INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

A Noise Robust Technique for Detecting Vowels in Speech Signals

Avinash Kumar[‡], S. Shahnawazuddin[†] and Waquar Ahmad[*]

[‡]Department of Electronics and Communication Engineering, NIT Sikkim, India
[†]Department of Electronics and Communication Engineering, NIT Patna, India
[*]Department of Electronics and Communication Engineering, NIT Calicut, India
avinash_ece@nitsikkim.ac.in, s.syed@nitp.ac.in, waquar@nitc.ac.in

Abstract

In this work, we propose a novel and noise robust method for the detection of vowels in speech signals. The proposed approach combines variational mode decomposition (VMD) and non-local means (NLM) estimation for the detection of vowels in a speech sequence. The VMD algorithm is used to determine a number of variational mode functions (VMFs). The lower-order VMFs represent the frequency contents corresponding to vowel regions. Thus by combining the lower-order VMFs and reconstructing the speech signal back, the energy corresponding to the vowel regions is enhanced while the non-vowel regions are suppressed. At the same time, the ill-effect of noise is also reduced. Finally, as reported in an earlier work, application of NLM followed by convolution with first-order difference of Gaussian window is performed on the reconstructed signal to determine the vowel region. The performance of proposed approach for the task of detecting vowels in speech is compared with three existing techniques and observed to be superior under clean as well as noisy test conditions.

Index Terms: Vowel detection, noise robust, VMD, NLM.

1. Introduction

In a speech signal, vowels are the primary voiced regions. The instants of starting and ending of a vowel are known as vowel onset point (VOP) and vowel end point (VEP), respectively [1]. The frequency response of the vocal tract system as well as the source of excitation information are better manifested within the vowel region [2, 3]. Effective detection of vowels and VOPs/VEPs has been used in the earlier reported works on developing robust speaker recognition systems [3, 2, 4, 5, 6]. In addition to that, information about vowel regions and corresponding VOPs/VEPs was also explored for the identification of consonant-vowel units [7, 8], speech segmentation [9], keyword spotting [10], dialect classification [11] and prosody modification [12, 13, 14]. Therefore, several front-end speech parameterization techniques and statistical modeling methods have been studied for detecting vowels and their corresponding VOPs and VEPs [1, 2, 3, 15, 16, 17, 18, 19, 20].

It is well known that the vowels are long duration, periodic and high-energy sound units [21, 1, 15]. These attributes have been exploited in the above mentioned works for extracting front-end acoustic features that enhance the energy and periodicity information. Some of those front-end features such as the energy difference of each of the peaks and their respective valleys in the short-term discrete Fourier transform (DFT) magnitude spectrum [22], largest peaks in the DFT magnitude spectrum [1] and mel-frequency cepstral coefficients (MFCCs) were used for representing spectral energy in different frequency bands. Features representing excitation strength like Hilbert envelope of the linear prediction (LP) residual [23] and the rate of change in excitation strength obtained from the zero frequency filtered (ZFF) speech signal [3, 2] were also studied for detecting vowel regions. In addition to those, the zero-crossing rate, energy and pitch information of the speech signal [17], wavelet scaling coefficients of the speech signal [24], modulation spectrum energies [1], spectral energy present in the glottal closure regions [15], uniformity of the epoch intervals [16] and cumulative sum of the DFT magnitude spectrum of the non-local estimated speech signal [25] have also been used as the discerning acoustic features. Furthermore, several acoustic features have been combined to represent the complementary information present in the vowels [1, 2, 3, 16, 20].

Since the detection of vowel regions has immense application as already discussed, we present a novel technique for extracting robust front-end features that can be used for effective detection of vowel regions and their corresponding VOPs and VEPs. In the proposed approach, we first suppress the unvoiced sound units with high frequency content using variational mode decomposition (VMD) [26]. The VMD technique is employed to break the speech signal into several variational mode functions (VMFs) centered around distinct frequency bands. Those VMFs that are centered around 100-5000 Hz are chosen and then combined to reconstruct the speech signals. Next, non-local means estimation is employed to determine the sum of weight values (SWV) for each of the samples in the reconstructed speech signal as suggested in [27]. The SWVs are then convolved with first-order difference of Gaussian window. The peaks and valleys in the resulting output are finally used to detect the vowel regions.

The rest of the paper is organized as follows: In Section 2, the proposed technique is described in detail. The experimental evaluations are presented in Section 3. Finally, the paper is concluded in Section 4.

2. Proposed method

In the proposed method, the vowels are identified by processing any given speech signal through the following sequence of steps:

STEP-I First, the speech signal is decomposed into n numbers of variational mode functions using VMD. The VMFs with lower center frequency correspond to the predominantly high magnitude vowel regions whereas the unvoiced sound units are represented by the VMFs with higher center frequency. It is well known that, in the case of VMD, choosing a large number of modes leads to under-binning or loss of relevant information. On the other hand, choosing a lower number of modes results in over-binning of modes or mode duplication [26]. During the preliminary experiments performed on a development

http://dx.doi.org/10.21437/Interspeech.2020-1204

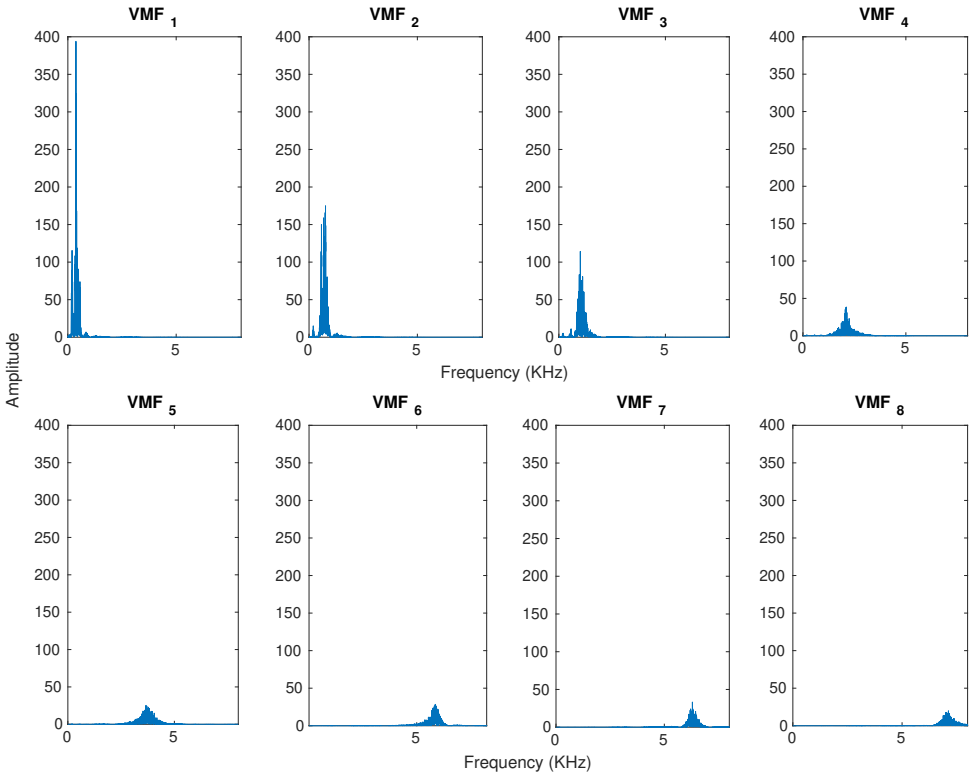

Figure 1: *Magnitude spectrum of VMFs for a speech signal. The modes are arranged from low- to high-frequency band (left to right).*

set, it was noted that a minimum of 8 levels of decomposition are required for efficient decomposition and reconstruction of the speech signal. The spectral magnitude for those 8 VMFs obtained by decomposing a speech signal collected from TIMIT database [28] are shown in Figure 1. It can be observed that by combining some of the VMFs one may be able to capture the energy of the vowel regions lying in between 100-5000 Hz depending on the location of their center frequencies. As evident from Figure 1, combining VMF-2 to VMF-6 leads to effectively capturing the desired frequency band. The energy of the vowel region is thus captured by this frequency band.

STEP-II Next, the selected m VMFs are summed and the speech signal is reconstructed so that the energy corresponding to the vowel regions is enhanced. At the same time, sound units such as fricatives are suppressed.

STEP-III Further to that, non-local means estimation is employed to determine the SWVs for each of samples in the reconstructed speech signal as suggested in [27].

STEP-IV The significant transition points in the SWV are then detected by convolving it with a first-order difference of Gaussian (FODG) window. In the convolved output, termed as the *vowel detection evidence*, the region between a valley and peak is selected as the vowel region. Similarly, the region between a peak and a valley is the non-vowel region.

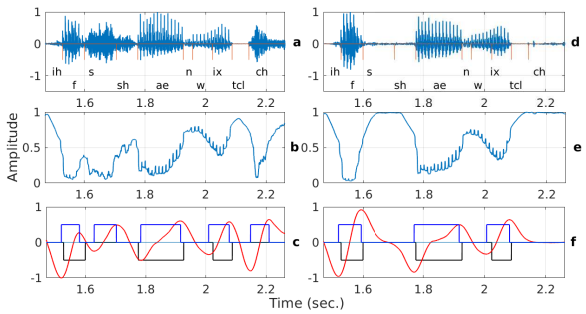

Figure 2: *Illustration of the suggested vowel detection method. (a) A clean speech segment from the TIMIT database with reference markings for the sound units, (b) sum of weight values obtained by using NLM estimation of the speech signal, (c) the solid red line shows the vowel detection evidence obtained after convolving SWV with FOGD window. The reference vowel regions are depicted by black lines while the blue lines represent the detected vowels regions. (d) Represents the speech signal reconstructed after selecting VMF-2 to VMF-6, (e) SWV obtained by processing VMF reconstructed speech signal and (f) vowel detection evidence, the reference and detected vowel regions.*

The outputs for each of the steps involved in the proposed vowel detection method, when applied on a clean speech sig-

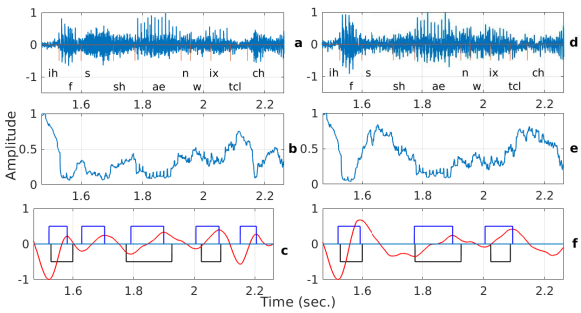

Figure 3: *(a) A segment of noisy speech with reference markings for the sound units, (b) sum of weight values obtained by using NLM estimation of the speech signal, (c) the solid red line shows the vowel detection evidence obtained after convolving SWV with FOGD window. The reference vowel regions are depicted by black lines while the blue lines represent the detected vowels regions. (d) Represents the noisy speech signal reconstructed after selecting VMF-2 to VMF-6, (e) SWV obtained by processing VMF reconstructed speech signal and (f) vowel detection evidence, the reference and detected vowel regions.*

nal, are shown in Figure 2. The top panes show the time domain waveforms and the corresponding sound units marked on it. On comparing Figure 2 (a) and Figure 2 (d), it is evident that the energy due to high frequency sound units is significantly deemphasized in the speech signal reconstructed by summing the VMFs. For example notice the differences in the regions corresponding to /s/, /sh/ and /ch/ sound units in Figure 2 (a) and Figure. 2 (d). Figure 2 (b) and Figure 2 (e), respectively, show the SWV computed using NLM estimation without and with VMD-based reconstruction. In Figure 2 (c) and Figure 2 (f), the red lines represent the *vowel detection evidence* obtained when the SWV is convolved with a FOGD window. As mentioned earlier, the region between a valley and peak is selected as the vowel region and the same is depicted using solid blue lines. The hand labeled markings available with TIMIT database are used as the reference vowel regions and those are represented using solid black lines. From Figure 2 (f), it is evident that the detected vowel regions almost match with the reference ones. On the other hand, from Figure 2 (c), it can be observed that the *vowel detection evidence* includes spurious detections for the long duration and high energy fricatives (/s/ and /c/) when the SWV are estimated directly from the speech signal. This is overcome when VMD-based reconstruction is performed. Similar trends are noted even in the case of noisy speech as shown in Figure 3 (a) - Figure 3 (f).

3. Experimental evaluations

In this section, the experimental evaluations demonstrating the efficacy of the proposed vowel detection algorithm is presented. TIMIT database was used for studying the effectiveness of the proposed approach. Since hand labeled reference markings are available with the TIMIT corpus, we chose it for determining the accuracy of detecting vowel region through the proposed approach. A test set was derived from TIMIT database that consisted of 400 utterances from 50 male/female speakers. A development set consisting of 200 utterances was used for selecting the optimal values for the parameters. The speech data used in this work was sampled at 16 kHz rate.

Table 1: *Performances, IR and SR, of the different explored approaches as well the technique proposed in this paper for the task of detecting vowels in a given speech signal under clean and noisy test conditions.*

SNR	Method	IR in %	SR in %		
			Semivowel	Nasal	Other
Clean	COMB-EVI	72.87	9.84	1.72	12.09
	SE-GCI	65.71	9.82	1.78	8.18
	NLM-SPE	69.23	10.44	1.41	6.45
	Proposed	**87.20**	**10.67**	**1.28**	**3.05**
15 dB	COMB-EVI	67.25	10.13	2.62	22.18
	SE-GCI	63.21	10.95	2.49	14.08
	NLM-SPE	67.57	10.98	2.14	7.16
	Proposed	**84.06**	**11.86**	**1.41**	**8.82**
10 dB	COMB-EVI	65.19	10.93	3.28	23.98
	SE-GCI	61.47	11.77	2.96	17.09
	NLM-SPE	66.04	11.47	2.83	8.82
	Proposed	**82.42**	**12.04**	**1.63**	**9.27**
5 dB	COMB-EVI	63.78	11.11	3.98	24.38
	SE-GCI	60.04	12.17	3.13	19.50
	NLM-SPE	65.07	11.98	3.08	9.84
	Proposed	**80.30**	**12.80**	**1.98**	**9.89**
0 dB	COMB-EVI	61.40	11.90	4.70	26.17
	SE-GCI	58.46	13.86	4.60	21.22
	NLM-SPE	63.63	12.45	3.83	11.85
	Proposed	**77.78**	**13.16**	**2.16**	**10.01**

Any given test speech signal was first decomposed into 8-levels using variational mode decomposition algorithm. For signal decomposition using VMD, the data fidelity constraint balancing parameter was chosen as 320. Further to that, the time-step and tolerance of convergence were set as 0 and 10^{-7}, respectively. For computing the sum of weight values, the half-width of the segment was fixed at 3 ms (48 samples for 16 kHz sampling rate) and while the neighborhood width was chosen to be and 50 ms (800 samples). The bandwidth parameter κ was fixed at 0.6σ. These values were determined by performing experiments on the development set as already stated earlier. Furthermore, it was noted that the performance did not change significantly by varying κ from 0.5σ to 0.9σ.

The accuracy of vowel detection technique were measured using the following parameters:

- *Identification rate (IR):* the percentage of reference vowels that match with the detected vowels.
- *Spurious rate (SR):* the percentage of detected vowels which lie outside the reference vowel regions.

The IR and SR values were computed under clean as well as noisy test conditions. Several noises were added to test set and performances reported are averaged over all the case. The SNR values were chosen to be 15dB, 10dB, 5dB and 0dB. Further to that, the proposed approach was compared with three existing vowel detection techniques reported in [1, 29, 30].

Table 2: *Performances of the proposed and explored techniques for the task of VOP and VEP detection. The terms IR and SR refer to identification rate and spurious rate, respectively. Performance is evaluated using different predefined deviations that are chosen to be either $\pm$ 10 ms or $\pm$ 20 ms in this study.*

SNR	Method	VOP detection			VEP detection		
		IR in %			IR in %		
		$\pm$ 10 ms	$\pm$ 20 ms	SR in %	$\pm$ 10 ms	$\pm$ 20 ms	SR in %
Clean	COMB-EVI	60.82	72.14	8.68	54.91	65.84	7.19
	SE-GCI	65.10	76.84	7.10	54.10	65.14	6.11
	NLM-SPE	65.13	76.15	4.01	55.85	67.18	4.28
	Proposed	**79.83**	**85.19**	**6.27**	**81.96**	**87.06**	**5.86**
1 5 dB	COMB-EVI	57.12	67.27	23.11	51.16	63.01	22.10
	SE-GCI	61.85	74.85	10.96	50.20	62.20	11.24
	NLM-SPE	63.15	75.44	14.04	53.15	65.14	13.87
	Proposed	**78.86**	**84.72**	**7.13**	**80.20**	**86.76**	**6.39**
10 dB	COMB-EVI	54.80	65.13	25.81	49.17	61.10	25.09
	SE-GCI	59.11	72.41	13.96	48.04	59.19	13.63
	NLM-SPE	62.48	73.98	16.15	51.18	63.33	17.64
	Proposed	**77.90**	**83.98**	**8.06**	**79.61**	**85.60**	**6.59**
5 dB	COMB-EVI	52.89	64.16	26.17	47.18	59.19	26.67
	SE-GCI	56.87	70.08	21.88	47.10	58.13	15.16
	NLM-SPE	61.16	72.23	18.45	50.86	61.94	20.11
	Proposed	**77.34**	**83.16**	**8.31**	**79.03**	**85.00**	**7.16**
0 dB	COMB-EVI	50.17	58.17	29.10	43.14	53.60	28.10
	SE-GCI	53.81	68.50	26.43	45.64	56.15	21.08
	NLM-SPE	58.23	70.13	21.90	48.99	60.01	22.92
	Proposed	**72.18**	**79.56**	**9.42**	**73.81**	**80.11**	**8.91**

Those approaches are referred to as **COMB-EVI**, **SPE-GCI** and **NLM-SE**, respectively. The performances of the existing techniques as well as proposed approach in terms of IR and SR are given in Table 1. As already mentioned, the IR and SR values were computed under clean as well as noisy test scenarios. It is evident from Table 1 that the IR values are significantly higher for the proposed approach not only in clean but also in noisy cases. At the same time, the spurious rates are lower.

Finally robustness of proposed method in term of vowel onset points and vowel end points was studied. The following two metrics were employed for performance evaluations.

- *Identification rate (IR):* the percentage of the reference VOPs/VEPs that match with the detected VOPs/VEPs within the predefined deviation (in ms).
- *Spurious rate (SR):* the percentage of detected VOPs/VEPs which are detected outside the reference vowel regions.

The IR and SR values for the task of VOPs and VEPs detection using the existing techniques as well as the proposed method are given in Table 2. The predefined deviations are chosen to be $\pm$10 ms and $\pm$20 ms. Even in this case the proposed approach is noted to outperform the existing ones under clean as well as noisy test conditions.

4. CONCLUSION

A novel and noise robust technique to detect vowels within as speech signal has been proposed in this paper. The proposed approach effectively exploits VMD algorithm to detect vowels. The VMD algorithm is used to break the speech signal into several modes. Few of the lower-order modes are then combined to reconstruct the speech signal. Consequently, the energy corresponding to the vowel regions is enhanced while those for the non-vowel sound units are suppressed. At the same the ill-effects of noise is also reduced. The experimental evaluations presented in this paper demonstrate the efficacy of the proposed approach under clean as well as noisy test conditions. Furthermore, the proposed technique is compared with three of the existing methods for vowel detection and is noted to be superior to those.

5. References

[1] S. R. M. Prasanna, B. V. S. Reddy, and P. Krishnamoorthy, "Vowel onset point detection using source, spectral peaks, and modulation spectrum energies," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 556–565, Mar. 2009.

[2] S. M. Prasanna and G. Pradhan, "Significance of vowel-like regions for speaker verification under degraded conditions," *IEEE Trans. Audio, Speech, Lang. Process*, vol. 19, no. 8, pp. 2552–2565, Nov. 2011.

[3] G. Pradhan and S. M. Prasanna, "Speaker verification by vowel and nonvowel like segmentation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 4, pp. 854–867, Apr. 2013.

[4] N. Almaadeed, A. Aggoun, and A. Amira, "Text-independent speaker identification using vowel formants," *J. Signal Process. Syst.*, vol. 82, no. 3, pp. 345–356, May 2015.

[5] N. Fakotakis, E. Tsopanoglou, and G. KokkinaKis, "A text independent speaker recognition system based on vowel spotting," *Speech Communication*, vol. 12, pp. 57–68, March 1993.

[6] K. Daqrouq and T. A. Tutunji, "Speaker identification using vowels features through a combined method of formants, wavelets, and neural network classifiers," *Appl. Soft. Comput.*, vol. 27, pp. 231–239, Feb. 2015.

[7] A. K. Vuppala, K. S. Rao, and S. Chakrabarti, "Spotting and recognition of consonant-vowel units from continuous speech using accurate detection of vowel onset points," *Circuits, Systems, and Signal Processing*, vol. 31, no. 4, pp. 1459–1474, Feb. 2012.

[8] ——, "Improved consonant-vowel recognition for low bit-rate coded speech," *Int. J. Adapt. Control Signal Process.*, vol. 26, no. 4, pp. 333–349, Oct. 2011.

[9] S. P. Panda and A. K. Nayak, "Automatic speech segmentation in syllable centric speech recognition system," *Int. J. Speech Technol.*, vol. 19, no. 1, pp. 9–18, Nov. 2016.

[10] B. S. Reddy, K. V. Rao, and S. M. Prasanna, "Keyword spotting using vowel onset point, vector quantization and hidden markov modeling based techniques," in *Proc. TENCON*, Nov. 2008, pp. 1–4.

[11] C. Themistocleous, "Dialect classification using vowel acoustic parameters," *Speech Commun.*, vol. 92, pp. 13–22, Sep. 2017.

[12] H. K. Vydana, S. R. Kadiri, and A. K. Vuppala, "Vowel-based non-uniform prosody modification for emotion conversion," *Circuits, Systems, and Signal Process.*, vol. 35, no. 5, pp. 1643–1663, May 2016.

[13] K. S. Rao and B. Yegnanarayana, "Duration modification using glottal closure instants and vowel onset points," *Speech Communication*, vol. 51, no. 12, pp. 1263–1269, December 2009.

[14] K. S. Rao and A. K. Vuppala, "Non-uniform time scale modification using instants of significant excitation and vowel onset points," *Speech Communication*, vol. 55, no. 6, pp. 745–756, July 2013.

[15] A. Vuppala, J. Yadav, S. Chakrabarti, and K. S. Rao, "Vowel onset point detection for low bit rate coded speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 6, pp. 1894–1903, Apr. 2012.

[16] A. K. Vuppala, K. S. Rao, and S. Chakrabarti, "Improved vowel onset point detection using epoch intervals," *AEU-Int. J. Electron. Commun.*, vol. 66, no. 8, pp. 697–700, Aug. 2012.

[17] J. Wang, C. Hu, S. Hung, and J. Lee, "A hierarchical neural network based C/V segmentation algorithm for Mandarin speech recognition," *IEEE Trans. Signal, Process.*, vol. 39, no. 9, pp. 2141–2146, Sep. 1991.

[18] J. Rao, C. C. Sekhar, and B. Yegnanarayana, "Neural network based approach for detection of vowel onset points," in *Proc. Int. Conf. Adv. Pattern Recognition Digital Tech.*, vol. 1, Dec. 1999, pp. 316–320.

[19] A. Kumar, S. Shahnawazuddin, and G. Pradhan, "Exploring different acoustic modeling techniques for the detection of vowels in speech signal," in *Proc. National Conf. on Communication (NCC)*, Mar. 2016, pp. 1–5.

[20] ——, "Improvements in the detection of vowel onset and offset points in a speech sequence," *Circuits, Systems, Signal Process.*, vol. 36, pp. 1–26, Sept. 2016.

[21] K. N. Stevens, *Acoustic Phonetics*. The MIT Press Cambridge, Massachusetts, London, England, 2000.

[22] D. J. Hermes, "Vowel onset detection," *J. Acoust. Soc. Amer*, vol. 87, no. 2, pp. 866–873, Feb. 1990.

[23] S. R. M. Prasanna and B. Yegnanarayana, "Detection of vowel onset point events using excitation source information," in *Proc. Interspeech*, Sept. 2005, pp. 1133–1136.

[24] J. H. Wang and S. H. Chen, "A C/V segmentation algorithm for Mandarin speech using wavelet transforms," in *Proc. Int. Conf. Acoust. Speech, Signal Process.*, vol. 1, Mar. 1999, pp. 417–420.

[25] A. Kumar, S. Shahnawazuddin, and G. Pradhan, "Non-local estimation of speech signal for vowel onset point detection in varied environments," in *Proc. INTERSPEECH*, Aug. 2017, pp. 429–433.

[26] K. Dragomiretskiy and D. Zosso, "Variational mode decomposition," *IEEE Transactions on Signal Processing*, vol. 62, no. 3, pp. 531–544, February 2014.

[27] A. Kumar and G. Pradhan, "Detection of vowel onset and offset points using non-local similarity between dwt approximation coefficients," *Electronics Letters*, vol. 54, no. 11, pp. 722–724, June 2018.

[28] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, *TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1*. Linguistic Data Consortium, Dec. 1993, vol. 33.

[29] A. Vuppala, J. Yadav, S. Chakrabarti, and K. S. Rao, "Vowel onset point detection for low bit rate coded speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1894–1903, August 2012.

[30] A. Kumar, S. Shahnawazuddin, and G. Pradhan, "Detection of vowel offset points using non-local similarity between speech samples," in *2018 International Conference on Signal Processing and Communications (SPCOM)*. IEEE, 2018, pp. 252–256.