IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 22, 2021, accepted October 22, 2021, date of publication November 8, 2021,
date of current version November 16, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3126280

Formant Tracking Using Quasi-Closed Phase Forward-Backward Linear Prediction Analysis and Deep Neural Networks

DHANANJAYA N. GOWDA[1], (Member, IEEE), BAJIBABU BOLLEPALLI[2], (Member, IEEE), SUDARSANA REDDY KADIRI[3], (Member, IEEE), AND PAAVO ALKU[3], (Fellow, IEEE)

[1]Samsung Research, Seoul 06765, South Korea
[2]Amazon, Slough SL3 7RS, U.K.
[3]Department of Signal Processing and Acoustics, Aalto University, 00076 Espoo, Finland

Corresponding author: Sudarsana Reddy Kadiri (sudarsana.kadiri@aalto.fi)

This work was supported in part by the Academy of Finland under Project 313390.

Dhananjaya Gowda, Bajibabu Bollepalli, and Sudarsana Reddy Kadiri contributed equally to this work.

ABSTRACT Formant tracking is investigated in this study by using trackers based on dynamic programming (DP) and deep neural nets (DNNs). Using the DP approach, six formant estimation methods were first compared. The six methods include linear prediction (LP) algorithms, weighted LP algorithms and the recently developed quasi-closed phase forward-backward (QCP-FB) method. QCP-FB gave the best performance in the comparison. Therefore, a novel formant tracking approach, which combines benefits of deep learning and signal processing based on QCP-FB, was proposed. In this approach, the formants predicted by a DNN-based tracker from a speech frame are refined using the peaks of the all-pole spectrum computed by QCP-FB from the same frame. Results show that the proposed DNN-based tracker performed better both in detection rate and estimation error for the lowest three formants compared to reference formant trackers. Compared to the popular Wavesurfer, for example, the proposed tracker gave a reduction of 29%, 48%, and 35% in the estimation error for the lowest three formants, respectively.

INDEX TERMS Speech analysis, formant tracking, linear prediction, dynamic programming, deep neural net.

I. INTRODUCTION

Estimation and tracking of formant frequencies is an important research topic in several areas of speech science and technology [1]–[6]. During the past few decades, many techniques have been proposed for formant tracking [7]–[10]. These algorithms typically consist of two parts, the estimation stage and the tracking stage. In the former, initial estimates of the vocal tract resonances (VTRs) are computed in short frames (e.g., 25 ms) using spectral estimation methods such as linear prediction (LP). In the latter, the formants estimated from individual frames are expressed using contours which cover a longer unit (e.g., word or sentence) [7], [8]. In addition, estimation and tracking can be done simultaneously using an initial representation of the vocal tract system [9], [10]. In both approaches, accurate estimation of VTRs is an important and necessary computational block.

LP is the most widely used technique to estimate VTRs from speech [11] and therefore many variants of LP have been proposed (e.g. [12], [13]). In formant estimation and tracking, the most popular variants are the autocorrelation and covariance methods [7], [8]. The closed phase (CP) analysis is known to improve VTR estimates by avoiding the contribution of the speech samples in the open phase of the glottal cycle thereby decoupling the effect of the trachea more effectively [14]. CP analysis, however, works better for low-pitched voices which typically have a larger number of samples in the closed phase of the glottal cycle compared to high-pitched voices which might have just a few samples in the closed phase. To reduce problems caused by having a small number of closed phase samples, LP can be computed over multiple neighboring cycles [14].

Weighted linear prediction (WLP) is an all-pole modeling method based on temporally weighting the prediction error [13], [15]–[20]. Temporal weighting of the prediction error has been shown to be beneficial in computing vocal

The associate editor coordinating the review of this manuscript and approving it for publication was Shaikh Anowarul Fattah.

tract models which are robust with respect to noise and the selection of analysis window as well as the biasing effect of high fundamental frequency. Formant estimation of high-pitched vowels was studied using WLP in [18] by developing a simple weighting function, called the attenuated main excitation (AME) function, to downgrade the strong effect of the glottal source in the computation of the vocal tract model. Based on [18], the quasi-closed phase (QCP) method was proposed for glottal inverse filtering (GIF) in [19]. In QCP, a more generalized AME-type of weighting function is used. Recently, a new formant estimation method based on QCP, called quasi-closed phase forward-backward (QCP-FB) LP analysis, was proposed in [21]. QCP-FB combines two approaches: (1) QCP analysis in which the residual is temporally weighted, and (2) forward-backward (FB) analysis in which the number of samples is increased in LP by using two prediction directions simultaneously. In addition, WLP methods have been proposed recently based on using stochastic approaches in the computation of the weighting function [22].

In this article, formant tracking is studied by investigating different all-pole modeling methods in formant estimation. The all-pole formant estimation methods are used with two formant tracking approaches, a dynamic programming (DP) -based approach and a deep neural net (DNN) -based approach. As the first part of the study, six different LP-based and WLP-based formant estimation methods are compared in formant tracking using a DP-based tracker. The novelty of this part is in studying how the potential new method, QCP-FB, which was investigated solely in formant *estimation* in [21], works in formant *tracking*. In the second part of the study, two most potential all-pole modeling methods from the first part are used with a modern DNN-based tracker by proposing a novel formant tracking approach, which combines benefits of the data-driven deep learning approach and benefits of the model-driven all-pole modeling approach. In this novel tracking approach, the formants, which are predicted by the DNN from a given speech frame, are refined using the spectral peaks, which are indicated by the spectrum, which is computed from the same frame with a model-based parametric all-pole spectral estimation method. Altogether five known formant trackers (Wavesurfer [8], PRAAT [7], MUST [23], KARMA [10], and Deep Formants [24]) are used as reference methods in this study.

The contributions of the study are as follows:

- The potential new formant estimation method, QCP-FB, is evaluated in formant *tracking* and its performance is compared with existing LP-based and WLP-based formant estimation methods using a DP-based tracker.
- A novel formant tracking technique is proposed by combining the data-driven DNN-based approach and the model-driven all-pole approach. In this technique, the formants predicted from a speech frame by a DNN are refined using the spectral peaks that are extracted from an all-pole model, which is computed from the same frame.
- A systematic investigation is carried out by comparing the novel formant tracking method described above with five reference formant trackers (Wavesurfer [8], PRAAT [7], MUST [23], KARMA [10], and Deep Formants [24]).

The paper is organised as follows. The QCP-FB method, which was introduced as a new formant estimation method recently in [21], is first described in section II. The other formant estimation methods and the formant trackers used in the study are described in section III. The results of the formant tracking experiments are reported in section IV. Finally, conclusions are drawn in section V.

II. QUASI-CLOSED PHASE FORWARD-BACKWARD ANALYSIS

The traditional formulation of LP is based on forward prediction in which the current speech sample is predicted from the past p samples. It is, however, also possible to use backward prediction in which the current sample is predicted from the future p samples. The filter coefficients computed using forward and backward predictions are inter-convertible, and therefore they do not carry any additional information when computed separately. However, by simultaneously using both backward and forward prediction, a prediction model different from that of traditional LP is obtained by using forward-backward (FB) analysis, where the current sample is predicted based on past and future samples using a common set of p coefficients. The combined error to be minimized is given by

$$\mathcal{E} = \mathcal{E}^f + \mathcal{E}^b, \tag{1}$$

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

$$\text{and } \mathcal{E}^b = \sum_n \left(x_n + \sum_{k=1}^{p} a_k x_{n+k} \right)^2 \tag{3}$$

denote the forward and backward errors, respectively, x_n denotes the current speech sample, and a_k denotes the prediction coefficients. The prediction coefficients can be computed by minimizing the combined error ($\partial\mathcal{E}/\partial a_i = 0, \quad 1 \leq i \leq p$) which results in the following normal equations

$$\sum_{k=1}^{p} c_{i,k} a_k = -c_{i,0}, \quad 1 \leq i \leq p \tag{4}$$

$$\text{where } c_{i,k} = \sum_n x_{n-i} x_{n-k} + \sum_n x_{n+i} x_{n+k}. \tag{5}$$

Previous studies have shown that FB analysis reduces the dependency of spectral estimates on the initial sinusoidal phase, shifting of frequency estimates due to additive noise and the so called line-splitting problem (see [21] for a review). The line-splitting problem refers to obtaining spectral models which show a single sinusoidal component incorrectly as two distinct peaks. By taking advantage of FB analysis, two benefits are achieved: (1) the estimated spectral peak locations are less sensitive to the window position, and (2) the combination of the two prediction directions gives more samples to compute correlations for the given frame.

Quasi-closed phase forward-backward (QCP-FB) analysis involves the use of FB analysis within the framework of QCP in order to combine the benefits of both techniques. The resulting method imposes the temporal QCP weighting function w_n, defined by [19], on the forward and backward errors individually. The combined error to be minimized is given by

$$\mathcal{F} = \mathcal{F}^f + \mathcal{F}^b, \tag{6}$$

$$\text{where } \mathcal{F}^f = \sum_n w_n \left(x_n + \sum_{k=1}^{p} a_k x_{n-k} \right)^2 \tag{7}$$

$$\text{and } \mathcal{F}^b = \sum_n w_n \left(x_n + \sum_{k=1}^{p} a_k x_{n+k} \right)^2 \tag{8}$$

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

$$\sum_{k=1}^{p} d_{i,k} a_k = -d_{i,0}, \quad 1 \le i \le p \tag{9}$$

$$\text{where } \; d_{i,k} = \sum_n w_n x_{n-i} x_{n-k} + \sum_n w_n x_{n+i} x_{n+k}. \tag{10}$$

Appropriate choice of range for the variable n results in the autocorrelation or covariance methods for QCP-FB.

QCP-FB is used in formant tracking in the current study and it is expected to show improved performance compared to existing formant tracking methods due to the following two main reasons. First, FB analysis helps to improve the formant estimation by providing more samples for prediction, and by reducing the problems of window positioning and line splitting. Second, QCP analysis exploits the WLP framework of sample selective prediction by designing a temporal weighting function that gives more emphasis on closed phase regions and deemphasizes the open phase as well as the region immediately after the main excitation. This results in more accurate closed phase estimates of the vocal tract system with a reduced influence from the glottal source.

III. FORMANT TRACKERS

Several formant tracking algorithms have been proposed in the literature [7]–[10], [24]. It is worth emphasising that a formant tracking algorithm will most likely show varying performance when combined with different formant estimation methods and this makes it difficult to compare different tracking algorithms. In principle, most of the tracking algorithms can be combined with any formant estimation method. Therefore, formant tracking is studied in this paper using trackers which are based on both DP and DNN.

A. DP-BASED FORMANT TRACKERS

Using the DP-based tracking algorithm proposed in [8], formant tracking performance was investigated by comparing six different formant estimation methods that all use all-pole modeling. These methods, listed in Table 1, are as follows: (1) conventional LP based on the autocorrelation method (LP-ACOR), (2) conventional LP based on the covariance method (LP-COV), (3) LP based on forward-backward prediction and the covariance method (LP-FBCOV), (4) QCP analysis based on the autocorrelation method (QCP-ACOR), (5) QCP analysis based on the covariance method (QCP-COV) and (6) QCP analysis based on forward-backward prediction and the covariance method (QCP-FBCOV). All these methods were computed using a frame length of 25 ms, a frame shift of 10 ms and an all-pole model order $p = 12$. Speech signals, sampled using 8 kHz, were pre-emphasised using an FIR filter ($P(z) = 1 - 0.5z^{-1}$). In the autocorrelation methods, the Hamming window was used. In the covariance methods, the rectangular window was used. The peaks in the spectrum were detected by convolving the spectrum using a Gaussian derivative window with a width of 100 Hz and picking the negative zero-crossings. Five most energetic peaks of the spectrum were selected as the formant candidates. A verbatim MATLAB implementation of the tracking algorithm [8] was used to track the best four contours from the underlying formant candidates estimated by the all-pole methods.

B. DNN-BASED FORMANT TRACKERS

In order to study the possible limitations of the DP-based tracker, a deep neural network (DNN) -based formant tracker was developed as an alternative. A simple four-layer feed-forward DNN was used to capture the nonlinear mapping between the spectrum and the formant frequencies. The DNN had 300 units with tangent-hyperbolic activation in each of the three hidden layers [25]. The input dimension of 143 units corresponded to 13 RASTA-PLP [26] cepstral coefficients with an 11-frame neighborhood, and the three linear output units corresponded to the first (F_1), second (F_2) and third (F_3) formant to be predicted.

A common input feature was deliberately used to have a common baseline performance, and to study the incremental improvement provided by different spectrum estimation methods when used for refinement. 300 utterances from the train subset of the VTR-TIMIT database [27] were used to train the models. Mean square error between the estimated and actual formant values was used as the objective function. All parameters of the network were initialized randomly.

The stochastic gradient descent algorithm with standard backpropagation of error was used to learn the network parameters. The dropout regularization method was used to prevent overfitting the network. Input values were normalized to the range of [0.1, 0.9] and output values were normalized to have zero mean and unit variance. The DNN-based tracker was used in three modes: (1) by predicting the lowest three formants directly, (2) by refining the formants predicted by the DNN by replacing them with the frequencies of the corresponding nearest peaks in the LP-FBCOV spectrum, (3) by refining the formants predicted by the DNN

TABLE 1. Formant tracking performance of the DP-based tracker using six all-pole modeling methods in formant estimation and performance of six reference trackers. The numbers in parentheses denote the potential performance of the underlying formant estimation method if any of the five formant candidates is found within the allowed deviation from the ground truth. The results are reported by averaging over of all the 192 utterances of the VTR test database.

Method	FDR (%)			FEE (Hz)		
	F_1	F_2	F_3	δF_1	δF_2	δF_3
All-pole modeling methods						
LP-ACOR	84.3 (92.2)	72.3 (90.9)	69.0 (87.4)	92 (66)	296 (135)	325 (178)
LP-COV	86.0 (92.3)	75.4 (91.5)	71.3 (87.9)	89 (64)	292 (131)	319 (174)
LP-FBCOV	86.0 (92.4)	75.4 (91.9)	71.3 (88.2)	89 (64)	292 (129)	319 (172)
QCP-ACOR	86.8 (91.6)	75.5 (91.4)	71.6 (88.5)	87 (69)	292 (132)	317 (167)
QCP-COV	89.7 (91.6)	86.1 (91.6)	79.6 (89.0)	73 (69)	187 (130)	228 (165)
QCP-FBCOV	90.0 (93.4)	82.1 (93.9)	77.0 (92.1)	73 (63)	233 (114)	258 (130)
Reference trackers						
PRAAT	86.0	70.0	63.1	88	268	340
MUST	81.1	86.3	76.9	91	152	230
WSURF-0	84.1	78.2	77.3	93	239	245
WSURF-1	86.6	82.7	80.8	87	223	228
KARMA	91.5	89.4	74.7	62	146	250
Deep Formants	91.7	92.3	89.7	85	120	143

TABLE 2. Performance of the DNN-based formant trackers and performance of three reference trackers. The results are reported by averaging over of all the 192 utterances of the VTR test database.

Method	FDR (%)			FEE (Hz)		
	F_1	F_2	F_3	δF_1	δF_2	δF_3
DNN-based formant trackers						
DNN	90.5	91.6	82.1	74	128	184
DNN-LP-FBCOV	92.3	91.9	87.0	64	127	182
DNN-QCP-FBCOV	**93.3**	**93.5**	**89.9**	**62**	**113**	**142**
Reference trackers						
WSURF-1	86.6	82.7	80.8	87	223	228
KARMA	91.5	89.4	74.7	**62**	146	250
Deep Formants	91.7	92.3	89.7	85	120	143

by replacing them with the frequencies of corresponding nearest peaks in the QCP-FBCOV spectrum. (Note that with the model order $p = 12$, the LP-FBCOV spectrum and the QCP-FBCOV spectrum can show maximally six peaks). These three trackers will be referred to as DNN, DNN-LP-FBCOV and DNN-QCP-FBCOV, respectively. It is worth emphasizing that the latter two modes combine a data-driven approach and a model-driven approach in formant tracking in a novel way: formants F_1–F_3 are first predicted using a data-driven *deep learning approach* from a given frame with the DNN after which the predicted formants are refined using a model-driven *signal processing approach* using the all-pole spectrum extracted from the frame.

C. REFERENCE FORMANT TRACKERS

The DP-based and the DNN-based formant tracking algorithms were compared to known formant trackers. These reference trackers include algorithms used in two popular speech analysis tools (Wavesurfer [8] and PRAAT [7]), the adaptive filter bank (AFB) -based formant tracking algorithm (denoted as MUST) [23], KARMA (based on Kalman filtering) [10], and Deep Formants (based on DNNs) [24],

TABLE 3. The formant tracking performance of KARMA, Deep Formants, DNN and DNN-QCP-FBCOV in terms of FDR and FEE for different phonetic categories. The results are reported by averaging over of all the 192 utterances of the VTR test database.

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
Deep Formants						
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
DNN						
Vowels (V)	91.6	93.3	82.9	70.0	121.5	176.5
Diphthongs (D)	92.9	92.6	84.6	71.5	120.9	175.2
Semivowels (S)	84.5	85.2	76.6	89.2	155.9	214.5
V+D+S	90.5	91.6	82.1	73.9	127.9	183.6
DNN-QCP-FBCOV						
Vowels (V)	**94.2**	**94.2**	90.6	**57.1**	**105.2**	136.2
Diphthongs (D)	**94.5**	**95.2**	**93.0**	**61.9**	**109.4**	**119.5**
Semivowels (S)	**88.8**	**89.2**	**84.6**	76.6	**143.1**	183.8
V+D+S	**93.3**	**93.5**	**89.9**	**61.7**	**113.3**	**141.8**

[28]. Both Wavesurfer and PRAAT use LP analysis followed by DP-based tracking. Wavesurfer was used in two forms corresponding to autocorrelation LP and stabilized covariance LP which are referred to as WSURF-0 and WSURF-1, respectively [8]. The PRAAT algorithm uses the BURG method in LP analysis [7]. All the algorithms tracked four formants from the top five formant candidates derived from the underlying spectrum at a frame rate of 100 Hz.

IV. EXPERIMENTS AND RESULTS

A. DATABASE AND PERFORMANCE METRICS

The formant tracking performance was evaluated using the VTR database, which is one of the most widely used speech databases in the areas of formant estimation and tracking [27]. The test data of the database was used for the evaluation. This data consists of 192 utterances (produced by 8 female and 16 male speakers, each pronouncing 8 utterances). The duration of each utterance varies between 2 and 5 s. The ground truth (i.e., formant frequencies) have been derived using a semi-supervised LP–based method [29]. The values of F_1–F_3 have been corrected manually using spectrograms. The ground truth values for formants are provided for every 10 ms interval. The formant tracking performance was evaluated using two known metrics: the formant detection rate (FDR) and the formant estimation error (FEE) as defined in [21]. During the performance evaluation, the reference ground truth for each of the lowest three formants was

TABLE 4. Formant tracking performance of different methods for male and female speakers separately. The results are reported by averaging over all the utterances of the male and female speakers of the VTR test database.

Method	FDR (%)			FEE (Hz)		
	F_1	F_2	F_3	δF_1	δF_2	δF_3
Male						
Deep Formants	93.1	96.1	93.9	76	97	115
DNN	89.4	91.8	83.3	75	126	177
DNN-QCP-FBCOV	92.6	94.3	90.5	60	109	137
Female						
Deep Formants	94.0	94.1	87.0	93	110	163
DNN	92.7	91.0	79.6	72	133	196
DNN-QCP-FBCOV	94.5	91.9	88.8	65	123	151

associated with the nearest formant candidate lying within a specified relative (τ_p) and absolute (τ_a) deviations.

The FDR is computed in terms of the percentage of frames for which a hypothesized formant occurs within a specified deviation from the ground truth formant. The FDR for the i^{th} formant over M analysis frames is computed as

$$D_i = \frac{1}{M} \sum_{n=1}^{M} I(\Delta F_{i,n}), \tag{11}$$

$$I(\Delta F_{i,n}) = \begin{cases} 1 & \text{if}(\Delta F_{i,n}/F_{i,n} \\ & < \tau_r \ \& \ \Delta F_{i,n} < \tau_a) \\ 0 & \text{otherwise}, \end{cases} \tag{12}$$

where $I(.)$ denotes a binary formant detector function. $\Delta F_{i,n} = |F_{i,n} - \hat{F}_{i,n}|$ is the absolute deviation of the hypothesized formant frequency ($\hat{F}_{i,n}$) from the reference ground truth ($F_{i,n}$) at the n^{th} frame for the i^{th} formant. The FEE is computed in terms of the average absolute deviation of the hypothesized formant from the ground truth formant. The FEE for the i^{th} formant over M analysis frames is computed as

$$R_i = \frac{1}{M} \sum_{n=1}^{M} \Delta F_{i,n}. \tag{13}$$

The FEE values in conjunction with FDR values give a better sense of the performance of a formant tracker.

B. DP-BASED FORMANT TRACKING

The FDRs (within $\tau_p = 30\%$ and $\tau_a = 300$ Hz deviation) and FEEs for the different formant estimation methods are given in Table 1. In this table, the two metrics are computed by associating the three hypothesized formant tracks with the lowest three reference tracks.

The scores in the parentheses, however, denote the best scores which were obtained by identifying each formant as the spectral peak (among the detected five candidates) that was closest to the corresponding reference formant. The scores in the parentheses describe the performance of the underlying formant estimation method when used with an ideal formant tracker. It can be seen from the results that the DP-based tracker gave scores inferior to the detection potential of the underlying spectral estimates.

It can be seen from Table 1 that the QCP-based methods performed consistently better than all of their LP-based counterparts. The covariance method performed better than the autocorrelation method for both LP and QCP. However, LP-FBCOV showed no improvement over LP-COV, and QCP-FBCOV seems to be inferior to QCP-COV, despite the detection potential being highest (the scores in parentheses) for QCP-FBCOV. This behavior can be attributed to the inherent limitations of the DP-based tracker with a possibility of tracking a spurious candidate instead of the best candidate (which is otherwise not known without the ground truth).

C. DNN-BASED FORMANT TRACKING

A comparison of the performance of the DNN-based formant tracker using LP-FBCOV and QCP-FBCOV for refinement is given in Table 2 along with the performance of the DNN predictor.

Three reference trackers (Wavesurfer (WSURF-1), KARMA, and Deep Formants) were chosen for comparison based on their performance shown in Table 1. It can be seen that the DNN-QCP-FBCOV tracker performed best, almost realizing the full potential of the QCP-FBCOV method (the scores in parentheses in Table 1). The improvement given by DNN-QCP-FBCOV compared particularly to the popular Wavesurfer tracker is large showing a reduction of 29%, 48%

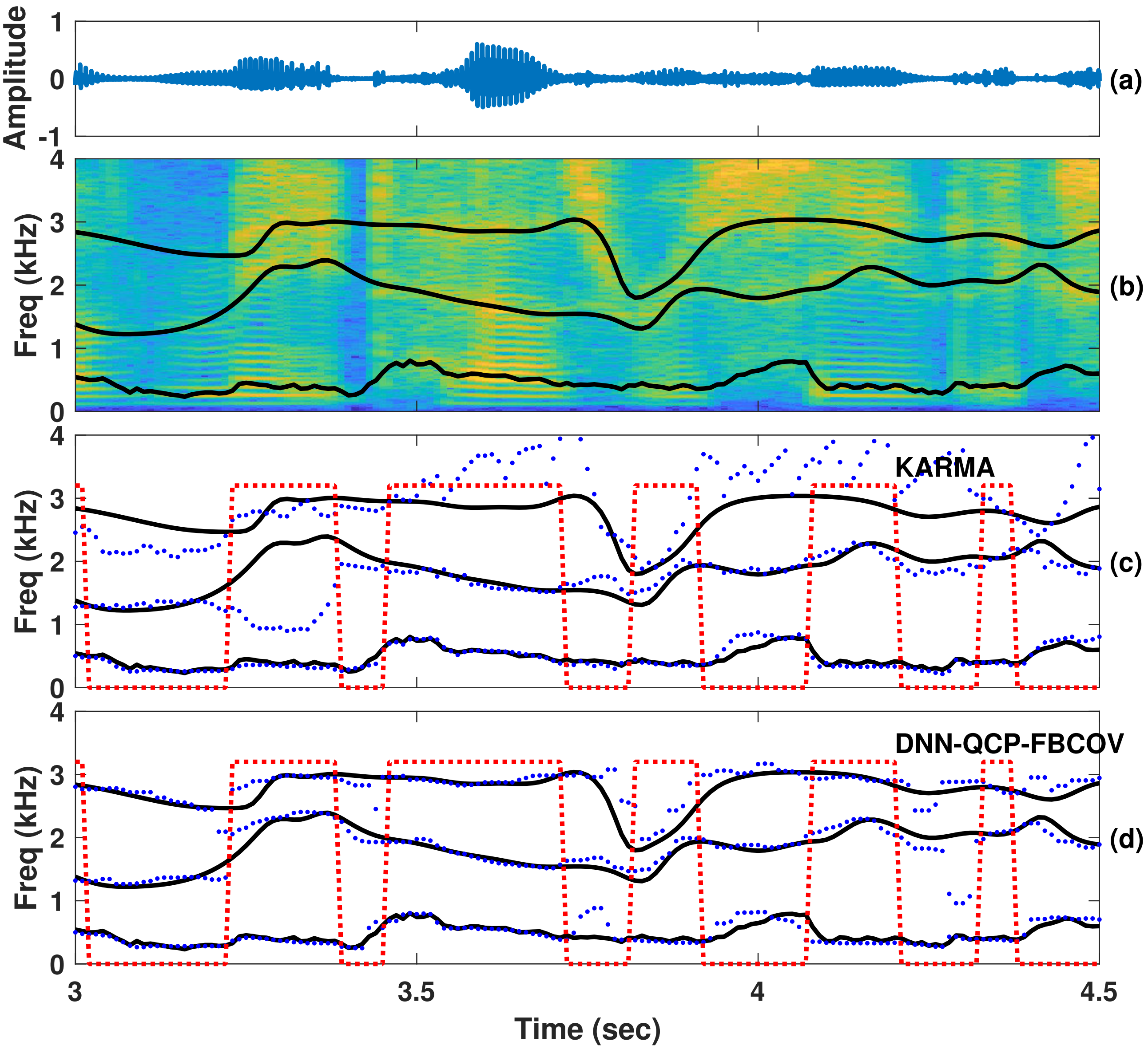


FIGURE 1. Formant tracking by KARMA and DNN-QCP-FBCOV for an utterance produced by a male speaker: (a) the time-domain speech signal, (b) the narrowband spectrogram with reference ground truth formant contours, (c) the formant track estimates of KARMA along with the voiced-unvoiced regions shown by a dotted rectangular-wave plot, and (d) the formant track estimates of DNN-QCP-FBCOV.

and 35% in the estimation error for the lowest three formants, respectively. These results demonstrate that the QCP-FBCOV method can be a good replacement for the popularly used LP-COV analysis in formant estimation and tracking tools and applications.

A detailed comparison in the formant tracking performance of KARMA, Deep Formants, DNN and DNN-QCP-FBCOV is given in Table 3 for different phonetic categories (vowels, dipthongs, and semovowels). It can be seen that the proposed DNN-QCP-FBCOV method performed clearly better for all the phonetic categories in both FDR and FEE. Formant tracking performance of different methods analyzed separately for male and female speakers is given in Table 4. From the table it can be observed that the performance of Deep Formants is better for male speakers (except in δF_1, where the DNN-QCP-FBCOV method is better) but for female speakers the DNN-QCP-FBCOV method is better.

Formant tracking performance for different methods using speech degraded with white and babble noise at signal-to-noise ratio (SNR) levels of 10 dB and 5 dB are given in Table 5. From the table it can be observed that the proposed DNN-QCP-FBCOV method performed better in the case of speech degraded with white noise. In the case of speech degraded with babble noise, Deep Formants and DNN methods seems to perform better.

An illustration of formant tracking by KARMA and DNN-QCP-FBCOV for an utterance produced by a male speaker is shown in Fig. 1. It can be seen from the figure

TABLE 5. Formant tracking performance for different methods using speech degraded with white and babble noise at SNR levels of 10 dB and 5 dB. The results are reported by averaging over of all the 192 utterances of the VTR test database.

Method	FDR (%)			FEE (Hz)		
	F_1	F_2	F_3	δF_1	δF_2	δF_3
White at 10 dB						
KARMA	86.2	80.1	68.8	75.5	191.3	256.5
Deep Formants	89.8	80.8	71.6	99.2	184.3	**238.7**
DNN	84.6	79.6	69.5	88.5	193.4	245.6
DNN-QCP-FBCOV	**91.1**	**86.1**	**71.8**	**69.0**	**162.2**	251.8
White at 5 dB						
KARMA	80.1	72.5	64.0	91.6	232.5	279.2
Deep Formants	**89.2**	71.7	64.5	101.1	238.7	274.3
DNN	84.8	79.6	**69.1**	88.1	**193.3**	**247.4**
DNN-QCP-FBCOV	87.8	**80.6**	65.3	**83.7**	**196.9**	282.8
Babble at 10 dB						
KARMA	90.3	83.8	71.8	**65.1**	176.1	246.0
Deep Formants	**91.1**	86.6	**81.7**	88.4	145.9	**182.7**
DNN	88.8	**89.1**	78.8	77.7	**141.7**	198.7
DNN-QCP-FBCOV	90.7	87.1	81.1	66.0	153.9	203.8
Babble at 5 dB						
KARMA	88.2	78.9	68.7	**70.9**	200.9	260.3
Deep Formants	**89.8**	81.4	76.1	89.9	177.3	**209.1**
DNN	87.5	**86.5**	**76.2**	80.5	**155.1**	211.0
DNN-QCP-FBCOV	87.7	81.7	74.9	72.4	187.9	239.7

that the formants tracked by the DNN-QCP-FBCOV method match closely the ground truth in voiced segments. Furthermore, it can be clearly seen that DNN-QCP-FBCOV is better than KARMA in tracking all the formants.

V. CONCLUSION

Formant tracking was studied in this paper based on the widely used two-stage approach consisting of the estimation stage and the tracking stage. In the former, six different all-pole modeling methods were first compared with a DP-based tracker. In addition, five known formant trackers were used as references. Two most potential all-pole modeling methods (LP-FBCOV and QCP-FBCOV) were then used with a modern DNN-based tracker by proposing a novel formant tracking technique which combines benefits of data-driven and model-driven approaches: the formants predicted with the data-driven DNN were refined using the frequencies of the peaks in the all-pole spectra computed by the model-driven LP-FBCOV and QCP-FBCOV methods.

The DNN-based formant trackers using the LP-FBCOV and QCP-FBCOV refinement were further compared to two conventional formant trackers (Wavesurfer and KARMA) and to one recently published DNN-based tracker (Deep Formants). With the QCP-FBCOV refinement, the DNN-based tracker outperformed the conventional reference formant trackers in all metrics. Compared to Deep Formants, the proposed DNN-tracker gave better performance in all other metrics except for FDR and FEE in F_3 where Deep Formants was just slightly better. In addition to these encouraging objective results, it is worth emphasising that the proposed QCP-FBCOV refinement technique can be used in principle to improve the performance of any existing DNN-based formant tracker which has been trained to map a speech signal frame into formants, that is, there is no need to re-train the DNN-based tracker used. However, it should be noted that the performance of the proposed method might depend on the accuracy of the estimated glottal closure instants, which are needed to generate the QCP weighting function [19]. Therefore, the robustness of the proposed method in various noisy conditions needs to be studied further. One way of improving the performance under degradations could be by training the DNN models for all noisy conditions of interest. Nevertheless, under clean conditions, the current study shows that the QCP-FBCOV method is a potential all-pole modeling technique to be used in formant tracking instead of the widely used conventional LP methods.

It is worth noting that the DNN-based formant tracking methods studied in this investigation (i.e., Deep Formants, DNN, DNN-LP-FBCOV, and DNN-QCP-FBCOV) are based on supervised learning and their computational complexity is relatively high compared to traditional model-based approaches. It is known that DNNs are resource hungry due to their need for training data and the architecture of the neural network adds more computational complexity when the trained network is used in formant tracking. The LSTM-based Deep Formants architecture has approximately 4M parameters, while the FFNN DNN architecture we propose has around 0.3M parameters. It is worth emphasizing, however, that the main contribution of this study, the QCP-FBCOV based refinement of formants, can be plugged into any existing pre-trained DNN-based tracker, which results only in a marginal increase in complexity. Compared to the conventional autocorrelation based LP, which has a computational complexity of $O(n^2)$, our proposed QCP-FBCOV-based tracker has $O(n^3)$ complexity, where n denotes the size of the covariance matrix (which is equal to the LP order, which was $p = 12$ in the experiments of the current study). However, the order of LP analysis being small, our proposed DNN-QCP-FBCOV/-based formant tracking results only in a negligible increase in the overall computational complexity. There is an added computation complexity due to the computation of the temporal weighting function (w_n in Eqs. 7 and 8), which calls for estimating glottal closure instants, which also requires LP inverse filtering and is proportional to $O(n^3)$ computations.

ACKNOWLEDGMENT

Dhananjaya Gowda and Bajibabu Bollepalli were with Aalto University when this work was initiated.

REFERENCES

[1] P. F. Assmann, "The role of formant transitions in the perception of concurrent vowels," *J. Acoust. SoC. Amer.*, vol. 97, no. 1, pp. 575–584, Jan. 1995.

[2] L. Welling and H. Ney, "Formant estimation for speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 1, pp. 36–48, Jan. 1998.

[3] T. Smit, F. Tärckheim, and R. Mores, "Fast and robust formant detection from LP data," *Speech Commun.*, vol. 54, no. 7, pp. 893–902, Sep. 2012.

[4] I.-C. Yoo, H. Lim, and D. Yook, "Formant-based robust voice activity detection," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 12, pp. 2238–2245, Dec. 2015.

[5] R. Singh, D. Gencaga, and B. Raj, "Formant manipulations in voice disguise by mimicry," in *Proc. 4th Int. Conf. Biometrics Forensics (IWBF)*, Limassol, Cyprus, Mar. 2016, pp. 1–6.

[6] H. K. Kathania, S. R. Kadiri, P. Alku, and M. Kurimo, "Study of formant modification for children ASR," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 7429–7433.

[7] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot Int.*, vol. 5, nos. 9–10, pp. 341–345, 2002.

[8] K. Sjolander and J. Beskow, "Wavesurfer—An open source speech tool," in *Proc. Int. Conf. Spoken Lang. Process.*, Beijing, China, Oct. 2000, pp. 464–467.

[9] 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/ACM Trans. Audio, Speech, Language Process.*, vol. 15, no. 1, pp. 13–23, Jan. 2007.

[10] D. D. Mehta, D. Rudoy, and P. J. Wolfe, "Kalman-based autoregressive moving average modeling and inference for formant and antiformant tracking," *J. Acoust. SoC. Amer.*, vol. 132, no. 3, pp. 1732–1746, Sep. 2012.

[11] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, Apr. 1975.

[12] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen, "Sparse linear prediction and its applications to speech processing," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 20, no. 5, pp. 1644–1657, Jul. 2012.

[13] C. Magi, J. Pohjalainen, T. Bäckström, and P. Alku, "Stabilized weighted linear prediction," *Speech Commun.*, vol. 51, no. 5, pp. 401–411, 2009.

[14] 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.

[15] 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.*, Paris, France, May 1982, pp. 1573–1576.

[16] C.-C. Lee, "On robust linear prediction of speech," *IEEE Trans. Acoust., Speech Signal Process.*, vol. ASSP-36, no. 5, pp. 642–650, May 1988.

[17] 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, Mar. 1993.

[18] 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. Acoust. SoC. Amer.*, vol. 134, no. 2, pp. 1295–1313, Aug. 2013.

[19] 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, Language Process.*, vol. 22, no. 3, pp. 596–607, Mar. 2014.

[20] D. Gowda, S. Reddy Kadiri, B. Story, and P. Alku, "Time-varying quasi-closed-phase analysis for accurate formant tracking in speech signals," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 1901–1914, 2020.

[21] D. Gowda, M. Airaksinen, and P. Alku, "Quasi-closed phase forward-backward linear prediction analysis of speech for accurate formant detection and estimation," *J. Acoust. SoC. Amer.*, vol. 142, no. 3, pp. 1542–1553, Sep. 2017.

[22] A. Rao and P. K. Ghosh, "Glottal inverse filtering using probabilistic weighted linear prediction," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 1, pp. 114–124, Jan. 2019.

[23] K. Mustafa and I. C. Bruce, "Robust formant tracking for continuous speech with speaker variability," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 2, pp. 435–444, Mar. 2006.

[24] Y. Dissen, J. Goldberger, and J. Keshet, "Formant estimation and tracking: A deep learning approach," *J. Acoust. SoC. Amer.*, vol. 145, no. 2, pp. 642–653, Feb. 2019.

[25] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org

[26] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. SoC. Amer.*, vol. 87, no. 4, pp. 1738–1752, 1990.

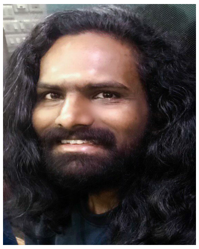
[27] L. Deng, X. Cui, R. Pruvenok, J. Huang, and S. Momen, "A database of vocal tract resonance trajectories for research in speech processing," in *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Toulouse, France, 2006, pp. I369–I372.

[28] Y. Dissen and J. Keshet, "Formant estimation and tracking using deep learning," in *Proc. Interspeech*, San Francisco, CA, USA, Sep. 2016, pp. 958–962.

[29] L. Deng, L. J. 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. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Montreal, QC, Canada, May 2004, pp. 1–9.

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

BAJIBABU BOLLEPALLI (Member, IEEE) received the bachelor's and master's degrees from IIIT-Hyderabad, India, 2011 and 2012, respectively, the Licentiate of Engineering from the KTH Royal Institute of Technology, Stockholm, Sweden, in 2017, and the Ph.D. degree from the Department of Signal Processing and Acoustics, Aalto University, Finland, in 2020. He worked as a Postdoctoral Fellow at Verisk Analytics, Munich, Germany, before joining in Amazon. He is currently working as an Applied Scientist at Amazon, Alexa AI, Cambridge, U.K. His research interests include speech synthesis, speech processing, natural language processing, and machine learning. He was a recipient of two best student paper awards one in ICASSP 2016 and another in Interspeech 2016.



SUDARSANA REDDY KADIRI (Member, IEEE) received the B.Tech. degree from Jawaharlal Nehru Technological University (JNTU), Hyderabad, India, in 2011, with a specialization in electronics and communication engineering (ECE), the M.S. (Research) 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), India, in 2018. He was awarded the Tata Consultancy Services (TCS) fellowship for his Ph.D. He is currently a Postdoctoral Researcher with the Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland. He was a Teaching Assistant for several courses at IIIT-H, from 2012 to 2018, and he has been involving in teaching and mentoring activities at Aalto University, since 2019. His research interests include signal processing, speech analysis, speech synthesis, paralinguistics, affective computing, voice pathologies, machine learning, and auditory neuroscience. He has published over 50 research papers in peer-reviewed journals and conferences in these areas.

PAAVO ALKU (Fellow, IEEE) received the M.Sc., Lic.Tech., and Dr.Sc.(Tech) degrees from the 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 at Aalto University, Espoo. He has published over 220 peer-reviewed journal articles and over 210 peer-reviewed conference papers. His research interests include the 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 is a fellow of ISCA. He is an Associate Editor of *The Journal of the Acoustical Society of America*. He served as an Academy Professor assigned by the Academy of Finland, from 2015 to 2019.