and after respectively [9]. The volume velocity in tube k, at time n, can be written as

$$u_k[n] = f_k[n-1] - b_k[n+1]. \tag{24}$$

Thus, one can obtain the air-volume velocity through any tube section, at any instant of time, by combining appropriate terms of the state vector $\underline{x}$. For instance, to get the air volume velocity in tube 1 at time n, the 1st element of $\underline{x}$ (of Eq. (17)) at time n should be added to the negative of 45th element of $\underline{x}$ at time $n+1$. This will be same as the subtraction of $b_1[n+1]$ from $f_1[n-1]$, resulting in $u_1[n]$. However, as the lip radiation effect has been introduced after the glottis, this will directly give the pressure value. The air pressure variation at five different sections of the vocal tract has been shown in Fig. 9. As can be seen from the figures, a time lag exists between the actual and estimated signal. The lag can be seen to be increasing while moving from lips towards the glottis. This lag comes into picture due to the finite amount of time consumed by the air volume velocity to travel from one end to the other.

To the best of our knowledge, this is the first method for GIF that can estimate the air pressure distribution in the vocal tract. Therefore, no comparison of the air pressure distribution results could be made with other existing methods. In [21], a lattice-structured digital inverse filter is formulated to solve the LP problem. The filter is claimed to be equivalent to the concatenated tube model under certain assumptions. As the reflection coefficients of the concatenated tube model are unknown, the filter is implemented using the reflection coefficients from the Levinson–Durbin recursion [9]. Therefore, the estimated error signal at every stage of the cascaded digital filter represents the LP residuals. However, as the glottal input is not pure impulse, the method fails to estimate the actual flow variation at different sections.

F. Computational Complexity Analysis

Considering computational time as the complexity measure, the computational complexity of the proposed technique has been compared with those of the IAIF and QCP method. The algorithms were run on a Dell Optiplex desktop having 64-bit OS, Intel Core i5 processor, 3.30 GHz speed. The methods were applied to a speech signal of duration 0.23 s. The algorithms were run 100 times on the same speech data, and average of the computational times was taken. Before testing the time

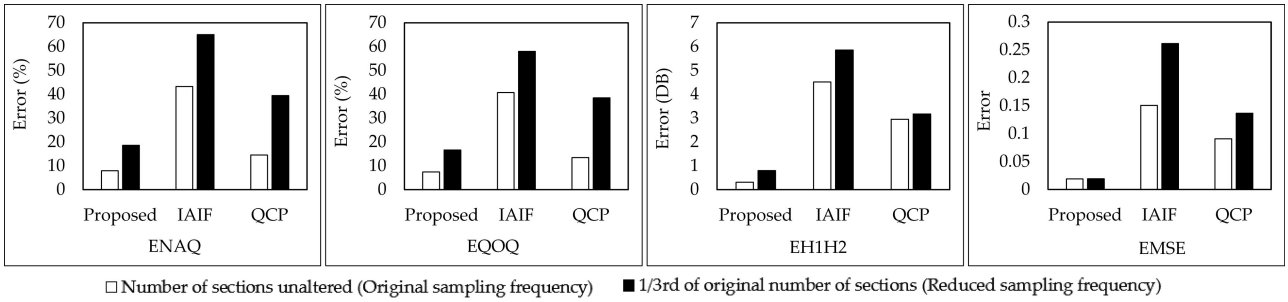

Fig. 6. Performance comparison between all the three methods for speech signals with original and reduced sampling frequency. The number of vocal tract sections for the proposed method is less for the reduced sampling frequency case.

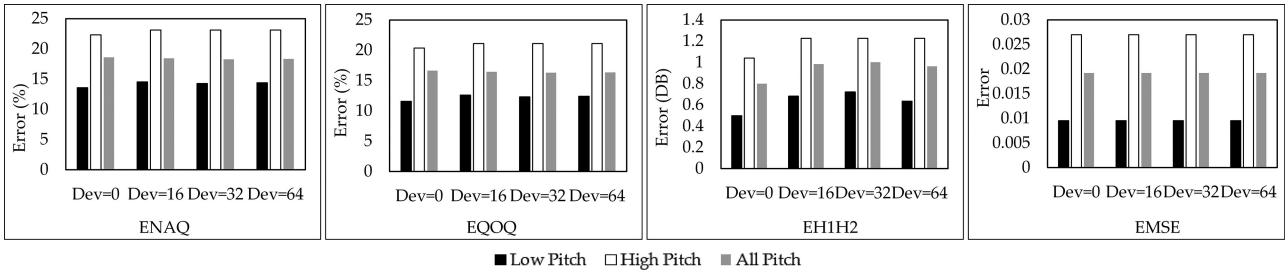

Fig. 7. Errors in glottal signal parameters for different sample deviations added to the GCIs and GOIs.

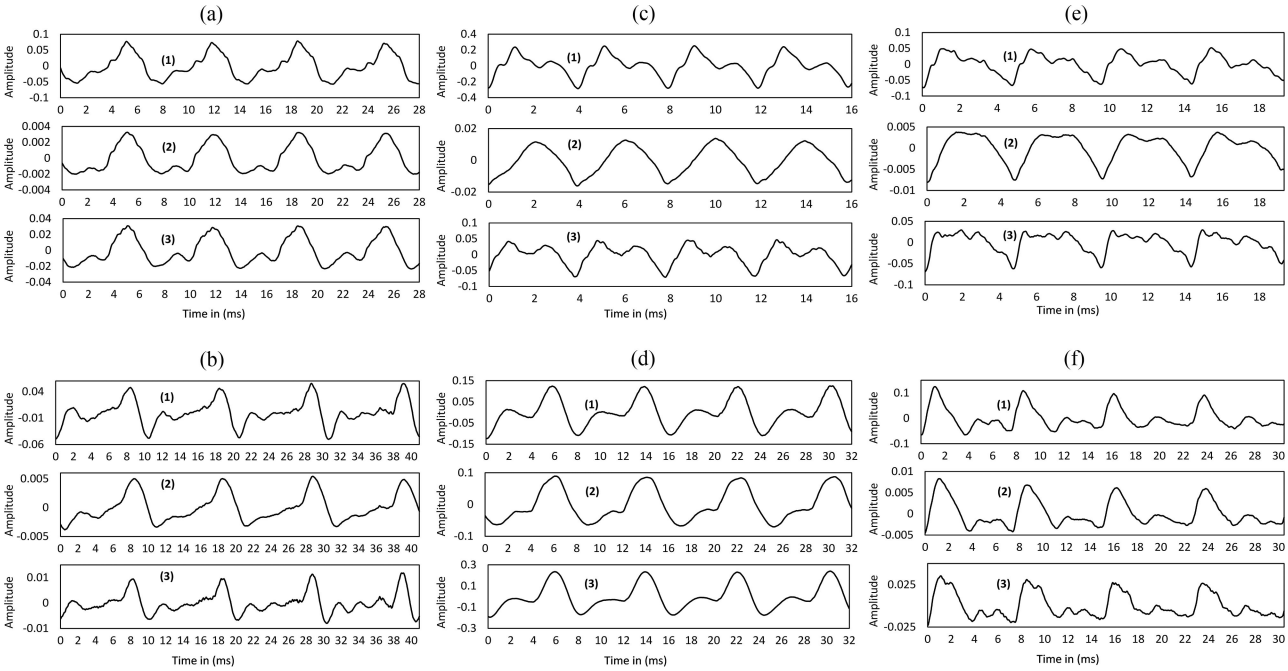

Fig. 8. GIF results for real speech. (a) /i/ vowel with pressed voice. (b) /i/ vowel with soft voice. (c) /e/ vowel with pressed voice. (d) /e/ vowel with soft voice. (e) /a/ vowel with pressed voice. (f) /a/ vowel with soft voice. In each figure (1) is for proposed method, (2) is for IAIF method and (3) is for the QCP method.

complexity, all other background programs were terminated to eliminate fluctuations in processor usage. The average computational time of the proposed method, QCP, and IAIF were found to be 0.68, 0.28, and 0.02 s, respectively. This testing has been carried out with reduced number of vocal tract sections in the proposed method. As observed from the computational time analysis, the proposed technique is computationally complex as compared to other methods.

G. Possible Applications

The proposed GIF technique is used for estimation of the signal produced due to vibration of vocal folds during speech

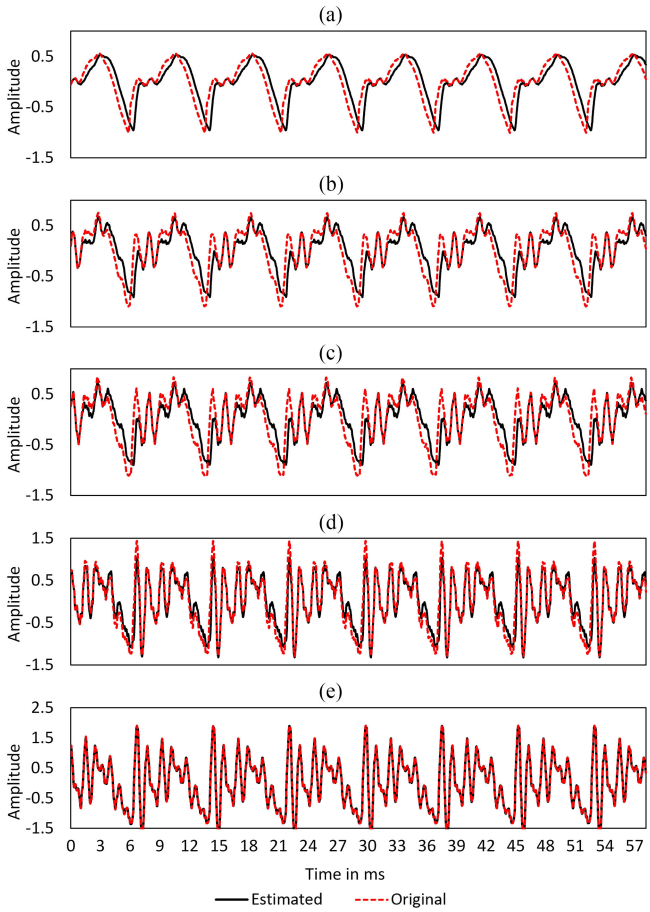

Fig. 9. Actual and estimated air pressure variation at different tube sections during the vowel /a/. (a) Section 1 (adjacent to glottis), (b) Section 11, (c) Section 22, (d) Section 33, (e) Section 44, (adjacent to lips).

production. The method could be applied to any application dealing with the glottal signal. Several examples of these applications have been given in [1] and [2]. In [1], the applications of the glottal source signal have been divided into three broad categories: The study of voice communication, medical applications, and speech technology. The applications in voice communication include the study of phonation type, voice quality evaluation, vocal emotion detection, analysis of prosodic features, etc. The medical applications include the analysis of pathological voices and vocal loading (effect of prolonged voice use, background noise, and air quality, on the pitch, the loudness of voice and vocal fold vibration patterns). Speech technology applications include speaker identification, word recognition, speech coding, speech synthesis, etc.

The proposed method also estimates the air pressure variations at different sections of the vocal tract during speech production. This could help in analyzing pathological voices more efficiently, by visualizing the variation. The air pressure variations could also be used as a prosodic feature to identify the human affective states from the speech signal. For instance, the air pressure variation at different vocal tract sections of the vowel /e/ in soft and angry male voice has been shown in Fig. 10(a) and (b), respectively. As can be seen from the figures, the air pressure distribution is significantly smooth or uniform in case of the soft voice, as compared to the angry voice.

V. Conclusion

A new GIF method has been introduced in this paper. The concatenated tube model for the vocal tract and LF model for the glottis have been used for state space modeling of the speech production system. The glottal air-volume-velocity as well as intermediate pressure values within the vocal tract have been estimated using EKF. Because of the piecewise differentiable nature of the glottal signal in the LF model, two separate set of system equations has been used for Mode 1 and Mode 2 of the glottal cycle (see Fig. 3). The opening of the glottis has been estimated by tracking formant modulation which occurs due to the interaction between glottis and vocal tract at the onset of opening. The GCI have been determined by selecting the peaks—resulting due to abrupt closing of glottis—from LP residual. The Kalman filter switches its operation between the model equations of Mode 1 and Mode 2, at GCIs and GOIs, and estimates the glottal waveform as well as the air pressure at different vocal tract sections.

From the results presented in Section IV-A and IV-E, it is evident that the proposed technique is superior to other existing methods with respect to accuracy, robustness to pitch, and extracted information content. As it is based on distributed model of the vocal tract, unlike other existing methods that are based on lumped model, it is able to estimate the glottal signal with better accuracy. The proposed method is least sensitive to pitch. The existing LP based methods perform well for low pitched speech signals due to a long duration of closed phase. However, at a higher pitch when the closed phase duration reduces, the performance of these methods also degrades. The proposed technique uses an iterative procedure to estimate the glottal signal at every speech sample. Also, the difference between the two systems of equations for Mode 1 and Mode 2 is very less. Therefore, the proposed method is able to estimate the glottal signal more efficiently for high-pitched speech signals. The other existing linear predication based methods consider the vocal tract to be a linear filter in a small time interval, and represent it by the LP coefficients in that interval. The proposed method considers the speech production system to be distributed. Hence, it also gives the provision to visualize the air pressure variation across the vocal tract during continuous speech. However, because of the iterative estimation procedure used in the proposed technique, it is computationally complex as compared to other methods.

The proposed method is applicable only after the vowel being spoken is detected. Incorrect vowel detection causes initialization of the unknown vocal tract parameters (reflection coefficients) to values that are not close to the actual value. This may cause divergence of the algorithm by increasing the Kalman gain beyond the computational limits. For continuous speech, where switching between the vowels occurs rapidly, difficulty arises in the detection of the exact boundary between the vowels. This leads to fluctuations in the estimated glottal signal when it becomes unable to track the changes. A possible solution to these problems could be achieved by applying constraints on the

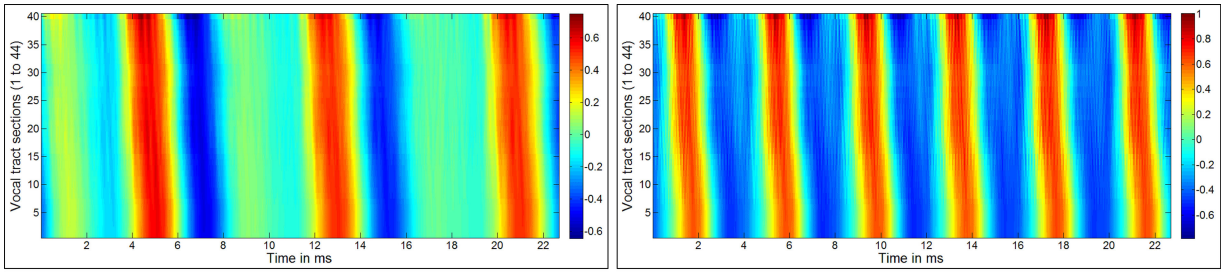

Fig. 10. Air pressure distribution for vowel /e/. (a) In soft voice. (b) In angry voice.

unknown parameters of the state space model. The constraints on the model parameters can be applied by putting constraints on the vocal tract area values at different sections. In this case, constrained Kalman filter [34] would be used for state estimation. The constraints will help to track the states while keeping the gain within computational limits.

Appendix A
Steps for Adaptive System Identification Using EKF [24]

For estimation of parameters along with the states in the state space model of (16), the state vector is augmented with the unknown parameters

$$\begin{bmatrix} \underline{x}_{n+1} \\ \underline{\theta}_{n+1} \end{bmatrix} = \begin{bmatrix} A_n(\underline{\theta}_n)\underline{x}_n \\ \underline{\theta}_n \end{bmatrix} + \begin{bmatrix} \underline{\xi}_n \\ \underline{\zeta}_n \end{bmatrix}$$
$$u_n = \begin{bmatrix} C_n(\underline{\theta}_n)\ 0 \end{bmatrix} \begin{bmatrix} \underline{x}_n \\ \underline{\theta}_n \end{bmatrix} + \eta_n \tag{25}$$

where $\underline{\zeta}$ represent the Gaussian white noise sequences associated with the parameters.

(1) Initialize the states and parameters to $\underline{x}_0$ and $\underline{\theta}_0$. Also, initialize the error covariance matrix P to $P_{0,0}$.

(2) For a new input sample n of the measured speech, compute the following:

$$\begin{bmatrix} \underline{x}_{n|n-1} \\ \underline{\theta}_{n|n-1} \end{bmatrix} = \begin{bmatrix} A_{n-1}(\underline{\theta}_{n-1})\underline{x}_{n-1} \\ \underline{\theta}_{n-1} \end{bmatrix} \tag{26}$$

$$P_{n,n-1} = \begin{bmatrix} A_{n-1}(\underline{\theta}_{n-1}) & \frac{\partial}{\partial \underline{\theta}}\left[A_{n-1}(\underline{\theta}_{n-1})\underline{x}_{n-1}\right] \\ 0 & I \end{bmatrix} P_{n-1,n-1} \begin{bmatrix} A_{n-1}(\underline{\theta}_{n-1}) & \frac{\partial}{\partial \underline{\theta}}\left[A_{n-1}(\underline{\theta}_{n-1})\underline{x}_{n-1}\right] \\ 0 & I \end{bmatrix}^T + \begin{bmatrix} Q_{n-1} & 0 \\ 0 & S_{n-1} \end{bmatrix} \tag{27}$$

where $Q_n = Var(\underline{\xi}_n)$ and $S_n = Var(\underline{\zeta}_n)$.

$$G_n = P_{n,n-1}\begin{bmatrix} C_n(\underline{\theta}_{n|n-1}) & 0 \end{bmatrix}^T \begin{bmatrix} \begin{bmatrix} C_n(\underline{\theta}_{n|n-1}) & 0 \end{bmatrix} & P_{n,n-1}\begin{bmatrix} C_n(\underline{\theta}_{n|n-1}) & 0 \end{bmatrix}^T \\ & +R_n \end{bmatrix}^{-1}. \tag{28}$$

(3) Update the state matrix, the parameter matrix, and the error covariance matrix

$$P_{n,n} = \left[I - G_n\begin{bmatrix} C_n(\underline{\theta}_{n|n-1}) & 0 \end{bmatrix}\right] P_{n|n-1}, \tag{29}$$

$$\begin{bmatrix} \underline{x}_n \\ \underline{\theta}_n \end{bmatrix} = \begin{bmatrix} \underline{x}_{n|n-1} \\ \underline{\theta}_{n|n-1} \end{bmatrix} + G_n(u_n - C_n(\underline{\theta}_{n|n-1})\underline{x}_{n|n-1}). \tag{30}$$

For $n = 1, 2, \ldots$ repeat steps 2 and 3.

For Mode 1, the Jacobian matrix $\frac{\partial}{\partial \underline{\theta}}\left[A_n\left(\underline{\theta}_n\right)\underline{x}_n\right]$ of Equation (27) is

$$\left[\begin{array}{cccccc|c} b_1[n] + \frac{1}{2}u_g[n] & 0 & 0 & \cdots & 0 & 0 & 0\ 0 \\ 0 & f_1[n] + b_2[n] & 0 & \cdots & 0 & 0 & 0\ 0 \\ 0 & 0 & f_2[n] + b_3[n] & \cdots & 0 & 0 & 0\ 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots\ \vdots \\ 0 & 0 & 0 & \cdots & f_{N-1}[n] + b_N[n] & 0 & 0\ 0 \\ \hline 0 & -(f_1[n] + b_2[n]) & 0 & \cdots & 0 & 0 & 0\ 0 \\ 0 & 0 & -(f_2[n] + b_3[n]) & \cdots & 0 & 0 & 0\ 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots\ \vdots \\ 0 & 0 & 0 & \cdots & -(f_{N-1}[n] + b_N[n]) & 0 & 0\ 0 \\ 0 & 0 & 0 & \cdots & 0 & -f_N & 0\ 0 \\ \hline 0 & 0 & 0 & \cdots & 0 & 0 & 0\ 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & x\ y \end{array}\right]$$

where $x = -2\omega_g\tau u_1[n]$ and $y = -2\alpha\tau u_1[n] + 2\tau u_2[n]$.

The Jacobian matrix for Mode 2 of the glottal cycle is

$$\left[\begin{array}{cccccc|c} b_1[n] & & & & & & \\ +\frac{1}{2}u_g[n] & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & f_1[n] & 0 & \cdots & 0 & 0 & 0 \\ & +b_2[n] & & & & & \\ 0 & 0 & f_2[n] & \cdots & 0 & 0 & 0 \\ & & +b_3[n] & & & & \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & f_{N-1}[n] & 0 & 0 \\ & & & & +b_N[n] & & \\ \hline 0 & -(f_1[n] & 0 & \cdots & 0 & 0 & 0 \\ & +b_2[n]) & & & & & \\ 0 & 0 & -(f_2[n] & \cdots & 0 & 0 & 0 \\ & & +b_3[n]) & & & & \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -(f_{N-1}[n] & 0 & 0 \\ & & & & +b_N[n]) & & \\ 0 & 0 & 0 & \cdots & 0 & -f_N & 0 \\ \hline 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & -\tau u_2[n] \end{array}\right].$$

REFERENCES

[1] P. Alku, "Glottal inverse filtering analysis of human voice production-a review of estimation and parameterization methods of the glottal excitation and their applications," *Sadhana*, vol. 36, no. 5, pp. 623–650, 2011.

[2] T. Drugman, P. Alku, A. Alwan, and B. Yegnanarayana, "Glottal source processing: From analysis to applications," *Comput. Speech Lang.*, vol. 28, no. 5, pp. 1117–1138, 2014.

[3] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, "HMM-based speech synthesis utilizing glottal inverse filtering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 153–165, Jan. 2011.

[4] M. D. Plumpe, T. F. Quatieri, D. Reynolds *et al*.,"Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 5, pp. 569–586, Sep. 1999.

[5] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris *et al*.,"Automatic speech recognition and speech variability: A review," *Speech Commun.*, vol. 49, no. 10, pp. 763–786, 2007.

[6] A. I. Iliev, M. S. Scordilis, J. P. Papa, and A. X. Falcão, "Spoken emotion recognition through optimum-path forest classification using glottal features," *Comput. Speech Lang.*, vol. 24, no. 3, pp. 445–460, 2010.

[7] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: A review," *Int. J. Speech Technol.*, vol. 15, no. 2, pp. 99–117, 2012.

[8] T. Drugman, B. Bozkurt, and T. Dutoit, "A comparative study of glottal source estimation techniques," *Comput. Speech Lang.*, vol. 26, no. 1, pp. 20–34, 2012.

[9] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*. Delhi, India: Pearson Education, 2002.

[10] L. R. Rabiner and B. Gold, *Theory and Application of Digital Signal Processing*, vol. 1. Englewood Cliffs, NJ, USA: Prentice-Hall, 1975, p. 777.

[11] D. Y. Wong, J. D. Markel, and A. H. Gray Jr, "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Trans. Acoust., Speech Signal Process.*, vol. 27, no. 4, pp. 350–355, Aug. 1979.

[12] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Commun.*, vol. 11, no. 2, pp. 109–118, 1992.

[13] M. Airaksinen, T. Raitio, B. Story, and P. Alku, "Quasi closed phase glottal inverse filtering analysis with weighted linear prediction," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 3, pp. 596–607, Mar. 2014.

[14] M. Cookey, H. J. Trussell, and I. Won, "Seismic deconvolution by multipulse methods," *IEEE Trans. Acoust., Speech Signal Process.*, vol. 38, no. 1, pp. 156–160, Jan. 1990.

[15] C. Ma, Y. Kamp, and L. F. Willems, "Robust signal selection for linear prediction analysis of voiced speech," *Speech Commun.*, vol. 12, no. 1, pp. 69–81, 1993.

[16] B. Bozkurt, B. Doval, C. D'Alessandro, and T. Dutoit, "Zeros of Z-transform representation with application to source-filter separation in speech," *IEEE Signal Process. Lett.*, vol. 12, no. 4, pp. 344–347, Apr. 2005.

[17] N. Sturmel, C. d'Alessandro, and B. Doval, "A comparative evaluation of the zeros of Z transform representation for voice source estimation," presented at 8th Annu. Conf. Int. Speech Communication Association, Antwerp, Belgium, Aug. 2007.

[18] A. V. Oppenheim, R. W. Schafer, and T. G. Stockham Jr, "Nonlinear filtering of multiplied and convolved signals," *IEEE Trans. Audio Electroacoust.*, vol. 16, no. 3, pp. 437–466, Sep. 1968.

[19] G. Fant, J. Liljencrants, and Q.-G. Lin, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 4, no. 1985, pp. 1–13, 1985.

[20] H. Li, "Glottal source parametrisation by multi-estimate fusion," Ph.D. dissertation, Dublin City Univ., Dublin, Ireland, 2013.

[21] H. Wakita, "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms," *IEEE Trans. Audio Electroacoust.*, vol. 21, no. 5, pp. 417–427, 1973.

[22] A. H. Jazwinski, *Stochastic Processes and Filtering Theory*. North Chelmsford, MA, USA: Courier Corporation, 2007.

[23] S. J. Julier and J. K. Uhlmann, "Unscented filtering and nonlinear estimation," *Proc. IEEE*, vol. 92, no. 3, pp. 401–422, Mar. 2004.

[24] C. K. Chui and G. Chen, *Kalman Filtering: With Real-Time Applications*. New York, NY, USA: Springer Science & Business Media, 2008.

[25] Y. Bar-Shalom, X. R. Li, and T. Kirubarajan, *Estimation With Applications to Tracking and Navigation: Theory Algorithms and Software*. New York, NY, USA: Wiley, 2004.

[26] I. Hwang, H. Balakrishnan, and C. Tomlin, "State estimation for hybrid systems: applications to aircraft tracking," *IEE Proc. Control Theory Appl.*, vol. 153, no. 5, 2006, Art. no. 556.

[27] S. Mathur, B. H. Story, and J. J. Rodríguez, "Vocal-tract modeling: Fractional elongation of segment lengths in a waveguide model with half-sample delays," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 5, pp. 1754–1762, Sep. 2006.

[28] J. Mullen, D. M. Howard, and D. T. Murphy, "Real-time dynamic articulations in the 2-D waveguide mesh vocal tract model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 577–585, Feb. 2007.

[29] B. H. Story, I. R. Titze, and E. A. Hoffman, "Vocal tract area functions from magnetic resonance imaging," *J. Acoust. Soc. Amer.*, vol. 100, no. 1, pp. 537–554, 1996.

[30] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarepa collaborative voice analysis repository for speech technologies," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 960–964.

[31] P. Alku, T. Bäckström, and E. Vilkman, "Normalized amplitude quotient for parametrization of the glottal flow," *J. Acoust. Soc. Amer.*, vol. 112, no. 2, pp. 701–710, 2002.

[32] M. Airas and P. Alku, "Comparison of multiple voice source parameters in different phonation types," presented at the 8th Annu. Conf. Int. Speech Commun. Assoc., Antwerp, Belgium, Aug. 2007.

[33] G. Fant, "The LF-model revisited. Transformations and frequency domain analysis," *Speech Trans. Lab. Quart. Rep., Roy. Inst. Technol. Stockholm*, vol. 2, no. 3, 1995, Art. no. 40.

[34] D. Simon, *Optimal State Estimation: Kalman, H Infinity, and Nonlinear Approaches*. New York, NY, USA: Wiley, 2006.

Subhasmita Sahoo has received the B.Tech. degree from Biju Pattnaik University, Rourkela, India, in 2011. She received the Master's degree from the Indian Institute of Technology (IIT) Kharagpur, Kharagpur, India, in 2014. She is currently working toward the Ph.D. degree at the Department of Electrical Engineering, IIT Kharagpur. Her research interests include speech modeling, audio–visual signal processing, and human emotion analysis.

Aurobinda Routray received the Master's degree from the Indian Institute of Technology Kanpur, Kanpur, India, in 1991 and the Ph.D. degree from Sambalpur University, Sambalpur, Odisha, India, in 1999. He has also worked as a Postdoctoral Researcher at Purdue University, West Lafayette, IN, USA, during 2003–2004. He is currently working as a Professor at the Department of Electrical Engineering, Indian Institute of Technology Kharagpur, Kharagpur India. His research interests include nonlinear and statistical signal processing, signal-based fault detection and diagnosis, real time and embedded signal processing, numerical linear algebra, and data driven diagnostics.