INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Identification of F1 and F2 in speech using modified zero frequency filtering

RaviShankar Prasad, Mathew Magimai.-Doss

Idiap Research Institute, Martigny CH-1920, Switzerland
{ravi.prasad, mathew}@idiap.ch

Abstract

Formants are major resonances in the vocal tract system. Identification of formants is important for study of speech. In the literature, formants are typically identified by first deriving formant frequency candidates (e.g., using linear prediction) and then applying a tracking mechanism. In this paper, we propose a simple tracking-free formant identification approach based on zero frequency filtering. More precisely, formants F1-F2 are identified by modifying the trend removal operation in zero frequency filtering and picking simply the dominant peak in the short-term discrete Fourier transform spectra. We demonstrate the potential of the approach by comparing it against state-of-the-art formant identification approaches on a typical speech data set (TIMIT-VTR) and an atypical speech data set (PC-GITA).
Index Terms: Formant identification, Zero frequency filtering, Speech analysis, atypical speech

1. Introduction

Speech is produced by excitation of a time varying vocal tract system by a time varying excitation signal. Formants are the major resonances in the vocal tract system. Identification of formants is interesting from the perspective of study of speech as well as technology development [1, 2, 3, 4]. Formant identification typically involves two steps: first step involves identification of formant candidates from a reliable spectral estimate, and second step selects an estimate from the candidate set based on continuity and optimality criteria.

The most popular approaches to derive candidate formant frequency locations are linear prediction (LP) spectra [5] and cepstrally smoothed spectra. One of the limitations with LP is to resolve closely appearing spectral peaks, or masking of poles due to spectral harmonics or nulls. To circumvent that, methods such as, pitch synchronous covariance formulation [6] and temporal weighting of samples [7] have been proposed. More recently, it has been shown that use of a weighting function based on the glottal flow signal, followed by a time–varying LP analysis with sparsity constraints, resulting in improved formant candidates [8]. Beside LP or cepstral based methods, AM–FM decomposition based methods have been proposed to generate formant candidates. Method based on group delay are proposed to extract formant candidates in shorter segments of speech [9].

The candidate derivation step is followed by a tracking mechanism which ensures smooth formant contours in speech. Continuity constraints ascertain smooth formant contours and help alleviate tracking errors, especially in phone transition regions. There are different approaches that have been employed for tracking formants such as, (a) dynamic programming-based [10, 11], (b) hidden Markov model-based [12], (c) Kalman filtering-based [13], (d) t-distribution-based [14] and (e) neural networks-based [15, 16, 17].

Zero frequency filtering (ZFF) is a signal processing approach where the signal is filtered through a heavily damped resonator centered at 0 Hz. This approach has orginally been developed in the context of extracting voice source characteristics. In this paper, we extend this approach for formant identification. Specifically, we show that by modifying the trend removal window duration formant frequencies can be highlighted and identified by peak picking in the short-term discrete Fourier transform (DFT) spectra of the filtered signal. We evaluate the proposed approach against the state-of-the-art approaches on TIMIT-VTR dataset and PC-GITA dataset. It is worth mentioning that modification in the trend removal step has previously been employed to study excitation source characteristics in expressive voices and non–verbal sounds [18]. To the best of our knowledge, this is the first work that shows that F1-F2 can be estimated by modifying the trend removal window duration in ZFF.

The remainder of the paper is organized as follows. Section 2 provides a background on ZFF. Section 3 presents the proposed approach based on ZFF. Section 4 presents the datasets and performance metrics adopted to evaluate the proposed method. Section 5 presents the results. Finally, Section 6 concludes the paper.

2. Zero frequency filtering

The zero frequency filter is implemented as a heavily decaying digital resonator centered at 0 Hz [19]. The ZFF method was originally proposed to identify location of glottal closure instants (GCIs) in the vibrating vocal fold source signal in speech. The underlying motivation of the method being that the spectral characteristics of the temporal discontinuities are evenly spread across all bands, including very low frequencies such as in the vicinity of 0 Hz. The contribution of the vocal tract system response is negligible at frequencies near the 0 Hz. The signal when filtered using the resonator, exhibits a source dominant behaviour, with the GCIs located at the zero crossing locations in the output.

The zero frequency filter is implemented as a cascaded resonators centered at 0 Hz. The impulse response of the filter is $x[n]$=$s[n] + 2x[n-1] - x[n-2]$, and the equivalent transfer function $H(z)$ is given by,

$$H(z) = \frac{1}{1 - 2z^{-1} + z^{-2}}, \tag{1}$$

where $s[n]$ is the input to the resonator and $x[n]$ is the filtered signal output. The zero frequency filtering is implemented as an integrator and therefore $x[n]$ shows a trend of polynomial growth with time. The trend in $x[n]$ is removed using a local mean removal operation across a duration comparable to the pitch period, given by

$$y[n] = x[n] - \frac{1}{2N+1} \sum_{k=n-N}^{n+N} x[k]; \quad N+1 \leq n \leq L-N, \tag{2}$$

http://dx.doi.org/10.21437/Interspeech.2021-1598

where L is the net length of the signal $x[n]$, and $2N+1$ is the trend removal window duration.

3. Proposed method

In this section, we first provide a theoretical understanding of modification in the trend removal step of ZFF in Sec. 3.1. We then present the proposed formant identification method based on ZFF in Section 3.2.

3.1. Modification in trend removal operation

The trend removal operation imposes its response over the heavily decaying response of the ZFF signal. The cumulative response of both these operations results in a peak at the fundamental frequency of the signal, while other components stay suppressed, when the trend removal duration is comparable to an estimate of the pitch period. The output $y(n)$ of the ZFF method therefore is periodic with each cycle resulting from source excitation.

The frequency domain response $H(\omega)$ of the trend removal operation (Eqn. 2) is given by,

$$H(\omega) = 1 - \frac{1}{2N+1} \sum_{k=-N}^{N} e^{-j\omega k} \tag{3}$$

$$H(\omega) = 1 - \frac{1}{M} \left\{ \frac{sin(\frac{\omega M}{2})}{sin(\frac{\omega}{2})} \right\}, \tag{4}$$

where $M = (2N+1)$ is the trend removal duration. $H(\omega)$ of the trend removal filter given in Eqn. 4 is an inverted moving average filter. Fig. 1 shows the spectral response $H(\omega)$ for different values of M (order). It can be observed that there are peaks and nulls in $H(\omega)$, which varies with the choice of M. $H(\omega)$ when imposed on to the heavily decaying response of the ZFF, results in normalization of the high gain of the zero frequency filter at 0 Hz and an emphasis of the first lobe of the inverted MA response and decay of other components. The dominant peak location and bandwidth in the response of the filter is in inverse proportion to the window duration. A smaller duration shifts the peak response towards mid to higher frequency range with a higher bandwidth, whereas a longer window duration shifts it closer to the origin with a sharper bandwidth.

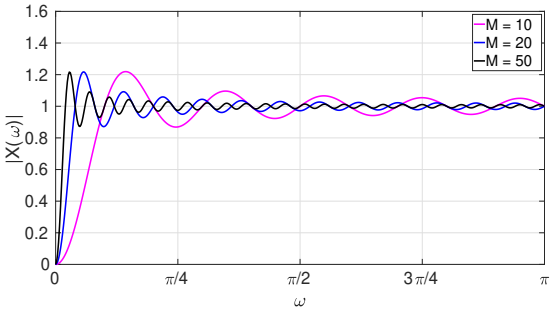

Figure 1: *Spectral response of the trend removal window. The response is plotted for window lengths of* 10, 20 *and* 50 *samples.*

3.2. F1–F2 identification using ZFF

In the case of voice source information extraction, as done in the original work [19], the trend removal duration $2N+1$ is comparable to the pitch period in the signal. This results in a sharper peak or emphasis in the range of fundamental frequency in speech. However, as shown in the previous section, the emphasis can be altered by changing M. For almost all voiced sounds, the first formant in sounds usually occurs beyond the first harmonic. Hence, modifying the trend removal window duration to 25% of the estimated fundamental period (T_0) results in shift in the peak response in the range of first formant frequency region. Similarly, a smaller window duration than that (e.g., $\sim 10-12\%$ of the pitch period) results in shifting the peak response within the range of the second formant.

Fig. 2 illustrates that aspect. Fig. 2(a) shows the spectrogram of a speech signal corresponding to an utterance of sustained vowel /i/. Fig. 2(b) shows the spectrogram of the filtered signal obtained from ZFF with a trend removal window size of $T_0/5$. It can be observed that the first formant region around 0.3–0.4 kHz is emphasized clearly. The higher frequency regions are de-emphasized due to the sharp decay introduced by the cascaded resonator at 0 Hz in the zero frequency filter. Fig. 2(c) shows the spectrogram of the the filtered signal obtained from ZFF with a trend removal window size of $T_0/10$. It can be observed that the second formant region around 2–2.5 kHz is well emphasized compared to other frequency regions including the high energy first formant region.

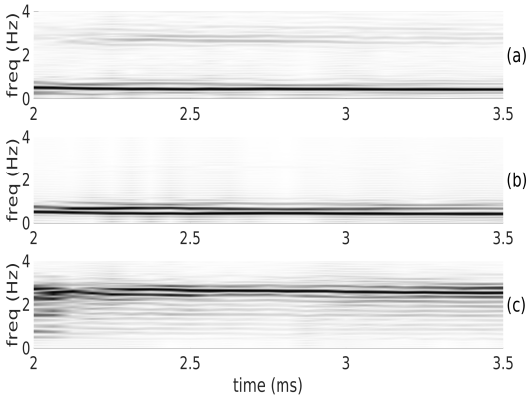

Figure 2: *Spectrograms of (a) signal corresponding utterance containing vowel /i/, (b) the filtered signal output after ZFF with trend removal duration of $T_0/5$, and (c) the filtered signal output after ZFF with trend removal duration of $T_0/10$.*

Based on this understanding, Algorithm 1 presents a formant identification method using ZFF.

Algorithm 1 F1–F2 identification using ZFF

1: Compute ZFF signal $x(n)$ from $s(n)$ by applying Eqn. 1.
2: Compute an estimate of fundamental periodicity of the signal (T_0) using autocorrelation.
3: Using a trend removal duration $2N+1 = T_0/5$, filter $x(n)$ to obtain $y_1(n)$.
Obtain $y_2(n)$ from $x(n)$ using $2N+1 = T_0/10$.
4: Compute short-term DFT spectra $Y_1(\omega)$ and $Y_2(\omega)$ for $y_1(n)$ and $y_2(n)$ over a desired resolution.
5: Identify the global peaks in $Y_1(\omega)$ and $Y_2(\omega)$ as $F1$ and $F2$ locations, respectively.

4. Experimental setup

Section 4.1 presents the data sets used to validate the proposed formant identification approach. Section 4.2 briefly presents different baseline formant extraction approaches. Finally, Section 4.3 presents the evaluation measures.

4.1. Datasets

We investigated the proposed approach on two data sets. TIMIT-VTR corpus that contains typical speech and PC-GITA Spanish language dataset that contains atypical speech, more precisely, from Parkinson's disease (PD) patients. The motivation being that PD patients speech result in the expansion/shrinkage in the vowel space area (VSA) owing to difficulties in exercising regular motor control. Formants in speech corresponding to PD are therefore more spread as compared to controlled speech [20]. So, whether the proposed approach is able to capture such differences?

TIMIT Vowel tract resonance (TIMIT–VTR) dataset [21]: The TIMIT–VTR dataset contains manually annotated formant ($F1$, $F2$, and $F3$) locations and their respective bandwidths, over a subset of utterances in the TIMIT corpus. Location and bandwidth values for the 3 formants are manually annotated for these utterances. The fourth formant is annotated using an automatic tracking algorithm [22]. The values are computed every 10 ms over a frame width of 10 ms. The third and fourth formants are not used for evaluation in this study. The TIMIT–VTR dataset contains a total of 376 sentences, recorded by 173 male and female speakers, grouped into 346 utterances for training and 192 utterances for testing purpose.

PC-GITA Spanish language dataset [23]: The dataset consists of a Spanish speech signal recording of 50 speakers with Parkinson's disease (PD) and 50 healthy control speakers. The dataset includes recordings of 25 male and 25 female speakers. The dataset belongs to Spanish language and contains recording of vowels, monologues and read text. Each recording consists of a variety over phonation, prosody and articulation. The proposed method is evaluated over sustained vowel segments recorded by control and PD speakers corresponding to the vowels /æ/, /i/ and /u/, which effectively illustrate the contrast in articulation dynamics for control and PD speakers [24].

4.2. Baseline formant extraction methods

We validate the proposed approach by comparing it against different formant extraction methods, namely, (a) Wavesurfer [11], (b) Praat [10], (c) KARMA method [13], (d) MUST method [25] and (e) time–varying quasi closed phase analysis (TVQCP) method [26]. The KARMA method applies Kalman filtering based state-space optimization routine over the auto–regressive moving average (ARMA) cepstral coefficients for formant identification. The MUST methods filters the analytic signal using adaptive band–pass filters centered around the estimate of formants. A dynamic tracking filter (DTF) is used to update the estimates of location of pole in the formant region while tracking over previous values. TVQCP is based on pitch synchronous processing to derive system response using weighted LP analysis, with the LP coefficient optimization routine acting as a tracking mechanism.

4.3. Performance metrics

Formants values are derived across non–overlapping frames of duration 10 ms, and F1 and F2 values derived over voiced regions are used for comparison, across different methods. As done in the literature [26], on TIMIT-VTR dataset the proposed method and baseline methods are evaluated based on the formant estimation error (FEE) and the formant detection rate (FDR). The FEE computed for the formant $i \in \{1, 2\}$, across a segment of N analysis frames, given by

$$FEE_i = \frac{1}{C}\sum_{c=1}^{C} |F_i^D(c) - F_i^G(c)|, \tag{5}$$

gives the average absolute deviation,where $F_i^D(c)$ and $F_i^G(c)$ are the estimated and the annotated values for the formant i for frame c, respectively and C is the number of frames.

FDR gives the proportion of frames with formants detected correctly. This is done as follows:

$$d(c) = 1 \text{ if } |F_i^D(c) - F_i^G(c)| \leq B_i^G(c) \text{ else } 0, \tag{6}$$

where $B_i^G(c)$ is the annotated or ground truth bandwidth value of formant i at frame c. The FDR is finally obtained as,

$$\frac{\sum_{c=1}^{C} d(c)}{C} \times 100. \tag{7}$$

In the case of PC-GITA, we analyze the different methods by estimating the VSA.

5. Results and analysis

Following section discusses the performance of the proposed method over typical and atypical speech.

5.1. TIMIT–VTR

We evaluated the proposed method and baseline methods on segments corresponding to vowels, nasals, and diphthongs, across all speakers. The corresponding segment boundaries are obtained from the annotations provided in the TIMIT dataset. There are a total of 10569 segments in the VTR dataset for evaluation.

Fig. 3 shows the formant contours derived using the proposed method, along with the annotated values, across vowel and nasal segments in an utterance obtained from the VTR dataset. The ground truth (●), are plotted along with the derived values (●) for F1 (Fig. 3(a)) and F2 (Fig. 3(b)) contours in the spectrographic representation of the segment. The ability of the proposed method to track the dynamic variation of first and second formants can be noted in the derived F1 and F2 contours. The figure illustrates the that formant information thus derived is independent of boundary conditions or continuity constraints.

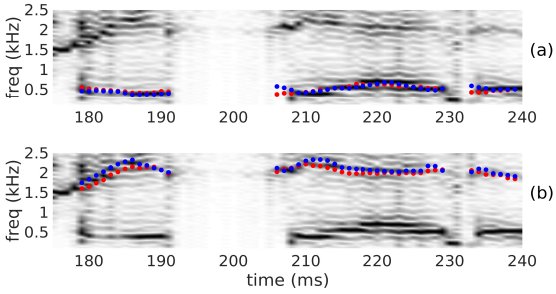

Figure 3: *Tracking of (a) $F1$ and (b) $F2$ peaks in natural speech (ground truth (· · ·), derived values (· · ·)).*

Tab. 1 presents the results obtained over the TIMIT–VTR dataset using different methods. It can be observed that FEE_1 value obtained using the proposed method, denoted as modZFF, is comparable to or better than the FEE_1 value obtained using the baseline methods. On the other hand, FEE_2 is lower than Wavesurfer; comparable to Praat; and higher than KARMA, MUST and TVQCP methods. Having said that, a high FDR_2 value signifies that the derived formant peaks are within the bandwidth of the second formant.

Table 1: *FEE (in Hz) and FDR (in %) for F1 and F2 obtained on the TIMIT-VTR database.*

Method	FDR_1	FDR_2	FEE_1	FEE_2
modZFF	97	87	70	185
KARMA	97	90	67	130
MUST	89	87	100	135
TVQCP	95	93	70	116
Wavesurfer	88	65	130	265
Praat	81	75	140	192

5.2. PC–GITA

We analyzed the sustained vowel segments of vowels /æ/, /i/, and /u/ uttered by healthy control speakers and PD patients by estimating the VSA (vowel space area) using the proposed modZFF method and the baseline methods. Fig. 4 shows the distribution of F1 values against F2 values obtained for sustained vowel segments. Figs. 4(a1)–(e1) show the F1 and F2 values obtained using the proposed method, KARMA, TVQCP, and MUST methods, for control speakers respectively. Formant contours are smoothed using a 5–point median filtering. The formants for vowels /i/ (●), /æ/ (●), and /u/ (●), appear in non–overlapping distinct clusters. The VSA can be noted across different clusters. The centroids of the clusters are marked and connected (●–). The axes for figures are not equal and hence the distribution appears skewed. Figs. 4(a2)–(e2) show the F1 and F2 values obtained using the different methods for PD patients speech. The clusters for all the methods appear expanding, and the VSA also changes. It can also be observed that the clusters for other methods appear in tighter bounds, which could be attributed to the tracking routine. The proposed method does not employ such a routine and hence the clusters illustrate the formant dynamics in an unconstrained manner. Whether this difference is desirable or undesirable is open for future research.

Tab. 3 gives the VSA (in Hz) for control (VSA_C) and PD (VSA_P) speakers. It can be seen that in all cases the VSA increases for PD patient's speech. The proportion increase of VSA for modZFF, KARMA and TVQCP is similar, while for MUST and Praat it is slightly more.

6. Conclusion

This paper presented a method for F1-F2 identification based on ZFF. It was shown that by modifying the trend removal window duration, F1 and F2 contours can be obtained by simply picking the dominant peaks in the short-term DFT spectra. Beside not requiring any post-processing or a tracking method, a distinguishing aspect of the proposed method is that, unlike the methods proposed in the literature, this approach does not requires source-system decomposition or AM-FM decomposition. Our investigations on TIMIT-VTR data sets showed that the proposed approach can reliably identify F1–F2. Analysis on

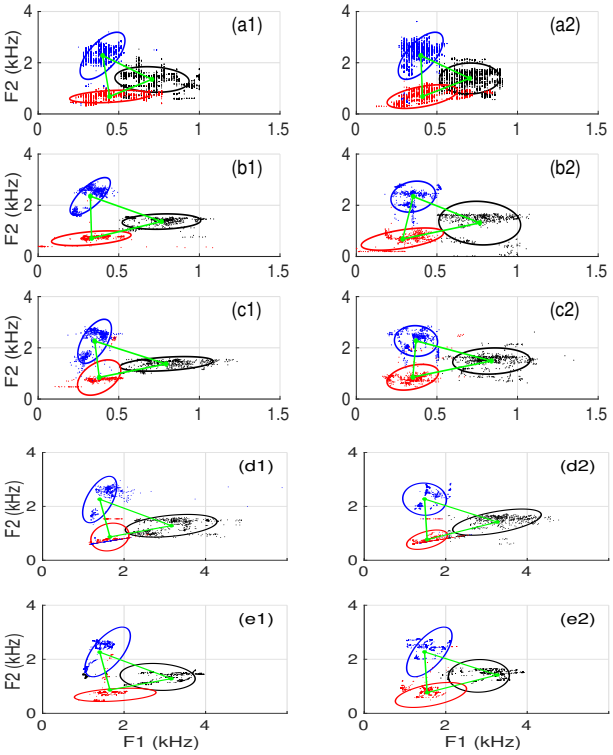

Figure 4: F1 vs. F2 values for different vowels (/u/,/i/,/æ/) in control and PD speakers, obtained using (a1, a2) proposed method, (b1, b2) KARMA, (c1, c2) MUST, (d1, d2) TVQCP, and (e1, e2) Praat methods, respectively.

Table 3: *VSA (Hz^2) for control and PD speakers in PC–GITA for /i/,/u/, and /æ/.*

Method	modZFF	KARMA	MUST	TVQCP	Praat
VSA_C	448.3	716.7	601.9	436.6	582.1
VSA_P	474.7	751.8	678.3	475.2	658.2

PC-GITA studies revealed that the proposed F1–F2 identification method can be applied to study pathological speech similar to the baseline methods. Our future work will focus on combining the proposed F1–F2 estimation capability with voice source information extraction capability of ZFF to assess PD patient's speech.

7. Acknowledgement

This work was funded by the Swiss National Science Foundation (SNSF) through the project Towards Integrated processing of Physiological and Speech signals (TIPS), grant no. 200021_188754. The authors would like to thank Dr. Sudarsana Reddy, Aalto University, for providing the implementation for KARMA, MUST and TVQCP methods, also used in [26].

8. References

[1] A. Suni, T. Raitio, M. Vainio, and P. Alku, "The glotthmm speech synthesis entry for blizzard challenge 2010," in *The Blizzard Challenge 2010 workshop*. Language Technologies Institute LTI, 2010, pp. 1–6.

[2] T.-C. Zorila, V. Kandia, and Y. Stylianou, "Speech-in-noise in-

telligibility improvement based on spectral shaping and dynamic range compression," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[3] G. S. Bhat, C. K. Reddy, N. Shankar, and I. M. Panahi, "Smartphone based real-time super gaussian single microphone speech enhancement to improve intelligibility for hearing aid users using formant information," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018, pp. 5503–5506.

[4] I.-C. Yoo, H. Lim, and D. Yook, "Formant-based robust voice activity detection," *IEEE/ACM Transactions on audio, speech, and language Processing*, vol. 23, no. 12, pp. 2238–2245, 2015.

[5] S. McCandless, "An algorithm for automatic formant extraction using linear prediction spectra," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 22, no. 2, pp. 135–141, 1974.

[6] B. Yegnanarayana and R. N. Veldhuis, "Extraction of vocal-tract system characteristics from speech signals," *IEEE transactions on Speech and Audio Processing*, vol. 6, no. 4, pp. 313–327, 1998.

[7] P. Alku, J. Pohjalainen, M. Vainio, A.-M. Laukkanen, and B. H. Story, "Formant frequency estimation of high-pitched vowels using weighted linear prediction," *The Journal of the Acoustical Society of America*, vol. 134, no. 2, pp. 1295–1313, 2013.

[8] 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 *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4980–4984.

[9] J. M. Anand, S. Guruprasad, and B. Yegnanarayana, "Extracting formants from short segments of speech using group delay functions," in *Ninth International Conference on Spoken Language Processing*, 2006.

[10] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer (version 5.3. 82)[computer software]," *Amsterdam: Institute of Phonetic Sciences*, 2012.

[11] K. Sjölander and J. Beskow, "Wavesurfer-an open source speech tool," in *Sixth International Conference on Spoken Language Processing*, 2000.

[12] M. Lee, J. Van Santen, B. Mobius, and J. Olive, "Formant tracking using context-dependent phonemic information," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 741–750, 2005.

[13] D. D. Mehta, D. Rudoy, and P. J. Wolfe, "Kalman-based autoregressive moving average modeling and inference for formant and antiformant tracking," *The Journal of the Acoustical Society of America*, vol. 132, no. 3, pp. 1732–1746, 2012.

[14] H. Sundar, C. S. Seelamantula, and T. V. Sreenivas, "A mixture model approach for formant tracking and the robustness of student's-t distribution," *IEEE transactions on audio, speech, and language processing*, vol. 20, no. 10, pp. 2626–2636, 2012.

[15] Y. Dissen and J. Keshet, "Formant estimation and tracking using deep learning." in *INTERSPEECH*, 2016, pp. 958–962.

[16] Y. Dissen, J. Goldberger, and J. Keshet, "Formant estimation and tracking: A deep learning approach," *The Journal of the Acoustical Society of America*, vol. 145, no. 2, pp. 642–653, 2019.

[17] 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," *arXiv preprint arXiv:2005.10803*, 2020.

[18] S. R. Kadiri and B. Yegnanarayana, "Epoch extraction from emotional speech using single frequency filtering approach," *Speech Communication*, vol. 86, pp. 52–63, 2017.

[19] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1602–1613, 2008.

[20] J. R. Orozco-Arroyave, J. C. Vásquez-Correa, J. F. Vargas-Bonilla, R. Arora, N. Dehak, P. S. Nidadavolu, H. Christensen, F. Rudzicz, M. Yancheva, H. Chinaei *et al.*, "Neurospeech: An open-source software for parkinson's speech analysis," *Digital Signal Processing*, vol. 77, pp. 207–221, 2018.

[21] L. Deng, X. Cui, R. Pruvenok, J. Huang, S. Momen, Y. Chen, and A. Alwan, "A database of vocal tract resonance trajectories for research in speech processing," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 1. IEEE, 2006, pp. I–I.

[22] 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 *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 2004, pp. I–557.

[23] J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla, M. C. Gonzalez-Rativa, and E. Nöth, "New spanish speech corpus database for the analysis of people suffering from parkinson's disease." in *LREC*, 2014, pp. 342–347.

[24] P. Gómez, J. Mekyska, A. Gómez, D. Palacios, V. Rodellar, and A. Álvarez, "Characterization of parkinson's disease dysarthria in terms of speech articulation kinematics," *Biomedical Signal Processing and Control*, vol. 52, pp. 312–320, 2019.

[25] 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, 2006.

[26] 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 Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1901–1914, 2020.