Aalto University

OPEN ACCESS

This is an electronic reprint of the original article.
This reprint may differ from the original in pagination and typographic detail.

Gowda, Dhananjaya; Airaksinen, Manu; Alku, Paavo

Quasi-closed phase forward-backward linear prediction analysis of speech for accurate formant detection and estimation

Published in:
Journal of the Acoustical Society of America

DOI:
10.1121/1.5001512

Published: 25/09/2017

Document Version
Publisher's PDF, also known as Version of record

Please cite the original version:
Gowda, D., Airaksinen, M., & Alku, P. (2017). Quasi-closed phase forward-backward linear prediction analysis of speech for accurate formant detection and estimation. *Journal of the Acoustical Society of America*, *142*(3), 1542-1553. https://doi.org/10.1121/1.5001512

This material is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the repository collections is not permitted, except that material may be duplicated by you for your research use or educational purposes in electronic or print form. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone who is not an authorised user.

Quasi-closed phase forward-backward linear prediction analysis of speech for accurate formant detection and estimation

Dhananjaya Gowda, Manu Airaksinen, and Paavo Alku

Citation: The Journal of the Acoustical Society of America **142**, 1542 (2017); doi: 10.1121/1.5001512

View online: https://doi.org/10.1121/1.5001512

View Table of Contents: http://asa.scitation.org/toc/jas/142/3

Published by the Acoustical Society of America

Articles you may be interested in

Speech intelligibility is best predicted by intensity, not cochlea-scaled entropy
The Journal of the Acoustical Society of America **142**, EL264 (2017); 10.1121/1.5002149

Maintaining intelligibility at high intensities with arrays of subcritical width speech bands and interpolated noise
The Journal of the Acoustical Society of America **142**, EL299 (2017); 10.1121/1.5002735

Measuring vocal motor skill with a virtual voice-controlled slingshot
The Journal of the Acoustical Society of America **142**, 1199 (2017); 10.1121/1.5000233

Frequency domain synthetic aperture focusing technique for variable-diameter cylindrical components
The Journal of the Acoustical Society of America **142**, 1554 (2017); 10.1121/1.5003650

Speech recognition error patterns for steady-state noise and interrupted speech
The Journal of the Acoustical Society of America **142**, EL306 (2017); 10.1121/1.5003916

Spectral integration of English speech for non-native English speakers
The Journal of the Acoustical Society of America **142**, 1646 (2017); 10.1121/1.5003933

Quasi-closed phase forward-backward linear prediction analysis of speech for accurate formant detection and estimation

Dhananjaya Gowda,[a] Manu Airaksinen, and Paavo Alku
Department of Signal Processing and Acoustics, Aalto University, Otakaari 5, FI-00076 Espoo, Finland

(Received 16 May 2016; revised 6 July 2017; accepted 24 August 2017; published online 22 September 2017)

Recently, a quasi-closed phase (QCP) analysis of speech signals for accurate glottal inverse filtering was proposed. However, the QCP analysis which belongs to the family of temporally weighted linear prediction (WLP) methods uses the conventional forward type of sample prediction. This may not be the best choice especially in computing WLP models with a hard-limiting weighting function. A sample selective minimization of the prediction error in WLP reduces the effective number of samples available within a given window frame. To counter this problem, a modified quasi-closed phase forward-backward (QCP-FB) analysis is proposed, wherein each sample is predicted based on its past as well as future samples thereby utilizing the available number of samples more effectively. Formant detection and estimation experiments on synthetic vowels generated using a physical modeling approach as well as natural speech utterances show that the proposed QCP-FB method yields statistically significant improvements over the conventional linear prediction and QCP methods. © *2017 Acoustical Society of America*.

[http://dx.doi.org/10.1121/1.5001512]

[CYE] Pages: 1542–1553

I. INTRODUCTION

Accurate detection, estimation, and tracking of formant frequencies from speech signals has potentially many applications in acoustic-phonetic analysis (Fant, 1960; Assmann, 1995), voice morphing (Singh *et al.*, 2016), speech recognition (Welling and Ney, 1998; Smit *et al.*, 2012), speech or signing voice synthesis (Pinto *et al.*, 1989; Chan *et al.*, 2015), voice activity detection (Yoo *et al.*, 2015) and in designing hearing aids for people suffering from sound-induced hearing loss (Schilling *et al.*, 1998; Bruce, 2004). Several algorithms have been proposed for tracking formant frequencies (Boersma, 2001; Sjolander and Beskow, 2000; Deng *et al.*, 2007; Mehta *et al.*, 2012). Most of these algorithms have a detection stage, where an initial estimate of the vocal tract resonances (VTRs) (manifested as spectral peaks or formants) is obtained, followed by a tracking stage (Boersma, 2001; Sjolander and Beskow, 2000). However, some algorithms try to do a simultaneous estimation and tracking from an initial representation of the vocal tract system (Deng *et al.*, 2007; Mehta *et al.*, 2012). In either approach, analysis of the signal for accurate estimation (or modeling) of the vocal tract system is an important and necessary computational block.

Linear prediction (LP) analysis is one of the widely used techniques for modeling the vocal tract system and estimating the VTRs from speech signals (Atal and Schroeder, 1967; Itakura and Saito, 1968; Makhoul, 1975). Several variants or formulations of LP analysis have been proposed to improve the accuracy of these estimations (Kay, 1988). However, the autocorrelation (ACOR) and covariance (COV) based analyses are the most popularly used methods for formant estimation and tracking (Boersma, 2001; Sjolander and Beskow, 2000). Covariance based LP analysis is known to provide more accurate formant estimates than the popular ACOR analysis, but does not ensure stability of the estimated filter (Makhoul, 1975; Wong *et al.*, 1979). The instability of the filter in itself is not a serious problem as long as the task on hand is only formant tracking with no need for reconstructing the signal such as in speech synthesis or coding. Also, a closed phase analysis of the speech signal 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 (Steiglitz and Dickinson, 1977; Yegnanarayana and Veldhuis, 1998). But the closed phase analysis works better for low-pitched (male) voices which has more samples in the closed phase of the glottal cycle as against high-pitched female and child voices. One way to counter this problem is to do a selective prediction of speech samples and over multiple glottal cycles.

The family of weighted linear prediction (WLP) methods performs a selective prediction by giving a different temporal weighting on the prediction error at each sample (Mizoguchi *et al.*, 1982; Yanagida and Kakusho, 1985; Lee, 1988; Ma *et al.*, 1993; Magi *et al.*, 2009; Pohjalainen *et al.*, 2010; Alku *et al.*, 2012, 2013; Airaksinen *et al.*, 2014). A sample selective linear prediction analysis with a hard rejecting weighting function to eliminate outlier samples was proposed by Mizoguchi *et al.* (1982) for better modeling of the vocal tract area function. A more generalized version of WLP was proposed by Yanagida and Kakusho (1985) with a continuous weighting function on the prediction residual. Lee (1988) proposed a robust linear prediction algorithm using iterative solutions by utilizing the non-Gaussian nature

[a] Current address: DMC R&D Center, Samsung Electronics, Seoul, Korea. Electronic mail: dhananjaya.gowda@aalto.fi

0001-4966/2017/142(3)/1542/12/$30.00 © 2017 Acoustical Society of America

of the excitation source to derive a weighting function based on the magnitude of the residual samples.

A non-iterative solution to WLP was proposed by Ma *et al.* (1993), where the short-time energy (STE) computed over 1–2 ms was used as the weighting function to improve the robustness of spectrum estimation. An STE weighting function gives more weight to high energy regions within a glottal cycle, which roughly correspond to the closed phase regions as well. Several variants of the STE weighting function have been explored in order to improve the robustness of WLP based features in the face of degradations (Ma *et al.*, 1993; Pohjalainen *et al.*, 2010), and to ensure stability of the estimated filter (Magi *et al.*, 2009). However, it was shown by Alku *et al.* (2012) that the accuracy of the vocal tract estimates suffers if the designed weighting function gives more weight to the region around the glottal closure instant (GCI) during which maximum excitation is imparted to the vocal tract. An attenuated main excitation (AME) weighting function was proposed to improve the accuracy of formant estimation, especially in the case of high-pitched voices (Alku *et al.*, 2012, 2013). Based on these ideas, a quasi-closed phase (QCP) analysis of speech signals for accurate glottal inverse filtering was proposed by the present authors (Airaksinen *et al.*, 2014). A more generalized AME weighting function with slant edges instead of vertical ones was used. However, the paper was focused primarily on estimating the glottal source parameters, without any evaluation of its performance in formant detection and estimation.

One drawback with the QCP analysis proposed by Airaksinen *et al.* (2014) is that the net effective number of samples over which the prediction error is minimized within a fixed window frame is reduced due to the use of an almost binary weighting function. This selective minimization of error over a reduced number of samples does not seem to result in data insufficiency problems when evaluating the method for formant estimation accuracies (deviation from the ground truth) using synthetic as well as long sustained natural vowel data (Alku *et al.*, 2013). However, its impact on natural continuous speech utterances and the possible remedies need to be studied carefully. Also, while computing formant estimation accuracies for natural speech signals, the authenticity of the ground truth formant locations is always questionable. Nevertheless, one thing that can be safely assumed is that the reference formant locations marked are within a reasonable deviation from an otherwise unknown absolute ground truth. In such a scenario, formant estimation accuracy is therefore not a good metric in evaluating the performance of vocal tract estimation. In view of this, the authors propose to evaluate vocal tract estimation methods using a formant detection rate (FDR) defined as the percentage of frames where the estimated formant location is within a reasonable deviation from the ground truth.

In order to address the above limitation of QCP due to selective optimization over reduced samples, the authors propose to combine, for the first time, two different approaches in linear predictive analysis of speech signals: (1) the framework of QCP analysis in which temporally weighted LP is used, and (2) the forward-backward (FB) analysis. The proposed combination of the two approaches gives rise to a new algorithm, quasi-closed phase forward-backward (QCP-FB) analysis. This new predictive algorithm provides two major advantages over the conventional LP methods. First, a weighting function which is used in QCP is based on the knowledge of GCIs. This temporal weighting function emphasizes the closed phase region of the glottal cycles, at the same time de-emphasizing the open phase region as well as the region immediately after the main excitation. This provides a more accurate closed phase estimate of the vocal tract model, and with a reduced effect of the glottal source. Second, FB analysis reduces considerably the problems of data insufficiency, spectral line splitting, and sensitivity of spectral peaks to window positioning as well as additive noise, commonly associated with conventional LP analyses. Formant detection experiments on natural as well as synthetic speech signals show that the proposed QCP-FB method improves considerably the formant detection accuracies as compared to conventional LP and WLP methods.

II. QCP ANALYSIS

QCP analysis is a variant of WLP with a specially designed weighting function based on the knowledge of GCIs (Airaksinen *et al.*, 2014). An overview of WLP and the design of QCP weighting function is given in this section.

A. WLP analysis

In conventional LP, the current sample x_n is predicted based on the past p samples given by

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

where $\{a_k\}_{k=0}^{p}$ with $a_0 = 1$ denotes the prediction coefficients. $H(z) = 1/A(z)$ denotes the estimated transfer function of the vocal tract system, where $A(z)$ is the z-transform corresponding to the prediction coefficients $\{a_k\}_{k=0}^{p}$. The optimal prediction coefficients are required to reduce the overall prediction error given by the cost function

$$E = \sum_{n} e_n^2, \tag{2}$$

where $e_n = x_n - \hat{x}_n$ is the sample-wise prediction error. The prediction coefficients are computed by minimizing the cost function ($\partial E/\partial a_i = 0,\ 1 \le i \le p$) and solving the resulting normal equations

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

$$\text{where } r_{i,k} = \sum_{n} x_{n-i} x_{n-k}. \tag{4}$$

In the above formulation, it can be seen that the prediction error is minimized in the least-square sense with equal temporal weight on predicting every sample or reducing the error at each sample. However, in WLP the cost function gives a differential weight to the prediction error at each sample. The cost function in WLP is given by

$$E_w = \sum_n w_n e_n^2, \tag{5}$$

where w_n denotes the weighting function on the sample-wise prediction error e_n. It should be noted here that the weighting in WLP methods is on the error signal, and should not be mixed up with the traditional short-time windowing (e.g., Hamming) used for reducing truncation effects. The prediction coefficients can be computed in a similar way by minimizing the cost function ($\partial E_w / \partial a_i = 0,\ 1 \le i \le p$) and solving the resulting normal equations

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

$$\text{where } b_{i,k} = \sum_n w_n x_{n-i} x_{n-k}. \tag{7}$$

B. Choice of weighting function

As mentioned earlier in Sec. I, several weighting functions have been proposed for a sample selective WLP. In this section, we discuss two most relevant weighting functions for this study, namely, the STE weighting function and the QCP weighting function. An illustration of the speech signal, electroglottogram (EGG) signal, derivative of electroglottogram (dEGG) signal, STE weight, and the QCP weighting functions along with rough closed and open phase markings is shown in Fig. 1.

1. STE weighting function

STE is one of the popular weighting functions used. The STE weighting function is computed as

$$w_n = \sum_{k=(D+1)}^{(D+M)} x_{n-k}^2, \tag{8}$$

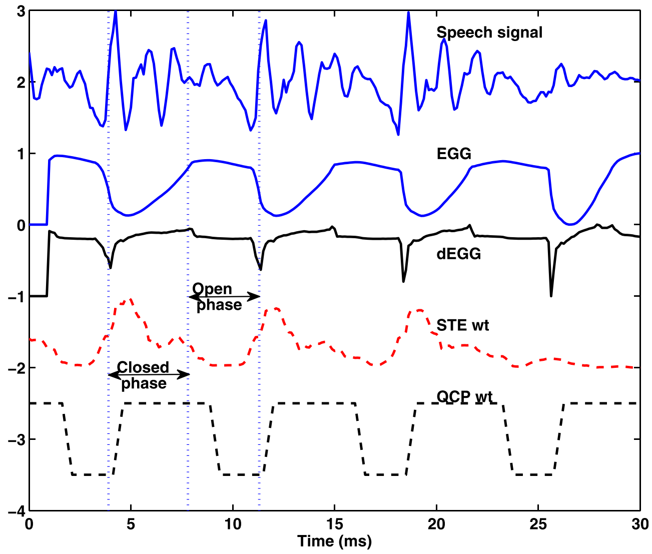

FIG. 1. (Color online) Speech signal, EGG, dEGG, STE, and QCP weighting function. The y-axis is relative and the signal offsets are only for clarity.

with $M = 12$ samples corresponding to 1.5 ms at 8 kHz sampling rate, and $D = 0$. The delay parameter D controls the peak position (or emphasis) of the weighting function within the glottal cycle. The length parameter M controls the peak or pulse width as well as the dynamic range and smoothness of the weighting function. It can be seen that the STE weighting function gives more weight to the high energy closed phase regions of the glottal cycle. However, it can be seen that the degree of suppression of the open phase and the main excitation depends on the signal decay within the glottal cycle and need not necessarily suppress these regions completely.

2. QCP weighting function

A detailed illustration of the QCP weighting function w_n along with the glottal flow derivative signal u_n for about one glottal cycle is shown in Fig. 2. The QCP weighting function is characterized by three parameters, namely, the position quotient ($Q_p = t_p/T_0$), duration quotient ($Q_d = t_d/T_0$), and the ramp duration t_r, where T_0 is the fundamental period. A small non-zero value, $d_w = 10^{-5}$, is used to avoid any possible singularities in the weighted ACOR matrices. It can be seen that the weighting function emphasizes the closed phase region of the glottal cycle, while at the same time de-emphasizes the region immediately after the main excitation as well as the open phase region.

The QCP weighting function provides two distinct advantages over the traditional LP or WLP methods. (1) Emphasis on the closed phase region provides for a more accurate modeling of the vocal tract by reducing the effect of coupling between subglottal and supraglottal cavities. (2) De-emphasizing the region immediately after the main excitation reduces the effect of glottal source on vocal tract modeling. 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 fundamental periods. This QCP analysis has been shown to yield more accurate estimates of the glottal source parameters compared to some of the popular glottal inverse methods (Airaksinen *et al.*, 2014). However, the performance of the QCP analysis in accurately modeling the vocal tract system parameters such as formants was not studied in Airaksinen *et al.* (2014).

III. QCP-FB ANALYSIS

In the traditional LP formulation, also referred to as forward prediction, the current sample is predicted based on the past p samples. At the same time, the current sample can also be predicted based on future p samples, referred to as backward prediction. It should be noted here that the forward and backward coefficients are inter-convertible, and hence do not carry any additional information when computed separately. However, a FB analysis combines both forward and backward predictions, where the current sample is predicted based on past as well as future samples using a common set of p coefficients. The combined error to be minimized is given by

$$\mathcal{E} = \mathcal{E}^f + \mathcal{E}^b, \tag{9}$$

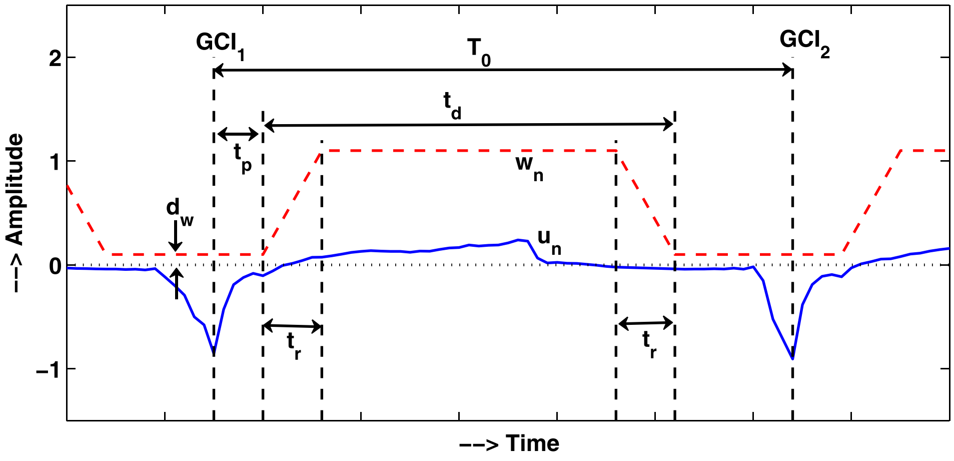

FIG. 2. (Color online) QCP weighting function w_n (dotted line) along with the glottal flow derivative (dEGG) signal u_n (solid line) for about one glottal cycle.

where

$$\mathcal{E}^f = \sum_n \left(x_n + \sum_{k=1}^{p} a_k x_{n-k} \right)^2 \quad (10)$$

and

$$\mathcal{E}^b = \sum_n \left(x_n + \sum_{k=1}^{p} a_k x_{n+k} \right)^2, \quad (11)$$

are the forward and backward errors, respectively. The predictor coefficients are computed by minimizing the combined error ($\partial \mathcal{E}/\partial a_i = 0,\ 1 \le i \le p$) and solving the resulting normal equations

$$\sum_{k=1}^{p} c_{i,k} a_k = -c_{i,0}, \quad 1 \le i \le p, \quad (12)$$

$$\text{where } c_{i,k} = \sum_n x_{n-i} x_{n-k} + \sum_n x_{n+i} x_{n+k}. \quad (13)$$

FB analysis is known to reduce substantially the dependence of traditional autoregressive spectral estimators on the initial sinusoidal phase (Chen and Stegen, 1974; Ulrych and Clayton, 1976), and the shifting of true frequency locations in the face of additive noise (Swingler, 1979). It is also known to reduce the line-splitting problem, where a single sinusoidal component appears as two distinct peaks in the estimated spectra, often encountered with the conventional ACOR or COV based LP analysis (Fougere *et al.*, 1976). The estimated spectral peak locations are therefore less sensitive to window positioning, and the use of both forward and backward predictions provides more samples to compute the correlations for a given window size. In view of this, the authors propose to use a QCP-FB analysis that combines the advantages of QCP and FB analyses for accurate formant detection and estimation.

QCP-FB analysis involves the use of FB analysis within the framework of WLP. The resulting FB-WLP imposes a temporal weighting function w_n on the forward and backward errors individually, and the combined error to be minimized is given by

$$\mathcal{F} = \mathcal{F}^f + \mathcal{F}^b, \quad (14)$$

where

$$\mathcal{F}^f = \sum_n w_n \left(x_n + \sum_{k=1}^{p} a_k x_{n-k} \right)^2 \quad (15)$$

and

$$\mathcal{F}^b = \sum_n w_n \left(x_n + \sum_{k=1}^{p} a_k x_{n+k} \right)^2 \quad (16)$$

are the weighted forward and backward errors, respectively. The resulting normal equations are given by

$$\sum_{k=1}^{p} d_{i,k} a_k = -d_{i,0}, \quad 1 \le i \le p, \quad (17)$$

$$\text{where } d_{i,k} = \sum_n w_n x_{n-i} x_{n-k} + \sum_n w_n x_{n+i} x_{n+k}. \quad (18)$$

Equations (17) and (18) form the main backbone for the formant detection experiments in the rest of the paper. An appropriate choice of range for the variable n results in ACOR or COV based FB-WLP (Makhoul, 1975; Kay, 1988). Also, note that there are multiple choices for the temporal weighting function that can be used for w_n in Eq. (18). The choice of QCP weighting function shown in Fig. 2 for w_n results in a special case of FB-WLP that will be referred to as QCP-FB analysis.

IV. FORMANT DETECTION EXPERIMENTS

The formant detection and estimation accuracies of the proposed QCP-FB method are evaluated using both synthetic as well as natural speech signals. Two different types of synthetic signals are considered, one generated using the Liljencrants-Fant (LF) source-filter model, and the other generated using a physical modeling approach of the speech production system. At this point, it would be good to discriminate the task of formant detection and estimation from formant tracking. In principle, most of the tracking algorithms can be applied on the initial estimates of formant locations derived using any underlying spectral representation. Therefore, in this section, the ability of different spectral representations in providing evidence for formant detection and estimation is evaluated, without the use of any tracking algorithm.

A. Experimental setup

Performance of different variants of the QCP method along with their LP and WLP counterparts is studied using different LP formulations, namely, ACOR, COV and forward-backward covariance (FBCOV) analyses. It is to be mentioned here that it is possible to have a FB ACOR formulation as well. However, only results for the FBCOV formulation are presented here as the COV formulations are traditionally known to perform better than ACOR formulations in formant estimation. In view of this, the proposed QCP-FB method will also be referred to as the QCP-FBCOV method in the remaining part of the paper.

A common framework is employed for detecting formants using each of the spectral representations, and for evaluating the performance of formant detection. All methods process the pre-emphasized speech signal [using a finite impulse response filter $P(z) = 1 - 0.97z^{-1}$] over 30 ms window segments and a frame rate of 100 frames per second. A Hamming window is used for ACOR analysis, whereas rectangular windowing is used for COV and FBCOV analyses. A position quotient of $Q_p = 0.05$, duration quotient of $Q_d = 0.7$, and a ramp duration of $t_r = 7$ samples is used for the QCP weighting function. The weighting function is derived based on the knowledge of GCIs detected using the SEDREAMS algorithm (Drugman *et al.*, 2012). A prediction order of $p = 13$ is used for all the methods at a sampling rate of 8 kHz, unless otherwise specified. The peaks in the spectrum are detected by convolving the spectrum with a Gaussian derivative window of width 100 Hz and picking the negative zero-crossings. The top five peaks are picked as the formant candidates. During performance evaluation, the reference ground truth for each of the first three formants is associated with the nearest formant candidate lying within a specified deviation.

B. Performance metrics

The performance of the methods is evaluated in terms of FDR and formant estimation error (FEE). 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 ith formant over N analysis frames is computed as

$$D_i = \frac{1}{N}\sum_{n=1}^{N} I(\Delta F_{i,n}), \tag{19}$$

$$I(\Delta F_{i,n}) = \begin{cases} 1 & \text{if } (\Delta F_{i,n}/F_{i,n} < \tau_p \text{ and } \Delta F_{i,n} < \tau_a) \\ 0 & \text{otherwise,} \end{cases} \tag{20}$$

where $I(\cdot)$ 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 ith formant at the nth frame from the reference ground truth $F_{i,n}$. The thresholds τ_p and τ_a denote the percentage deviation and absolute deviation, respectively.

On a linear frequency scale, using a single detection threshold, either a percentage threshold or an absolute threshold is problematic. The percentage deviation for higher formants 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 address this issue, two thresholds, one on percentage deviation and the other on absolute deviation, are used in order to define a common detection strategy for all formants. The percentage threshold is set to control the detection rates of lower formants, whereas the absolute threshold controls the detection rates of higher formants.

FEE is measured in terms of the average absolute deviation of the hypothesized formants from the ground truth. The FEE for the ith formant over N analysis frames is computed as

$$E_i = \frac{1}{N_i}\sum_{n=1}^{N} \Delta F_{i,n}. \tag{21}$$

Mean absolute deviation is chosen over the root-mean-square error measure so as to reduce the domineering effect of the outliers on the average score.

C. Experiments on LF model based synthetic vowels

1. Dataset

The effect of QCP-FB analysis on formant detection and estimation accuracy is studied using synthetic speech signals generated using an LF glottal source signal (Fant *et al.*, 1985) and an all-pole vocal tract filter (Makhoul, 1975). The synthetic signals are generated for six different vowels ([a], [i], [u], [e], [o], [ae]) for four different phonation types (creaky, modal, breathy, and whispered), and four different fundamental frequencies (80, 150, 250, and 350 Hz) (Gobl, 2003; Airaksinen *et al.*, 2014). The LF model is a parametric model for the glottal flow derivative waveform (Fant *et al.*, 1985), where t_e, t_p, t_a, and E_e constitute the LF parameters that define the waveform. The LF parameters can be represented in an alternate dimensionless form of E_e, R_a, R_g, and R_k with the advantage that they can be interpolated within their respective ranges obtained from Gobl (2003) to get a thorough range of possible excitations. The all-pole filter is generated using the typical first four formant frequency and bandwidth values for all the vowels (Gold and Rabiner, 1968). The LF source parameter values used to synthesize the vowels are given in Table I.

2. Results

Formant detection rates for the compared methods are given in Fig. 3(a). A stringent detection threshold (within $\tau_p = 10\%$ and $\tau_a = 100$ Hz deviation) is used for FDR computation owing to the availability of absolute ground truth. Two thresholds are used, one based on percentage to

TABLE I. The standard LF parameters used to synthesize the glottal flow derivative signals with different phonation types.

Phonation type	E_e	R_a	R_g	R_k
Modal	1	0.01	1.17	0.34
Breathy	$10^{(0.7/20)}$	0.025	0.88	0.41
Whisper	$10^{(-4.6/20)}$	0.07	0.94	0.32
Creaky	$10^{(-1.8/20)}$	0.008	1.13	0.2

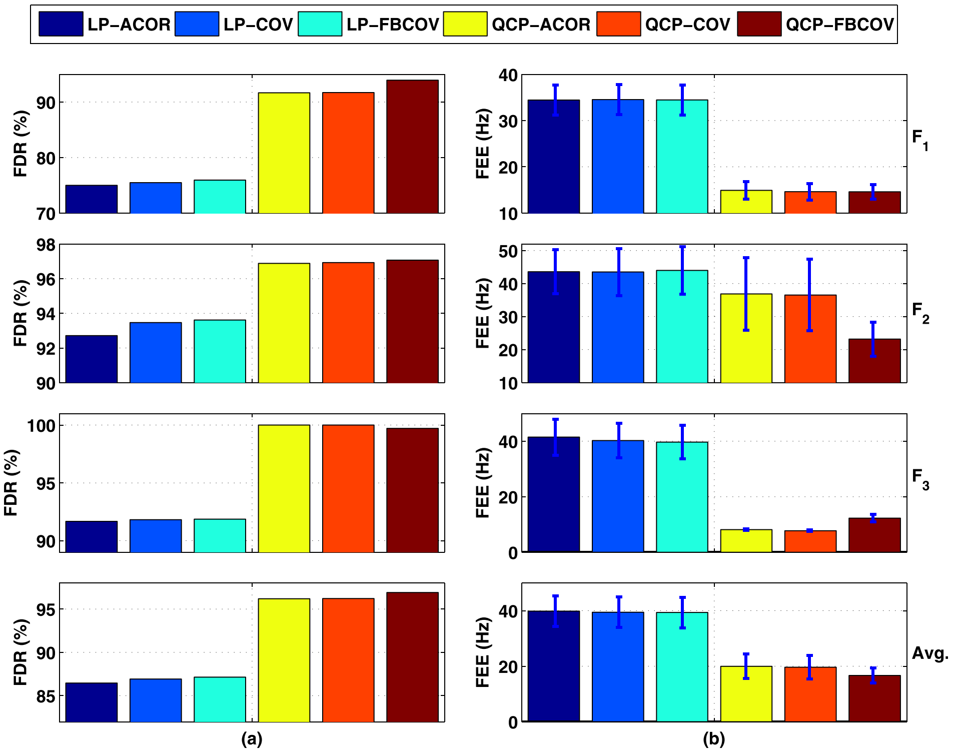

FIG. 3. (Color online) (a) FDR and (b) FEE for the first three formants (F_1, F_2, and F_3) and the average performance across all three formants (fourth row) on synthetic vowels with the LF glottal source averaged over all four phonation types and all four fundamental frequencies. In all subsequent figures, the error bars on the FEE scores denote the 95% confidence interval for the mean absolute errors, and the legend denotes the list of methods being compared in the order of plotting.

constrain lower formants and the other based on the absolute value to constrain higher formants. However, it is important to note that all formants have to satisfy both conditions to be considered as detected. It can be seen that all the QCP methods outperform the conventional LP based methods. The improvements in FDR are around 16–18 percentage points (pp) for the first formant, 3–4 pp for the second, and around 8 pp for the third.

Similarly, formant estimation accuracy of the methods measured in terms of FEE is shown in Fig. 3(b). There is a corresponding decrease in FEEs by around 20, 21, and 27 Hz for the first three formants, respectively, between the LP-FBCOV and QCP-FBCOV methods. It should be noted that all the QCP methods perform significantly better than their LP counterparts. Within the QCP family, the FBCOV method performs marginally better (overall when averaged over all three formants) compared to ACOR and COV formulations, in spite of a marginal decline in the already high F_3 detection and estimation performance. It should however be noted that the performance of all the QCP methods are already quite high on the synthetic data, and the data may not be challenging enough to demonstrate any decisive superiority of the FBCOV method.

The average FDRs and FEEs of the different methods for different phonation types and fundamental frequency are given in Figs. 4 and 5. The QCP variants perform consistently better than their LP counterparts by around 10 pp across all phonation types both in terms of detection as can be seen from Fig. 4(a). Similarly, the QCP methods perform better than their LP counterparts at high fundamental frequencies (250 and 350 Hz mean F_0) by around 18–20 pp, as can be seen from Fig. 4(b). This demonstrates the effectiveness of the QCP weighting function. However, the LP and QCP methods perform almost similar at low fundamental frequencies (80 and 150 Hz mean F_0).

The FEEs for the QCP methods are around 15–20 Hz lower than their LP counterparts across all phonation types as can be seen from Fig. 5(a). Similarly, the FEEs for the QCP methods are around 1–2, 10, 25, and 40 Hz lower than their LP counterparts for the four different mean F_0 values, respectively. In terms of FDR, the QCP-FBCOV method shows marginal improvement in performance compared to the ACOR or COV formulations across all phonation types and across all fundamental frequencies (Fig. 4). In terms of FEE, the QCP-FBCOV method when compared to their ACOR and COV counterparts has 2–10 Hz lesser error across different phonation types and fundamental frequencies, except for some degradation for creaky phonation (by around 12 Hz) and at 350 Hz F_0 (by around 2 Hz). However, the QCP-FBCOV method performs consistently and significantly better than the conventional LP methods across all phonation types and fundamental frequencies.

D. Experiments on physical model based synthetic vowels

One of the main issues with comparative experiments using synthetic data is the possibility of an inherent bias in the performance metrics toward any method with a modeling technique similar to that used for synthesizing the data. The experiments in Sec. IV C using LF source and all-pole vocal tract filter model are inherently biased toward any LP based method. However, the bias is not very serious considering the fact that the methods being compared are all based on LP. Nevertheless, one way to address this bias issue is to use data synthesized using a different modeling technique.

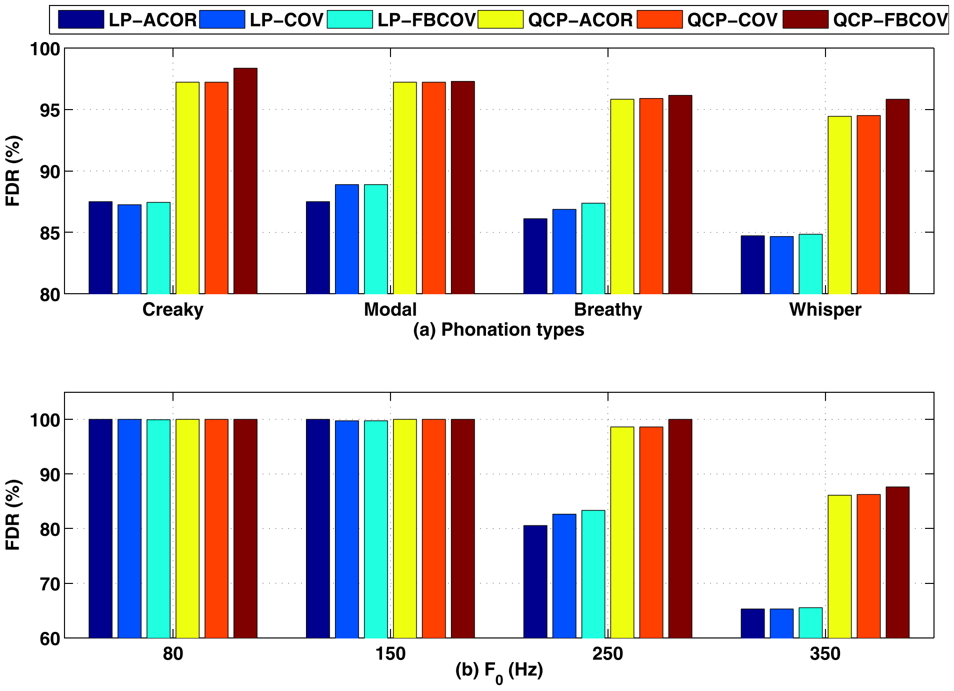

FIG. 4. (Color online) FDRs of different methods on synthetic LF vowels averaged over the first three formants for different (a) phonation types and (b) fundamental frequency (F_0).

1. Dataset

A computational model of the physical speech production system was used to generate synthetic vowel samples for this purpose. The data consists of four different vowel types ([a],[i],[ae],neutral vowel) synthesized at eight different fundamental frequencies (100 to 450 Hz in steps of 50 Hz) for three different representative speakers (an adult male, adult female, and a child aged approximately 5 yrs). The dataset consists a total of 96 ($4 \times 8 \times 3$) steady vowels of duration 0.4 s each at a sampling frequency of 44.1 kHz, and later downsampled to 10 kHz. More details of the dataset on the physical modeling of the vocal source and tract can be found in Alku *et al.* (2013).

2. Results

Performance of the different LP and QCP methods with ACOR, COV, and FBCOV formulations in formant detection and estimation is given in Table II. A prediction order of 12 with a pre-emphasis filter $[1 - 0.97z^{-1}]$ was used for all methods. The results in Table II are shown with and without (in parentheses) inclusion the 8 female vowel utterances [i] as almost all methods (except LP-ACOR) seem to have a problem detecting the third formant at 4909 Hz (with 10 kHz sampling rate) in these utterances. It is to be noted here that the third formant frequency used for the female vowel [i] by Alku *et al.* (2013) is a bit higher than the average $F3$ of 3372 Hz reported for adult female [i] by Hillenbrand *et al.*

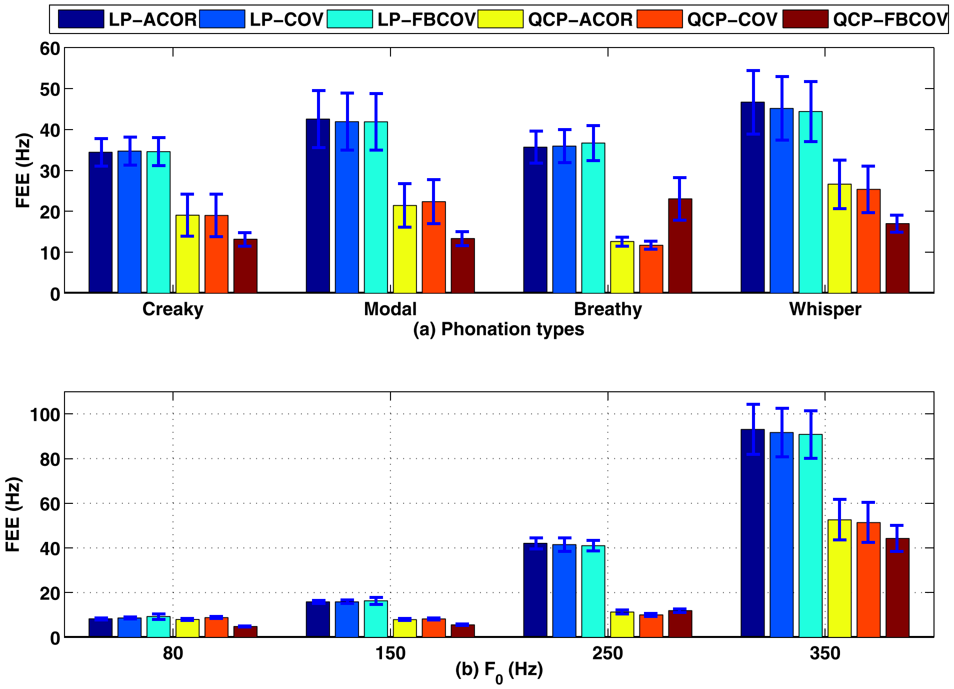

FIG. 5. (Color online) FEEs of different methods on synthetic LF vowels averaged over the first three formants for different (a) phonation types and (b) fundamental frequency (F_0).

TABLE II. Formant detection and estimation performance of different LP and QCP methods on synthetic data generated using a physical model. The numbers within the parentheses denote the performance excluding 8 female vowel ([i]) utterances with a difficult to detect third formant close to half the sampling frequency.

	FDR (%)			FEE (Hz)		
Method	F_1	F_2	F_3	δF_1	δF_2	δF_3
LP-ACOR	94.7 (94.3)	94.1 (93.6)	**93.4** (92.8)	76 (80)	102 (107)	**100** (104)
LP-COV	**95.0** (**94.5**)	94.1 (93.5)	85.5 (93.2)	74 (77)	96 (101)	219 (98)
LP-FBCOV	94.9 (94.4)	94.5 (94.0)	85.5 (93.2)	74 (77)	96 (101)	219 (98)
QCP-ACOR	91.3 (90.5)	91.9 (91.2)	89.2 (95.4)	72 (76)	82 (86)	173 (67)
QCP-COV	91.6 (91.4)	91.0 (90.2)	89.4 (95.8)	84 (74)	77 (82)	175 (63)
QCP-FBCOV	93.4 (92.8)	**94.9** (**94.5**)	91.2 (**95.5**)	**70** (**74**)	**73** (**76**)	143 (**66**)

(1995). However, with the exception of this third formant, QCP-FBCOV shows a consistent improvement over the other methods. The FDR scores evaluated at a threshold of $\tau_p = 30\%$ and $\tau_a = 300\,\text{Hz}$ deviation show that the QCP-FBCOV can improve upon the QCP-COV method even though the LP-ACOR method gives the best scores in some cases. The FEE scores for the different methods point toward a trend where an FBCOV analysis can in general improve upon the ACOR analysis. However, the scores for COV and FBCOV show a bit of a conflicting trend with LP-COV being better than LP-FBCOV and QCP-FBCOV being better than QCP-COV, though the differences are small.

E. Experiments on natural speech

1. Dataset

Performance of formant detection is evaluated on natural speech signals using the VTR-TIMIT database (Deng *et al.*, 2006). The test data of the VTR-TIMIT database which has 192 utterances, 8 utterances each from 24 different speakers (8 female and 16 male), are used for the evaluation. The first three reference formant frequencies provided in the database have been obtained in a semi-supervised manner, where the formant tracks derived using an LP based algorithm (Deng *et al.*, 2004) is verified and corrected manually based on spectrographic evidence. Performance on natural speech data is evaluated only using FDR, since the reference formant locations cannot be taken as absolute ground truth for FEE computation. All the speech data, originally recorded at 16 kHz sampling rate, is downsampled to 8 kHz before processing.

2. Results and discussions

Performance of the LP and QCP methods for a detection threshold of within $\tau_p = 30\%$ and $\tau_a = 300\,\text{Hz}$ deviation is given in Fig. 6. The FDRs are computed only for the regions of vowels, semivowels, and diphthongs. In general, it can be seen that the COV analysis performs better than ACOR, and FBCOV performs better than both ACOR and COV. Also, the QCP methods perform better than their LP counterparts, with the exception of QCP-ACOR and QCP-COV in detecting the first formant. This may be due to the availability of a less number of samples for prediction, mostly from the closed phase regions, in the case of QCP methods within a glottal cycle. However, it can be seen that the use of FBCOV analysis improves the detection performance of QCP-FBCOV method by 2–3 pp compared to the ACOR and COV methods.

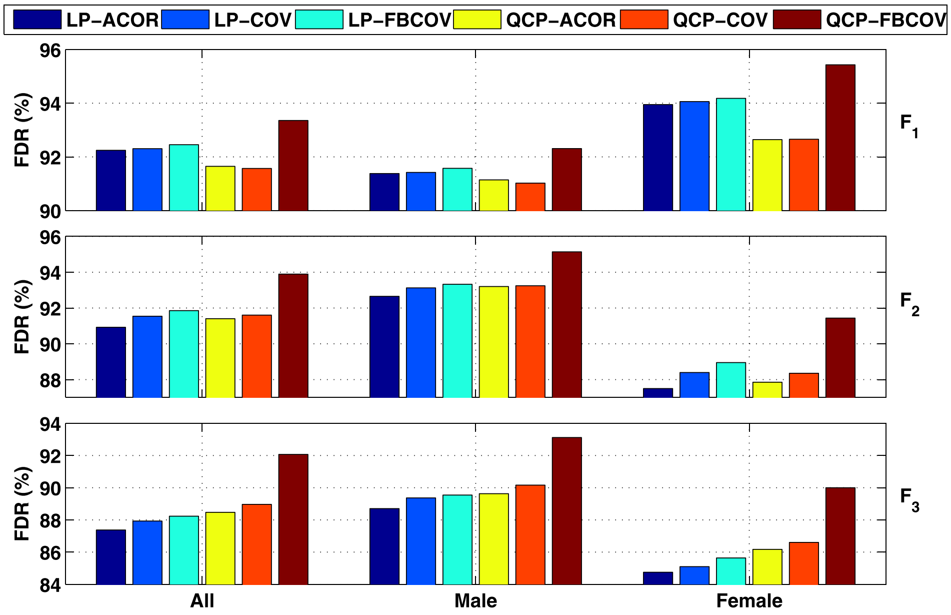

FIG. 6. (Color online) FDR for the first three formants (F_1, F_2, and F_3) on natural speech data.

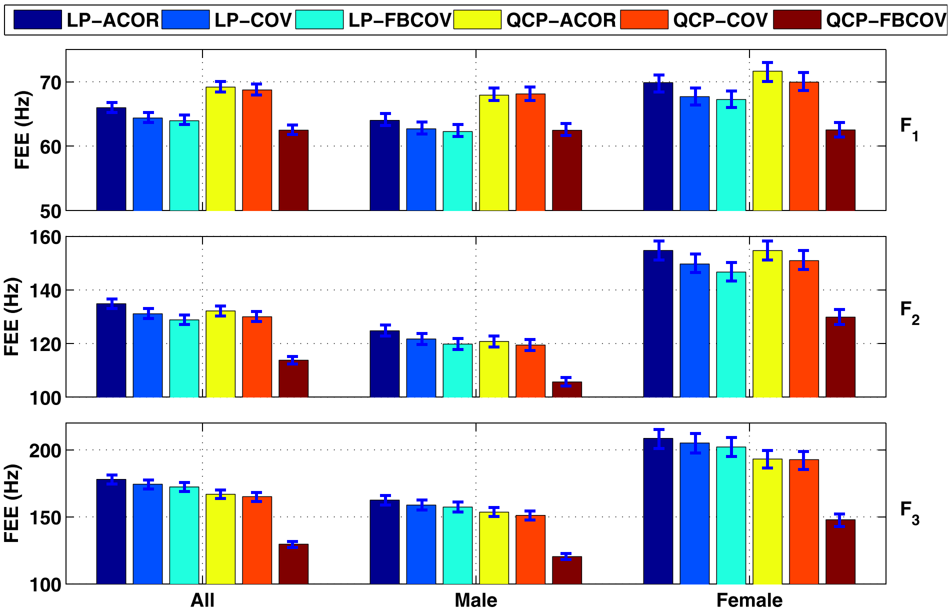

FIG. 7. (Color online) FEE for the first three formants (F_1, F_2, and F_3) on natural speech data.

A male–female gender analysis of the results shows a similar trend. All methods detect the first formant better in female voices than male, while the trend is opposite in the case of second and third formants. Also, the quantum of improvement in FDRs by QCP-FBCOV over QCP-COV for female voices (around 3–4 pp) is larger than that for male voices (around 1–2 pp). Similarly, the FEE scores in Fig. 7 shows that the overall estimation error for QCP-FBCOV is around 5 Hz (male 5 Hz, female 5 Hz), 16 Hz (male 12 Hz, female 20 Hz), and 35 Hz (male 30 Hz, female 40 Hz) lesser for the first three formants, respectively, as compared to the QCP-COV method.

Performance of different methods averaged over all three formants for three different phonetic classes, namely, vowels, diphthongs, and semivowels, is shown in Figs. 8 and 9. It can be seen from Fig. 8 that the QCP-FBCOV method consistently yields around 2%–4% improvement in detection rates over other methods across all three phonetic classes. The QCP-FBCOV reduces the estimation errors by 5–8 Hz for the first formant, 10–25 Hz for the second, and 20–40 Hz for the third, across the three phonetic classes. In general, the improvements in FEE for the dynamic semivowel class are larger compared to that for the steady vowel or diphthong classes.

One important question that arises when evaluating a new method is the significance of the improvements achieved. The answer to this question mainly depends on the application in which the proposed method would be used, e.g., auditory perception of vowels, speaker identification, automatic speech recognition, neurological speech disorder assessment. This improved performance of QCP-FBCOV in FEE might not be considered very meaningful from the point of view of human auditory perception for which the difference limen for $F1$ and $F2$ is known to be on the order of 3%–5% of the formant

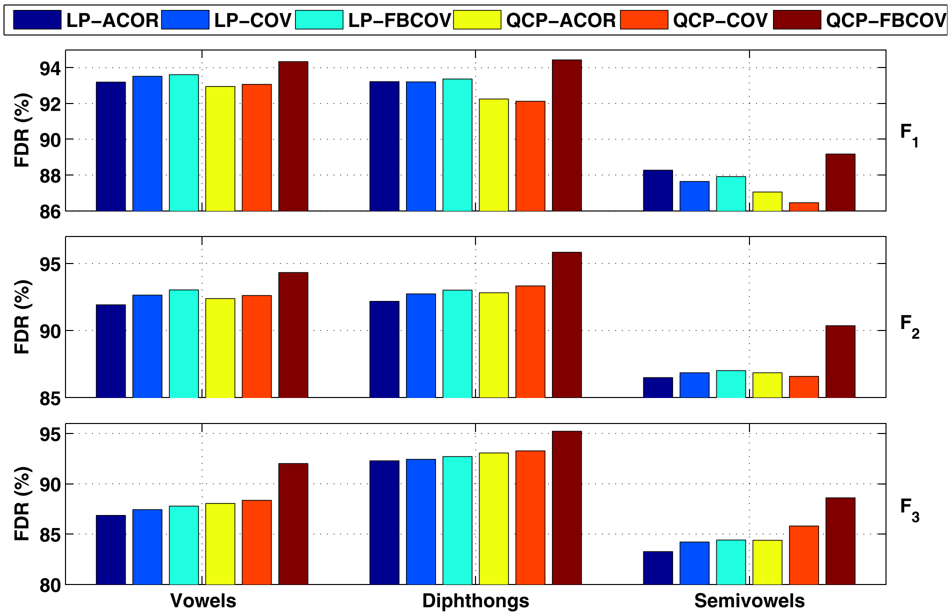

FIG. 8. (Color online) Average FDRs (in %) of different methods for different phonetic classes vowels, diphthongs, and semivowels.

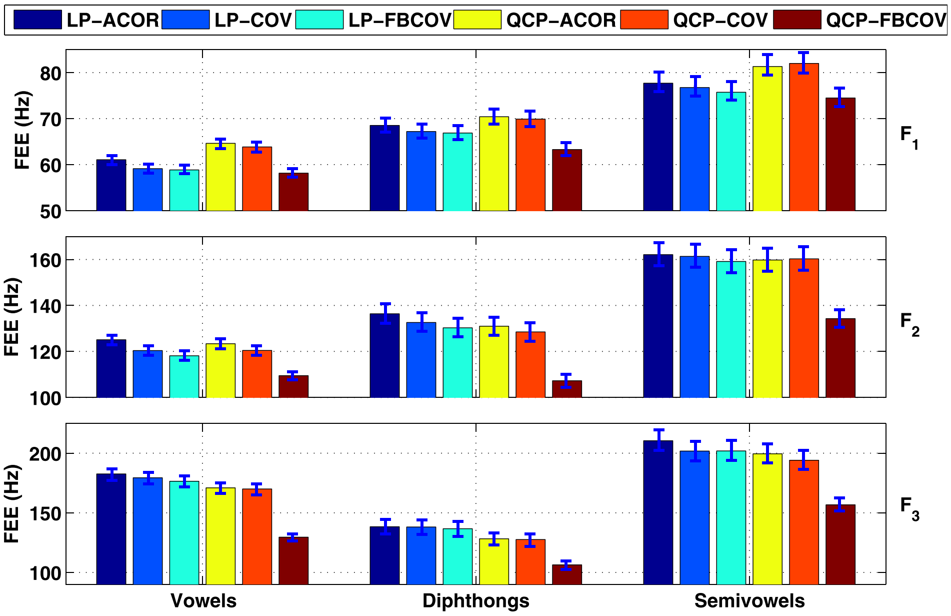

FIG. 9. (Color online) Average FEE (in Hz) of different methods for different phonetic classes vowels, diphthongs, and semivowels.

frequency (Flanagan, 1972). In acoustic-phonetic speech analysis, however, the lesser estimation error achieved by QCP-FBCOV can be considered meaningful. Nevertheless, a statistical significance test of these improvements using repeated measures analysis of variance (RM-ANOVA) is presented in Sec. IV E 3.

3. Repeated measures ANOVA

A RM-ANOVA followed by a *post hoc* Newman-Keuls pair-wise *t*-tests was performed comparing the means of absolute errors for the methods QCP-ACOR, QCP-COV, and QCP-FBCOV. Significant differences ($p < 0.001$) were observed in the means of the three methods for all three formants (F_1: [$F(2,\ 53\,796) = 427.55$; $p < 0.001$; $\eta_p^2 = 0.02$], F_2: [$F(2,\ 53\,796) = 568.05$; $p < 0.001$; $\eta_p^2 = 0.02$], and F_3: [$F(2,\ 53\,796) = 545.01$; $p < 0.001$; $\eta_p^2 = 0.02$]). A pair-wise comparison of the methods showed that the QCP-FBCOV method differed significantly ($p < 0.001$) from each of the other two methods for all three formants. However, a comparison of QCP-ACOR and QCP-COV methods showed a reduced significance ($p < 0.05$) for the first formant, a high significance ($p < 0.001$) for the second formant, and no significance ($p > 0.05$) for the third formant. Error bars denoting 95% confidence intervals for the mean absolute errors are shown in Figs. 7 and 9.

A similar comparison of the methods QCP-ACOR, QCP-COV, and QCP-FBCOV was performed on LF synthetic vowels and physical model based synthetic vowels. In the case of LF synthetic data, the means of all three methods showed significant differences ($p < 0.05$) for all formants, except for the first formant between QCP-ACOR and QCP-FBCOV as well as the second formant between QCP-ACOR and QCP-COV. Similarly, in the case of physical models data, all the means showed significant difference ($p < 0.05$) for all formants, except for the third formant between QCP-ACOR and QCP-COV.

4. Comparison with other WLP methods

A comparison of performance of the proposed QCP-FBCOV method within the family of popular WLP methods

TABLE III. FDRs (in %) of different spectral representations for the first three formants (F_1, F_2, and F_3) on natural speech data.

Method (→)	LP			WLP			XLP			QCP		
Analysis type (↓)	F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3
FDR within 20% and 200 Hz dev												
ACOR	**83.2**	80.5	78.9	81.5	79.7	77.7	83.1	79.9	76.9	81.1	**81.0**	**79.9**
COV	84.1	81.4	79.8	82.7	81.1	79.4	**84.2**	81.1	78.6	81.6	**81.8**	**80.7**
FBCOV	84.2	81.9	80.1	81.8	79.9	74.5	84.0	80.8	77.7	**84.9**	**85.0**	**83.9**
FDR within 25% and 250 Hz dev												
ACOR	88.9	87.0	84.1	88.0	86.6	83.8	**89.1**	86.4	82.4	88.0	**87.4**	**85.3**
COV	89.3	87.7	85.0	88.6	87.5	85.0	**89.7**	87.2	83.7	88.1	**87.8**	**86.0**
FBCOV	89.4	88.2	85.2	88.8	86.5	80.9	89.7	86.9	82.9	**90.4**	**90.5**	**89.1**
FDR within 30% and 300 Hz dev												
ACOR	92.2	90.9	87.4	91.5	90.9	87.4	**92.5**	90.6	86.0	91.6	**91.4**	**88.5**
COV	92.3	91.5	87.9	91.8	91.5	88.2	**92.7**	91.1	86.9	91.6	**91.6**	**89.0**
FBCOV	92.4	91.9	88.2	92.2	90.5	84.3	92.7	90.8	86.1	**93.4**	**93.9**	**92.1**

in formant detection on natural speech from the VTR-TIMIT database is given in Table III. It should be noted here that the conventional LP analysis can also be considered as a special case of WLP methods with equal temporal weight on the prediction of each sample. The other two methods compared are the conventional WLP using the STE weighting function (Ma *et al.*, 1993) and the extended linear prediction (XLP) using a more generalized weighting function that allows different weights at different lags (Pohjalainen *et al.*, 2010). The WLP and XLP methods have been shown to provide more robust spectral representations compared to conventional LP under degradations and vocal effort mismatch.

The performance in Table III is provided for three different decision thresholds, and using all three formulations, namely, ACOR, COV, and FBCOV, for each of the WLP methods. It can be seen that the QCP-ACOR or QCP-COV methods perform poorer than the conventional LP and XLP methods in detecting the first formant. However the QCP-FBCOV method performs better in detecting all three formants compared among all WLP methods, and against all analysis types. Results show that the proposed QCP-FBCOV method performs better than the widely used LP-COV with an improvement of ∼2.7 pp averaged across all formants and thresholds. The improvements are much higher in the case of second (∼3 pp) and third formants (∼4.1 pp) compared to the first formant (∼1 pp). Also, QCP-FBCOV improves the detection rate of first formant by ∼2.7 pp compared to the QCP-ACOR method. It can also be seen from the results in Table III that the COV formulation performs better than ACOR for all LP methods, and FBCOV improves upon COV for LP and QCP methods. However, the quantum of improvement by FBCOV over COV (or ACOR) is much higher for the QCP method, demonstrating the effectiveness of FB analysis in addressing the data-insufficiency problem of the QCP analysis.

V. CONCLUSIONS

In this paper, a modified QCP analysis of speech signals for accurate formant detection and estimation was proposed that combines several advantages of WLP in the form of QCP analysis and FB analysis. QCP analysis exploits the WLP framework of sample selective prediction by designing a weighting function that gives more emphasis on closed phase regions and de-emphasizes the open phase as well as the region immediately after the main excitation. The result is a more accurate closed phase estimate of the vocal tract system with a reduced influence from the glottal source. A FBCOV analysis within the framework of WLP was utilized for the first time. The FB analysis helps improve the FDRs by providing more samples for prediction and by reducing the problems of window positioning and line splitting.

Results from the formant detection experiments on natural speech data show that the proposed QCP-FBCOV method performs significantly better than the conventional LP, WLP, and QCP methods. QCP-FBCOV gives a FDR 2–3 pp better than the QCP-COV method, and a reduction of average estimation error in the range of 6–35 Hz for the three formants. QCP-FBCOV performs 1–4 pp better than the LP-COV method in formant detection with a reduction of average estimation error in the range of 1–44 Hz. The quantum of improvement is higher for female voices as compared to male voices underlining the significance of the method for tracking formants from high-pitched voices.

However, it should be noted that the performance of the proposed method is dependent on the accuracies of the estimated GCIs. The robustness of the proposed method to inaccuracies in GCI estimation in the face of degradations still needs to be studied. Nevertheless, under clean conditions the QCP-FBCOV method is clearly a better choice over conventional LP based methods for formant detection, even more so for high-pitched female voices.

ACKNOWLEDGMENTS

This work has been funded by the Academy of Finland (Project Nos. 256961 and 284671). We would like to thank Dr. Santeri Yrttiaho for his assistance in designing and running the repeated measures ANOVA experiments.

Airaksinen, M., Raitio, T., Story, B., and Alku, P. (**2014**). “Quasi closed phase glottal inverse filtering analysis with weighted linear prediction,” IEEE/ACM Trans. Audio, Speech, Lang. Process. **22**, 596–607.

Alku, P., Pohjalainen, J., Vainio, M., Laukkanen, A.-M., and Story, B. (**2012**). “Improved formant frequency estimation from high-pitched vowels by downgrading the contribution of the glottal source with weighted linear prediction,” in *Proceedings of Interspeech*, Portland, Oregon, pp. 1610–1613.

Alku, P., Pohjalainen, J., Vainio, M., Laukkanen, A.-M., and Story, B. H. (**2013**). “Formant frequency estimation of high-pitched vowels using weighted linear prediction,” J. Acoust. Soc. Am. **134**, 1295–1313.

Assmann, P. F. (**1995**). “The role of formant transitions in the perception of concurrent vowels,” J. Acoust. Soc. Am. **97**, 575–584.

Atal, B. S., and Schroeder, M. R. (**1967**). “Predictive coding of speech signals,” in *Proceedings of the 1967 Conference Communication and Processing*, Cambridge, MA, pp. 360–361.

Boersma, P. (**2001**). “Praat, a system for doing phonetics by computer,” Glot Int. **5**, 341–345.

Bruce, I. C. (**2004**). “Physiological assessment of contrast-enhancing frequency shaping and multiband compression in hearing aids,” Physiol. Measure. **25**, 945–956.

Chan, P. Y., Dong, M., Lim, Y. Q., Toh, A., Chong, E., Yeo, M., Chua, M., and Li, H. (**2015**). “Formant excursion in singing synthesis,” in *Proceedings of the 2015 IEEE International Conference on Digital Signal Processing* (*DSP*), Singapore, pp. 168–172.

Chen, W. Y., and Stegen, G. R. (**1974**). “Experiments with maximum entropy power spectra of sinusoids,” J. Geophys. Res. **79**, 3019–3022, doi:10.1029/JB079i020p03019.

Deng, L., Cui, X., Pruvenok, R., Huang, J., and Momen, S. (**2006**). “A database of vocal tract resonance trajectories for research in speech processing,” in *Proceedings of the International Conference on Acoustics Speech and Signal Processing* (*ICASSP*), Toulouse, France, pp. I369–I372.

Deng, L., Lee, L., Attias, H., and Acero, A. (**2004**). “A structured speech model with continuous hidden dynamics and prediction-residual training for tracking vocal tract resonances,” in *Proceedings of the International Conference on Acoustics Speech and Signal Processing* (*ICASSP*), Vol. 1, Montreal, Quebec, Canada, pp. I-557–I-560.

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

Drugman, T., Thomas, M., Gudnason, J., Naylor, P., and Dutoit, T. (**2012**). “Detection of glottal closure instants from speech signals: A quantitative review,” IEEE Trans. Audio, Speech, Lang. Process. **20**, 994–1006.

Fant, G. (**1960**). *Acoustic Theory of Speech Production* (Mouton & Co., The Hague, Netherlands), pp. 1–328.

Fant, G., Liljencrants, J., and Lin, Q. G. (**1985**). “A four-parameter model of glottal flow,” Q. Prog. Stat. Rep. **4**, 1–17.

Flanagan, J. L. (**1972**). *Speech Analysis, Synthesis and Perception* (Springer-Verlag, New York), pp. 279–280.

Fougere, P. F., Zawalick, E. J., and Radoski, H. R. (**1976**). "Spontaneous line splitting in maximum entropy power spectrum analysis," Phys. Earth Planetary Int. **12**, 201–207.

Gobl, C. (**2003**). "The voice source in speech communication—production and perception experiments involving inverse filtering and synthesis," Ph.D. thesis, Stockholm, Sweden.

Gold, B., and Rabiner, L. (**1968**). "Analysis of digital and analog formant synthesizers," IEEE Trans. Audio Electroacoust. **16**, 81–94.

Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (**1995**). "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. **97**, 3099–3111.

Itakura, F., and Saito, S. (**1968**). "Analysis synthesis telephony based upon the maximum likelihood method," in *Proceedings of the 6th International Congress on Acoustics*, edited by Y. Kohasi, Tokyo, Japan, pp. C5–C5, C17–C20.

Kay, S. M. (**1988**). *Modern Spectral Estimation: Theory & Application* (Prentice Hall, Englewood Cliffs, NJ), pp. 1–543.

Lee, C.-H. (**1988**). "On robust linear prediction of speech," IEEE Trans. Acoust. Speech Signal Process. **36**, 642–650.

Ma, C., Kamp, Y., and Willems, L. F. (**1993**). "Robust signal selection for linear prediction analysis of voiced speech," Speech Commun. **12**, 69–81.

Magi, C., Pohjalainen, J., Bäckström, T., and Alku, P. (**2009**). "Stabilized weighted linear prediction," Speech Commun. **51**, 401–411.

Makhoul, J. (**1975**). "Linear prediction: A tutorial review," Proc. IEEE **63**, 561–580.

Mehta, D. D., Rudoy, D., and Wolfe, P. J. (**2012**). "Kalman-based autoregressive moving average modeling and inference for formant and antiformant tracking," J. Acoust. Soc. Am. **132**, 1732–1746.

Mizoguchi, R., Yanagida, M., and Kakusho, O. (**1982**). "Speech analysis by selective linear prediction in the time domain," in *Proceedings of the International Conference on Acoustics Speech and Signal Processing* (*ICASSP*), Vol. 7, Paris, France, pp. 1573–1576.

Pinto, N. B., Childers, D. G., and Lalwani, A. L. (**1989**). "Formant speech synthesis: Improving production quality," IEEE Trans. Acoust. Speech Signal Process **37**, 1870–1887.

Pohjalainen, J., Saeidi, R., Kinnunen, T., and Alku, P. (**2010**). "Extended weighted linear prediction (XLP) analysis of speech and its application to speaker verification in adverse conditions," in *Proceedings Interspeech*, Makuhari, Japan, pp. 1477–1480.

Schilling, J. R., Miller, R. L., Sachs, M. B., and Young, E. D. (**1998**). "Frequency-shaped amplification changes the neural representation of speech with noise-induced hearing loss," Hear. Res. **117**, 57–70.

Singh, R., Gencaga, D., and Raj, B. (**2016**). "Formant manipulations in voice disguise by mimicry," in *Proceedings of the 4th International Conference on Biometrics and Forensics* (*IWBF*), Limassol, Cyprus, pp. 1–6.

Sjolander, K., and Beskow, J. (**2000**). "Wavesurfer—An open source speech tool," in *Proceedings of the International Conference on Spoken Language Processing*, Beijing, China, pp. 464–467.

Smit, T., Trckheim, F., and Mores, R. (**2012**). "Fast and robust formant detection from LP data," Speech Commun. **54**, 893–902.

Steiglitz, K., and Dickinson, B. (**1977**). "The use of time-domain selection for improved linear prediction," IEEE Trans. Acoust. Speech Signal Process. **25**, 34–39.

Swingler, D. N. (**1979**). "A comparison between Burg's maximum entropy method and a nonrecursive technique for the spectral analysis of deterministic signals," J. Geophys. Res. **84**, 679–685, doi:10.1029/JB084iB02p00679.

Ulrych, T. J., and Clayton, R. W. (**1976**). "Time series modeling and maximum entropy," Phys. Earth Planetary Int. **12**, 188–200.

Welling, L., and Ney, H. (**1998**). "Formant estimation for speech recognition," IEEE Trans. Speech Audio Process. **6**, 36–48.

Wong, D., Markel, J., and Gray, A. (**1979**). "Least squares glottal inverse filtering from the acoustic speech waveform," IEEE Trans. Acoust. Speech Signal Process. **27**, 350–355.

Yanagida, M., and Kakusho, O. (**1985**). "A weighted linear prediction analysis of speech signals by using the Givens reduction," in *Proceedings of the International Symposium on Applied Signal Processing and Digital Filtering* (*IASTED*), Paris, France, pp. 129–132.

Yegnanarayana, B., and Veldhuis, R. (**1998**). "Extraction of vocal-tract system characteristics from speech signals," IEEE Trans. Speech Audio Process. **6**, 313–327.

Yoo, I. C., Lim, H., and Yook, D. (**2015**). "Formant-based robust voice activity detection," IEEE/ACM Trans. Audio, Speech, Lang. Process. **23**, 2238–2245.