INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Time-varying quasi-closed-phase weighted linear prediction analysis of speech for accurate formant detection and tracking

Dhananjaya Gowda and Paavo Alku

Dept. of Signal Processing and Acoustics, Aalto University, Finland
dhananjaya.gowda@aalto.fi, paavo.alku@aalto.fi

Abstract

In this paper, we propose a new method for accurate detection, estimation and tracking of formants in speech signals using time-varying quasi-closed phase analysis (TVQCP). The proposed method combines two different methods of analysis namely, the time-varying linear prediction (TVLP) and quasi-closed phase (QCP) analysis. TVLP helps in better tracking of formant frequencies by imposing a time-continuity constraint on the linear prediction (LP) coefficients. QCP analysis, a type of weighted LP (WLP), improves the estimation accuracies of the formant frequencies by using a carefully designed weight function on the error signal that is minimized. The QCP weight function emphasizes the closed-phase region of the glottal cycle, and also weights down the regions around the main excitations. This results in reduced coupling of the subglottal cavity and the excitation source. Experimental results on natural speech signals show that the proposed method performs considerably better than the detect-and-track approach used in popular tools like Wavesurfer or Praat.

Index Terms: Quasi-closed-phase (QCP) analysis, weighted linear prediction (WLP), time-varying linear prediction (TVLP), time-varying weighted linear prediction (TVWLP), time-varying quasi-closed phase (TVQCP) analysis

1. Introduction

Accurate tracking of formants in speech signals has potential applications in acoustic-phonetic analysis of speech signals, speech enhancement, formant-based speech synthesis, pronunciation correction [1–5]. Many algorithms of varying complexity have been proposed in the literature for tracking formants in speech signals [6–10]. A dynamic programming (DP) based tracking with a heuristic cost function on the initial formant candidates estimated using a conventional LP analysis is used in [6, 7]. An integrated approach towards tracking is adopted in [8–10] using state-space methods such as Kalman filtering (KF) and factorial hidden Markov model (FHMM). Most of these algorithms use an underlying linear prediction (LP) based modeling of speech signals, except in [10] which uses a non-negative matrix factorization (NMF) based source-filter modeling of speech signals.

Linear prediction (LP) analysis of speech signals is widely used to model the vocal tract system [11, 12]. Many refinements to the conventional LP analysis have been proposed such as temporally weighted linear prediction (WLP) and sparse linear prediction (SLP) for accurate modeling of the vocal tract as well as the excitation source [13–16]. However, the performance of these alternative LP models in tracking formants has not been studied extensively. The conventional LP analysis is still widely used in popular speech analysis tools for formant tracking [6, 7].

Temporally weighted LP algorithms give differential emphasis on the samples by defining a weight function on the error signal being minimized [13–15, 17–19]. Different weight functions have been proposed in the literature for WLP analysis based on different criterion and for different purposes. Weight functions that follow the short-time energy of the speech signal within a glottal cycle have been used to increase the robustness of the analysis against degradations [13,17,19]. Another weight function with an attenuated main excitation (AME) reduces the effect of glottal source on the vocal tract estimation [14]. A generalized AME weight function also known as the quasi-closed phase (QCP) weight function was proposed for accurate estimation of glottal source parameters by glottal inverse filtering [15].

The conventional least squares solution to the LP problem involves minimizing the L_2 norm of the prediction error signal with an inherent assumption that the excitation source signal is Gaussian process [20, 21]. Sparsity constraints based on the theory of compressed sensing may be used to utilize the super Gaussian nature of the excitation signal [16, 22]. This is achieved by approximating a non-convex L_0 norm optimization problem by a more tractable convex L_1 norm optimization [16]. Also, it has been shown that an iterative reweighted minimization of the norm can achieve increased sparsity of error signal and thereby yielding a solution more closer to L_0 norm optimization [22].

Speech signal is conventionally analyzed over short segments (5-50 ms) with an inherent assumption of quasi stationarity [11]. This conventional short-time analysis can only give a piecewise approximation to the slowly but continuously varying vocal tract system. Also, the conventional methods for tracking formants based on short-time LP analysis typically use a two-stage detect-and-track approach [6, 7]. Even the advanced formant tracking algorithms which directly track formants from the cepstral coefficients use this piecewise approximation of the vocal tract system [8, 9]. Time varying linear prediction (TVLP) tries to bridge this gap by modeling the speech signal over longer intervals of time by defining the vocal tract model parameters as a function of time [23–25].

In this paper, we propose a new time-varying quasi-closed phase (TVQCP) linear prediction analysis of speech signals for accurate modeling and tracking of the vocal tract resonances which integrates the advantages of temporally weighted LP, time varying LP and sparse LP.

2. Time-varying weighted linear prediction

Time-varying weighted linear prediction combines the ideas of sample selective prediction from weighted linear prediction and the time-continuity constraint from time-varying linear prediction.

Copyright © 2016 ISCA

http://dx.doi.org/10.21437/Interspeech.2016-153

2.1. Conventional least squares linear prediction

In conventional linear prediction (LP) analysis the current sample $x[n]$ is predicted as a linear weighted sum of the past p samples given by

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

where $\{a_k\}_{k=1}^{p}$ denotes the predictor coefficients. The predictor coefficients can be estimated as a solution to the convex optimization problem given by

$$\hat{\boldsymbol{a}} = \underset{\boldsymbol{a}}{\operatorname{argmin}} ||\boldsymbol{x} - \boldsymbol{X}\boldsymbol{a}||_m^m \tag{2}$$

$$\text{where} \quad \boldsymbol{x} = [x[0], x[1], \ldots, x[N-1]]_{N\times 1}^T \tag{3}$$

$$\boldsymbol{a} = [a_1, a_2, \ldots, a_p]_{p\times 1}^T \tag{4}$$

$$\boldsymbol{X} = [X_0, X_1, \ldots, X_{N-1}]_{N\times p}^T \quad \text{and} \tag{5}$$

$$X_n = [x[n-1], \ldots, x[n-p]]_{p\times 1}^T. \tag{6}$$

Here, N denotes the window length or the number of samples over which the predictor coefficients are optimized. Minimization of the L_2 norm of the error signal leads to the least square solution of the conventional LP analysis. However, a sparsity constraint imposed on the error signal is known to provide a better modeling of the excitation source and vocal tract system. This is achieved by minimizing the L_1 norm of the error signal which gives a convex approximation to the solution of an L_0 norm optimization problem, also referred to as sparse linear prediction (SLP) [16, 22].

2.2. Weighted linear prediction

Weighted linear prediction (WLP) differs from the conventional LP in the sense that it uses a sample selective prediction. It gives differential emphasis to different regions of the speech signal within a glottal cycle towards their contributions in estimating the predictor coefficients. This is achieved by minimizing a weighted error signal given by

$$\hat{\boldsymbol{a}} = \underset{\boldsymbol{a}}{\operatorname{argmin}}\, \boldsymbol{W}||\boldsymbol{x} - \boldsymbol{X}\boldsymbol{a}||_m^m \tag{7}$$

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

2.3. Time-varying linear prediction

Time-varying linear prediction (TVLP) is a generalization of the conventional LP analysis where the predictor filter coefficients are continuous functions of time. TVLP introduces a time-continuity constraint on the vocal tract (VT) system and is a better approximation of the slowly varying VT system than the piecewise constant quasi-stationary approximation used in the conventional LP analysis. The current sample is predicted using the past p samples as

$$\hat{x}[n] = \sum_{k=1}^{p} a_k[n] x[n-k] \tag{8}$$

where $a_k[n]$ denotes the k^{th} time-varying prediction filter coefficient at time instant n. Different approximations of $a_k[n]$ are possible using a power series or trigonometric series or Legendre polynomials. In this paper, we use a simple power series or polynomial approximation of q^{th} order given by

$$a_k[n] = \sum_{i=0}^{q} b_{k_i} n^i. \tag{9}$$

The TVLP coefficients are estimated by minimizing the L_m norm of the error signal and represented as a convex optimization problem given by

$$\hat{\boldsymbol{b}} = \underset{\boldsymbol{b}}{\operatorname{argmin}} ||\boldsymbol{x} - \boldsymbol{Y}\boldsymbol{b}||_m^m \tag{10}$$

$$\text{where} \quad \boldsymbol{x} = [x[0], x[1], \ldots, x[N-1]]_{N\times 1}^T \tag{11}$$

$$\boldsymbol{b} = [b_{1_0}, \ldots, b_{1_q}, \ldots, b_{p_0}, \ldots, b_{p_q}]_{p(q+1)\times 1}^T \tag{12}$$

$$\boldsymbol{Y} = [Y_0, Y_1, \ldots, Y_{N-1}]_{N\times p(q+1)}^T \quad \text{and} \tag{13}$$

$$Y_n = [x[n-1], nx[n-1], \ldots, n^q x[n-1],$$
$$\ldots, x[n-p], nx[n-p], \ldots, n^q x[n-p]]_{p(q+1)\times 1}^T. \tag{14}$$

Again, an L_2 or L_1 norm minimization leads to a least square solution or a sparse solution to the convex optimization problem, respectively [16, 22, 25].

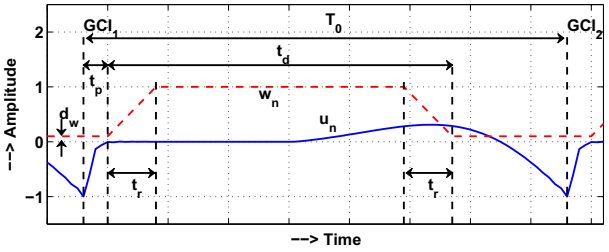

Figure 1: *QCP weight function w_n (dotted line) along with the LF glottal flow derivative signal u_n (solid line) for about one glottal cycle.*

2.4. Time-varying weighted linear prediction

Time varying weighted linear prediction (TVWLP) is analogous to WLP where the predictor coefficients are estimated by minimizing a weighted error signal given by

$$\hat{\boldsymbol{b}} = \underset{\boldsymbol{b}}{\operatorname{argmin}}\, \boldsymbol{W}||\boldsymbol{x} - \boldsymbol{Y}\boldsymbol{b}||_m^m \tag{15}$$

where $\boldsymbol{W}_{N\times N}$ is a diagonal matrix with its diagonal elements corresponding to the weight function w_n defined on the error signal. In this paper, we propose to use the QCP weight function within the TVWLP framework which provides a more accurate closed phase estimate of the vocal tract and also imposes a limited sparsity constraint on the excitation signal.

3. Time-varying quasi-closed phase analysis

Time-varying quasi-closed phase (TVQCP) analysis is a combination of QCP based WLP analysis and the time-varying linear prediction analysis. Quasi closed phase analysis of speech signals belongs to the family of weighted linear prediction (WLP) methods with a specially designed weight function based on the knowledge of GCIs [15].

The QCP analysis uses a weight function that combines the advantages of WLP, sparse LP, and AME weight function. The

Table 1: *Formant tracking performance in terms of FDRs and FEEs for different methods on natural speech data. The best scores are shown in boldface.*

Method	FDR (%)			FEE (%)		
	F_1	F_2	F_3	δF_1	δF_2	δF_3
PRAAT-BURG	85.8	72.0	64.8	17.8	21.3	16.8
MUST-AFB	81.1	87.6	74.0	20.2	9.5	12.4
WSURF-SCOV	90.1	92.7	79.5	15.0	8.6	14.2
TVLP-L2	88.9	92.7	88.0	14.7	8.7	6.5
TVLP-L1	90.4	94.1	90.2	14.1	8.3	6.0
TVQCP-L2	90.9	94.1	90.8	14.1	8.5	5.8
TVQCP-L1	**91.0**	**94.5**	**91.0**	**13.9**	**8.2**	**5.7**

weight function is designed so as to emphasize the closed phase region of the glottal cycle while at the same time attenuate the region around the main excitation [15]. Attenuation of the main excitation automatically imposes a limited sparsity constraint on the error signal and also reduces the influence of the excitation source on the vocal tract estimates. By defining a continuous weight function on the error signal the QCP analysis provides a flexible framework to approximate a closed phase analysis over multiple glottal cycles using either an autocorrelation-based or a covariance-based formulation.

An illustration of the QCP weight function w_n along with the Liljencrants-Fant (LF) glottal flow derivative signal u_n for about one glottal cycle is shown in Fig. 1. It 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 pitch period. A small non-zero value of deemphasis factor $d_w = 10^{-5}$ is used to avoid any possible singularities in the weighted autocorrelation matrices.

4. Formant tracking experiments

Performance of most formant tracking algorithms depends on two aspects: (1) the tracking algorithm used and (2) the underlying spectrum estimation method. In principle, most of the tracking algorithms can be combined with any underlying spectral representation including the one derived using a QCP based WLP analysis. However, the main focus of this paper is to study the improvements provided by TVLP and TVWLP over the popular two-stage detect-and-track approach that uses conventional LP. In view of this, we compare the performance of our proposed TVQCP method in formant tracking with that of the popular speech analysis tools Wavesurfer and Praat [6, 7].

4.1. Database

Performance of different methods in formant tracking is evaluated on natural speech signals using the vocal tract resonance (VTR) database [26]. The test data of the VTR database which has 192 utterances, 8 utterances each from 24 different speakers (8 female and 16 male), is 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 [27] is verified and corrected manually based on spectrographic evidence. All the speech data, originally recorded at 16 kHz sampling rate, is downsampled to 8 kHz before processing.

4.2. Performance metrics

Formant tracking performance of the methods is evaluated in terms of formant detection rate (FDR) and formant estimation error (FEE). Formant detection rate is measured in terms of the percentage of frames where a formant is hypothesized within a specified deviation from the ground truth. Formant estimation error is measured in terms of the average absolute deviation of the hypothesized formants from the ground truth. The FEE for a single frame of analysis and for the i^{th} formant is computed as $FEE_i = |F_i - \hat{F}_i|/F_i * 100$, where F_i is the reference ground truth and $\hat{F}_i$ is the hypothesized formant frequency.

4.3. Experiments and results

Based on our earlier experiments on formant tracking using synthetic speech signals we propose to use a window size of 100 ms, an LP order of 8, and a polynomial order of 3 for the TVLP and TVWLP methods [28]. A preemphasis filter of $P(z) = 1 - 0.97z^{-1}$ is used to preprocess the speech signals.

4.3.1. Comparison with other methods

Performance of the TVLP and TVQCP methods for different norm minimizations as compared to some of the popular formant tracking methods is given in Table 1. TVQCP-L2 and TVQCP-L1 denote the least squares and sparse TVQCP methods with L_2 and L_1 norm minimization, respectively. Similarly, TVLP-L2 and TVLP-L1 denote the least squares and sparse TVLP methods. WSURF-SCOV denotes the LP type-1 of Wavesurfer which uses a stabilized covariance analysis over 25 ms Hamming window. PRAAT-BURG denotes the Burg method of LP analysis with a 50 ms Gaussian-like window function. MUST-AFB denotes an adaptive filter-bank (AFB) based method proposed by Mustafa et al. [29].

It can be seen from Table 1 that the TVLP and TVQCP methods perform better than the popular methods using a two-stage detect-and-track approach. The improvement of performance in tracking (both FDRs and FEEs) the third formant by the time-varying methods is considerably high. It can be seen from the results for TVLP-L2 and TVQCP-L2, that use of QCP based WLP analysis seem to improve the performance of formant tracking. However, the improvement provided by QCP analysis on top of the sparsity constraint is only marginal, at least on the dataset used.

4.3.2. Effect of window length, LP order, and polynomial order

The effect of the choices for window size, LP order p, and the order of the polynomial q on the tracking performance is provided in Table 2. As the performance of TVQCP-L2 and

Table 2: *Effect of window length, LP order and polynomial order on formant tracking performance of the least-squares TVQCP method (TVQCP-L2). The best scores are shown in boldface.*

	FDR (%)			FEE (%)		
	F_1	F_2	F_3	δF_1	δF_2	δF_3
Win. length (N)	Effect of window length (p=8, q=3)					
50 ms	90.8	93.9	89.8	14.4	**8.2**	5.9
100 ms	**90.9**	**94.1**	**90.8**	**14.1**	8.5	**5.8**
200 ms	**90.9**	94.0	90.6	14.6	8.8	**5.8**
LP order (p)	Effect of LP order (N=100 ms, q=3)					
7	76.1	79.3	53.4	34.9	21.7	21.8
8	90.9	**94.1**	**90.8**	14.1	**8.5**	**5.8**
9	**91.8**	89.2	84.1	**13.0**	10.2	7.7
Poly. order (q)	Effect of polynomial order (N=100 ms, p=8)					
0	87.8	90.6	87.2	15.9	9.3	6.6
1	90.6	**94.5**	**91.4**	14.1	**8.3**	**5.6**
2	90.7	94.3	91.2	14.2	8.4	5.7
3	**90.9**	94.1	90.8	**14.0**	8.5	5.8

TVQCP-L1 are almost comparable the effects of varying these parameters are studied using only the TVQCP-L2 method. It can be seen that the performance of the TVQCP method is quite stable over a range of values for the window length, and the polynomial order of the predictor coefficients. However, the performance seems to be a bit sensitive to the choice of LP order, which needs further investigations.

5. Conclusions

In this paper, we proposed a time-varying quasi-closed phase linear prediction analysis for accurate tracking formants in natural speech signals. The proposed method combines the advantages of multi-cycle closed phase analysis of WLP, continuity constraints of TVLP, and the sparsity constraints of sparse LP. Formant tracking experiments on natural speech signals show that the TVQCP method performs better than the conventional two-stage detect-and-track approaches used in popular tools for speech analysis. The time-varying constraints on the vocal tract filter and the QCP analysis greatly improves our ability to track the third formant. Notwithstanding a fairly stable performance over different analysis window lengths and polynomial orders of the coefficients, the sensitivity of TVQCP to the choice of LP order needs to be addressed. Also, the robustness of the proposed method against additive noise and reverberations needs to be investigated.

6. Acknowledgements

This work has been funded by the Academy of Finland (project no. 256961 and 284671).

7. References

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

[2] N. B. Pinto, D. G. Childers, and A. L. Lalwani, "Formant speech synthesis: improving production quality," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 12, pp. 1870–1887, Dec 1989.

[3] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, "Acoustic characteristics of american english vowels," *The Journal of the Acoustical Society of America*, vol. 97, no. 5, 1995.

[4] L. Deng, D. Yu, and A. Acero, "Structured speech modeling," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1492–1504, Sept 2006.

[5] Y. Iribe, S. Manosavan, K. Katsurada, R. Hayashi, C. Zhu, and T. Nitta, "Improvement of animated articulatory gesture extracted from speech for pronunciation training," in *Proc. Int. Conf. Acoustics Speech and Signal Processing*, March 2012, pp. 5133–5136.

[6] 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.

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

[8] 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 Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 13–23, 2007.

[9] D. D. Mehta, D. Rudoy, and P. J. Wolfe, "Kalman-based autoregressive moving average modeling and inference for formant and antiformant trackinga)," *The Journal of the Acoustical Society of America*, vol. 132, no. 3, 2012.

[10] J. L. Durrieu and J. P. Thiran, "Source/filter factorial hidden markov model, with application to pitch and formant tracking," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 12, pp. 2541–2553, Dec 2013.

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

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

[13] 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.

[14] P. Alku, J. Pohjalainen, M. Vainio, A.-M. Laukkanen, and B. H. Story, "Formant frequency estimation of high-pitched vowels using weighted linear predictiona)," *The Journal of the Acoustical Society of America*, vol. 134, no. 2, 2013.

[15] M. Airaksinen, T. Raitio, B. Story, and P. Alku, "Quasi closed phase glottal inverse filtering analysis with weighted linear prediction," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, no. 3, pp. 596–607, March 2014.

[16] D. Giacobello, M. Christensen, M. Murthi, S. Jensen, and M. Moonen, "Sparse linear prediction and its applications to speech processing," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 5, pp. 1644–1657, July 2012.

[17] J. Pohjalainen, H. Kallasjoki, K. J. Palomäki, M. Kurimo, and P. Alku, "Weighted linear prediction for speech analysis is noisy conditions," in *Proc. Interspeech*, Brighton, UK, September 2009.

[18] R. Saeidi, J. Pohjalainen, T. Kinnunen, and P. Alku, "Temporally weighted linear prediction features for tackling additive noise in speaker verification," *IEEE Signal Processing Letters*, vol. 17, no. 6, pp. 599–602, 2010.

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

[20] D. Wong, J. Markel, and J. Gray, A., "Least squares glottal inverse filtering from the acoustic speech waveform," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 27, no. 4, pp. 350–355, Aug 1979.

[21] S. M. Kay, *Modern Spectrum Estimation: Theory and Application*. Prentice Hall NJ, USA, 1988.

[22] D. Wipf and S. Nagarajan, "Iterative reweighted l_1 and l_2 methods for finding sparse solutions," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 4, no. 2, pp. 317–329, April 2010.

[23] M. G. Hall, A. V. Oppenheim, and A. S. Willsky, "Time-varying parametric modeling of speech," *Signal Processing*, vol. 5, no. 3, pp. 267 – 285, 1983.

[24] K. Schnell and A. Lacroix, "Time-varying linear prediction for speech analysis and synthesis," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, March 2008, pp. 3941–3944.

[25] S. Chetupalli and T. Sreenivas, "Time varying linear prediction using sparsity constraints," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 6290–6293.

[26] 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*, Toulouse, France, 2006, pp. I–369–I–372.

[27] L. Deng, L. Lee, H. Attias, and A. Acero, "A structured speech model with continuous hidden dynamics and prediction-residual training for tracking vocal tract resonances," in *Proc. Int. Conf. Acoustics Speech and Signal Processing*, vol. 1, May 2004, pp. I–557–60 vol.1.

[28] D. Gowda, M. Airaksinen, and P. Alku, "Quasi closed phase analysis of speech signals using time varying weighted linear prediction for accurate formant tracking," in *Proc. Int. Conf. Acoustics Speech and Signal Processing*, Shanghai, China, 2016, pp. 4980–4981.

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