Received XX Month, XXXX; revised XX Month, XXXX.

Formant Tracking by Combining Deep Neural Network and Linear Prediction

Sudarsana Reddy Kadiri[1] (Member, IEEE),, Kevin Huang[1], Christina Hagedorn[2], Dani Byrd[1], Paavo Alku[3] (Fellow, IEEE), Shrikanth Narayanan[1] (Fellow, IEEE)

[1]University of Southern California, USA. [2]City University of New York, College of Staten Island & The Graduate Center, USA. [3]Aalto University, Finland.

CORRESPONDING AUTHORS: Sudarsana Reddy Kadiri and Paavo Alku (e-mails: skadiri@usc.edu; paavo.alku@aalto.fi).

This study was funded by National Science Foundation (Grant: 2311676) and Academy of Finland (Project: 330139)

ABSTRACT Formant tracking is an area of speech science that has recently undergone a technology shift from classical model-driven signal processing methods to modern data-driven deep learning methods. In this study, these two domains are combined in formant tracking by refining the formants estimated by a data-driven deep neural network (DNN) with formant estimates given by a model-driven linear prediction (LP) method. In the refinement process, the three lowest formants, initially estimated by the DNN-based method, are frame-wise replaced with local spectral peaks identified by the LP method. The LP-based refinement stage can be seamlessly integrated into the DNN without any training. As an LP method, the study advocates the use of quasi-closed phase forward-backward (QCP-FB) analysis. Three spectral representations are compared as DNN inputs: mel-frequency cepstral coefficients (MFCCs), the spectrogram, and the complex spectrogram. Formant tracking performance was evaluated by comparing the proposed refined DNN tracker with seven reference trackers, which included both signal processing and deep learning based methods. As evaluation data, ground truth formants of the Vocal Tract Resonance (VTR) corpus were used. The results demonstrate that the refined DNN trackers outperformed all conventional trackers. The best results were obtained by using the MFCC input for the DNN. The proposed MFCC refinement (MFCC-DNN$_{\text{QCP-FB}}$) reduced estimation errors by 0.8 Hz, 12.9 Hz, and 11.7 Hz for the first (F1), second (F2), and third (F3) formants, respectively, compared to the Deep Formants refinement (DeepF$_{\text{QCP-FB}}$). When compared to the model-driven KARMA tracking method, the proposed refinement reduced estimation errors by 2.3 Hz, 55.5 Hz, and 143.4 Hz for F1, F2, and F3, respectively. A detailed evaluation across various phonetic categories and gender groups showed that the proposed hybrid refinement approach improves formant-tracking performance across most test conditions.

INDEX TERMS Formant tracking, linear prediction, MFCCs, spectrogram, machine learning, deep learning

I. Introduction

Vocal tract resonances, commonly referred to as formants, are essential parameters in characterizing speech signals. In continuous speech, formants exhibit temporal variation, appearing as contours over time. These time-domain formant contours have been extensively studied across various fields in speech science, including acoustic phonetics [1], [2], pathological speech analysis [3]–[5], automatic speech recognition [6], [7], and auditory research [8]–[10]. The automatic estimation of formant contours from speech signals requires sophisticated formant tracking techniques. Given the complexity of formant tracking, this task presents a considerable engineering challenge and has driven the development of numerous methodologies over recent decades [11]–[15]. Typically, these methodologies involve two main stages. The first stage, known as estimation, computes preliminary formant values over short time-segments (e.g., 25 ms) using techniques such as linear prediction (LP) [16] or cepstral analysis [17]. The second stage, tracking, integrates the formant estimates from individual frames into continuous contours that span longer speech units, such as syllables, words, or phrases [11], [12]. Some advanced methods merge these two stages, using an initial model of the vocal tract system to estimate and track formants simultaneously [13], [14].

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

Formant trackers can be broadly classified into two categories based on their estimation methodologies: (1) model-driven and (2) data-driven approaches. Model-driven trackers, the traditional type, rely on signal processing techniques grounded in all-pole spectral estimation, primarily using various linear prediction (LP) methods to estimate formants from short speech segments. In these methods, formant frequencies are typically derived either by peak-picking the power spectrum of the LP-based all-pole model [2] or by calculating the roots of the denominator polynomial of the all-pole model [18]. One key characteristic of model-driven LP-based trackers is that they generate formant estimates directly from the speech signal without requiring prior training on formant data. Notable model-driven methods include autocorrelation and covariance-based LP techniques [11], [12], along with closed-phase (CP) analysis, which improves formant estimation accuracy by excluding speech samples from the open phase of the glottal cycle, thereby reducing the influence of tracheal effects [19]. Further advancements in model-driven formant estimation involve the use of weighted linear prediction (WLP), where temporal weighting is applied to the LP prediction error, increasing robustness against noise and high fundamental frequencies [20]–[22]. A standout WLP-based method, quasi-closed phase forward-backward (QCP-FB) analysis, has shown superior accuracy in formant estimation [23].

In contrast, data-driven formant trackers leverage deep learning (DL) neural networks trained to directly map selected acoustic features to formant frequencies. This process involves training a network on a data set of acoustic features paired with corresponding formant data, followed by estimating formants in test utterances by extracting relevant features from the speech signal and feeding them into the trained model. A known example was published in [24], where two DL models, a multi-layer perceptron (MLP) and a convolutional neural network (CNN), were evaluated for formant estimation using supervised learning on the manually annotated VTR speech corpus [25]. Similarly, [26], [27] explored a supervised learning approach using a bilinear network and a temporal attention-augmented bilinear network for formant prediction. Recent studies have also examined unsupervised DL-based formant trackers. For instance, [28] proposed an autoencoder-type network where latent features are interpreted as formants through a specialized loss function. Another study [29] utilized a CNN to map spectrograms into a latent representation, which is processed by multiple decoders to predict formants hierarchically.

In this study, we explore a hybrid formant tracking approach that integrates data-driven and model-driven methods, referred to as the *refinement* of a data-driven formant tracker. The process begins with a deep neural network (DNN)-based system, which initially estimates formants from the speech signal. These estimates are then refined by replacing them, frame-by-frame, with formants predicted by a model-driven, LP-based signal processing method. This combined approach addresses two key challenges in formant estimation:

- Overfitting in data-driven methods: Data-driven formant trackers can suffer from overfitting to the training data, resulting in reduced accuracy when applied to unseen test data [29]. We hypothesize that refining the DNN-predicted formants with a model-driven approach, such as QCP-FB, which showed superior accuracy in formant estimation in [23] and which is independent of training data, can mitigate the overfitting issue.
- Spurious peaks in model-driven methods: All-pole spectra computed by LP-based model-driven methods often show artifacts known as spurious peaks [30]. In the approach proposed in this study, the effect of spurious spectral peaks is reduced, because the LP-based refinement stage only considers the spectral peaks closest to the formants detected by the DNN-based tracker (further details provided in Section IIA).

This study is a sequel to our previous works on formant tracking [31], [32]. In the current study, formant tracking experiments are extended to a broader range of phonetic categories, going beyond the vowels, diphthongs, and semivowels, which were examined in our earlier works, and by conducting gender-wise analysis of all the phonetic categories. We propose a simple yet **accurate** DNN-based tracker that maps commonly used speech features to formant estimates. This approach raises an important, previously unexplored question: Can the accuracy of formant tracking using a modern, simple DNN-based data-driven tracker be improved by incorporating a model-driven signal processing module to refine the formant estimates? Our investigation seeks to address this question, highlighting the potential benefits of this combined approach. We compare the performance of the refined formant trackers with several reference methods, including both traditional signal processing-based trackers and more recent DL-based trackers.

The proposed formant tracker is described in Section II by first presenting the DNN-based and the LP-based formant estimation stages, after which the proposed combination of these two stages is presented. Section III describes the experimental setup, including data set details, performance metrics, and the formant trackers used for comparison. The results of the study are presented in Section IV, with conclusions drawn in Section V.

II. Formant Tracking by Combining a DNN-based and an LP-based Formant Estimation Method

This section describes the proposed formant tracker, which combines a DNN-based and an LP-based formant estimation method. The section begins with an overview of the DNN-based formant estimation technique, followed by a description of the LP-based method to refine the formants estimated by the DNN-based approach. The section concludes with a detailed explanation of the proposed refinement technique in formant tracking.

A. DNN-based Formant Estimation

A schematic block diagram of the DNN-based formant estimation method is shown in Figure 1. The method uses a four-layer feed-forward neural network to map an input spectral representation into formant frequencies. The network has 300 units with the tanh (tangent-hyperbolic) activation function in each of the three hidden layers. The output layer comprises three linear units corresponding to the first (F_1), second (F_2), and third (F_3) formants to be predicted. In this study, we compare the effectiveness of three different spectral representations as input of the DNN: (1) spectrogram, (2) complex spectrogram, and (3) MFCCs.

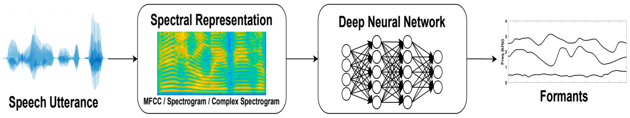

FIGURE 1: A schematic block diagram the DNN-based formant estimation approach.

- Spectrogram: A representation, that captures the power spectral density of speech over time based on the short-time Fourier transform (STFT). We computed STFT with an FFT size of 256, resulting in a 129-dimensional vector per time frame. The spectrogram effectively represents time-varying frequency characteristics of speech such as formants.
- Complex Spectrogram: A representation, which retains both magnitude and phase information of speech spectrum by including the real and imaginary components of the STFT. With an FFT size of 256, the complex spectrogram gives a 258-dimensional vector per time frame (129 real and 129 imaginary components), potentially enhancing formant tracking accuracy by capturing fine temporal and frequency details.
- Mel-Frequency Cepstral Coefficients (MFCCs): A widely used speech feature that models the speech spectrum using a compressed set of coefficients. Computed via a discrete cosine transform (DCT) on the log of the mel-spectrogram, MFCCs emphasize perceptually relevant features and formants with a low-dimensional representation. In the current study, we used the popular 13-dimensional vector per time frame.

Each representation has distinct advantages: the spectrogram models the general spectro-temporal properties of speech, the complex spectrogram preserves phase information, and MFCCs provide a compact speech feature whose computation uses the mel-filterbank which improves frequency resolution at lower frequencies where the lowest formants are located.

To train the DNNs for all the inputs described above, we utilized the **training subset of the VTR-TIMIT database** [25]. A total of 300 utterances were used for training, while an additional 24 utterances were set aside for validation during the training process. The mean squared error (MSE) between the estimated and actual formants was used as the objective function to optimize the model. The network parameters were initialized randomly, and the Adam optimizer with a learning rate of 0.0001 was employed to update the model parameters. Early stopping was implemented with a patience of 15 epochs, with a total of 50 epochs. To mitigate overfitting, a dropout rate of 0.2 was applied as a regularization technique, and a batch size of 1 was used during training. Input values were normalized to the range [0.1, 0.9], and output values were normalized to have zero mean and unit variance.

B. LP-based Formant Estimation

As the model-driven, LP-based formant estimation method for refining the formants predicted by the DNN-based estimation stage, the current study uses the QCP-FB method. This method was chosen because it showed the best performance in a comparison between six model-driven formant estimation methods [23]. A brief description of QCP-FB is provided below.

The traditional formulation of LP, as used in methods such as the covariance method, is based on forward prediction, where the current speech sample is predicted from the past p samples. However, backward prediction is also possible, where the current sample is predicted from the future p samples. The QCP-FB method utilizes the combination of these two approaches—referred to as forward-backward analysis. The combined error to be minimized in this approach is given by:

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

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

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

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 ($\partial\mathcal{E}/\partial a_i = 0,\ 1 \le i \le p$), which results in the following normal equations

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

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

QCP-FB analysis applies forward-backward analysis within the framework of weighted linear prediction to leverage the strengths of both techniques. The weighting is performed using a temporal weighting function known as the QCP function, as defined in [22]. In this approach, both the

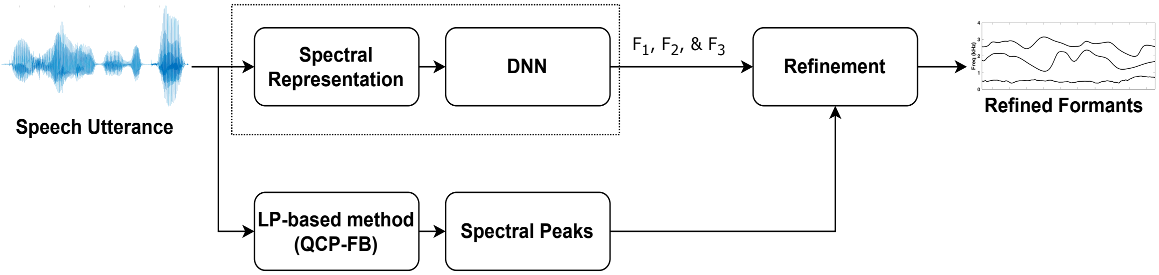

FIGURE 2: A schematic block diagram of the proposed formant tracker based on the combination of the DNN-based and the LP-based formant estimation methods. In the upper path, the DNN estimates the three formants on a frame-by-frame basis. In the lower path, up to six local spectral peaks are extracted from the all-pole spectrum, computed per frame using QCP-FB analysis. The outputs from both paths are then used for refinement, as shown in Fig. 3, to define the final refined formant tracks.

forward and backward prediction errors are weighted using the QCP function. Letting w_n represent the QCP weighting function, the combined error to be minimized is expressed as:

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

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

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

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

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

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

An appropriate choice of the range for the variable n determines whether the autocorrelation or covariance method is used in QCP-FB. In this study, we adopt the covariance method for QCP-FB. The analysis is performed using a rectangular windowed frame of 25 ms with a frame shift of 10 ms and an all-pole model order of p=13. Speech signals, sampled at 8 kHz, are pre-emphasized using a first-order FIR filter $P(z) = 1 - 0.97z^{-1}$. To detect formant peaks, the resulting spectrum is convolved with a Gaussian derivative window of 100 Hz width, and negative zero-crossings are identified as spectral peaks.

C. Combining the DNN-based and LP-based Formant Estimation Methods

The new formant tracking method proposed in this study takes advantage of the DNN-based and the LP-based formant estimation methods described in Section IIA and Section IIB, respectively. The new tracker is based on refining the formants predicted by the data-driven DNN-based estimator in each time frame with spectral peaks shown by the all-pole spectrum computed by the model-driven LP-based estimator. The flow diagram of the tracker is shown in Fig. 2. First, the DNN maps frame-wise acoustic features, as described in Section IIA, into initial estimates of F_1, F_2, and F_3. Next, an all-pole spectral model is computed using QCP-FB for each input frame, and the local peaks of the all-pole spectrum are identified. With a model order of p = 13, QCP-FB can display up to six local peaks or resonances (each resonance is represented by a pair of complex conjugate poles). Third, each of the three initial formants predicted by the DNN is replaced by the local peak in the all-pole spectrum that is nearest to the corresponding predicted formant. A graphical representation of this refinement process is shown in Fig. 3.

As described in Section IIA, we wanted to compare the performance of the new formant tracker between three DNN input presentations. In order to distinguish these three tracker versions, we will denote in the following sections the new tracker using the MFCC, spectrogram and complex spectrogram input by MFCC-DNN$_{\text{QCP-FB}}$, SPEC-DNN$_{\text{QCP-FB}}$ and CSPEC-DNN$_{\text{QCP-FB}}$, respectively. In order to evaluate the effect of refinement, we also included these same trackers by excluding the LP-based refinement block and by using the formants estimated by the DNN as such. The corresponding three trackers will be denoted for the MFCC, spectrogram and complex spectrogram inputs by MFCC-DNN, SPEC-DNN and CSPEC-DNN, respectively.

III. Experimental Protocol

A. Database

The formant tracking performance was evaluated using the test subset of the VTR-TIMIT database, a widely recognized resource for formant estimation and tracking [25]. This test subset comprises 192 utterances, produced by 8 female and 16 male speakers, with each speaker recording eight utterances. The duration of each utterance varies between two and five seconds. Ground truth formant frequencies were obtained using a semi-supervised LP-based method [33], with manual corrections made to the lowest three formants (F_1, F_2, and F_3) using spectrograms. These ground truth values are provided at 10 ms intervals. Table 1 provides

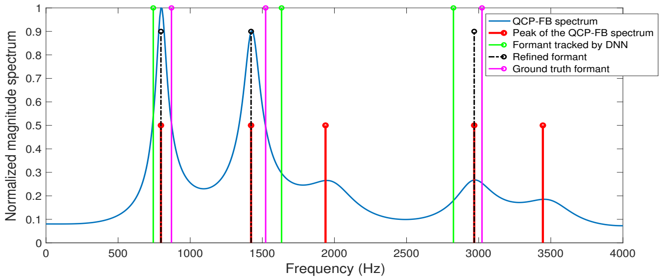

FIGURE 3: An illustration of the refinement process. In this example, the QCP-FB spectrum shows five local peaks, indicated by red lines. The three formants initially estimated by the DNN are represented by green lines. The refinement process selects those three local peaks in the QCP-FB spectrum which are closest to the formants predicted by the DNN. The resulting refined formants are marked by black lines.

a detailed categorization of phonemes used in this study, marked according to their phonetic properties.

TABLE 1: Phonetic categories of the VTR database used in this study.

Category	Phonemes (Arpabet)
Vowels	{'iy', 'ih', 'eh', 'ae', 'aa', 'ah', 'uh', 'uw', 'ux', 'er', 'ax', 'ix', 'axr', 'ax-h'}
Diphthongs	{'ey','aw','ay','ao','oy','ow'}
Semivowels	{'y','r','l','w','hh','hv','el'}
Nasals	{'m','n','ng','em','en','eng','nx'}
Voice Bars	{'bcl','dcl','gcl','dx'}
Fricatives & Affricates	{'s','sh','f','th','jh','ch','z','zh','v','dh'}
Stops	{'p','t','k','q','b','d','g'}

B. Performance metrics

The formant tracking performance was evaluated using two metrics: the formant detection rate (FDR) and the formant estimation error (FEE), both of which have been utilized in previous studies [23], [32]. The FDR measures the percentage of frames in which a formant is detected within a specified deviation from the ground truth. The FDR for the i^{th} formant over K analysis frames is:

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

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

where $I(.)$ denotes a binary 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 i^{th} formant at the n^{th} frame from the reference ground truth $F_{i,n}$. The thresholds τ_r and τ_a denote the relative deviation and absolute deviation, respectively. As in [23], [34], these parameters were set as $\tau_r = 30\%$ and $\tau_a = 300$ Hz. FEE is measured as the average absolute deviation of the hypothesized formants from the ground truth. The FEE for the i^{th} formant over K analysis frames is computed as: $E_i = \frac{1}{K}\sum_{n=1}^{K} \Delta F_{i,n}$. A formant tracker's performance is considered better if it yields a higher FDR and a lower FEE. The FEE values, when combined with FDR values, provide a comprehensive index of formant tracking accuracy. In this study, FDR and FEE values were computed specifically for frames within the phonetic categories of interest (see Section IV).

C. Reference trackers

The reference trackers evaluated in this study include:

- **PRAAT** [11]: BURG method-based LP analysis tracker (denoted as PBURG).
- **AFB-based** [35]: Adaptive filter bank formant tracker (denoted as MUST).
- **Wavesurfer (Autocorrelation)** [12]: LP-based tracker using the autocorrelation method (denoted as WSURF-0).
- **Wavesurfer (Covariance)** [12]: LP-based tracker using the covariance method (denoted as WSURF-1).
- **Kalman Filter** [14]: Tracker using Kalman filtering (denoted as KARMA).
- **DeepF** [24]: Default DL-based tracker (denoted as DeepF).
- **DeepF$_{\text{QCP-FB}}$** [32]: Refined DeepFormants tracker using QCP-FB (denoted as DeepF$_{\text{QCP-FB}}$).

The first five are classical trackers, that are based on signal processing methods. The latter two are more recent DL-based trackers: DeepF is the default DL-based tracker and DeepF$_{\text{QCP-FB}}$ is the same tracker supplemented with formant refinement. All trackers provide formant estimates at a frame rate of 100 Hz (10 ms).

IV. Results

Performance of the different trackers described in Section IIIC were evaluated in formant tracking by running three experiments for the VTR test set. In the first experiment, we evaluated all trackers by combining the vowels, diphthongs, and semivowels produced by the female and male speakers of the test set into one comprehensive evaluation data set. In the second experiment, two best DNN-based trackers were compared to the default DL-based tracker (DeepF) separately for female and male speech. In the third experiment, performance was further analyzed for fine-grained phonetic categories separately for female and male speech. The results of the first experiment are summarized in Table 2 by presenting the performance metrics for the seven reference trackers (separately for the five conventional signal processing trackers and for the two DL trackers) as well as for the six trackers that all used the DNN-based formant estimation (either without or with the LP-based refinement). From the table, it can be observed that the formant detection

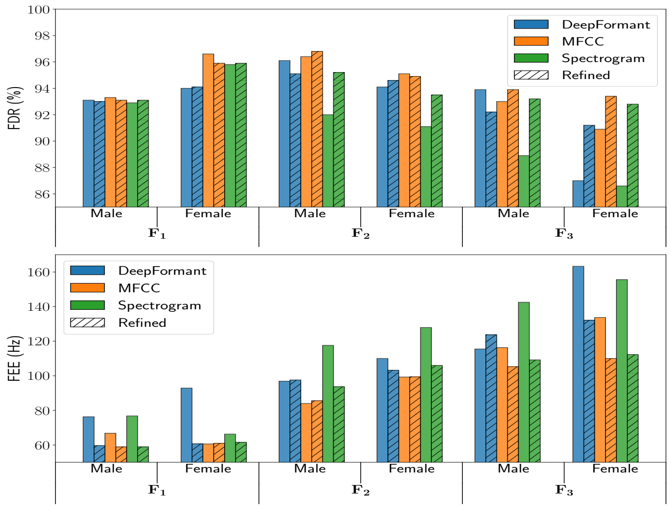

FIGURE 4: Formant tracking performance for male and female speakers, evaluated using different methods.

TABLE 2: Formant tracking results obtained using the comprehensive VTR test set for all formant trackers. Altogether 13 trackers are included consisting of five conventional trackers, two Deep Formant- trackers (DeepF and its refined version DeepF$_{\text{QCP-FB}}$), and six DNN-based trackers. FDR denotes formant detection rate and FEE denotes formant estimation error.

Method	FDR (%)			FEE (Hz)		
	F_1	F_2	F_3	δF_1	δF_2	δF_3
Conventional trackers [32]						
PBURG	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	61.9	145.8	250.3
Deep Formant trackers [24], [32]						
DeepF	91.7	92.3	89.7	85.1	119.6	142.8
DeepF$_{\text{QCP-FB}}$	93.9	94.4	92.6	60.4	103.2	118.6
Proposed DNN-based trackers (without and with the LP-based refinement)						
MFCC-DNN	94.4	96.0	92.3	64.7	89.2	122.1
MFCC-DNN$_{\text{QCP-FB}}$	94.0	96.2	93.7	59.6	90.3	106.9
SPEC-DNN	93.9	91.7	88.1	73.3	121.0	146.9
SPEC-DNN$_{\text{QCP-FB}}$	94.0	94.6	93.1	59.8	97.8	110.2
CSPEC-DNN	88.8	73.6	68.9	88.0	214.0	239.1
CSPEC-DNN$_{\text{QCP-FB}}$	93.9	91.6	88.0	59.9	115.1	146.8

rate (FDR) for all three formants (F_1, F_2, and F_3) is generally higher for the deep learning (DL)-based trackers compared to conventional signal processing methods. The only exception is the CSPEC-DNN tracker, which performs worse in formant detection compared to the other DL trackers. However, the remaining DL-based trackers demonstrate superior detection capabilities, showcasing their advantage over the conventional approaches. In terms of estimation error (FEE), the DL-based trackers once again exhibit better performance, with smaller estimation errors for all formants. Notably, while KARMA, the best conventional tracker, does well in estimating F_1—outperforming the default DL tracker (DeepF)—it falls behind in the estimation of F_2 and F_3. This suggests that KARMA is particularly strong in tracking F_1 but less effective for higher formants. The results in Table 2 indicate that MFCC-DNN$_{\text{QCP-FB}}$, one of the proposed combined trackers, achieves the best FDR and FEE values for all three formants among the 13 trackers compared, except for F1 (FDR) and F2 (FEE), where MFCC-DNN outperforms it. The proposed MFCC refinement (MFCC-DNN$_{\text{QCP-FB}}$) reduced estimation errors by 0.8 Hz, 12.9 Hz, and 11.7 Hz for F1, F2, and F3, respectively, compared to the Deep Formants refinement (DeepF$_{\text{QCP-FB}}$). When compared to the model-driven KARMA tracking method, the proposed MFCC-DNN$_{\text{QCP-FB}}$ refinement reduced estimation errors by 2.3 Hz, 55.5 Hz, and 143.4 Hz for F1, F2, and F3, respectively. MFCC-based formant trackers outperform those using spectrogram or complex spectrogram inputs primarily due to their reduced sensitivity to harmonics. While spectrograms preserve detailed harmonic information, which can cause DNN-based trackers to overfit, MFCCs focus on capturing the spectral envelope while attenuating harmonic influences. This characteristic makes MFCCs more effective for accurate and reliable formant tracking.

Based on the above overall results, we selected for the second experiment two best DNN trackers (MFCC-DNN and SPEC-DNN) and the best reference tracker (DeepF), all three with and without refinement, to be compared in tracking formants of female and male speech. The results of this second experiment are given in Table 3 (and also in Figure 6). The findings indicate that the detection of F_1 is more accurate for female speakers for all trackers, while the detection of F_2 and F_3 is more accurate for male speakers. This suggests a gender-based variation in formant detection accuracy for all the compared trackers, with F_1 being more reliably detected in female speakers and the higher formants, F_2 and F_3, being better identified in male speakers. In formant estimation error, male speakers show consistently lower errors for all three formants for all trackers. This trend implies that, regardless of the tracking method used, formant tracking is generally more accurate for male speakers. This is most likely due to the lower fundamental frequency (F0) of male speech, which results in harmonics that are more closely spaced. Consequently, when formant frequencies fall between these harmonics, the error introduced by selecting the nearest large-amplitude harmonic is smaller for male speakers compared to female (or child) speakers, whose harmonics are more widely spaced. This phenomenon leads to a more precise estimation of formant frequencies in male speech [36]. Among the different trackers, the DNN-based trackers, especially those using the MFCC feature and the refinement approach, stand out with superior performance in both formant detection and estimation error for both male

TABLE 3: Formant tracking results for the Deep Fromant-trackers and for the selected four DNN-based trackers separately for male and female speakers. The results are presented as averages across all utterances from both male and female speakers in the VTR test set.

	FDR (%)			FEE (Hz)		
Method	F_1	F_2	F_3	δF_1	δF_2	δF_3
Male						
DeepF	93.1	96.1	93.9	76.3	96.9	115.5
DeepF$_{\text{QCP-FB}}$	93.0	95.1	92.2	59.7	97.6	123.7
MFCC-DNN	93.3	96.4	93.0	66.8	84.0	116.3
MFCC-DNN$_{\text{QCP-FB}}$	93.1	96.8	93.9	58.9	85.6	105.3
SPEC-DNN	92.9	92.0	88.9	76.8	117.6	142.5
SPEC-DNN$_{\text{QCP-FB}}$	93.1	95.2	93.2	58.9	93.7	109.2
Female						
DeepF	94.0	94.1	87.0	92.9	110.0	163.3
DeepF$_{\text{QCP-FB}}$	94.1	94.6	91.2	60.7	103.2	132.1
MFCC-DNN	96.6	95.1	90.9	60.6	99.3	133.7
MFCC-DNN$_{\text{QCP-FB}}$	95.9	94.9	93.4	61.0	99.4	110.0
SPEC-DNN	95.8	91.1	86.6	66.3	127.9	155.6
SPEC-DNN$_{\text{QCP-FB}}$	95.9	93.5	92.8	61.6	106.0	112.3

and female speakers. These results confirm the advantage of the DNN-based formant tracking methods (either as such or in combination with the LP-based refinement approach) in providing more accurate and reliable formant measurements across genders.

In the third experiment, formant tracking performance was further investigated for fine-grained phonetic categories, focusing separately on female (8 speakers in the VTR test set) and male (16 speakers in the VTR test set), with the over-representation of males noted. The results, as shown in Tables 4 and 5, demonstrate a consistent trend for both genders. For vowels and diphthongs, the MFCC-DNN$_{\text{QCP-FB}}$ tracker achieves the highest formant detection rates and lowest estimation errors across all formants. For categories such as semivowels, nasals, and voice bars, MFCC-DNN$_{\text{QCP-FB}}$ still outperforms DeepF and SPEC-DNN, which exhibit slightly higher errors in some cases. However, fricatives & affricates, voice bars, and stops present more challenging conditions. In particular, male speakers show lower performance for these categories, both in formant detection and estimation errors, compared to female speakers. The performance further degrades for male speakers when using the combined MFCC-DNN$_{\text{QCP-FB}}$ approach. This could be due to the fact that fricatives & affricates include more aperiodic noise and turbulence, which degrades the capability of QCP-FB analysis to show clear all-pole spectral peaks, especially in male speech where formants are lower and closer together. It is also important to note that the manual formant tracking of the VTR data, which gives the ground-truth in our study, may not always be accurate. Since human annotators determined formant contours by visually inspecting spectrograms, there may be some deviation in the annotated data from the actual formant values [33], [36]. This subjectivity could introduce inconsistencies in the data, as seen in Fig. 3, where the ground-truth formants (marked by magenta lines) deviate from the spectral peaks detected by the LP method (marked by red lines). These deviations underscore the inherent challenge in establishing a precise ground-truth, especially for complex speech phonetic categories. Despite these challenges, the results indicate that incorporating QCP-FB into the MFCC-DNN and SPEC-DNN estimators generally improves formant tracking across many phonetic categories.

TABLE 4: Formant tracking results for different phonetic categories of the *female speakers* data of the VTR test set.

	FDR (%)			FEE (Hz)		
Method	F_1	F_2	F_3	δF_1	δF_2	δF_3
Vowels						
MFCC-DNN	97.0	94.6	91.1	57.1	102.9	131.6
MFCC-DNN$_{\text{QCP-FB}}$	96.7	94.8	94.0	58.2	99.6	102.7
SPEC-DNN	96.3	92.1	87.8	63.7	123.6	147.3
SPEC-DNN$_{\text{QCP-FB}}$	96.7	94.2	93.6	58.7	102.3	103.1
Diphthongs						
MFCC-DNN	97.2	97.2	93.7	62.7	87.1	117.0
MFCC-DNN$_{\text{QCP-FB}}$	96.7	97.2	97.0	63.2	87.9	89.0
SPEC-DNN	97.4	95.3	90.4	66.6	111.1	133.2
SPEC-DNN$_{\text{QCP-FB}}$	96.7	96.1	96.3	63.7	91.9	93.8
Semivowels						
MFCC-DNN	94.4	94.0	87.4	68.3	102.8	159.0
MFCC-DNN$_{\text{QCP-FB}}$	92.6	92.7	87.7	66.6	112.3	155.7
SPEC-DNN	92.1	83.1	78.8	73.5	159.9	205.7
SPEC-DNN$_{\text{QCP-FB}}$	92.6	88.4	86.5	67.5	132.9	160.5
Nasals						
MFCC-DNN	84.2	84.1	84.9	74.7	172.1	167.1
MFCC-DNN$_{\text{QCP-FB}}$	85.0	86.1	85.5	80.8	169.0	167.2
SPEC-DNN	84.9	74.0	80.4	74.2	226.1	185.8
SPEC-DNN$_{\text{QCP-FB}}$	85.6	83.4	84.4	80.0	189.4	173.2
Fricatives & Affricates						
MFCC-DNN	73.4	91.3	89.0	125.5	131.4	157.3
MFCC-DNN$_{\text{QCP-FB}}$	69.4	88.1	84.3	142.3	140.7	169.4
SPEC-DNN	69.3	88.0	78.8	135.4	154.7	191.3
SPEC-DNN$_{\text{QCP-FB}}$	69.0	88.1	81.6	141.4	144.0	180.8
Voice Bars						
MFCC-DNN	70.5	90.6	87.4	86.5	132.9	151.5
MFCC-DNN$_{\text{QCP-FB}}$	72.5	88.9	86.1	81.9	135.4	166.0
SPEC-DNN	70.5	81.4	84.3	80.8	185.8	169.9
SPEC-DNN$_{\text{QCP-FB}}$	71.5	84.4	85.0	86.0	159.2	172.9
Stops						
MFCC-DNN	75.9	87.1	85.3	131.3	145.6	164.3
MFCC-DNN$_{\text{QCP-FB}}$	74.5	87.3	84.9	130.8	153.1	165.7
SPEC-DNN	74.8	84.9	85.0	133.7	165.4	168.6
SPEC-DNN$_{\text{QCP-FB}}$	74.6	85.8	87.7	129.3	159.6	159.0

TABLE 5: Formant tracking results for different phonetic categories of the *male speakers* data of the VTR test set.

Method	FDR (%)			FEE (Hz)		
	F_1	F_2	F_3	δF_1	δF_2	δF_3
Vowels						
MFCC-DNN	94.0	97.9	94.4	62.7	76.9	108.6
MFCC-DNN$_{\text{QCP-FB}}$	93.9	97.6	94.5	54.7	80.4	98.0
SPEC-DNN	94.1	94.8	90.9	74.0	104.4	133.3
SPEC-DNN$_{\text{QCP-FB}}$	93.9	96.5	94.0	54.7	84.9	101.7
Diphthongs						
MFCC-DNN	95.5	97.1	94.4	67.1	76.2	107.3
MFCC-DNN$_{\text{QCP-FB}}$	94.3	97.5	95.4	58.2	80.6	95.8
SPEC-DNN	94.2	93.9	92.8	78.9	106.4	121.9
SPEC-DNN$_{\text{QCP-FB}}$	94.3	96.7	95.3	58.2	85.4	97.6
Semivowels						
MFCC-DNN	88.4	91.1	87.0	79.0	114.7	149.7
MFCC-DNN$_{\text{QCP-FB}}$	89.1	93.3	90.0	72.2	107.4	138.2
SPEC-DNN	87.8	81.6	78.4	82.6	170.3	194.0
SPEC-DNN$_{\text{QCP-FB}}$	89.1	89.3	88.5	72.4	129.9	145.3
Nasals						
MFCC-DNN	84.0	84.3	91.7	74.3	159.1	128.6
MFCC-DNN$_{\text{QCP-FB}}$	86.2	84.4	89.7	74.3	164.1	126.1
SPEC-DNN	84.4	73.3	88.9	77.6	225.0	153.0
SPEC-DNN$_{\text{QCP-FB}}$	86.6	77.0	88.8	73.2	211.7	139.2
Fricatives & Affricates						
MFCC-DNN	69.1	92.8	88.6	128.0	118.8	150.8
MFCC-DNN$_{\text{QCP-FB}}$	57.6	88.8	85.5	175.8	138.8	159.3
SPEC-DNN	66.0	85.5	82.3	135.2	162.7	178.5
SPEC-DNN$_{\text{QCP-FB}}$	57.0	85.5	82.7	175.4	154.6	169.5
Voice Bars						
MFCC-DNN	78.1	91.9	86.9	71.4	124.1	144.4
MFCC-DNN$_{\text{QCP-FB}}$	75.1	94.5	85.0	80.9	114.5	154.1
SPEC-DNN	76.0	79.6	78.7	76.7	182.6	201.2
SPEC-DNN$_{\text{QCP-FB}}$	74.7	87.0	79.2	81.2	149.4	195.6
Stops						
MFCC-DNN	67.4	93.0	87.5	137.3	118.9	151.5
MFCC-DNN$_{\text{QCP-FB}}$	66.1	90.6	83.6	144.7	130.7	166.8
SPEC-DNN	67.0	85.0	79.6	138.5	154.4	190.5
SPEC-DNN$_{\text{QCP-FB}}$	66.1	86.0	79.5	145.1	156.9	193.2

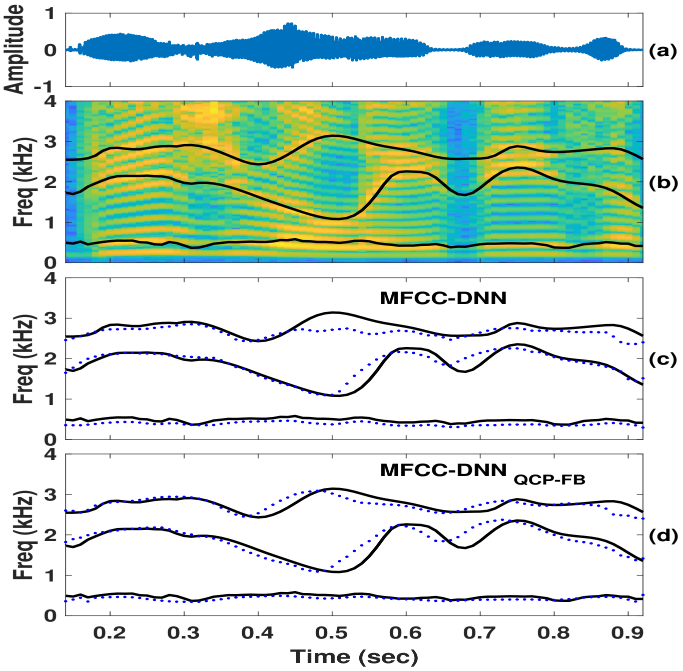

FIGURE 5: Formant frequencies for MFCC-DNN and MFCC-DNN$_{\text{QCP-FB}}$ for a phrase ("visually these") produced by a female speaker: (a) the time-domain speech signal, (b) the narrowband spectrogram with reference ground truth formants, (c) the formant track estimates of MFCC-DNN, and (d) the formant track estimates of MFCC-DNN$_{\text{QCP-FB}}$ along with ground truth formants. The improved performance of MFCC-DNN$_{\text{QCP-FB}}$ over MFCC-DNN is evident when comparing panels (c) and (d), particularly in tracking F_3 between 0.5 and 0.6 seconds.

An illustration of formant frequencies tracked by MFCC-DNN and MFCC-DNN$_{\text{QCP-FB}}$ for an utterance produced by a female speaker is shown in Fig. 5. The figure demonstrates that MFCC-DNN$_{\text{QCP-FB}}$ aligns more closely with the ground truth formant contours compared to the non-refined MFCC-DNN, highlighting the improvement in accuracy achieved by the proposed refinement process.

V. Conclusions

This paper presented a novel formant tracking technique that combines a modern DL-based formant estimation method (DNN) with a classical signal processing-based estimation method (QCP-FB). By refining the DNN-predicted formants using spectral peaks shown by the QCP-FB spectrum, the method effectively combines data-driven and model-driven approaches. Among the three spectral representations evaluated—MFCCs, spectrogram, and complex spectrogram—MFCCs yielded the best performance. The proposed DNN tracker outperformed seven reference trackers, including both signal processing and deep learning methods, and demonstrated improved performance across various phonetic categories and gender groups.

In terms of computational complexity, the proposed DNN architecture uses only 0.3M parameters, which is considerably smaller than the number of parameters in the LSTM-based Deep Formants tracker (4M parameters) [24]. Finally, we would like to point out that the proposed refinement approach based on QCP-FB is flexible because the refinement procedure can in principle by applied not only with a DNN as in the current study but with *any* trained DL-based formant estimation method. The proposed method requires further evaluation under challenging noisy conditions and across diverse databases to assess its generalizability and robustness. Furthermore, accurate formant tracking holds significant potential for atypical speech attributes in various speech disorders, making it a valuable tool for pathological speech analysis.

REFERENCES

[1] P. F. Assmann, “The role of formant transitions in the perception of concurrent vowels,” *J. Acoust. Soc. Am.*, vol. 97, no. 1, pp. 575–584, 1995.

[2] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, “Acoustic characteristics of American English vowels,” *J. Acoust. Soc. Am.*, vol. 97, no. 5, p. 3099–3111, May 1995.

[3] J. Rusz, R. Cmejla, T. Tykalova, H. Ruzickova, J. Klempir, V. Majerova, J. Picmausova, J. Roth, and E. Ruzicka, “Imprecise vowel articulation as a potential early marker of Parkinson’s disease: Effect of speaking task,” *J. Acoust. Soc. Am.*, vol. 134, no. 3, p. 2171–2181, September 2013.

[4] R. D. Kent, G. Weismer, J. F. Kent, H. K. Vorperian, and J. R. Duffy, “Acoustic studies of dysarthric speech: Methods, progress, and potential,” *J. Commun. Disord.*, vol. 32, no. 3, p. 141–186, May-June 1999.

[5] M. Kodali, S. R. Kadiri, and P. Alku, “Automatic classification of the severity level of parkinson’s disease: A comparison of speaking tasks, features, and classifiers,” *Computer Speech & Language*, vol. 83, p. 101548, 2024.

[6] H. K. Kathania, S. R. Kadiri, P. Alku, and M. Kurimo, “Study of formant modification for children asr,” in *ICASSP*. IEEE, 2020, pp. 7429–7433.

[7] A. Potamianos and S. Narayanan, “Robust recognition of children’s speech,” *IEEE Transactions on speech and audio processing*, vol. 11, no. 6, pp. 603–616, 2003.

[8] J. R. Schilling, R. L. Miller, M. B. Sachs, and E. D. Young, “Frequency-shaped amplification changes the neural representation of speech with noise-induced hearing loss,” *Hearing Research*, vol. 117, no. 1-2, pp. 57 – 70, 1998.

[9] I. C. Bruce, “Physiological assessment of contrast-enhancing frequency shaping and multiband compression in hearing aids,” *Physiological Measurement*, vol. 25, no. 4, pp. 945–956, 2004.

[10] T. Mellahi and R. Hamdi, “Lpc-based formant enhancement method in kalman filtering for speech enhancement,” *AEU-International Journal of Electronics and Communications*, vol. 69, no. 2, pp. 545–554, 2015.

[11] P. Boersma, “Praat, a system for doing phonetics by computer,” *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[12] K. Sjolander and J. Beskow, “Wavesurfer - An open source speech tool,” in *Proc. Int. Conf. Spoken Language Processing*, Beijing, China, October 2000, pp. 464–467.

[13] L. Deng, L. Lee, H. Attias, and A. Acero, “Adaptive Kalman filtering and smoothing for tracking vocal tract resonances using a continuous-valued hidden dynamic model,” *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 13–23, Jan. 2007.

[14] D. D. Mehta, D. Rudoy, and P. J. Wolfe, “Kalman-based autoregressive moving average modeling and inference for formant and antiformant tracking,” *J. Acoust. Soc. Am.*, vol. 132, no. 3, pp. 1732–1746, 2012.

[15] B. Story and K. Bunton, “Formant measurement in children’s speech based on spectral filtering,” *Speech Commun.*, vol. 76, p. 93–111, 2016.

[16] J. Makhoul, “Linear prediction: A tutorial review,” *Proc. IEEE*, vol. 63, pp. 561–580, Apr. 1975.

[17] A. V. Oppenheim and R. W. Schafer, “From frequency to quefrency: A history of the cepstrum,” *IEEE Signal Process. Mag.*, vol. 21, no. 5, p. 95–106, September 2004.

[18] T. Wang and T. Quatieri, “High-pitch formant estimation by exploiting temporal change of pitch,” *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, pp. 171–186, 2010.

[19] B. Yegnanarayana and R. Veldhuis, “Extraction of vocal-tract system characteristics from speech signals,” *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 4, pp. 313–327, Jul 1998.

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

[21] 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. Am.*, vol. 134, no. 2, pp. 1295–1313, 2013.

[22] 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, and Language Processing*, vol. 22, no. 3, pp. 596–607, March 2014.

[23] 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. Am.*, vol. 142, no. 3, pp. 1542–1553, 2017.

[24] Y. Dissen, J. Goldberger, and J. Keshet, “Formant estimation and tracking: A deep learning approach,” *J. Acoust. Soc. Am.*, vol. 145, no. 2, pp. 642–653, 2019.

[25] 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. Acoustics Speech and Signal Processing (ICASSP)*, Toulouse, France, 2006, pp. I369–I372.

[26] W. Dai, Z. Hua, J. Zhang, Y. Xie, and B. Li, “Gated bilinear networks for vowel formant estimation,” in *Proc. Int. Conf. on Asian Language Processing*, Kuala Lumpur, Malaysia, 2020, pp. 205–209.

[27] W. Dai, J. Zhang, Y. Gao, W. Wei, D. Ke, B. Lin, and Y. Xie, “Formant tracking using dilated convolutional networks through dense connection with gating mechanism,” in *Proc. Interspeech*, Shanghai, China, 2020,

pp. 150–154.

[28] J. Lilley and T. Bunnell, “Unsupervised training of a DNN-based formant tracker,” in *Proc. Interspeech*, Brno, Czech Republic, 2021, pp. 1189–1193.

[29] Y. Shrem, F. Kreuk, and J. Keshet, “Formant estimation and tracking using probabilistic heat-maps,” in *Proc. Interspeech*, Incheon, Korea, 2022, pp. 3563–3567.

[30] S. M. Kay and S. L. Marple, “Spectrum analysis – A modern perspective,” *Proc. IEEE*, vol. 69, no. 11, pp. 1380–1419, 1981.

[31] D. Gowda, B. Bollepalli, S. R. Kadiri, and P. Alku, “Formant tracking using quasi-closed phase forward-backward linear prediction analysis and deep neural networks,” *IEEE Access*, vol. 9, pp. 151 631–151 640, 2021.

[32] P. Alku, S. R. Kadiri, and D. Gowda, “Refining a deep learning-based formant tracker using linear prediction methods,” *Computer Speech & Language*, vol. 81, p. 101515, 2023.

[33] L. Deng, L. Lee, H. Attias, and A. Acero, “A structured speech model with continuous hidden dynamics and prediction-residual training for tracking vocal tract resonances,” in *ICASSP*, vol. 1, 2004, pp. I–557–60.

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

[35] K. Mustafa and I. C. Bruce, “Robust formant tracking for continuous speech with speaker variability,” *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 435–444, March 2006.

[36] C. H. Shadle, H. Nam, and D. Whalen, “Comparing measurement errors for formants in synthetic and natural vowels,” *J. Acoust. Soc. Am.*, vol. 139, no. 2, pp. 713–727, 2016.