Energy Separation Algorithm Based Spectrum Estimation for Very Short Duration of Speech

Hemant A. Patil[1], *Srikant Viswanath*[2]

[1]Speech Research Lab, DA-IICT, Gandhinagar
[2]Ex-student, Speech Research Lab, DA-IICT, Gandhinagar
hemant_patil@daiict.ac.in,srikantviswanath@hotmail.com

Abstract

In this paper, we propose a novel method of estimation of short-time spectrum for analysis of speech signals in the closed phase regions of glottal activity. This method uses Teager Energy Operator (TEO) and a related Energy Separation Algorithm (ESA) iteratively, along with the design of digital resonator to estimate formants from a very short duration of the speech. The spectrum of cascade of these four resonators is referred to as our proposed ESA spectrum of speech. The novelty of the proposed approach lies in using very short duration of analysis speech frame that is synchronized with glottal closure instant (i.e., about 1-2 ms) to estimate the proposed spectrum in order to ensure that the vocal tract system characteristics do not change much within this interval and to alleviate erroneous estimation of formants due to *nonlinear* interaction of excitation source with the vocal tract system. To demonstrate the effectiveness of proposed algorithm for formant estimation on speech data, we have used 1.5 ms speech signal corresponding to closed phase glottal cycles derived from a male speaker of CMU-ARCTIC database.

Index Terms: Teager Energy Operator, Energy Separation Algorithm, SESA, Digital Resonator, ESA Spectrum.

1. Introduction

The main objective of this paper is to propose an algorithm that could estimate short-time spectrum of a speech signal based on an energy separation algorithm (ESA) [1] in order to estimate the resonant frequencies of vocal tract system, which could aid in several speech analysis and other related applications [2-4]. These resonant frequencies known as formants play a significant role in establishing speech production-perception link for most of the speech processing applications [5-7]. Proposed ESA spectrum is obtained by combining two independent schools of thought, namely, iterative application of smoothened (namely, linear-phase FIR filtering) energy separation algorithm (SESA) and design of cascade of four 2^{nd} order digital resonators [3-4].

Vocal tract characteristics vary within one pitch period (T_0) although the articulators themselves do not move ([8-10]). This may be due to the fact that the vocal folds change state between an open and a closed phase, thereby changing the vocal tract characteristics within one glottal cycle. At the glottal closure instants (GCIs), glottal airflow becomes relatively lower (or ideally zero) and hence, acoustically decoupling supralaryngeal vocal tract system from the trachea. Thus, the speech signal in the closed phase regions, represents the *free* resonances of the supralaryngeal vocal tract system as opposed to the open phase, where trachea and the vocal tract system are acoustically coupled, which alters the free resonances of the system due to the change in length of vocal tract system. Furthermore, the first formant of vocal tract interacts nonlinearly with the glottal flow during the open phase of a glottal cycle, resulting in sudden pressure drop at the lips and sudden increase in lower formants and their corresponding -3 dB bandwidths (Section 4.5 pp. 153-161,[11]) [12-13]. In addition, the vocal tract system characteristics are not constant, but signal-dependent during the open phase of the glottis. We are interested in analyzing the speech signal in the closed phase regions, which is expected to provide an accurate estimate of the frequency response of the supralaryngeal vocal tract system [14-15].

It is very challenging to estimate formants from such small time interval of speech signal. One may use any of the existing epoch extraction techniques ([16-21]) for determination of epochs in case the Electroglottograph (EGG) (i.e., the ground truth) readings are not available. We would like to model the first four free resonances of the vocal tract system exhibited in the closed phase region by a cascade of four all-pole 2^{nd} order digital resonators. These closed phase glottal activity regions have been dynamically extracted from the voiced segments of speech signals. Previous traditional methods for finding formants, such as peak picking of the cepstrally-smoothed spectrum, linear prediction (LP) spectrum ([22-23]), and finding roots of the LP polynomial [24] require about 20-30 ms of block of speech segment to estimate formants (or LP analysis would require estimation of unreliable autocorrelation function from very short speech segments in the order of 1-2 ms). As a result, vocal tract characteristics which could be time-varying within the block are spread over the entire block. In addition, the fundamental frequency (i.e., F_0) significantly influences the vocal tract spectrum if the analysis frame contains more than one pitch period (i.e., To=1/Fo). Besides, block processing of speech signals involves its own set of limitations regarding the size and shape of the analysis frame [16].

Position of the analysis frame is very critical to the performance of high resolution techniques as signal properties can be significantly different nearby closed regions of glottis [16]. Therefore, we propose an ESA spectrum that can analyze speech signal of length as short as 1-2 ms of voiced speech closed phase regions for formants and other system-related characteristics. At the same time, one cannot determine true formants from one single speech sample as it requires atleast a considerable number of samples to produce a resonance (because resonances is not an instant phenomenon rather it builds over a period of time). Our aim is to use as less duration of speech signals as possible, just enough to estimate formant information so that the time-varying vocal tract system

978-9-0827-9703-9/19/$31.00 ©2019 IEEE

characteristics do not spread over the time duration under consideration.

2. Iterative Energy Separation Algorithm

For estimation of formants by application of ESA, one must isolate the vocal tract resonances by bandpass filtering the voiced speech signal. For this purpose, Gabor filter with an impulse response given by:

$$g(t) = exp(-\alpha^2 t^2) cos(\omega_c t)\,, \quad \omega_c = 2\pi f_c\,, \qquad (1)$$

and frequency response given by [24], [30]:

$$G(\omega) = \frac{\sqrt{\pi}}{2\alpha}(exp[-\frac{(\omega-\omega_c)^2}{(4\alpha^2)}] + exp[-\frac{(\omega+\omega_c)^2}{(4\alpha^2)}])\,, \qquad (2)$$

is used. It can be observed from eq. (1) and eq. (2) that Gabor filter have *optimal* time-frequency resolution as Fourier transform of a Gaussian is also a Gaussian [1]. The idea of IESA was implemented using the iteration [25],[32], i.e.,

$$f_c^{(i+1)} = \frac{1}{N}\sum_{n=0}^{N-1} f^{(i)}(n)\,, \qquad (3)$$

where $f(n)$ was computed using the DESA-1 version of discrete ESA using finite backward difference [25],

$$f(n) \approx \frac{1}{2\pi T}\cos^{-1}\left(1 - \frac{\psi_d[y(n)] + \psi_d[y(n+1)]}{4\psi_d[x(n)]}\right), \text{ and}$$

$$|a(n)| \approx \sqrt{\frac{\psi_d[x(n)]}{1 - \left(\frac{\psi_d[y(n)] + \psi_d[y(n+1)]}{4\psi_d[x(n)]}\right)^2}}\,, \qquad (4)$$

$$\text{where } \psi_d\{y(n)\} = y^2(n) - y(n-1)y(n+1)\,, (5)$$

is the Teager Energy Operator [29-31],and the center frequency, F_c of the Gabor bandpass filter on the $(i+1)^{th}$ iteration is set to the mean value of instantaneous frequency $f(n)$ obtained from DESA-1 in the i^{th} iteration. The algorithm was assumed to have converged when the center frequency f_c do not change by more than 5 Hz. The final converged frequency value in a particular subband was believed to be an estimate of the formant. It was reported that the initial estimate $f_c^{(1)}$ of Gabor filter was very important for accurate convergence of the iteration towards the formants. For this purpose, morphological peak picking has been implemented, where voiced speech segments of relatively larger duration of 20-30 ms are considered [25]. Other recent applications of ESA algorithm for Spoof Speech Detection (SSD) task are reported in [5-7].

2.1. 2nd Order Digital Resonator

Since the concept of *2nd* order digital resonator is very crucial to the development of proposed ESA spectrum in this paper, we describe it in brief here [3-4]. System function of a *2nd* order digital resonator with poles p_1 and p_1^* in z-domain, isgivenby:

$$H(z) = \frac{b_0}{(1-p_1 z^{-1})(1-p_1^* z^{-1})} = \frac{A_1}{(1-p_1 z^{-1})} + \frac{A_2}{(1-p_1^* z^{-1})},$$

where $p_1 = re^{j\omega_0}$ and $p_1^* = re^{-j\omega_0}$ (r= pole radius and = ω_0 pole angle). Solving for unknowns A_1 and A_2 using partial fractions, we get,

$$h(n) = \frac{b_0 r^n \sin((n+1)\omega_0)}{\sin(\omega_0)} u(n)\,. \qquad (6)$$

Using concept of Impulse Invariant Transformation (IIT), we have

$$r = e^{-\pi BT}\,, \qquad (7)$$

where B is-*3 dB*bandwidth (in Hz) and T is sampling interval (in seconds). Figure1 shows the impulse response and frequency response of a *2nd*order digital resonator for a pole radius $r = 0.975$ and $\omega_0 = 0.37$ rad (i.e., 471.09 Hz).

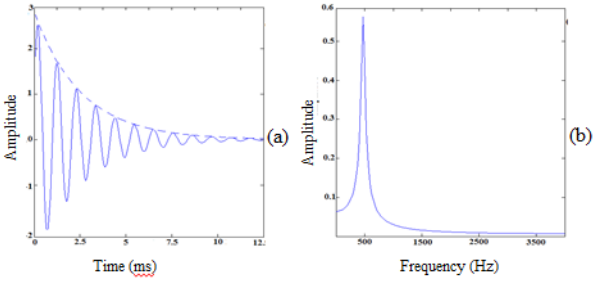

Figure 1:2nd order digital resonator with pole radius, r=0.975 and pole angle, ω_0 =0.37 rad. (a): impulse response $h(n)$. Dashed line indicates the amplitude envelope of the resonator. (b): corresponding magnitude response in the frequency-domain.

Dashed line in Figure1 (a) indicates the corresponding amplitude envelope of the impulse response which is governed by the value of pole radius (i.e., *r* is *inversely* related to the *-3 dB* bandwidth). In addition, closer the value of r to 1, sharper is the bandwidth of the resonance at pole angle ω_0. Furthermore, as $r \to 1$, $\omega_0 \approx \omega_r$, i.e., pole angle approximates well to the resonant frequency (i.e., formant in the present case). In this work, we propose to use the amplitude envelope of smoothened DESA-2 using 7-point binomial linear phase FIR filter, i.e., SESA (which uses two asymmetric difference as opposed to one sample difference in DESA-1) to obtain pole radius r of digital resonator while exploiting its instantaneous frequency output to extract formant from the pole angles. In the next section, we give a detailed description of estimation of our proposed ESA spectrum from real speech signals using SESA and digital resonator design.

3. Basis for Proposed ESA Spectrum

As a proof of concept, we would first like to generate an ESA spectrum for the synthetic case, i.e., cascade of four *2nd* order digital resonators (to approximate first four formants).

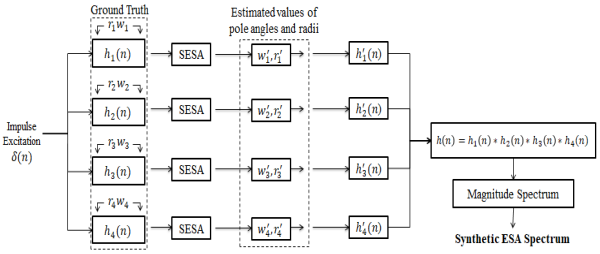

Figure 2. *Block diagram for generation of synthetic ESA spectrum form cascade of four 2nd order digital resonators.*

Table 1: *ESA-based estimation of formant frequencies compared to ground truth (adapted from [16]) for a case of synthetic resonators*

Formants	$F_1(Hz)$	$B_1(Hz)$	$F_2(Hz)$	$B_2(Hz)$	$F_3(Hz)$	$B_3(Hz)$	$F_4(Hz)$	$B_4(Hz)$
Ground Truth	473.5	62.3	1423.6	90.5	2372.8	114.5	3322.1	158.7
ESA estimation	489.7	70.5	1438	95.6	2375	116.1	3313	160.8

We cascade four 2^{nd} order digital resonators in four different subbands with the following predetermined values [11]. In particular, r_1=0.9879 (B_1=62.3 Hz), F_1=473.5 Hz; r_2=0.9824 (B_1=90.5 Hz), F_2= 1423.6 Hz; r_3=0.9778 (B_3=114.5 Hz), F_3= 2372.8 Hz and r_4=0.9694 (B_4=158.7 Hz), F_4=3322.1 Hz. Each of the subband impulse responses $\{h(n)\}_{i\in[1,4]}$ are given as input to SESA. Ratio of any two consecutive values in the amplitude envelope and mode of instantaneous frequency of SESA gives pole radius r and pole angle ω_0, respectively, of a particular subband. Figure 2 and Figure 3 show slightly different block diagram for implementations of this synthetic ESA spectrum. The impulse responses $\{h(n)\}_{i\in[1,4]}$ of 2^{nd}order digital resonator are shown in Figure 3 (corresponding to eq. (6)).

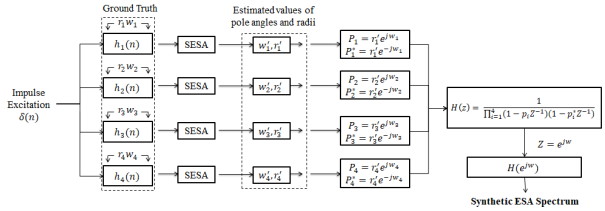

Figure 3. *A slightly different approach for implementation of synthetic ESA spectrum using four pairs of conjugate poles, without cascading of impulse responses of four 2^{nd} order digital resonators.*

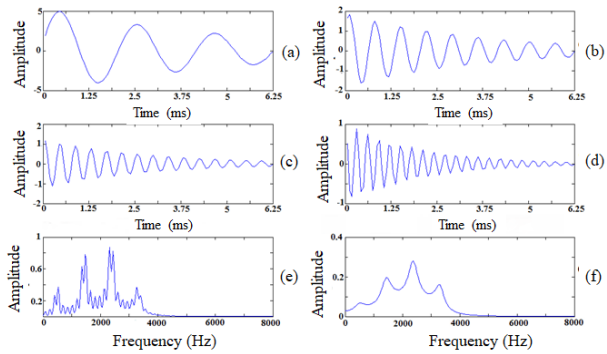

Figure 4. *Illustration of synthetic ESA spectrum generated from the following resonators' impulse responses. (a): r_1=0.9879 (B_1= 62.3 Hz) , F_1=473.5 Hz. (b): r_2=0.9824 (B_2=90.5 Hz), F_2=1423 Hz. (c): r_3=0.9778 (B_3= 114.5 Hz), F_3=2372.8 Hz. (d): r_4= 0.9694 (B_4= 158.7 Hz), F_4=3322.1 Hz. (e): Synthetic ESA spectrum generated according to Fig. 2. (f): Synthetic ESA spectrum generated according to Fig. 3. Note that the ESA spectra are generated corresponding to 1 ms of the input signal.*

Figure 4 shows each of the individual impulse responses corresponding to a particular formant, and its corresponding estimated ESA spectrum (for synthetic case) as depicted by the block diagrams in Figure 2 and Figure 3. Figure 4(f) shows the spectrum generated corresponding to *only* 1 *ms* duration of the original impulse responses (as shown in Figure 4 (a)-(d)) signifying the signal between a pair of zero-crossings of the impulse response corresponding to the first formant F_1. The formants along with corresponding $-3dB$ bandwidths of the synthetic ESA spectrum are shown in Table 1. The estimated formants are only slightly deviated from the ground truth, considering its high time resolution (as it is spectrum estimation for only *1 ms* of input signal).

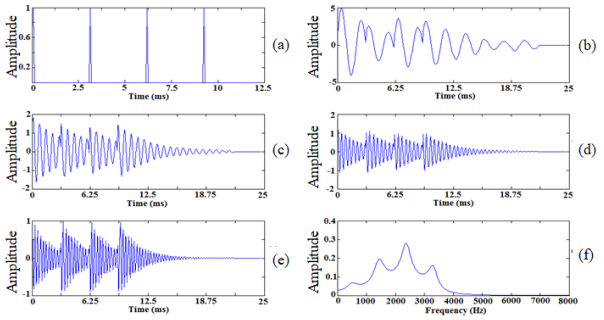

Figure 5: *Illustration of synthetic ESA spectrum (synchronized with synthetic GCI) generated by exciting the impulse responses of resonators by predefined impulses. (a): impulses at 3 ms pitch period (To) used to excite the impulse responses of resonators shown in Figure 4(a)-(d). (b), (c), (d) and (e): output of exciting the impulse responses of resonators in Figure 4(a)-(d) with impulses in Figure 5(a), respectively. (f): Synthetic ESA spectrum generated according to the block diagram in Figure 3. Note: This synthetic ESA spectrum has also been generated corresponding to 1 ms duration of the impulse responses in (b), (c), (d), and (e) as in Figure 4(f).*

Figure 5 shows a similar synthetic ESA spectrum generated using impulses at known locations of 3 *ms* pitch period (i.e., T_0) which may correspond to an infant's cry signal for exciting the impulse responses of the resonators (as shown in Figure 4(a)-(d)). This synthetic ESA spectrum could be considered analogous to proposed ESA spectrum of a synthetic speech. It can be observed from Figure 4 (e) and Figure 4 (f), that there is a significant difference between the spectral characteristics of ESA spectrum computed using cascading of impulse responses (followed by spectrum computation) than using z-domain system function, i.e., $H(e^{j\omega})$ from $H(z)$ (refer Figure 5). This difference is due to the finite-duration *truncation* effect which is explained in the next section. We can observe from Figure 5 (f) that the ESA spectrum is unaffected by the excitation using a sequence of impulses with a very short pitch period of *3 ms* (which may correspond to an infant's cry signal) compared to Figure 4 (f).

4. Experimental Results

In this section, we formally describe the steps involved in generating the proposed ESA spectrum followed by comparing its results with that of state-of-the-art LP spectrum. The following steps are involved in generating the proposed ESA spectrum:

Step 1: Extract the instants of significant excitation in a segment of voiced speech signal. In this work, differenced EGG recordings present in the publicly available CMU-ARCTIC databasehave been used for greater accuracy [26]. However, we may also use state-of-the-art epoch or GCI detection methods (such as *0*-Hz resonator [16-17], TEO profile[19], DYPSA[18], Plosion Index [20-21], SEDREAMS [27], etc.)

Step 2: Dynamically extract the regions of closed phase glottal activity using the information from the differenced EGG plots.

Step 3: In each subband, spectral peak picking of this closed phase input voiced speech signal (which is about 1 ms in duration) yields an initial estimate, f_c^1, for the center frequency of Gabor bandpass filter.

Step 4: Using this initial estimate of center cut-off frequency, SESA is implemented according to eq. (3). The algorithm is considered to have converged after 20 iterations or whenever the center frequency does not change by 5 Hz, whichever condition is satisfied first. Mean of the final converged instantaneous frequency is considered to be a formant value to be used as a pole angle, ω_i ; $1 \le i \le 4$, for four subbands corresponding to the first four formants.

Step 5: Amplitude envelope of the final iteration is used to approximate the pole radius, r_i ; $1 \le i \le 4$, for four subbands.

Step 6: Frequency response of a cascade of the four impulse responses corresponding to the four subbands generated using the pole angles, ω_i and pole radii, r_i is our proposed ESA spectrum$H(\omega)$ and it is given by:

$$h(n) = [\frac{b_1 r_1^n \sin\left((n+1)\omega_1\right)}{\sin(\omega_1)} * \frac{b_2 r_2^n \sin\left((n+1)\omega_2\right)}{\sin(\omega_2)} * \frac{b_3 r_3^n \sin\left((n+1)\omega_3\right)}{\sin(\omega_3)} * \frac{b_4 r_4^n \sin\left((n+1)\omega_4\right)}{\sin(\omega_4)}]u(n)$$

$$\therefore H\left(e^{j\omega}\right) = DTFT\{h(n)\}, \quad (8)$$

where '*' and $DTFT\{.\}$ are convolution and DTFT operations, respectively. For this work, we have taken the values of $b_i = 1$, for $1 \le i \le 4$. Figure6 illustrates the ESA spectrum generated from the closed phase glottal activity voiced speech signal of duration about 1.5 ms. From Figure6(h), we can observe that there are several inconsistencies present in the generated ESA spectrum, i.e., presence of several redundant bumps between formants. This is due to the fact that, the four impulse responses (i.e., $h_{i\in[1,4]}(n)$) which are theoretically supposed to be of infinite duration in length are truncated when cascaded in time-domain.Therefore, higher the number of samples used to generate the resonators' impulse responses, more accurate will be the estimate of ESA spectrum. To overcome this signal processing artifact, we have slightly modified the algorithm by directly feeding the four conjugate pole-pairs (corresponding to the four subbands) to the z-domain system function and then finding its corresponding frequency response (as shown in the block diagram of Figure3), i.e., computing $H(e^{j\omega})$ from $H(z)$. In particular, we have

$$H\left(e^{j\omega}\right) = \frac{1}{\prod_{i=1}^{i=4}(1 - p_i e^{-j\omega})(1 - p_i^* e^{-j\omega})}, \quad (9)$$

where $p_i = r_i e^{j\omega_i}$ and $p_i^* = r_i e^{-j\omega_i}$ are the conjugate pole pairs corresponding to i^{th} subband. We can observe from Fig. 6(i) that, this variation of ESA spectrum is much cleaner than that shown in Fig. 6(h). It is very interesting to note that the estimated ESA spectrum from 1.5 *ms* of speech resemble to state-of-the-art LP spectrum. Thus, the key difference between Fig. 6(h) and Fig. 6(i) is the time-domain convolution *vs.* frequency response from Z-domain system function.

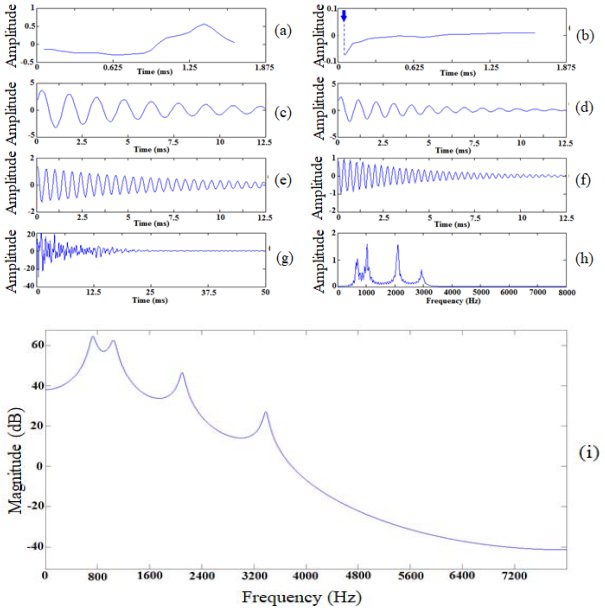

Figure 6. Illustration of the ESA spectrum generated for a 1.5 ms (synchronized with closed phase glottal region) male voiced speech vowel /a/. (a): voiced vowel waveform in the closed phase glottal cycle corresponding to 1.5 ms. (b): differenced EGG signal corresponding to the waveform in (a). Arrow in Figure5 (b) indicates the glottal closure instant. (c),(d),(e) and (f): impulse responses $h_{i\in[1,4]}(n)$ corresponding to the four subbands, respectively. (g): cascaded output of the impulse responses shown in (c)-(f). (h): ESA spectrum generated by taking the magnitude spectrum of the cascaded output in (g). (i): ESA spectrum generated without cascading individual impulse responses.

5. Summary and Conclusions

In this study, a novel spectrum estimation technique based on energy separation algorithm (ESA), and design of digital resonators was proposed that could estimate spectrum from voiced speech segment of duration of *1-2 ms* corresponding to closed phase glottal regions. Performance of the proposed spectrum was compared to that of the state-of-the-art LP spectrum for *30 ms* duration of voiced speech (as the latter was a block-based technique). The onlylimitation of the present work could be the need of ground truth, i.e., differenced EGG (DEGG) readings to detect the closed phase of the glottis. However, any one of the state-of-the-art GCI detection methods, such as *0*-Hz resonator, DYPSA, Hilbert envelope of LP residual, TEO profile, etc. could be used to detect the GCI in case DEGG is not available. Our future research efforts will be directed towards using effectiveness of proposed ESA spectrum for estimation of nasal formants from short duration of speech.

6. References

[1] P. Maragos, J. F. Kaiser, T.F. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Trans. Signal Processing,* 1993, vol. 41, no. 10, pp. 3024-3051.

[2] L. R. Rabiner, R. W. Schafer, Digital Processing of Speech Signals. Englewood Cliffs, NJ: Prentice Hall, 1978.

[3] J. G. Proakis, and D. G. Manolakis, Introduction to Digital Signal Processing, Prentice Hall, *1st* Ed., 1988.

[4] A. V. Oppenheim, R. W. Schafer, Discrete-Time Signal Processing. Pearson Education, *3rd* Ed, 2014.

[5] M. R. Kamble, H. A. Patil, "Novel energy separation based instantaneous frequency features for spoof speech detection," *IEEE European Signal Processing Conference (EUSIPCO),* Kos Island, Greece, 2017, pp. 106-110.

[6] H. A. Patil, M. R. Kamble, T. B. Patel, M. H. Soni, "Novel variable length Teager energy separation-based instantaneous frequency features for replay detection," in *INTERSPEECH,* Stockholm, Sweden, 2017, pp. 12-16.

[7] M. R. Kamble, H. A. Patil, "Effectiveness of Mel scale-based ESA-IFCC features for classification of Natural vs. spoofed speech," In B. Uma Shankar *et. al.* (Eds), Lecture Notes in Computer Science (LNCS), Pattern Recognition and Machine Intelligence (PReMI), Kolkata, India, 2017, vol. 10597, pp. 308-316.

[8] T. V. Ananthapadmanabha, G. Fant, "Calculations of true glottal volume-velocity and its components," *Speech Communication*, vol. 1, no. 1, 1982, pp. 167-184.

[9] B. Cranen, L. Boves, "On subglottal formant analysis," *J. Acoust. Soc. Amer.(JASA)*, Vol. 81, 1987, pp. 734-746.

[10] D. G. Childers, C. F Wong, "Measuring and modeling vocal source-tract interaction," *IEEE Trans. Biomed. Eng.*, vol. 41, 1994, pp. 663- 671.

[11] T. F. Quatieri, Discrete-Time Speech Signal Processing: *Principles and Practices*. Pearson Education, 2002.

[12] J. W. Van der Berg, "On the air response and the Bernoulli effect of the human larynx," *J. Acoust. Soc. Amer.(JASA)*, vol. 29, 1957, pp. 626-631.

[13] T. V. Ananthapadmanabha, "Aerodynamic and acoustic theory of voiced production," in: Neustein, A., and Patil, H.A. (Eds.), *Forensic Speaker Recognition, Law Enforcement and Counter terrorism, Springer*, New York, USA, 2011, pp. 309-364.

[14] D. Venneman, S. BeMent, "Automatic glottal inverse filtering from speech and electroglottographic signals," *IEEE Trans. Signal Process.*, vol. 33, no. 4, 1985, pp. 369-377.

[15] B. Yegnanarayana, R, N, J Veldhuis, "Extraction of vocal tract system characteristics from speech signals," *IEEE Trans. Speech Audio Process.*,vol. 6, no. 4, 1998, pp. 313-327.

[16] K. S. R Murty, B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. on Audio Speech and Lang. Process.,*vol. 16, no. 8, 2008, pp. 1602-1613.

[17] B. Yegnanarayana, K. S. R. Murty, "Event-based instantaneous fundamental frequency estimation from speech signals," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 17, no. 4, 2009, pp. 614-624.

[18] P. A. Naylor, A. Kounoudes, J. Gudnason, M. Brookes, "Estimation of glottal closure instants in voiced speech using the DYOSA algorithm," *IEEE Trans. on Audio Speech and Lang. Process.*, vol. 15, no. 1, 2007, pp. 34-43.

[19] H. A. Patil, S. Viswanath, "Effectiveness of Teager energy operator for epoch detection from speech signals," *Int. J. Speech Tech. (IJST), Springer, v*ol. 14, no. 4, 2011, pp. 321-337.

[20] A. P. Prathosh, T. V. Ananthapadmanabha, and A. G. Ramakrishnan, "Epoch extraction based on integrated linear prediction residual using plosion index," *IEEE Trans. on Audio, Speech, and Lang. Process*, vol. 21, no. 12, 2013, pp. 2471-2480.

[21] T. V. Ananthapadmanabha, A. P. Prathosh, A. G. Ramakrishnan, "Detection of the closure-burst transitions of stops and affricates in continuous speech using the plosion index," in *J. of the Acoust. Soc. of Amer. (JASA)*, 2014, vol. 135, no. 1, pp. 460-471.

[22] S. S McCandless, "An algorithm for automatic formant extraction using linear prediction spectra," *IEEE Trans. Acoust. Speech. Signal Process.*, vol. 22, 1974, pp. 135-141.

[23] R. W Schafer, L. R. Rabiner, "System for automatic formant analysis of voiced speech," *J. Acoust. Soc. Amer.(JASA)*, vol. 47, no. 2, 1970, pp. 634-648.

[24] B. S. Atal, S. U. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.(JASA),* vol. 50, 1971, pp. 637-655.

[25] H. M. Hanson, P. Maragos, A. Potamianos, "A system for finding speech formants and modulations via energy separation," *IEEE Trans. on Speech and Audio Process.,* vol. 2, no. 3, 1994, pp. 436-443.

[26] Kominek and A. Black, "The CMU-ARCTIC speech databases," in *5th ISCA Speech Synthesis Workshop* (SSW), Pittsburgh, PA, 2004, pp. 223–224.

[27] T. Drugman, M. Thomas, J.Gudnason, P.Naylor,T. Dutoit, "Detection of glottal closure instants from speech signals: A quantitative review," *IEEE Transactions on Audio, Speech, and Language Processing*, 2012,vol. 20, no. 3, pp. 994-1006

[28] H. M. Teager, "Some observations on oral air flow during phonation," *IEEE Trans. Acoust., Speech, Signal Processing,* vol.28, 1980, pp. 599-601.

[29] H. M. Teager, S. M. Teager, "Evidence for nonlinear sound production mechanisms in the vocal tract," in W. J. Hardcastle and A. Marchal (Eds.), *Speech Production and Speech Modeling,* Kluwer Academic Publishers, 1990, pp. 241-261.

[30] P. Maragos, J. F. Kaiser, T. F. Quatieri, "On amplitude and frequency demodulation using energy operators," *IEEE Trans. Signal Process.,* vol. 41, no. 4, 1993, 1532-1550.

[31] P. Maragos, J. F. Kaiser, T. F. Quatieri, "Speech nonlinearities, modulations, and energy operators," *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Toronto, Canada, 1993, pp. 421-424.

[32] A. Potamianos, P. Maragos, "A comparison of the energy operator and the Hilbert transform approach to signal and speech demodulation," *Signal Process., Elsevier*, vol. 37, 1994, pp. 95-120.