Time-Varying Quasi-Closed-Phase Analysis for Accurate Formant Tracking in Speech Signals

Dhananjaya Gowda, *Member, IEEE*, Sudarsana Reddy Kadiri, *Member, IEEE*, Brad Story, and Paavo Alku, *Fellow, IEEE*

Abstract*—In this paper, we propose a new method for the accurate estimation and tracking of formants in speech signals using time-varying quasi-closed-phase (TVQCP) analysis. Conventional formant tracking methods typically adopt a two-stage estimate-and-track strategy wherein an initial set of formant candidates are estimated using short-time analysis (e.g., 10–50 ms), followed by a tracking stage based on dynamic programming or a linear state-space model. One of the main disadvantages of these approaches is that the tracking stage, however good it may be, cannot improve upon the formant estimation accuracy of the first stage. The proposed TVQCP method provides a single-stage formant tracking that combines the estimation and tracking stages into one. TVQCP analysis combines three approaches to improve formant estimation and tracking: (1) it uses temporally weighted quasi-closed-phase analysis to derive closed-phase estimates of the vocal tract with reduced interference from the excitation source, (2) it increases the residual sparsity by using the L_1 optimization and (3) it uses time-varying linear prediction analysis over long time windows (e.g., 100–200 ms) to impose a continuity constraint on the vocal tract model and hence on the formant trajectories. Formant tracking experiments with a wide variety of synthetic and natural speech signals show that the proposed TVQCP method performs better than conventional and popular formant tracking tools, such as Wavesurfer and Praat (based on dynamic programming), the KARMA algorithm (based on Kalman filtering), and DeepFormants (based on deep neural networks trained in a supervised manner). Matlab scripts for the proposed method can be found at: https://github.com/njaygowda/ftrack

***Index Terms*—Time-varying linear prediction, weighted linear prediction, quasi-closed-phase analysis, formant tracking.**

I. Introduction

VOCAL tract resonances (VTRs), commonly referred to as *formant frequencies*, are speech parameters that are of fundamental importance in all areas of speech science and technology. The estimation and tracking of VTRs from speech signals is a challenging problem that has many applications in various areas: in acoustic and phonetic analysis [1], [2], in voice morphing [3], in speech recognition [4], [5], in speech and singing voice synthesis [6], [7], in voice activity detection [8], and in designing hearing aids [9], [10]. Many algorithms of varying complexity have been proposed in the literature for tracking formants in speech signals [11]–[15]. A dynamic programming (DP)–based tracking algorithm with a heuristic cost function on the initial formant candidates estimated using conventional linear prediction (LP) analysis was used in [11], [12]. This two-stage approach has a detection stage, where an initial estimate of the VTRs is obtained, followed by a tracking stage. An integrated approach towards tracking was adopted in [13]–[15] using state-space methods such as Kalman filtering (KF) and the factorial hidden Markov model (FHMM). In both approaches, analysis of the signal for the accurate estimation (or modeling) of the vocal tract system is an important and necessary computational block. However, it should be mentioned here that there are a few exceptions, such as [15], which uses a non-negative matrix factorization (NMF)–based source-filter modeling of speech signals. Recently, deep learning–based techniques [16]–[18] have also been studied as alternatives to conventional statistical signal processing–based formant estimation and tracking methods. These methods, however, are based on supervised machine learning, which calls for having annotated speech corpora with which to obtain the ground truth formant frequencies for system training.

LP analysis is one of the most widely used methods for estimating VTRs from speech signals [19]–[21]. To improve the accuracy of LP, several variants of this all-pole spectral modeling method have been proposed [22]. Among the different modifications, autocorrelation and covariance analyses are the most popular LP methods in formant estimation and tracking [11], [12]. Covariance analysis is known to give more accurate formant estimates than autocorrelation analysis, but the stability of the resulting all-pole filter is not guaranteed in covariance analysis [21], [23]. Even though the filter instability must be avoided in applications where the signal needs to be reconstructed (such as speech synthesis and coding), the instability in itself is not a serious problem in formant tracking. Compared to covariance analysis, closed-phase analysis is known to provide even more accurate VTR estimates by avoiding the open-phase regions of the glottal cycle, which are influenced by the coupling of the vocal tract with the trachea [24], [25]. Closed-phase analysis, however, works better for utterances such as those of low-pitched male voices, which have more samples in the closed

Manuscript received March 4, 2019; revised November 1, 2019 and May 4, 2020; accepted May 21, 2020. Date of publication June 4, 2020; date of current version June 29, 2020. This work was supported by the Academy of Finland (Project 284671, 312490). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Lei Xie. *(Corresponding author: Sudarsana Reddy Kadiri.)*

Dhananjaya Gowda was with Aalto University, 02150 Espoo, Finland. He is now with Samsung Research, Seoul R&D Campus, Seoul 06765, Republic of Korea (e-mail: njaygowda@gmail.com).

Sudarsana Reddy Kadiri and Paavo Alku are with the Department of Signal Processing and Acoustics, Aalto University, 02150 Espoo, Finland (e-mail: sudarsana.kadiri@aalto.fi; paavo.alku@aalto.fi).

Brad Story is with the University of Arizona, Tucson, AZ 85721 USA (e-mail: bstory@email.arizona.edu).

Digital Object Identifier 10.1109/TASLP.2020.3000037

phase of the glottal cycle compared to high-pitched female and child voices that might have just a few closed-phase samples per glottal cycle.

As a remedy for the lack of data samples in formant estimation, a selective prediction of speech samples can be conducted in spectral modeling. A sample-selective prediction is used in weighted linear prediction (WLP) methods by giving different temporal weighting to the prediction error at each discrete time instant [26]–[33]. One such method, called sample selective linear prediction (SSLP) analysis, was proposed in [26] for better modeling of the vocal tract area function. In SSLP, a hard rejecting weighting function is used to eliminate outlier samples in sample selection. A more generalized WLP algorithm was developed in [27] with a continuous weighting function for the prediction residual. In [28], an iterative LP algorithm, robust linear prediction was proposed by utilizing the non-Gaussian nature of the excitation signal to derive a temporal weighting function based on the magnitude of the residual samples.

To improve the robustness of linear predictive spectrum estimation, a simple non-iterative WLP method was studied in [29] based on the short-time energy (STE) weighting function. The STE weighting function is a temporal energy function that is computed, for example, in 1–2 ms frames of the speech signal waveform. The STE weighting function emphasizes the importance of the high-energy regions within a glottal cycle in computing the autocorrelation (or covariance) matrix. Therefore, this WLP method is similar to closed-phase LP analysis because the high-energy sections of voiced speech emphasized by the STE weighting correspond roughly to glottal closed-phase regions. Since the publication of WLP in [29], several variants of this all-pole modeling method have been developed and used, for example, in the robust feature extraction of speech [29], [31] and in glottal inverse filtering (GIF) [33], [34]. Some of these more recent WLP algorithms have also addressed the stability of the all-pole filter [30], [32]. In [32], a new weighting function, called the attenuated main excitation (AME) window, was studied to improve the accuracy of formant estimation, especially for high-pitched voices. The AME function is designed to attenuate the effect of prominent speech samples in the vicinity of the glottal closure instants (GCIs) on the autocorrelation function. This is justified because these high-energy speech samples are greatly contributed to by the glottal source, which results in distortion of the formant estimates by the biasing effect of the glottal source. As a sequel to using AME as a temporal weighting function in WLP, the quasi-closed-phase (QCP) analysis of speech signals was proposed in [33] for the estimation of glottal flow with GIF. QCP analysis uses a more generalized version of the AME weighting function, for example, with slanted edges instead of vertical ones. In addition, the weighting function of QCP analysis puts more emphasis on the closed-phase regions compared to the open-phase regions that are prone to subglottal coupling. However, the previous experiments with QCP analysis in [33] focused solely on GIF analysis of the voice source, without any evaluation of the QCP algorithm's performance in formant detection and estimation.

The spectral modeling of speech is conducted using conventional LP in short-time segments (5–50 ms) by assuming speech to be a quasi-stationary process [21]. This traditional short-time analysis models the real, continuously varying human vocal tract system in a piecewise manner. In addition, the conventional methods based on short-time LP analysis typically use a two-stage detect-and-track approach in tracking formants [11], [12]. It should be noted that even those formant tracking methods that directly track formants from the cepstral coefficients use this piecewise approximation of the vocal tract system [13], [14]. In order to take into account the inherent slowness of the real human vocal tract (i.e., the system being inertial), time-varying linear prediction (TVLP) provides an attractive method that models the vocal tract over longer time-spans by defining the model parameters as a function of time by using selected, low-order basis functions [35]–[37].

The solution to conventional LP involves minimizing the L_2 norm of the prediction error signal, the residual, with an inherent assumption that the excitation source signal is a Gaussian process [22], [38]. Based on the theory of compressed sensing, sparsity constraints can be used to utilize the super Gaussian nature of the excitation signal [39], [40]. This is achieved by approximating a non-convex L_0 norm optimization problem by using a more tractable convex L_1 norm optimization [39]. In addition, it was shown in [40] that an iterative reweighted minimization of the norm can achieve increased sparsity of the error signal, which yields a solution closer to L_0 norm optimization.

In this article, we propose a new time-varying quasi-closed-phase (TVQCP) linear prediction analysis of speech for accurate modeling and tracking of VTRs. The proposed method aims to improve the estimation and tracking of formants by combining three different ideas: QCP analysis, increased sparsity of the error signal and time-varying filtering. To the best of our knowledge, this combination has not been studied before in formant estimation and tracking and is justified as follows. First, in order to reduce the effect of the glottal source in formant estimation, it is justified to take advantage of QCP analysis to temporally weight the prediction error, which has been shown to improve the estimation of the vocal tract in voice source analysis [33], [34]. Second, filter optimization in previous QCP studies has been conducted using the L_2 norm which is known to result in less sparse residuals. Therefore, in order to further enhance the performance of temporal weighting, it is justified to increase the sparsity of the residual in QCP analysis by using the L_1 norm. Third, in order to take into account the fact that the natural human vocal tract is a slowly varying physiological system, we argue that formant tracking can be further improved by implementing the proposed L_1 norm -based QCP analysis using time-varying filtering. A preliminary investigation of TVQCP for formant tracking was published in a conference paper in [41]. In the current study, our preliminary experiments reported in [41] are expanded in many ways by, for example, including a larger number of evaluation datasets and a larger number of reference methods. In summary, the contributions of the current study are as follows.

- Combining the ideas of QCP analysis, L_1 norm optimization and TVLP analysis to create a new formant estimation and tracking method, TVQCP.

- Studying the advantages of sparsity by comparing the L_1 and L_2 norm optimization in TVQCP.
- Analysing the effects of the different parameters in TVQCP.
- Studying the formant tracking performance of TVQCP using synthetic vowels of varying fundamental frequency values and phonation types, using high-pitched child speech simulated with a physical modeling approach, and using natural speech.
- Comparing TVQCP with popular formant tracking methods (Wavesurfer, Praat and KARMA) and with a recently proposed deep neural network -based method (DeepFormants) that is based on supervised learning.
- Studying the noise robustness of TVQCP for different noise types and signal-to-noise ratio (SNR) scenarios.

In the following two sections, the optimization of the TVQCP model is described by first presenting the time-invariant (i.e., stationary) QCP analysis in Section II as background information. After this, the TVQCP (i.e., non-stationary QCP) analysis is presented in Section III. Formant tracking experiments are reported in Section IV and conclusions are drawn in Section V.

II. Quasi-Closed-Phase Analysis

QCP analysis belongs to the family of temporally weighted LP methods with a specially designed weighting function based on the knowledge of GCIs [33]. An overview of WLP and the design of the QCP weighting function is given in this section.

A. Weighted Linear Prediction

In conventional LP, the current speech sample $x[n]$ is predicted based on the past p speech samples as

$$\hat{x}[n] = -\sum_{k=1}^{p} a_k x[n-k], \tag{1}$$

where $\{a_k\}_{k=0}^{p}$ with $a_0 = 1$ denote the prediction coefficients and p is the prediction order. Let us denote the estimated transfer function of the vocal tract system as $H(z) = 1/A(z)$, where $A(z)$ is the $z-$ transform of the prediction coefficients $\{a_k\}_{k=0}^{p}$. The optimal prediction coefficients minimize the overall prediction error given by the cost function

$$E = \sum_{n} e^2[n], \tag{2}$$

where $e[n] = x[n] - \hat{x}[n]$ is the sample-wise prediction error, the residual. The optimal prediction coefficients are computed by minimizing the cost function ($\partial E/\partial a_i = 0, \quad 1 \le i \le p$), which results in the following normal equations

$$\sum_{k=1}^{p} r_{i,k} a_k = -r_{i,0}, \quad 1 \le i \le p, \tag{3}$$

$$\text{where} \quad r_{i,k} = \sum_{n} x[n-i]x[n-k]. \tag{4}$$

In the above formulation, it can be seen that the prediction error is minimized in the least-square sense by having equal temporal weighting for every sample. However, in WLP, a different (positive) weighting value is imposed on each squared residual sample, resulting in the following WLP cost function

$$E_w = \sum_{n} w[n] e^2[n], \tag{5}$$

where $w[n]$ denotes the weighting function on the sample-wise prediction error $e[n]$. It should be noted that the weighting in WLP methods is on the error signal and should not be confused with the traditional short-time windowing (e.g., Hamming) of the speech signal that is used for reducing truncation effects in spectral analysis. The prediction coefficients can be computed in a similar way to that of conventional LP by minimizing the cost function ($\partial E_w/\partial a_i = 0, \quad 1 \le i \le p$) and solving the resulting normal equations

$$\sum_{k=1}^{p} b_{i,k} a_k = -b_{i,0}, \quad 1 \le i \le p, \tag{6}$$

$$\text{where} \quad b_{i,k} = \sum_{n} w[n]x[n-i]x[n-k]. \tag{7}$$

B. The Choice of Weighting Function

As mentioned earlier in Section I, several weighting functions have been proposed for WLP. STE is one of the popular weighting functions used in WLP, and it is demonstrated in Fig. 1. The figure shows an example of a vowel utterance, an electroglottography (EGG) signal, and the derivative of the EGG signal (dEGG), along with rough markings for the closed phases

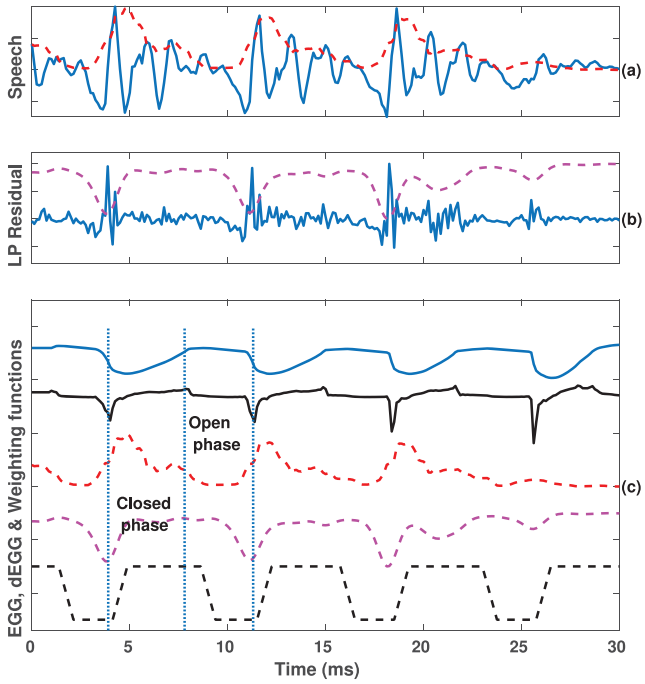


Fig. 1. An illustration of different weighting functions for use in WLP: (a) the speech signal (the solid line) and the short-time energy (STE) weighting function (the dashed line); (b) the LP residual (the solid line) and the weighting function (the dashed line) derived from the residual; and (c) EGG (the solid blue line), dEGG (the solid black line), and three different weighting functions: speech signal–based STE weighting (the dashed red line), residual based weighting (the dashed pink line), and QCP weighting (the dashed black line).

and open phases. The STE weighting function is computed as

$$w[n] = \sum_{k=(D+1)}^{(D+M)} x^2[n-k], \tag{8}$$

where the delay parameter D controls the peak position (or emphasis) of the weighting function within the glottal cycle and the length parameter M controls the peak width, as well as the dynamic range and smoothness of the function. Typical values for these two parameters are $D = 0$ and $M = 12$, the latter corresponding to 1.5 ms at an 8 kHz sampling rate. It can be seen that the STE function puts more weighting to the high-energy closed-phase regions of the glottal cycle. However, Fig. 1 also demonstrates that the degree of suppression in both the glottal open phase and at the instant of the main excitation depends on the decay of the speech signal waveform within the glottal cycle. Therefore, the STE weighting function does not necessarily suppress these regions completely. The effect of this problem of the STE weighting function was studied in our previous study on formant estimation of high-pitched vowels [32]. This previous study indicated that by changing the weighting function from STE to AME resulted in a clear improvement in formant estimation accuracy particularly for the first formant for which the average estimation accuracy improved by almost 10 percentage units.

A weighting function based on the residual signal energy can also be used. Fig. 1 shows a residual weighting function derived by inverting and normalizing (between 0 to 1) a zero-mean residual energy signal, computed similar to the STE function. As can be seen from the figure, the residual weighting function may not suppress some weaker glottal excitations (at around 25 ms) as well as the stronger ones. This effect can be more pronounced in the vowel beginning and ending frames with a highly transient signal energy. Also, the residual weighting function may not effectively down-weight the contributions from the open-phase regions of the glottal cycle. A QCP weighting function derived from knowledge of GCIs is also shown in Fig. 1. It can be seen that this weighting function emphasizes the closed-phase region of the glottal cycle, while at the same time the function de-emphasizes the region immediately after the main excitation as well as the open-phase region.

C. Quasi-Closed-Phase Weighting Function

An example of the QCP weighting function w_n is shown in Fig. 2, along with the Liljencrants-Fant (LF) glottal flow derivative waveform u_n for about one glottal cycle. The QCP weighting function can be expressed with three parameters: the position quotient ($PQ = t_p/T_0$), the duration quotient ($DQ = t_d/T_0$), and ramp duration t_r, where T_0 is the time-length of the glottal cycle. In order to avoid possible singularities in the weighted correlation matrix given in Eq. (6), a small positive value, $d_w = 10^{-5}$, is used (instead of zero) as the minimum value in the weighting function.

The parameters of the QCP weighting function were optimized in [33] using a set of about 65000 LF-excited synthetic vowels of different phonation types and different fundamental frequency values. Rather than aiming at a generic optimal weighting function, the optimization procedure adopted in [33] was based on using a simple, pre-defined waveform depicted in Fig. 2 whose parameters were optimized in a grid search. For more details about the optimization procedure, the reader is referred to Section IV-A in [33]. The optimization procedure reported in [33] gave both fixed QCP parameters and parameters where one of the values (DQ) was pitch-adaptive. In the current study, we used the pitch-adaptive QCP parameters of [33] and the values of the two fixed parameters were as follows: $PQ = 0.05$ and $t_r = 0.375$ ms (which corresponds to $N_{ramp} = 3$ samples using the notation of [33]). DQ was varied between 0.5 and 0.9 (as will be reported in Section IV-E4) and was set to $DQ = 0.8$.

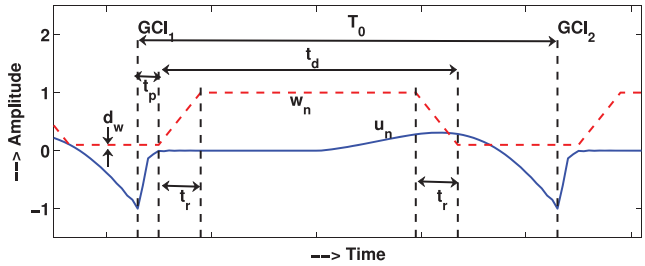


Fig. 2. The design of the quasi-closed-phase (QCP) weighting function w_n (the dotted line), along with the LF glottal flow derivative signal u_n (the solid line) for about one glottal cycle.

Using the QCP function as a temporal weighting waveform in WLP provides two distinct advantages when compared to conventional LP (i.e., giving equal weighting to all squared residual samples) or conventional WLP (i.e., weighting is given using the STE function). The first advantage is that the emphasis of the QCP weighting function is on the closed phase region, which provides more accurate modeling of the vocal tract by reducing the effect of coupling between subglottal and supraglottal cavities. The second is that the QCP weighting de-emphasizes the region immediately after the main excitation of the vocal tract, which reduces the biasing effect of the glottal source in the modeling of VTRs. De-emphasizing the main excitation can also be justified from the observation that this region typically shows large prediction errors that become increasingly dominant with short pitch periods. QCP analysis has previously been shown to be effective in estimating the voice source with GIF [33].

III. Time-Varying Quasi-Closed-Phase Analysis

The spectral estimation and tracking method proposed in this study, TVQCP analysis, combines the ideas of sample selective prediction (i.e., the underlying idea of QCP), sparsity of the prediction error, and long-time nonstationary analysis of the vocal tract system (i.e., the underlying idea of TVLP). In the following, the normal equations of the proposed TVQCP analysis are derived by starting from conventional LP. Note that the optimization schemes in Section II all used the L_2 norm of the error signal whereas this section uses more general optimization norms.

A. Linear Prediction

In conventional LP, the current sample $x[n]$ is predicted according to Eq. (1) as a linear weighted sum of the past p samples. By denoting the window size as N, the predictor coefficients can be estimated as a solution to the convex optimization problem of generic norm L_m given by

$$\hat{\boldsymbol{a}} = \arg\min_{a} ||\boldsymbol{x} - \boldsymbol{X}\boldsymbol{a}||_m^m, \tag{9}$$

$$\text{where} \quad \boldsymbol{x} = [x[0], x[1], \ldots, x[N-1]]_{N\times 1}^T, \tag{10}$$

$$\boldsymbol{a} = [a_1, a_2, \ldots, a_p]_{p\times 1}^T, \tag{11}$$

$$\boldsymbol{X} = [X_0, X_1, \ldots, X_{N-1}]_{N\times p}^T, \quad \text{and} \tag{12}$$

$$X_n = [x[n-1], \ldots, x[n-p]]_{p\times 1}^T. \tag{13}$$

The minimization of the L_2 norm of the residual leads to the least square solution of conventional LP. However, imposing a sparsity constraint on the residual provides better modeling of both the excitation and vocal tract system. This is achieved by minimizing the L_1 norm of the residual instead of its L_2 norm. This change in the optimization norm is known to give a convex approximation of the solution to the L_0 norm optimization problem, also referred to as sparse linear prediction (SLP) [39], [40].

B. Weighted Linear Prediction

WLP analysis uses sample-selective prediction and gives differential emphasis to different regions of the speech signal within a glottal cycle (as discussed earlier in Section II-A). Using a generic L_m norm, WLP can be expressed by minimizing the weighted error signal given by

$$\hat{\boldsymbol{a}} = \arg\min_{a} \boldsymbol{W}||\boldsymbol{x} - \boldsymbol{X}\boldsymbol{a}||_m^m, \tag{14}$$

where $\boldsymbol{W}_{N\times N}$ is a diagonal matrix with its diagonal elements corresponding to a weighting function w_n, imposed on the prediction error signal.

C. Time-Varying Linear Prediction

TVLP is a generalization of conventional LP where the predictor coefficients are continuous functions of time. Therefore, TVLP can be used in the spectral analysis of nonstationary speech signals using long-time (e.g., 100–200 ms) frames. TVLP imposes a time-continuity constraint on the vocal tract system in the form of low-order basis functions. Due to this time-continuity constraint, TVLP is capable of modeling the slowly varying vocal tract system better than conventional LP that is based on a piecewise constant quasi-stationary approximation. In TVLP, the current speech sample is predicted using the past p samples as

$$\hat{x}[n] = \sum_{k=1}^{p} a_k[n]x[n-k], \tag{15}$$

where $a_k[n]$ denotes the k^{th} time-varying prediction filter coefficient at time instant n. The time-variant predictor coefficient $a_k[n]$ can be expressed using different basis functions, such as polynomials (i.e., power series), trigonometric series, or Legendre polynomials [35]. In this study, we use the simple q^{th} order polynomial approximation given by

$$a_k[n] = \sum_{i=0}^{q} b_{k_i} n^i. \tag{16}$$

The TVLP coefficients are estimated by minimizing the L_m norm of the error signal. This can be presented as the convex optimization problem given by

$$\hat{\boldsymbol{b}} = \arg\min_{b} ||\boldsymbol{x} - \boldsymbol{Y}\boldsymbol{b}||_m^m, \tag{17}$$

$$\text{where} \quad \boldsymbol{x} = [x[0], x[1], \ldots, x[N-1]]_{N\times 1}^T, \tag{18}$$

$$\boldsymbol{b} = [b_{1_0}, \ldots, b_{1_q}, \ldots, b_{p_0}, \ldots, b_{p_q}]_{p(q+1)\times 1}^T, \tag{19}$$

$$\boldsymbol{Y} = [Y_0, Y_1, \ldots, Y_{N-1}]_{N\times p(q+1)}^T, \quad \text{and} \tag{20}$$

$$Y_n = [x[n-1], nx[n-1], \ldots, n^q x[n-1], \ldots, x[n-p], nx[n-p], \ldots, n^q x[n-p]]_{p(q+1)\times 1}^T. \tag{21}$$

Again, the L_2 and L_1 norm minimization lead to the least square solution and the sparse solution to the convex optimization problem respectively [37], [39], [40]. It is to be noted that the L_2 norm minimization can be solved in closed form whereas convex optimisation calls for an iterative approach and therefore its computational complexity is larger. The current study uses linear programming in convex optimization for the L_1 norm-based methods. Hence, the computational complexity of the L_1 norm-based LP methods studied in this article is clearly higher than in the L_2 norm-based LP methods.

D. Time-Varying Weighted Linear Prediction

As the final step of the model optimization, let us combine WLP, the technique described in Sections II-A and III-B, and TVLP, the approach presented in Section III-C. The combination of these two, time-varying weighted linear prediction (TVWLP) analysis, is analogous to WLP where the predictor coefficients are estimated by minimizing the weighted error signal given by

$$\hat{\boldsymbol{b}} = \arg\min_{b} \boldsymbol{W}||\boldsymbol{x} - \boldsymbol{Y}\boldsymbol{b}||_m^m, \tag{22}$$

where $\boldsymbol{W}_{N\times N}$ is a diagonal matrix with its diagonal elements corresponding to the weighting function $w[n]$, imposed on the error signal.

Based on Eq. (22), in this study we propose a new TVQCP analysis of speech signals that uses the QCP weighting function (described in Section II-C) in matrix $\boldsymbol{W}$ of the TVWLP framework above. By using the L_1 norm (i.e., assigning $m = 1$ in Eq. (22)), the TVQCP analysis enables imposing a sparsity constraint on the excitation signal.

IV. Formant Tracking Experiments

One of the main problems with evaluating the performance of a formant tracker and comparing it with other methods is the availability of absolute ground truth in formant frequency values.

It is possible to have such absolute ground truth in the case of synthetic speech signals. However, there are two limitations with using synthetic speech signals. The first is that the reference formant frequencies provided by synthetic utterances can be biased towards a particular method of formant tracking if there is a strong similarity in the synthesis model and the analysis model of the tracker. The second is that the formant trackers are ultimately required to process natural speech signals that do not have any reference ground truth. The problem with using natural speech signals is the need for a semi-supervised human annotation of the formant frequency value, which by itself can vary from one annotator to another [42]. Formant tracking from natural speech can also be biased by the tools and techniques used for the annotation, such as spectrographic representations and/or methods used for deriving some of the initial estimates. Also, it should be noted that actual resonance frequencies of the vocal tract cavities need not exactly coincide with the apparent peaks in speech spectra because these spectral peaks might also be harmonics that are a result of the glottal excitation.

In order to address the above problem with reference ground truth, the performance of formant tracking with the proposed TVQCP method is studied using both synthetic and natural speech signals. Two different types of synthetic signals were used. In one type, vowels are produced with conventional source-filter modeling of the speech production apparatus using the LF glottal source model and an all-pole vocal tract filter. In the other type, utterances are generated using physical modeling of the vocal tract and glottal source [43], [44]. The latter approach is different from the LF source-filter technique because the speech signal is generated based on physical laws, rather than by a digital parametric model similar to the model assumed in LP and it variants. The physical modeling approach is used to avoid any inherent bias that the LF source-filter technique may have towards the proposed TVQCP method, owing to the fact that both use LP-based methods in vocal tract modeling.

A. Performance Metrics

The formant tracking performance of different methods is evaluated in terms of two different metrics: the formant detection rate (FDR) and formant estimation error (FEE). Throughout this study, formants are identified by looking for the local peaks of the power spectrum. The FDR is measured in terms of the percentage of frames where a formant is hypothesized within a specified deviation from the ground truth. The FDR for the i^{th} formant over K analysis frames is computed as

$$D_i = \frac{1}{K}\sum_{n=1}^{K} I(\Delta F_{i,n}), \tag{23}$$

$$I(\Delta F_{i,n}) = \begin{cases} 1 & \text{if } (\Delta F_{i,n}/F_{i,n} < \tau_r \quad \& \quad \Delta F_{i,n} < \tau_a) \\ 0 & \text{otherwise,} \end{cases} \tag{24}$$

where $I(.)$ denotes a binary formant detector function and $\Delta F_{i,n} = |F_{i,n} - \hat{F}_{i,n}|$ is the absolute deviation of the hypothesized formant frequency $\hat{F}_{i,n}$ for the i^{th} formant at the n^{th} frame from the reference ground truth $F_{i,n}$. The thresholds τ_r and τ_a denote the relative deviation and absolute deviation respectively.

Using a single detection threshold, either a relative threshold or an absolute threshold, is problematic on a linear frequency scale. For higher formants, the relative deviation needs to be smaller than that for the lower formants. Similarly, the absolute deviation for lower formants needs to be smaller than that for the higher formants. In order to define a common detection strategy for all formants, two thresholds, one on relative deviation and the other on absolute deviation, must be used. The relative threshold controls the detection rates of lower formants whereas the absolute threshold controls the detection rates of higher formants.

The FEE is measured in terms of the average absolute deviation of the hypothesized formants from the ground truth. The FEE for the i^{th} formant over K analysis frames is computed as

$$R_i = \frac{1}{K}\sum_{n=1}^{K} \Delta F_{i,n}. \tag{25}$$

The FDR and FEE values are only computed for frames that are voiced or for some particular phonetic category of interest. One problem with accumulating FEEs over all frames is that a few large error outliers can dictate the overall score. This is even more severe for the root mean square error (RMSE) criterion that is a widely used metric for measuring formant estimation accuracy. In view of this, we propose using mean absolute error, which is less sensitive to outliers, as a measure for FEE. The reading of FEE scores in conjunction with FDR scores, which denote the number of frames detected within a fixed threshold, can give a better sense of the performance of a formant tracker.

B. The Choice of Window Size and Polynomial Order

As outlined in Section III-C, TVLP analysis involves two parameters (in addition to prediction order p) that need to be set: window size N and polynomial order q. Longer window sizes (e.g., 500 ms) are useful for the efficient parameterization of speech signals but would introduce longer delays. Moreover, longer window sizes require higher polynomial orders in order to model the time-varying characteristics of the vocal tract and can lead to computational problems due to the inversion of rank deficient matrices. Therefore, moderate window sizes (e.g., 100–200 ms) are a good overall compromise that enables the efficient parameterization of the slowly time-varying characteristics of the vocal tract using low-order polynomials (e.g., $q = 3$).

In order to study the choice of the window size and polynomial order in TVLP analysis, an initial experiment was conducted on a set of synthetic utterances. The effect of these two parameters on a larger dataset of natural speech utterances will be studied later. The synthetic speech utterances were generated starting with ten (5 male, 5 female) randomly chosen natural speech utterances from the TIMIT-VTR database [42]. The natural utterance was first inverse filtered using a high order ($p = 18$) short-time LP analysis (20-ms frame size, 10-ms frame shift, and a sampling rate of 8 kHz) to compute a spectrally flat residual signal that was void of any formant structure. This residual signal was then used to excite an 8^{th} order all pole model constructed using the

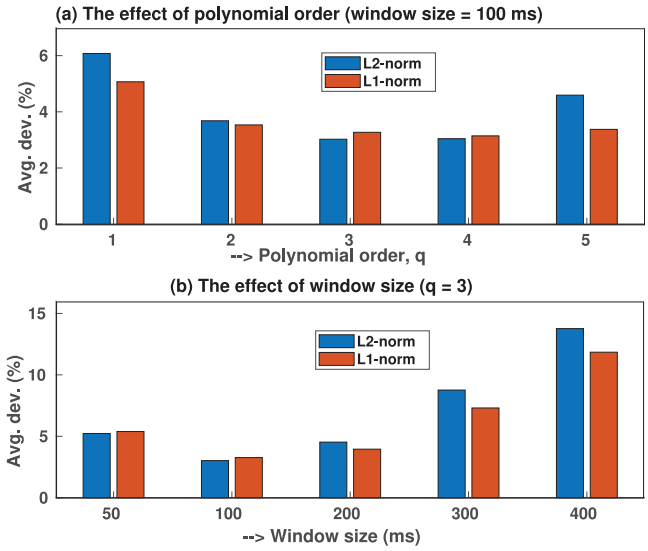


Fig. 3. Relative deviation (in percentage) of the TVLP-estimated formants from their ground truth, averaged over the three first formants as a function of (a) polynomial order and (b) window size.

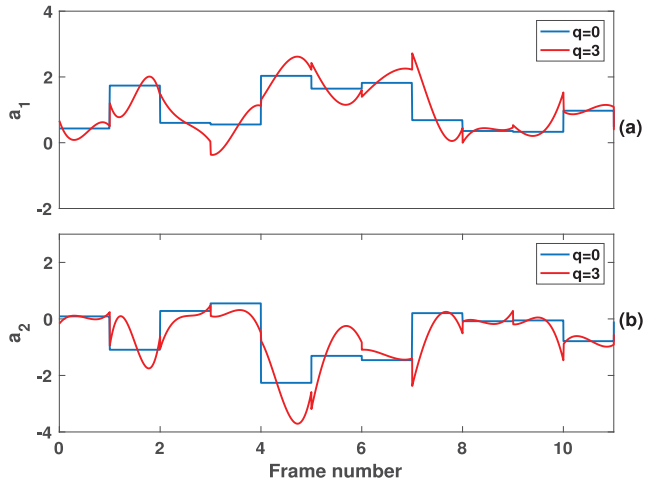


Fig. 4. The trajectories of $a_k[n]$ for q = 0 and q = 3 in TVLP-L2 for the first and second coefficients (a_1 and a_2) are shown in (a) and (b), respectively. The word $'materials'$ produced by a female talker is used for the illustration.

first four reference formants and bandwidths available for the utterances as part of the VTR database [42].

The results of the experiment are shown in Fig. 3, which depicts the relative deviation of the estimated formants from their ground truth, averaged over the first three formants (F_1, F_2, and F_3) for different values of polynomial order and window size. TVLP analyses computed using the L_2 norm are shown as blue bars and those computed with the L_1 norm are shown as red bars. Fig. 3(a) depicts the TVLP performance using a fixed window size of 100 ms but with the varying polynomial order q. It can be seen that the best performance is obtained for polynomial orders between $q = 2$ and $q = 4$ and the performance starts to deteriorate at the order of $q = 5$. Similarly, Fig. 3(b) shows the performance by varying the window size at a fixed polynomial order of $q = 3$. It can be seen that the performance is good with moderate window sizes of 100 ms and 200 ms, but the performance starts to deteriorate for longer window sizes. Therefore, in the experiments that follow in the remainder of the paper, we used a window size of 100 ms and a polynomial order of $q = 3$ in time-varying LP analyses. An example of using two different polynomial orders ($q = 0$ and $q = 3$) for an utterance produced by a female speaker is shown in Fig. 4. The figure depicts the contours of the two lowest coefficients (a_1 and a_2) computed using TVLP with the L_2 norm. It can be seen that the filter taps computed using $q = 0$ and $q = 3$ follow a similar general trend over the entire time-span shown in the figure but the contours computed using $q = 3$ are clearly more dynamic and their values change also during each frame.

C. Experiments on LF Model–Based Synthetic Data

The performance of the proposed TVQCP method in formant tracking is studied next in this section by analyzing how the method's performance is affected by variations in the glottal excitation (both in fundamental frequency and phonation type). Formant tracking provided by the TVQCP method is compared with that of TVLP using both the L_1 norm and L_2 norm. In addition, a comparison with the traditional LP covariance–based method (known as the entropic signal processing system (ESPS) method [45]) used in the popular open source tool Wavesurfer [11] (denoted by "WSURF") is also provided.

The TVQCP and TVLP analyses are carried out over non-overlapping 100-ms windows using a prediction order of $p = 8$ and a polynomial order of $q = 3$. The ESPS method used in Wavesurfer adopts a short-time (25-ms Hamming window, 10-ms frame shift) 12^{th} order stabilized covariance–based LP analysis followed by a dynamic programming–based tracking of formants [11].

1) The Dataset: Four different phonation types (creaky, modal, breathy, and whispery phonation) and four different ranges of fundamental frequency (mean utterance F_0 scaled by the factors 1.0, 1.5, 2.0, and 2.5) are considered for generating the synthetic speech test utterances. The phonation type and F_0 range are controlled by using the LF model for the glottal source [46]. The LF source parameter values used to synthesize the different phonation types in the current study are taken from [47], [48].

The four different fundamental frequency ranges are generated by scaling the original F_0 contour of a natural speech utterance (3–5 sec long) by different factors before synthesizing the speech signal. A modal LF excitation is generated based on the new F_0 contour while retaining the original rate of formants and hence keeping the speaking rate intact. Speech signals are synthesized by filtering the LF glottal flow derivative signal using an all-pole model with the first four semi-automatically derived formants and bandwidths of the natural utterance part of the VTR database [42]. Ten randomly selected utterances (5 male and 5 female) from the VTR database are synthesized for the four different phonation types and four different mean F_0 at a sampling rate of 8 kHz.

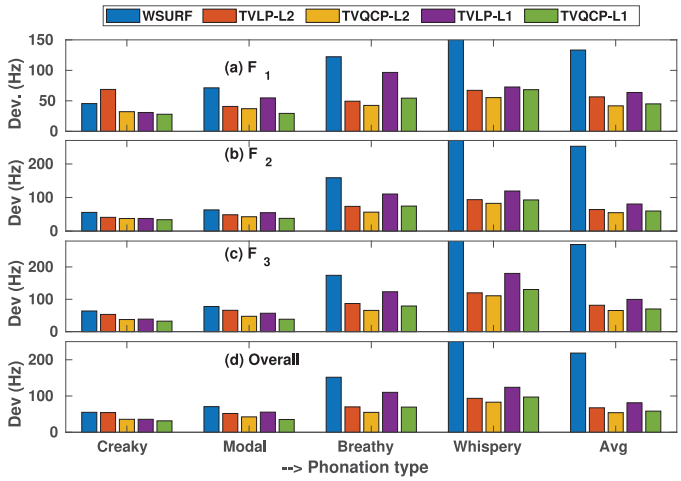


Fig. 5. The absolute deviation (FEE) of the estimated first three formants (F_1, F_2, and F_3) from their ground truth and their overall average for different phonation types of the LF model–based synthetic data.

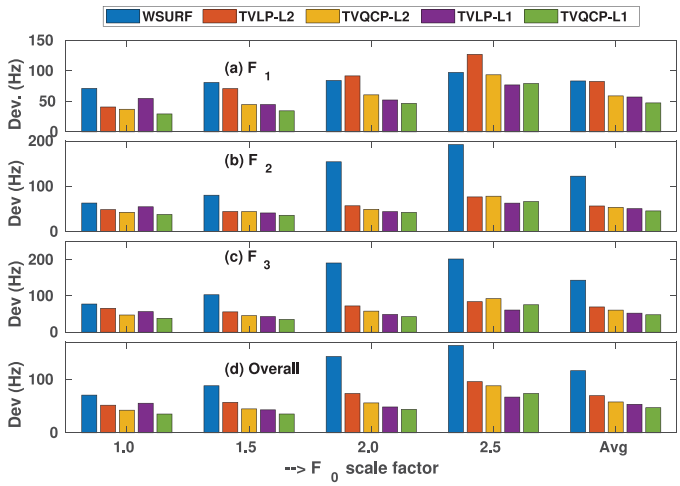


Fig. 6. The absolute deviation (FEE) of the estimated first three formants (F_1, F_2, and F_3) from their ground truth and their overall average for different mean F_0 values of the LF model–based synthetic data.

2) The Effect of Phonation Type: The performance of the TVQCP and TVLP methods are shown in Fig. 5 for the four different phonation types. The TVQCP method that minimizes the L_2 norm (denoted by TVQCP-L2) performed best overall, marginally better than the TVQCP method that minimizes the L_1 norm (denoted by TVQCP-L1). The L_1 norm minimization seemed to perform better than the L_2 norm for most cases in creaky and modal phonations while the L_2 norm performed better for breathy and whispery phonations that exhibit larger open quotients and higher spectral tilts. Overall, it can be seen that the TVQCP methods performed better than their TVLP counterparts across all formants and all phonation types. Moreover, the performance of the both TVLP and TVQCP methods is clearly better than that of the popular Wavesurfer tool.

3) The Effect of Fundamental Frequency: The performance of the TVQCP and TVLP methods are shown in Fig. 6 for all the four ranges of F_0 values. It can be seen that TVQCP optimized using both the L_1 and L_2 norms provided consistent

TABLE I
THE ABSOLUTE DEVIATION (FEE IN HZ) OF THE ESTIMATED FIRST THREE FORMANTS (F_1, F_2, AND F_3) FROM THEIR GROUND TRUTH AND THEIR OVERALL AVERAGE OVER ALL PHONATION TYPES AND FUNDAMENTAL FREQUENCIES OF THE LF MODEL–BASED SYNTHETIC DATA

	WSURF	TVLP-L2	TVQCP-L2	TVLP-L1	TVQCP-L1
(a) Avg. over all phonation types					
F_1	133.3	56.5	**41.8**	63.7	45.0
F_2	252.7	64.2	**54.8**	80.6	59.8
F_3	269.3	81.6	**65.4**	99.7	70.1
Avg	218.4	67.4	**54.0**	81.3	58.3
(b) Avg. over all F_0 range					
F_1	83.5	82.7	59.0	57.2	**47.5**
F_2	122.4	56.6	53.6	50.7	**45.7**
F_3	143.5	69.7	61.0	52.6	**48.1**
Avg	116.5	69.7	57.9	53.5	**47.1**
(c) Avg. over all phonation types and F_0 range					
Overall	167.5	68.6	56.0	67.4	**52.7**

improvements over TVLP up to a scale factor of 2.0. The mixed performance for the scale factor 2.5 may be due to the new F_0 values moving very close to F_1 in the synthetic utterances. However, it has been observed that this minor aberration gets corrected if F_1 is shifted upward by a small percentage. Also, the L_1 norm optimization seemed to perform better than the L_2 norm in most cases except for TVLP for F_1 and F_2 at F_0 scale factor 1.0. In terms of overall performance across all fundamental frequency ranges and formants, TVLP-L2, TVQCP-L2, TVLP-L1, and TVQCP-L1 showed a consistent improvement in this order.

The overall performance of the formant tracking methods is given in Table I by averaging over all phonation types and F_0 ranges.The general observation is that the FEE reduced considerably with the use of QCP analysis (TVQCP analysis vs. TVLP) and that there is a marginal reduction when using the sparsity constraint (L_1 norm vs. L_2 norm). Overall, both the TVLP and TVQCP methods provided large improvements over the popular Wavesurfer tool with 60 to 70 percentage unit reduction in the estimation error.

D. Experiments on Simulated High-Pitched Child Speech Using a Physical Modeling Approach

The formant estimation accuracy of the proposed TVQCP method is compared to that of TVLP using synthetic data generated by an alternate, physical modeling approach of the speech production apparatus [43]. The experiments in this section try to address two issues with the evaluation of formant estimation and tracking methods. One is the bias of the LF model–based synthetic data towards LP–based methods, and the other is the performance of these methods on speech signals at very high fundamental frequencies.

An 8^{th} order analysis is used for all the methods, and the original data at 44.1 kHz is downsampled to 16 kHz and passed through a preemphasis filter $P(z) = 1 - 0.97\ z^{-1}$ before further processing. The TVLP and TVQCP methods use a 100-ms window size and a polynomial order of $q = 3$. The final formant

TABLE II
THE ABSOLUTE DEVIATION (FEE IN Hz) OF THE ESTIMATED FIRST FOUR FORMANTS (F_1–F_4) FROM THEIR GROUND TRUTH ON CHILD SPEECH GENERATED USING THE PHYSICAL MODELING APPROACH

Method	F_1	F_2	F_3	F_4
TVLP-L2	70.8	163.9	69.3	**76.8**
TVLP-L1	52.8	105.0	61.4	106.4
TVQCP-L2	**32.9**	51.9	61.4	136.1
TVQCP-L1	33.0	**48.3**	**54.4**	148.4

estimates are evaluated at a 20-ms frame shift to match the reference formants rate.

1) The Dataset: The simulated data consists of eight short child speech utterances of a high pitch (as high as 600 Hz) used in [44]. The eight utterances include two steady vowels, [a] and [i], of 340 ms duration each with a constant F_0 of 400 Hz. The six simulations of 1.03 s each are three time-varying vocal tract shapes combined with two different time-dependent F_0 variations. The three time-varying vocal tract shapes correspond to the sequence of sounds {i.a.i.a.i.a}, {ae.u.ae.u.ae.u}, and {i.a.i}. The fundamental frequency of the utterances varies between 240 Hz to 500 Hz, one in a smooth increasing–decreasing pattern and the other in a reverse pattern over the entire length of the utterance. All the utterances have four vocal tract resonances and are stored at a 44.1 kHz sampling rate. More information on the formant and F_0 contours used and other details of the dataset can be found in [44].

2) The Results: FEEs computed using both the L_1 and L_2 norms in the TVLP and TVQCP methods are given in Table II. It is seen that the TVQCP method tends to give a consistent shift in estimating the fourth formant. This could be due to many reasons including the pre-emphasis, sampling rate, model limitations, limited synthetic data, and this needs further investigation. In view of this, further discussions in this section are limited to the first three formants.It can be seen from the table that imposing a sparsity constraint with the L_1 norm minimization clearly improves the accuracy of TVLP and TVQCP. The continuity constraint imposed by time-varying models (TVLP) do not seem to provide much improvement on their own. However, when combined with the QCP weighting, the continuity constraint seems to provide large improvements in the case of TVQCP-L1 and some marginal improvement in the case of TVQCP-L2. Owing to the limited availability of data, it may not be possible to draw too many inferences from this experiment. Nevertheless, it demonstrates the usefulness of combining the ideas of QCP analysis, time-varying linear predictive analysis, and the sparsity constraint for formant tracking applications.

E. Experiments on Natural Speech Data

One of the primary goals of this paper is to evaluate the performance of the proposed TVQCP–based formant tracker on real speech utterances. A detailed evaluation of the TVQCP method and a comparison with some of the state-of-the-art formant trackers is presented in this section.

TABLE III
THE EFFECT OF WINDOW SIZE (IN ms), PREDICTION ORDER, AND POLYNOMIAL ORDER ON THE FORMANT TRACKING PERFORMANCE OF TVQCP-L1

	FDR (%)			FEE (Hz)		
	F_1	F_2	F_3	δF_1	δF_2	δF_3
N_t	Effect of window size (p=8, q=3)					
50 ms	88.8	94.9	90.1	79.2	102.0	138.0
100 ms	**90.9**	**96.5**	92.2	67.6	**91.7**	123.9
200 ms	90.8	**96.5**	**92.6**	**66.6**	93.6	**123.0**
p	Effect of prediction order (N_t=100 ms, q=3)					
6	65.9	65.3	21.7	224.9	454.8	919.9
7	75.9	81.5	54.6	151.0	238.8	459.4
8	90.9	**96.5**	**92.2**	67.6	**91.7**	**123.9**
9	**92.0**	89.9	84.1	**63.0**	159.0	198.4
10	91.9	71.6	57.6	63.6	309.2	433.2
q	Effect of polynomial order (N_t=100 ms, p=8)					
0	88.3	92.5	89.4	72.2	114.4	143.3
1	**91.0**	96.5	92.8	**65.4**	90.3	**119.5**
2	**91.0**	**96.6**	**92.5**	66.3	**90.2**	121.4
3	90.9	96.5	92.2	67.6	91.7	123.9

1) The Dataset: The performance of different methods in formant tracking was evaluated on natural speech signals using the VTR database published in [42]. The test data of the VTR database is used for the evaluation and this data consists of 192 utterances (8 utterances pronounced by 8 female and 16 male speakers). The duration of each utterance varies between 2 and 5 s. The first four reference formant frequency and bandwidth values derived using a semi-supervised LP–based algorithm [49] are provided for every 10-ms interval. The first three reference formant frequency values have been verified and corrected manually based on spectrographic evidence. All the speech data, originally recorded at a 16 kHz sampling rate, are downsampled to 8 kHz before processing. A pre-emphasis filter of $P(z) = 1 - 0.97\ z^{-1}$ is used to preprocess the speech signals. Based on our earlier experiments on formant tracking using synthetic speech signals, we use a default window size of 100 ms, a prediction order of 8, and a polynomial order of 3 for the time-varying linear predictive methods unless otherwise mentioned. All the performance metrics presented in this section are average scores computed over vowels, diphthongs, and semivowels. These are phonetic categories whose manually corrected formant ground truths are more reliable compared to other categories.

2) The Effect of Window Size, Prediction Order, and Polynomial Order: The effect of the choices for the window size, prediction order, and polynomial order for the tracking performance of the TVQCP-L1 and TVQCP-L2 methods is provided in Tables III and VI by denoting window size in ms as N_t. It can be seen that the performance of the TVQCP methods is quite stable over a range of values for the window size and polynomial order. However, the performance seems to be slightly sensitive to the choice of prediction order, which needs further investigation.

3) The Choice of Weighting Function: The effect of using different weighting functions within the framework of TVWLP

TABLE IV
THE EFFECT OF DIFFERENT WEIGHTING FUNCTIONS ON THE FORMANT TRACKING PERFORMANCE OF TVWLP (L_1 NORM)

Weighting func.	FDR (%)			FEE (Hz)		
	F_1	F_2	F_3	δF_1	δF_2	δF_3
STE	89.1	95.1	90.5	73.6	105.1	141.3
Residual	90.6	96.3	91.8	68.2	93.2	126.9
QCP	**90.9**	**96.5**	**92.2**	**67.6**	**91.7**	**123.9**

TABLE V
THE EFFECT OF GCI DETECTION ERRORS ON FORMANT TRACKING PERFORMANCE WITH TVQCP-L1. $Rerr$ AND $Ferr$ REFER TO RANDOM AND FIXED ERROR, RESPECTIVELY. THE EFFECT OF THE DURATION QUOTIENT (DQ) OF THE QCP WEIGHTING FUNCTION IS ALSO PRESENTED

	FDR (%)			FEE (Hz)		
	F_1	F_2	F_3	δF_1	δF_2	δF_3
$Rerr$	The effect of GCI error					
0	**90.9**	**96.5**	**92.2**	**67.6**	**91.7**	**123.9**
4	**90.9**	96.3	92.0	68.4	92.7	125.3
8	90.7	96.1	91.7	69.1	94.6	127.6
12	90.6	96.2	91.9	69.1	94.1	127.7
16	90.6	96.0	91.5	69.2	94.9	129.9
$Ferr$	The effect of GCI error					
-16	90.1	95.8	91.4	71.0	98.7	133.8
-8	90.5	96.0	91.3	69.3	96.1	132.1
-4	90.3	96.0	91.4	68.8	95.1	131.2
0	**90.9**	**96.5**	**92.2**	**67.6**	**91.7**	**123.9**
4	90.8	96.1	91.9	71.0	95.9	125.5
8	90.5	95.9	91.3	72.2	97.7	130.9
16	88.5	95.6	90.1	74.4	100.4	143.2
DQ	The effect of DQ					
0.5	90.8	96.0	91.7	70.7	97.0	129.1
0.6	90.8	96.0	92.0	69.5	95.9	127.1
0.7	**90.9**	96.3	92.0	68.1	93.6	125.2
0.8	**90.9**	**96.5**	**92.2**	**67.6**	**91.7**	**123.9**
0.9	**90.9**	96.3	**92.2**	68.0	92.4	124.5

TABLE VI
THE EFFECT OF WINDOW SIZE (IN ms), PREDICTION ORDER, AND POLYNOMIAL ORDER ON THE FORMANT TRACKING PERFORMANCE OF TVQCP-L2

	FDR (%)			FEE (Hz)		
	F_1	F_2	F_3	δF_1	δF_2	δF_3
N_t	Effect of window size (p=8, q=3)					
50 ms	**90.9**	94.4	90.2	67.8	111.0	139.7
100 ms	90.6	**96.1**	**92.0**	68.4	**94.5**	**126.2**
200 ms	89.9	93.9	91.0	**68.3**	118.7	137.9
p	Effect of prediction order (N_t=100 ms, q=3)					
6	66.8	65.6	21.9	226.9	454.2	922.1
7	75.5	78.6	53.0	152.7	261.8	477.0
8	90.6	**96.1**	**92.0**	68.4	**94.5**	**126.2**
9	91.6	89.0	83.6	62.5	165.3	201.6
10	**92.2**	72.1	55.2	**61.3**	292.6	448.9
q	Effect of polynomial order (N_t=100 ms, p=8)					
0	87.8	90.6	87.3	73.1	125.4	157.8
1	90.6	94.5	91.4	**65.7**	112.6	132.8
2	**90.7**	94.3	91.2	66.5	113.3	135.5
3	90.6	**96.1**	**92.0**	68.4	**94.5**	**126.2**

TABLE VII
THE EFFECT OF DIFFERENT WEIGHTING FUNCTIONS ON THE FORMANT TRACKING PERFORMANCE OF TVWLP (L_2 NORM)

Weighting func.	FDR (%)			FEE (Hz)		
	F_1	F_2	F_3	δF_1	δF_2	δF_3
STE	87.4	93.0	88.1	80.0	121.9	161.2
Residual	89.6	95.9	91.6	**68.3**	96.8	132.1
QCP	**90.6**	**96.1**	**92.0**	68.4	**94.5**	**126.2**

TABLE VIII
THE EFFECT OF GCI DETECTION ERRORS ON FORMANT TRACKING PERFORMANCE WITH TVQCP-L2. $Rerr$ AND $Ferr$ REFER TO RANDOM AND FIXED ERROR, RESPECTIVELY. THE EFFECT OF THE DURATION QUOTIENT (DQ) OF THE QCP WEIGHTING FUNCTION IS ALSO PRESENTED

	FDR (%)			FEE (Hz)		
	F_1	F_2	F_3	δF_1	δF_2	δF_3
$Rerr$	The effect of GCI error					
0	**90.6**	**96.1**	**92.0**	**68.4**	**94.5**	**126.2**
4	90.3	96.0	91.4	68.9	95.5	130.2
8	89.9	95.5	91.0	69.7	99.3	134.8
12	89.6	95.5	90.9	69.8	98.3	136.7
16	89.3	95.4	90.7	70.6	99.8	138.0
$Ferr$	The effect of GCI error					
-16	87.7	94.4	89.1	75.5	109.8	150.8
-8	88.5	95.1	90.0	72.0	103.2	143.9
-4	89.0	95.3	90.5	71.1	100.9	141.1
0	90.6	**96.1**	**92.0**	**68.4**	**94.5**	**126.2**
4	**90.7**	95.7	91.4	71.8	97.8	128.6
8	89.8	95.6	91.0	73.6	99.8	132.1
16	87.0	92.6	85.5	80.9	121.4	171.3
DQ	The effect of DQ					
0.5	90.6	95.9	91.8	69.8	96.6	127.7
0.6	**90.8**	96.1	92.0	68.9	95.0	126.1
0.7	90.6	96.1	92.0	68.4	94.5	126.2
0.8	90.6	**96.2**	92.0	**68.2**	**94.0**	**125.9**
0.9	90.5	96.1	**92.1**	68.5	94.5	126.5

for L_1 norm and L_2 norm on formant tracking performance is given in Tables IV and VII. The different weighting functions studied include the signal energy–based STE function, the residual–based weighting function, and the QCP weighting function discussed earlier in Section II-B. It can be seen that the QCP weighting function performs best among the three compared weighting functions. Note that the TVWLP method with the QCP weighting in Tables IV and VII corresponds to TVQCP-L1 analysis and TVQCP-L2 analysis, respectively.

4) Robustness to GCI Detection Errors and the DQ Parameter: The robustness of the proposed TVQCP method to errors in GCI detection was studied by artificially inducing errors in the estimated GCI locations. Two types of errors were studied. In the first, a uniformly distributed random error ($Rerr$) was added to the estimated GCIs. In the second, there was a fixed error ($Ferr$) that gives a consistent bias to the estimated GCIs. The formant tracking results for random and fixed GCI errors is given in Tables V and VIII for TVQCP-L1 and TVQCP-L2, respectively. It can be seen that the performance of the proposed TVQCP methods is robust to GCI errors in the range of 1–2 ms.

TABLE IX
FORMANT TRACKING PERFORMANCE ON NATURAL SPEECH DATA IN TERMS OF FDR AND FEE FOR DIFFERENT METHODS

Method	FDR (%)			FEE (Hz)		
	F_1	F_2	F_3	δF_1	δF_2	δF_3
PRAAT	86.0	70.0	63.1	87.9	268.3	340.1
MUST	81.1	86.3	76.9	90.5	152.3	229.8
WSURF	86.6	82.7	80.8	87.3	222.5	228.2
KARMA	91.5	89.4	74.7	**61.9**	145.8	250.3
DeepF	**91.7**	92.3	89.7	85.1	119.6	142.8
TVLP-L2	88.8	94.9	89.3	70.1	104.9	149.8
TVQCP-L2	90.6	96.1	92.0	68.4	94.5	126.2
TVLP-L1	90.3	96.2	91.7	68.9	94.4	129.6
TVQCP-L1	90.9	**96.5**	**92.2**	67.6	**91.7**	**123.9**

Simulating a fixed GCI error is equivalent to altering the position quotient (PQ) of the QCP weighting function (described in Section II-C). The performance of TVQCP in relation to varying the duration quotient (DQ) of the QCP weighting function between 0.5 and 0.9 is given in Tables V and VIII using L_1 and L_2 norm minimization, respectively. It can be seen that TVQCP performed robustly over a range of DQ values, and the best performance was obtained with DQ=0.8, i.e., using a weighting function, which suppresses the residual energy in 20% of the samples during the fundamental period. Therefore, this value of DQ was used in all the analyses of the study.

5) A Comparison of Time-Variant Linear Predictive Methods and Other Formant Tracking Methods for Clean Speech: The performance of the TVLP and TVQCP methods with different norms are compared to some of the popular formant tracking methods in Table IX. "PRAAT" denotes the Burg method of LP analysis with a 50-ms Gaussian-like window function that is used in formant tracking in Praat, a widely used speech research tool [12]. "MUST" denotes an adaptive filter-bank based method proposed by Mustafa *et al.* [50]. "WSURF" denotes the formant tracker part of Wavesurfer [11] that uses a stabilized covariance analysis over a 25-ms Hamming window. "KARMA" denotes the state-of-the-art KF–based formant tracking method published in [14]. "DeepF" (DeepFormants) denotes the deep-learning based formant tracking method proposed recently in [16], [18], [51]. It is worth emphasizing that DeepF is based on supervised learning and calls for an annotated speech corpus to be trained.

It can be seen from Table IX that the TVLP and TVQCP methods clearly performed better (a 20–60% reduction in error across the three formants) compared to the popular formant tracking methods (Praat and Wavesurfer) that use a two-stage detect-and-track approach. The proposed TVQCP method provided an improvement in the performance (both FDRs and FEEs) of tracking the second and third formants (a reduction in the estimation error of around 30% and 50% respectively) compared to KARMA. The KARMA method performed slightly better than the TVQCP method (with a relative improvement of around 9%) in tracking the first formant. Compared to DeepF, the proposed TVQCP method provided an improvement in FEEs of around 20%, 21% and 12% for all the three formants, respectively. In terms of FDR, DeepF performed slightly better (around 1%) than TVQCP for the first formant. However, for the second and third formants, TVQCP improved the FDR by around 4% and 3%, respectively, compared to DeepF. Differences in performance within the family of time-varying methods were not as evident. However, it can be seen from the results that the use of TVQCP analysis seems to improve the performance of formant tracking. It can also be observed that TVLP-L1 is slightly better than TVLP-L2, and TVQCP-L1 is slightly better than TVQCP-L2. Between TVLP-L1 and TVQCP-L1, TVQCP-L1 is better than TVLP-L1 in both FDRs and FEEs for all the three formants (a reduction in the estimation error of around 2%, 3% and 4% for F1, F2 and F3, respectively).

TABLE X
THE FORMANT TRACKING PERFORMANCE OF KARMA, DEEPF, TVLP-L1 AND TVQCP-L1 IN TERMS OF FDR AND FEE FOR DIFFERENT PHONETIC CATEGORIES OF NATURAL SPEECH DATA

Phonetic category	FDR (%)			FEE (Hz)		
	F_1	F_2	F_3	δF_1	δF_2	δF_3
KARMA						
Vowels (V)	92.6	89.0	74.5	**57.1**	149.5	251.1
Diphthongs (D)	92.5	92.3	76.5	**62.8**	128.7	239.8
Semivowels (S)	86.9	86.9	73.6	**76.1**	154.8	258.3
V+D+S	91.5	89.4	74.7	**61.9**	145.8	250.3
All voiced sounds	87.9	88.4	75.0	**70.9**	151.7	248.0
DeepF						
Vowels (V)	**92.7**	93.7	91.0	81.5	112.9	135.4
Diphthongs (D)	**93.2**	93.8	90.6	84.8	112.2	132.9
Semivowels (S)	**87.0**	86.1	84.4	96.1	148.4	176.2
V+D+S	**91.7**	92.3	89.7	85.1	119.6	142.8
All voiced sounds	**88.8**	90.6	**88.7**	86.8	**129.7**	**147.4**
TVLP-L1						
Vowels (V)	91.9	97.0	92.8	63.3	86.4	117.8
Diphthongs (D)	92.0	98.1	93.7	64.9	81.4	114.9
Semivowels (S)	83.6	91.6	85.9	90.1	**133.4**	**182.0**
V+D+S	90.3	96.2	91.7	68.9	94.4	129.6
All voiced sounds	83.6	90.9	87.2	96.0	133.8	162.9
TVQCP-L1						
Vowels (V)	92.6	**97.4**	**93.6**	62.0	**82.5**	**110.6**
Diphthongs (D)	92.5	**98.3**	**94.3**	63.6	**77.4**	**107.7**
Semivowels (S)	84.2	**91.7**	**85.5**	89.0	135.7	182.5
V+D+S	90.9	**96.5**	**92.2**	67.6	**91.7**	**123.9**
All voiced sounds	84.1	**91.2**	87.5	95.1	131.4	159.2

A detailed comparison in the formant tracking performance of KARMA, DeepF, TVLP-L1 and TVQCP-L1 is given in Table X for different phonetic categories. It can be seen that the estimation error of TVQCP-L1 is 15–40% and 25–55% smaller than that of KARMA for F_2 and F_3 respectively. Likewise, KARMA gave an estimation error that was 1–15% smaller than that of TVQCP-L1 for F_1 across different phonetic categories. In comparison to DeepF, the estimation error of TVQCP-L1 was 7–25%, 9–30% and 13–20% smaller for F_1, F_2 and F_3, respectively (except in semivowels for F_3). The performance of DeepF for F_3 in semivowels was better (by around 4%) than that of TVQCP-L1. It can also be observed that the performance of TVLP-L1 for F_2 and F_3 in semivowels was slightly better (by

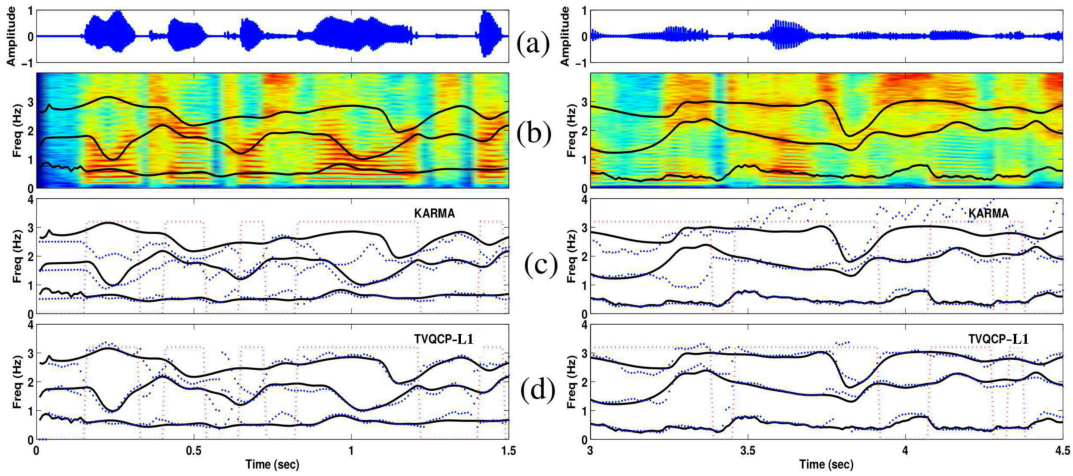


Fig. 7. Examples of formant tracks estimated by KARMA and TVQCP-L1 from utterances produced by a female (on the left) and male (on the right) talker: (a) time-domain speech signals, (b) a narrowband spectrogram with reference ground truth formant contours, (c) formant track estimates using KARMA along with the voiced–unvoiced regions shown by a dotted rectangular-wave plot, and (d) formant track estimates using TVQCP-L1.

around 1%) than that of TVQCP-L1. On the other hand, the performance of TVQCP-L1 for F_1, F_2 and F_3 was better (by around 2–6%) than that of TVLP-L1. When all the voiced sounds are considered, the performance of DeepF was better than the other methods reflecting the fact that DeepF benefits from supervised learning of the formant contours in the model training. Note that the reliability of the manually corrected reference ground truth is less for the other phonetic categories. In view of this, we can argue that the proposed TVQCP method provided the best overall formant tracking performance compared to the popular and state-of-the-art reference methods. A qualitative comparison of the formant tracking performances of the TVQCP-L1 and KARMA methods is demonstrated by Fig. 7 for utterances produced by a male and female speaker. It can be seen from the figure that the TVQCP-L1 method clearly performed better than KARMA in tracking F_2 and F_3, with a comparable performance for F_1.

6) A Comparison of Time-Variant Linear Predictive Methods and Other Formant Tracking Methods for Noisy Speech: In this section, the performance of the TVQCP-L1 method is compared to KARMA and DeepF in formant tracking of noisy speech, as these methods were shown to perform better than the other methods for clean speech. Noisy speech was obtained by corrupting the original clean signals of the VTR speech database with different types of additive noise. The results obtained for stationary and non-stationary noise of three types (volvo, babble, and white) in three SNR categories (20 dB, 10 dB and 5 dB) are given in Table XI. From the results, it can be observed that the performance of the methods drops as the SNR decreases from 20 dB to 5 dB. For signals corrupted by volvo noise, it is observed that the performance of TVQCP-L1 is better than that of the other methods in all SNR categories in both FDR and FEE for all the formants. In the case of the babble noise, the methods behave similarly as in clean speech. That is, FDR of F_1 is better for DeepF, FEE of F_1 is better for KARMA, and the proposed method is better in both FDR and FEE for F_2 and F_3 formants. In the case of white noise (without pre-emphasis), the performance

TABLE XI
FORMANT TRACKING PERFORMANCE FOR DIFFERENT METHODS USING SPEECH DEGRADED WITH VOLVO, BABBLE AND WHITE NOISE AT SNR LEVELS OF 20 dB, 10 dB, AND 5 dB

Method	FDR (%)			FEE (Hz)		
	F_1	F_2	F_3	δF_1	δF_2	δF_3
Volvo at 20 dB						
KARMA	90.1	88.5	73.3	68.1	153.8	266.4
DeepF	90.0	92.0	88.2	96.1	117.5	149.3
TVQCP-L1	**91.0**	**96.2**	**92.1**	**68.0**	**93.2**	**125.3**
Volvo at 10 dB						
KARMA	86.2	86.6	71.7	80.4	167.6	278.8
DeepF	89.4	91.5	87.4	97.3	120.4	153.7
TVQCP-L1	**90.7**	**95.8**	**91.5**	**70.9**	**96.9**	**130.2**
Volvo at 5 dB						
KARMA	80.8	84.9	70.6	96.4	182.9	299.3
DeepF	89.7	90.7	85.9	95.4	124.7	160.3
TVQCP-L1	**89.8**	**95.1**	**90.8**	**76.3**	**105.2**	**138.2**
Babble at 20 dB						
KARMA	**91.7**	88.0	74.2	**61.4**	152.6	247.8
DeepF	91.3	91.7	87.1	89.7	118.0	155.1
TVQCP-L1	89.6	**94.7**	**89.9**	68.6	**103.2**	**136.9**
Babble at 10 dB						
KARMA	90.3	83.8	71.8	**65.1**	176.1	246.0
DeepF	**91.1**	86.6	81.7	88.4	145.9	182.7
TVQCP-L1	84.3	**88.1**	**82.5**	78.1	**144.5**	**181.0**
Babble at 5 dB						
KARMA	88.2	78.9	68.7	**70.9**	200.9	260.3
DeepF	**89.8**	81.4	76.1	89.9	177.3	**209.1**
TVQCP-L1	80.9	**83.2**	**76.6**	86.4	**174.0**	212.7
White at 20 dB						
KARMA	90.4	87.6	73.6	**64.4**	150.5	240.6
DeepF	90.1	90.4	**84.4**	95.4	**125.9**	**167.9**
TVQCP-L1	**92.0**	**93.0**	79.6	68.3	129.1	205.7
White at 10 dB						
KARMA	86.2	80.1	68.8	75.5	191.3	256.5
DeepF	89.8	80.8	**71.6**	99.2	184.3	**238.7**
TVQCP-L1	**90.3**	**85.2**	66.1	**73.7**	**179.2**	280.7
White at 5 dB						
KARMA	80.1	72.5	64.0	91.6	232.5	279.2
DeepF	**89.2**	71.7	**64.5**	101.1	238.7	**274.3**
TVQCP-L1	88.3	**77.9**	59.6	**82.4**	**220.1**	314.5

of the proposed method for F_1 and F_2 is better than others in both FDR and FEE (except at 20 dB SNR). Whereas, the performance of the DeepF method for F_3 is better than others in both FDR and FEE.

V. Conclusion

In this paper, we proposed a new formant tracking method, TVQCP,[1] for speech signals. The TVQCP method combines the advantages of QCP analysis (reducing the effect of the glottal source in formant estimation by using temporal weighting of the prediction error), the increased sparsity of the prediction error due to the L_1 norm minimization and TVLP analysis (imposing a time-continuity constraint to take into account the slowness of movements in the real human vocal tract). The use of a time-continuity constraint on the vocal tract parameters eliminates the need for a two-stage detect-and-track strategy to combine them into one. Formant tracking experiments on synthetic speech utterances demonstrate the advantages of the proposed TVQCP method over TVLP. A comparison of performance on natural speech utterances shows that the TVQCP method performs better than some of the state-of-the-art formant trackers, such as Praat, Wavesurfer, KARMA and DeepFormants. One limitation of the proposed TVQCP method is its apparent sensitivity to the choice of prediction order, though a prediction order of 8 works consistently well in tracking formants over a large set of natural speech utterances of male and female talkers. In addition, there is a need for devising a better coasting strategy (such as the one that is used in KARMA) for tracking formants in non-speech and non-voiced sections or in less-reliable voiced regions.

Acknowledgment

The authors would like to thank the authors of Deep Formants for their helpful discussions.

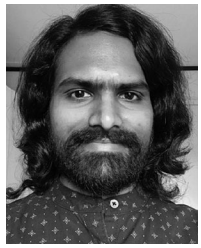
References

[1] G. Fant, *Acoustic Theory of Speech Production*. The Hague, Netherlands: Mouton & Co., 1960, pp. 1–328.
[2] P. F. Assmann, "The role of formant transitions in the perception of concurrent vowels," *J. Acoustical Soc. America*, vol. 97, no. 1, pp. 575–584, 1995.
[3] R. Singh, D. Gencaga, and B. Raj, "Formant manipulations in voice disguise by mimicry," in *Proc. 4th Int. Conf. Biometrics Forensics*, Mar. 2016, pp. 1–6.
[4] L. Welling and H. Ney, "Formant estimation for speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 1, pp. 36–48, Jan. 1998.
[5] T. Smit, F. Türckheim, and R. Mores, "Fast and robust formant detection from LP data," *Speech Commun.*, vol. 54, no. 7, pp. 893–902, 2012.
[6] N. B. Pinto, D. G. Childers, and A. L. Lalwani, "Formant speech synthesis: Improving production quality," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 12, pp. 1870–1887, Dec. 1989.
[7] P. Y. Chan *et al.*, "Formant excursion in singing synthesis," in *Proc. IEEE Int. Conf. Digit. Signal Process.*, Jul. 2015, pp. 168–172.
[8] I. C. Yoo, H. Lim, and D. Yook, "Formant-based robust voice activity detection," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 23, no. 12, pp. 2238–2245, Dec. 2015.
[9] J. R. Schilling, R. L. Miller, M. B. Sachs, and E. D. Young, "Frequency-shaped amplification changes the neural representation of speech with noise-induced hearing loss," *Hearing Res.*, vol. 117, no. 1-2, pp. 57–70, 1998.
[10] I. C. Bruce, "Physiological assessment of contrast-enhancing frequency shaping and multiband compression in hearing aids," *Physiological Meas.*, vol. 25, no. 4, pp. 945–956, 2004.
[11] K. Sjolander and J. Beskow, "Wavesurfer - An open source speech tool," in *Proc. Int. Conf. Spoken Lang. Process.*, Oct. 2000, pp. 464–467.
[12] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot Int.*, vol. 5, no. 9/10, pp. 341–345, 2001.
[13] L. Deng, L. J. Lee, H. Attias, and A. Acero, "Adaptive Kalman filtering and smoothing for tracking vocal tract resonances using a continuous-valued hidden dynamic model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 13–23, Jan. 2007.
[14] D. D. Mehta, D. Rudoy, and P. J. Wolfe, "Kalman-based autoregressive moving average modeling and inference for formant and antiformant tracking," *J. Acoustical Soc. America*, vol. 132, no. 3, pp. 1732–1746, 2012.
[15] J. L. Durrieu and J. P. Thiran, "Source/filter factorial hidden Markov model, with application to pitch and formant tracking," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 12, pp. 2541–2553, Dec. 2013.
[16] Y. Dissen and J. Keshet, "Formant estimation and tracking using deep learning," in *Proc. Interspeech*, Sep. 2016, pp. 958–962.
[17] F. Schiel and T. Zitzelsberger, "Evaluation of automatic formant trackers," in *Proc. 11th Int. Conf. Lang. Resour. Eval.*, 2018, pp. 2843–2848.
[18] Y. Dissen, J. Goldberger, and J. Keshet, "Formant estimation and tracking: A deep learning approach," *J. Acoustical Soc. America*, vol. 145, no. 2, pp. 642–653, 2019.
[19] B. S. Atal and M. R. Schroeder, "Predictive coding of speech signals," in *Proc. Conf. Commun. Process.*, 1967, pp. 360–361.
[20] F. Itakura and S. Saito, "Analysis synthesis telephony based upon the maximum likelihood method," in *Proc. Rep. 6th Int. Congr. Acoust.*, 1968, pp. C17–C20.
[21] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, Apr. 1975.
[22] S. M. Kay, *Modern Spectral Estimation: Theory Application*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1988.
[23] D. Wong, J. Markel, and J. Gray, A., "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Trans. Acoust., Speech Signal Process.*, vol. ASSP-27, no. 4, pp. 350–355, Aug. 1979.
[24] K. Steiglitz and B. Dickinson, "The use of time-domain selection for improved linear prediction," *IEEE Trans. Acoust., Speech Signal Process.*, vol. ASSP-25, no. 1, pp. 34–39, Feb. 1977.
[25] B. Yegnanarayana and R. Veldhuis, "Extraction of vocal-tract system characteristics from speech signals," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 4, pp. 313–327, Jul. 1998.
[26] R. Mizoguchi, M. Yanagida, and O. Kakusho, "Speech analysis by selective linear prediction in the time domain," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 1982, vol. 7, pp. 1573–1576.
[27] M. Yanagida and O. Kakusho, "A weighted linear prediction analysis of speech signals by using the givens reduction," in *Proc. IASTED Int. Symp. Appl. Signal Process. Digit. Filtering*, Jun. 1985, pp. 129–132.
[28] C.-H. Lee, "On robust linear prediction of speech," *IEEE Trans. Acoust., Speech Signal Process.*, vol. 36, no. 5, pp. 642–650, May 1988.
[29] 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.
[30] C. Magi, J. Pohjalainen, T. Bäckström, and P. Alku, "Stabilized weighted linear prediction," *Speech Commun.*, vol. 51, no. 5, pp. 401–411, 2009.
[31] J. Pohjalainen, R. Saeidi, T. Kinnunen, and P. Alku, "Extended weighted linear prediction (XLP) analysis of speech and its application to speaker verification in adverse conditions," in *Proc. INTERSPEECH*, Sep. 2010, pp. 1477–1480.
[32] P. Alku, J. Pohjalainen, M. Vainio, A.-M. Laukkanen, and B. H. Story, "Formant frequency estimation of high-pitched vowels using weighted linear prediction," *J. Acoustical Soc. America*, vol. 134, no. 2, pp. 1295–1313, 2013.
[33] 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.
[34] A. R. Mv and P. K. Ghosh, "Glottal inverse filtering using probabilistic weighted linear prediction," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 1, pp. 114–124, Jan. 2019.
[35] M. G. Hall, A. V. Oppenheim, and A. S. Willsky, "Time-varying parametric modeling of speech," *Signal Process.*, vol. 5, no. 3, pp. 267–285, 1983.
[36] K. Schnell and A. Lacroix, "Time-varying linear prediction for speech analysis and synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2008, pp. 3941–3944.

[1] [Online]. Available: https://github.com/njaygowda/ftrack

[37] S. Chetupalli and T. Sreenivas, "Time varying linear prediction using sparsity constraints," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2014, pp. 6290–6293.
[38] D. Wong, J. Markel, and A. Gray, "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.
[39] D. Giacobello, M. Christensen, M. Murthi, S. Jensen, and M. Moonen, "Sparse linear prediction and its applications to speech processing," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 5, pp. 1644–1657, Jul. 2012.
[40] D. Wipf and S. Nagarajan, "Iterative reweighted l_1 and l_2 methods for finding sparse solutions," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 2, pp. 317–329, Apr. 2010.
[41] D. Gowda, M. Airaksinen, and P. Alku, "Quasi closed phase analysis of speech signals using time varying weighted linear prediction for accurate formant tracking," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, 2016, pp. 4980–4981.
[42] L. Deng, X. Cui, R. Pruvenok, J. Huang, and S. Momen, "A database of vocal tract resonance trajectories for research in speech processing," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, 2006, pp. I–369–I–372.
[43] B. H. Story, "Phrase-level speech simulation with an airway modulation model of speech production," *Comput. Speech Lang.*, vol. 27, no. 4, pp. 989–1010, 2013.
[44] B. H. Story and K. Bunton, "Formant measurement in childrens speech based on spectral filtering," *Speech Commun.*, vol. 76, pp. 93–111, 2016.
[45] D. Talkin, "Speech formant trajectory estimation using dynamic programming with modulated transition costs," *J. Acoust. Soc. America*, vol. 82, no. S1, pp. S55–S55, 1987.
[46] G. Fant, J. Liljencrants, and Q. G. Lin, "A four-parameter model of glottal flow," *Speech Transmiss. Lab. Quart. Prog. Status Rep.*, vol. 4, pp. 1–17, 1985.
[47] C. Gobl, "The voice source in speech communication - production and perception experiments involving inverse filtering and synthesis," Ph.D. dissertation, Dept. Speech Transmiss. Music Acoust., Roy. Inst. Technol., Stockholm, Sweden, 2003.
[48] C. Gobl, "A preliminary study of acoustic voice quality correlates," *Speech Transmiss. Lab. Quart. Prog. Status Rep.*, vol. 4, pp. 9–22, 1989.
[49] L. Deng, L. Lee, H. Attias, and A. Acero, "A structured speech model with continuous hidden dynamics and prediction-residual training for tracking vocal tract resonances," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, May 2004, vol. 1, pp. I–557–60.
[50] K. Mustafa and I. C. Bruce, "Robust formant tracking for continuous speech with speaker variability," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 2, pp. 435–444, Mar. 2006.
[51] [Online]. Available: https://github.com/MLSpeech/DeepFormants, Accessed on: Oct. 30, 2019.

Dhananjaya Gowda (Member, IEEE) received the master's and Doctorate degrees in the area of speech signal processing from the Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai, India, in 2004 and 2011, respectively. He is currently a Principal Engineer with Speech Processing Lab, AI Center, Samsung Research at Seoul R&D Campus, Seoul, South Korea. He was a postdoctoral researcher with Aalto University, Espoo, Finland, from 2012 to 2017. His research interests include speech processing, signal processing, speech recognition, machine learning, and spoken language understanding.



Sudarsana Reddy Kadiri (Member, IEEE) received the Bachelor of Technology degree from Jawaharlal Nehru Technological University Hyderabad, Hyderabad, India, in 2011, with a specialization in Electronics and Communication Engineering (ECE), the M.S. (Research) degree during 2011–2014, and later converted to Ph.D., and received the Ph.D. degree from the Department of ECE, International Institute of Information Technology Hyderabad (IIIT-H), Hyderabad, India, in 2018. He was a Teaching Assistant for several courses with IIIT-H during 2012–2018. He is currently a Postdoctoral Researcher with the Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland. His research interests include signal processing, speech analysis, speech synthesis, paralinguistics, affective computing, voice pathologies, machine learning, and auditory neuroscience.

Brad Story is currently a Professor with the Department of Speech, Language, and Hearing Sciences, the University of Arizona, Tucson, AZ, USA. His research is concerned with development of computational, physically-based models that simulate the observed structure, movement, and acoustic characteristics of specific components of the speech production system. He has taught courses at both the undergraduate and graduate levels in Speech Science, Speech Perception, Acoustics, Hearing Science, and Anatomy and Physiology. He is a fellow of the Acoustical Society of America, recipient of the Rossing Prize in Acoustics Education and Willard R. Zemlin Lecture in Speech Science, and has served as an Associate Editor for the *Journal of the Acoustical Society of America.* He has authored more than 100 publications in the area of voice and speech science.

Paavo Alku (Fellow, IEEE) received the M.Sc., Lic.Tech., and Dr.Sc.(Tech) degrees from Helsinki University of Technology, Espoo, Finland, in 1986, 1988, and 1992, respectively. He was an assistant professor with the Asian Institute of Technology, Bangkok, Thailand, in 1993, and an assistant professor and a professor with the University of Turku, Finland, from 1994 to 1999. He is currently a professor of speech communication technology with Aalto University, Espoo, Finland. His research interests include analysis and parameterization of speech production, statistical parametric speech synthesis, spectral modeling of speech, speech-based biomarking of human health, and cerebral processing of speech. He has authored or coauthored more than 200 peer-reviewed journal articles and more than 200 peer-reviewed conference papers. He is an Associate Editor for the *Journal of the Acoustical Society of America.* He served as an Academy Professor assigned by the Academy of Finland in 2015–2019. He is a fellow of ISCA.