Interspeech 2018
2-6 September 2018, Hyderabad

Analysis of Variational Mode Functions for Robust Detection of Vowels

Surbhi Sakshi, Avinash Kumar and Gayadhar Pradhan

Department of Electronics and Communication Engineering
National Institute of Technology Patna, India.
(surbhi.ecepg16,k.avinash, gdp)@nitp.ac.in

Abstract

In this work, initially the speech signal is decomposed into variational mode functions (VMFs) with the aid of variational mode decomposition (VMD). Each decomposed VMF represents different frequency band of the input speech signal. An approximate speech signal is then reconstructed by using a set of selected VMFs whose center frequency predominantly corresponds to the frequency range of the vowels. In the reconstructed speech signal, energy due to the high frequency unvoiced sound units and noises is suppressed. Consequently, over an analysis frame, the mean of the square magnitude (MSM) of the sample points is significantly higher for the vowels than other sound units. Further, the MSM at each time instant is non-linearly mapped (NLM) using a negative exponential functions to enhance the transitions at the onset and the offset points of vowels, and suppress small fluctuations. The NLM-MSM is used as a front-end feature for discriminating vowels in a given speech signal. The experiments conducted on TIMIT database show that, the proposed approach outperforms the existing methods for the task of detecting vowels in a given speech signal under clean and noisy test scenarios.

Index Terms: Vowel, variational mode decomposition, variational mode functions, reconstructed speech.

1. Introduction

In a speech sequence, vowels are periodic, larger amplitude and longer duration sound units [1–3]. Vowel onset point (VOP) and vowel ending point (VEP) are refer to as the starting and ending points of a vowel, respectively [4–7]. Vowel's distinguishing features have led it to be considered as highly informative in extraction of different traits for development of speech based applications, such as speech recognition, speaker recognition, speech analysis, end point detection in utterances, emotion classification etc. [8–12]. The performance of these automated systems depend on the accurate detection of vowels.

In the earlier reported works, several front-end speech parameterization approach encapsulating the nature of excitation and vocal-tract system response have been proposed for the detection of vowels and their corresponding VOPs [1, 3, 13, 14]. Those front-end speech parameterization approaches include, the difference in energy of each of the peaks and their corresponding valleys in the the short-term discrete Fourier transform (DFT) magnitude spectrum [1], largest peaks in the DFT magnitude spectrum [3], mel-frequency cepstral coefficients (MFCC) and spectral energy in different frequency bands [15]. The feature representing excitation strength such as Hilbert envelope of the linear prediction (LP) residual [16] and rate of change of excitation strength extracted from the zero frequency filtered (ZFF) speech signal [9, 14] have also been explored. The zero-crossing rate, energy and pitch information of the speech signal [17], wavelet scaling coefficients of the speech signal [18], modulation spectrum energies [3], spectral energy present in the glottal closure regions [13], uniformity of the epoch intervals [19] and cumulative sum of the DFT magnitude spectrum of the non-local estimated speech signal [20] have also been used as the discriminating features.

In a continuous speech signal, discriminative features present within the vowels such as average energy, duration, magnitude of transitions at VOP and VEP, etc., varies in accordance with the utterance as well as the environmental noises [14]. In case of noisy speech signal, most of the features used for the detection of vowels and VOPs deviate with respect to the noise and input signal to noise ratio (SNR) [21]. As stated in various works, the major issue regarding vowel detection is the confusion between vowels and other similar sound units like semivowels [9, 14, 15]. The transition at the VEPs also does not show resemblance with VOPs [5, 15]. The speech signal dies down at a much slower rate at VEPs. Therefore, in order to improve the detection accuracy of vowel and reduce ill effects of noise an improved method is highly desirable.

The ill effects of noise can be suppressed to a great extent by analyzing the frequency bands of the speech signal corresponding to the vowels. The variational mode decomposition (VMD) decompose the signal into predefined number of narrow-band variational mode functions (VMFs). Each VMF represents a smaller portion of the overall frequency band of the signal. Therefore, by proper selection of VMFs a signal can be reconstructed to enhance the energy of the vowels and suppress the ill effect of noises. Through this motivation, a vowel detection approach is proposed in this paper by utilizing the efficacy of VMD. Further, a simple and efficient approach is proposed to enhance the weak transition at the VEPs.

The rest of the paper is organized as follows: Section 2 discusses the proposed method for detecting vowels. The existing front-end speech parametrization methods used for performance comparison and the experimental results are discussed in Section 3. Finally, the paper is concluded in Section 4.

2. Proposed approach for detecting vowels

The block diagram representing the proposed method for detecting vowels in a speech signal is shown in Fig 1. In the proposed approach, the vowels are detected by processing the speech signal through the following sequence of steps:

i) The speech signal is first decomposed into n number of VMFs using VMD. The VMFs having lower center frequency predominantly represent the high magnitude vowel regions where as the VMFs having higher center frequency represent the unvoiced sound units and noise frequency components.

ii) Then, m number of VMFs whose center frequency corresponds to the $500 - 2500$ frequency band are selected. The energy due to vowel sound units in a speech signal is

10.21437/Interspeech.2018-1947

predominately present in that frequency range. [13]. The m number of selected VMFs are summed to enhance the energy corresponding to the vowel regions and to suppress high energy fricatives and noise components.

iii) Over an analysis frame of length l, the mean of the square magnitude (MSM) of the sample points is computed and non-linearly mapped (NLM). The NLM-MSM is used as the front-end feature for detecting the vowels. The reconstructed speech signal will have significantly higher value of MSM for vowel regions than other sound units. The fluctuations present in MSM reduced significantly by non-linear mapping.

iv) The significant transition points in the NLM-MSM are then detected by convolving it with a first-order difference of Gaussian window (FODG). In the convolved output, termed as the *vowel detection evidence*, the valleys and peaks correspond to the VOPs and VEPs, respectively. The regions between them are selected as the vowels.

2.1. Variational mode decomposition of speech signal

The VMD is a non-recursive, concurrent signal decomposition method that decomposes the given input signal $(s(t))$ into different modes termed as VMFs [22]. Each VMF (c_n) represents a narrow-band frequency region of the input signal. The VMD also estimates the center frequency (ω_n) of each VMF as H^1-norm. The center frequencies are sparsity prior which helps to reconstruct back the input signal $s(t)$. The v_n and ω_n are computed by solving the constrained variational problem as follows:

$$\min_{\{c_n\},\{\omega_n\}} \left\{ \sum_n \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * c_n(t) \right] e^{-j\omega_n t} \right\|_2^2 \right\} \quad (1)$$

such that $\sum_n c_n(t) = s(t)$. Where, $\{c_n\} = \{c_1, c_2, ...c_k\}$, $\{\omega_n\} = \{\omega_1, \omega_2, ...\omega_n\}$, n, $\delta(t)$ and $*$ represent the VMFs (modes), the center frequencies for each of the VMFs, total number of modes, Dirac distribution and convolution operator, respectively.

The signal reconstruction constraint is addressed by using Lagrangian multipliers (λ) and the quadratic penalty factor (α). The convergence properties of the penalty term at a finite weight value and strict enforcement of constraint by the Lagrangian multiplier are being utilized. The augmented Lagrangian $\mathcal{L}$ is represented as follows:

$$\mathcal{L}(\{c_n\},\{\omega_n\},\lambda) = \alpha \sum_n \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * c_n(t) \right] e^{-j\omega_n t} \right\|_2^2 + \left\| s(t) - \sum_n c_n(t) \right\|_2^2 + \left\langle \lambda(t), s(t) - \sum_n c_n(t) \right\rangle \quad (2)$$

By using augmented Lagrangian and the alternate direction method of multipliers optimization framework, the VMFs and its corresponding center frequencies can be computed. After optimization, the resultant updated modes $\{\hat{c}_n\}$ in frequency domain are computed as follows:

$$\hat{c}_n^{p+1}(\omega) = \frac{\hat{s}(\omega) - \sum_{i \neq j} \hat{c}_i(\omega) + \frac{\hat{\lambda}(\omega)}{2}}{1 + 2\alpha(\omega - \omega_n)^2} \quad (3)$$

where $\hat{c}(w)$, $\hat{s}(w)$ and $\hat{\lambda}(w)$ are the frequency domain representations of $c_n(t)$, $s(t)$ and $\lambda(t)$, respectively. The modes in time domain, $c_n(t)$ can be obtained from $\hat{c}_n(\omega)$ using the inverse Fourier transform. Similarly, the updated center frequencies are optimized in Fourier domain as follows:

$$\omega_n^{p+1} = \frac{\int_0^\infty \omega |\hat{c}_n(\omega)|^2 d\omega}{\int_0^\infty |\hat{c}_n(\omega)|^2 d\omega} \quad (4)$$

It locates the updated frequency which is at the center of the n^{th} mode power spectrum.

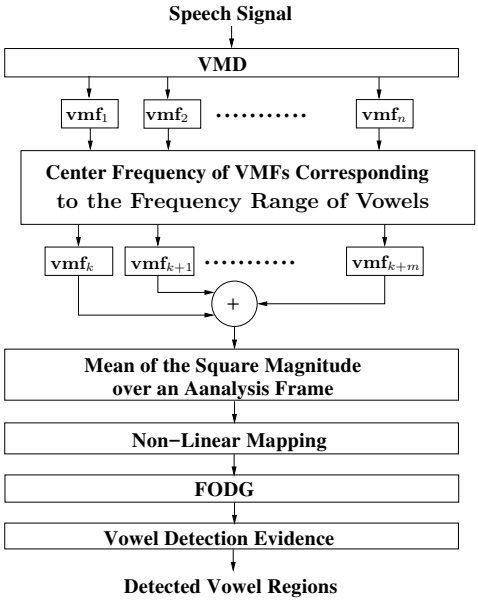

Figure 1: *The block diagram representing proposed method for detecting vowels in a given speech signal.*

2.2. Selection of VMFs corresponding to the frequency range of vowels

If a large number of modes are selected for decomposition, under-binning of modes (loss of information) occurs. On the other hand, lower number of modes results in over-binning of modes (mode duplication) [22]. During the preliminary experiments performed on the development set, it was observed that for effective decomposition and reconstruction of speech signal, a minimum of 8 levels of decomposition is required. The magnitude spectra for the 8 VMFs derived from a speech signal taken from the TIMIT database are shown in Figure 2. The magnitude spectra are shown from left to right in ascending order of VMFs. It can be observed that, depending upon the location of their center frequency, some of the VMFs can be combined together to represent the signal components corresponding to the vowels $(500 - 2500)$ Hz. In the present case, $VMF - 2$ to $VMF - 6$ can be combined to reconstruct the signal.

2.3. Extraction of proposed feature and detection of vowels

The MSM of the reconstructed signal $(r(k))$ over an analysis frame of length $(2l + 1)$ is computed at each time instants by centering the sample point k as follows:

$$M(k) = \frac{1}{2l+1} \sum_{q=-l}^{l} |r(k+q)|^2 \quad (5)$$

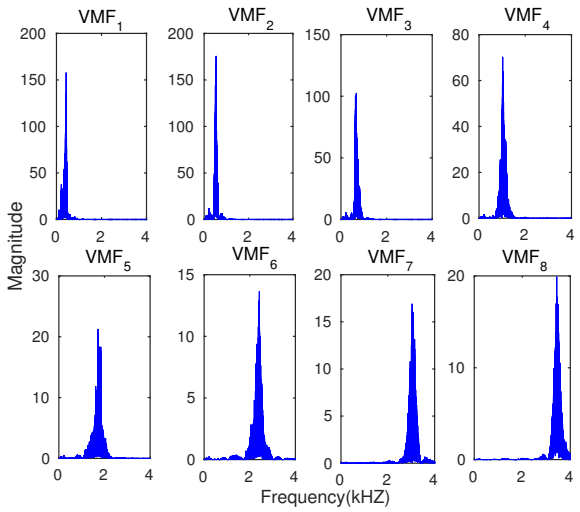

Figure 2: *Magnitude spectrum of VMFs for a speech signal. The modes are arranged from low- to high-frequency band (left to right).*

where, $M(k)$ and $2l+1$ represents the MSM at sample point k and the length of the analysis frame, respectively.

The transitions in MSM at VOPs and VEPs differ depending on the context of speech signal. In noisy speech signal, MSM also fluctuates depending on the type of noise and SNR. To reduce the fluctuations and sharpen the transitions at the VOPs and VEPs, MSM at each time instant is non-linearly mapped using negative exponential as follows:

$$M_{nl}(k) = \exp\left(-\frac{M(k)}{\zeta}\right) \tag{6}$$

where, $M_{nl}(k)$ represents the NLM-MSM) and ζ is a real constant.

The significant transition points in the NLM-MSM are then detected by convolving it with a 100 ms long FODG having a standard deviation one sixth of the window length. In the convolved output, the valleys and peaks correspond to the VOPs and VEPs, respectively. The VOPs and VEPs are detected by considering the sum of the magnitude of the valley and the corresponding peak. If sum of the magnitudes is above a predefined threshold, the valleys and corresponding peaks are hypothesized as the VOPs and VEPs, respectively. The regions between them are selected as the vowels.

The proposed approach for detecting vowels in clean and noisy speech signals are illustrated in left and right panes of Fig. 3, respectively. In the left pane, the VMFs reconstructed speech for a segment of clean speech (Fig. 3 (a)) is shown in Fig. 3 (e). The MSM, NLM-MSM and vowel detection evidence are shown in Fig. 3 (b)-Fig. 3 (d), respectively, when the MSM is directly computed from the speech signal. The MSM, NLM-MSM and vowel detection evidence are shown in Fig. 3 (f)-Fig. 3 (h), respectively, when the MSM is computed from the VMFs reconstructed speech signal following the proposed approach. It is evident from the figures that, the energy due to high frequency sound unit /ch/ is significantly de-emphasized in VMFs reconstructed speech signal. As discussed earlier, it can also be observed that, the fluctuations of the MSM within the vowel regions are significantly reduced in NLM-MSM. This also sharpen the transitions at the VOPs and the VEPs. This in turn, reduces the spurious detection of vowels and improve the detection accuracy of VOPs and VEPs. The same trends are also observed when the speech signal is corrupted by 0 dB white noise (right pane). From the *vowel detection evidences*, it may be observed that, the detected vowels (solid blue lines) nearly matches the reference ones (black dash-dot lines). By comparing the detected and reference vowels for the clean and noisy speech signals, it may be noted that the proposed feature is highly robust towards the ill-effects of environmental noise.

3. Experimental results and discussions

In this work, TIMIT database [23] is used for performance evaluation. Test dataset consists of 300 randomly selected utterances, equally divided between male and female speakers. Performance of the proposed and existing methods are evaluated on this dataset. A development set consisting of 50 utterances is used for optimizing the tunable parameters of the proposed and existing methods. To simulate noisy test conditions, three different noises namely White, Factory and Babble noise from the NOISEX-92 database [24] are added to the speech files. The energy level of the noise is varied so that the SNR of the noisy speech is either 0, 5 or 10 dB.

In this study, we have applied 8-level decomposition of speech signal using VMD technique. For VMD, the data fidelity constraint balancing parameter α, time-step τ and tolerance of convergence were set as 320, 0 and 10^{-7}, respectively. For computing the NLM-MSM, the length of the analysis frame (l) and value of the constant (ζ) are selected as 100 to 0.02, respectively. All these values are selected empirically using the development dataset. For all the experiments, these values are kept constant to simulate a realistic testing conditions.

The system performance are governed by using the following parameters: *Sample identification rate (SIR):* the percentage of reference vowel samples that completely match with the detected vowel samples and *Sample spurious rate (SSR):* the percentage of detected vowel samples which lie outside the reference vowel regions. For a more detailed analysis, SSR is further broken into following three categories: *SSR for semivowels:* The percentage of the detected vowel samples that exactly match with reference semivowel samples; *SSR for nasals:* The percentage of the detected vowel samples that exactly match with reference nasal samples; *SSR for others:* The percentage of the detected vowel samples that match with other speech samples (excluding vowels, semivowels and nasals).

The proposed approach is compared with two existing vowel detection techniques reported in [13, 20]. In [13], the glottal closure instants (GCIs) were detected using the zero frequency filter (ZFF) speech signal. Anchoring the GCIs, short-term DFT magnitude spectrum was computed for the speech samples present in the 30% of glottal cycle. The spectral energies for those regions in the frequency band of 500-2500 Hz was used as the feature. The features were further smoothed over 50 ms regions to suppress the fluctuations. Next, the slope values were computed using the first order difference and the peaks having lower slope values were eliminated. The desired peaks are then locally enhanced by finding corresponding zero crossings. The enhanced feature was convolved with FODG and the vowels are detected following similar procedure as explained in the proposed approach. In rest of the paper this approach is termed as *SPE-GCI*. In [20], an estimate of the speech signal at each time instant was obtained using the NLM. Then the cumulative sum of the short-term DFT spectrum was used

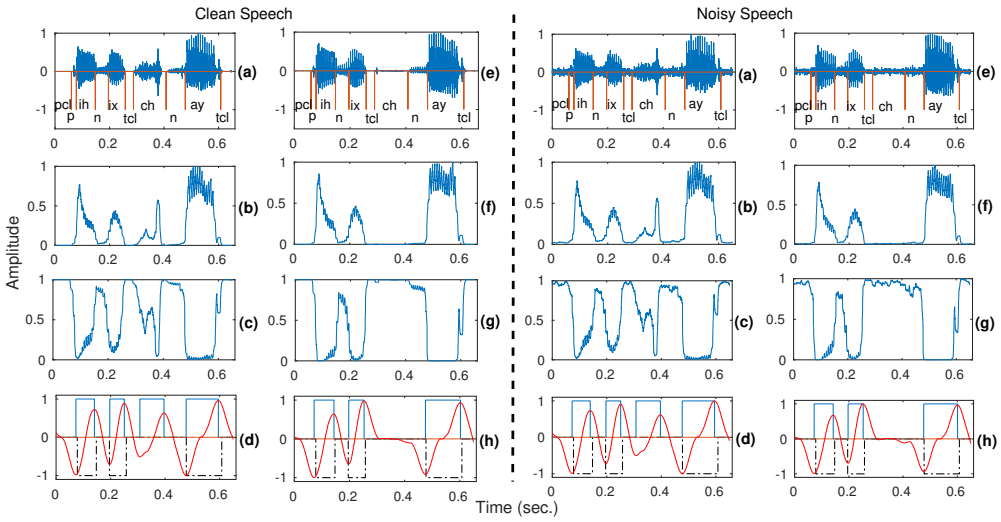

Figure 3: *Plots illustrating the proposed approach for detecting vowels in a given speech signal under clean and noisy conditions. Left pane (a) A segment of speech taken from the TIMIT database with reference marking of sound units as given in the database. (b)-(d) MSM, NLM-MSM, vowel detection evidence with reference (black dash-dot lines) and detected (blue solid line) vowels,respectively, when the MSM directly computed from the speech signal. (e) Speech signal reconstructed from the slected VMFs. (f)-(h) MSM, NLM-MSM, vowel detection evidence,respectively, when MSM computed from the reconstructed speech signal. Right pane represent respective plots when the speech signal is corrupted by* 0 *dB white noise.*

Table 1: *Performances of the proposed and explored methods for the task of detecting vowels for clean as well as noisy speech signals.*

SNR	Method	SIR in %	SSR in %		
			Semivowel	Nasal	Other
Clean speech					
	SPE-GCI	65.63	10.87	1.95	3.86
	NLM-SPE	67.67	10.19	1.12	6.42
	Proposed	88.31	9.13	1.45	3.02
White Noise					
	SPE-GCI	65.14	10.95	2.13	12.20
10 dB	NLM-SPE	66.49	10.85	1.55	7.05
	Proposed	81.05	11.65	2.08	11.50
	SPE-GCI	64.10	11.14	2.20	13.45
5 dB	NLM-SPE	66.46	10.88	2.01	7.45
	Proposed	80.75	11.89	2.42	12.20
	SPE-GCI	63.03	12.39	2.29	15.12
0 dB	NLM-SPE	65.42	11.09	2.29	7.93
	Proposed	78.34	12.32	3.78	13.48
Factory Noise					
	SPE-GCI	66.00	10.43	2.18	14.58
10 dB	NLM-SPE	68.41	10.85	1.10	7.55
	Proposed	85.47	10.52	1.71	10.33
	SPE-GCI	64.55	10.88	2.21	15.14
5 dB	NLM-SPE	67.11	10.96	1.83	8.77
	Proposed	83.30	10.88	1.94	11.63
	SPE-GCI	62.90	11.50	2.88	16.02
0 dB	NLM-SPE	65.74	11.17	2.28	13.11
	Proposed	81.19	11.05	2.24	12.06
Babble Noise					
	SPE-GCI	62.69	11.83	2.17	21.72
10 dB	NLM-SPE	68.04	10.54	1.87	10.88
	Proposed	81.89	10.00	2.01	12.11
	SPE-GCI	61.24	12.49	2.87	25.55
5 dB	NLM-SPE	67.58	11.08	2.15	18.65
	Proposed	78.43	11.04	3.20	13.15
	SPE-GCI	59.88	12.82	3.20	29.32
0 dB	NLM-SPE	66.41	11.33	2.66	23.94
	Proposed	76.23	11.07	3.45	14.57

as the front-end feature. The features was smoothed over 50 ms regions to suppress the fluctuations. The smoothed feature was convolved with FODG and the vowels are detected following similar procedure as explained in proposed approach. In rest of the paper this approach is termed as *NLM-SPE*.

The SIR and SSR obtained by proposed and explored methods are enlisted in Table 1. On comparing the SIR and SSR it can be observed that, performance of the proposed approach is superior than the existing methods for clean as well as noisy speech signals. As discussed earlier from these experimental results it is evident that, the VMFs having higher center frequencies mostly contains energy due to high frequency sound units and noise components.

4. Conclusions

In this work, a novel vowel detection algorithm has been proposed. The VMD is performed on the speech data to decompose it into narrow band VMFs. Then, the signal is reconstructed using a set of selected VMFs whose center frequency corresponds to the frequency range of vowels. The NLM-MSM computed from the reconstructed speech signal is used as the front-end feature for detecting vowels in a given speech signal. The experimental results presented in this study show that, the proposed approach of detecting vowel provides significantly improved performance when compared with the state of the art methods.

5. References

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

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

[3] 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.

[4] 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.

[5] J. Yadav and K. S. Rao, "Detection of vowel offset point from speech signal," *IEEE Signal Processing Letters*, vol. 20, no. 4, pp. 299–302, Apr. 2013.

[6] R. A. Fox and E. Jacewicz, "Cross-dialectal variation in formant dynamics of american english vowels," *The Journal of the Acoustical Society of America*, vol. 126, no. 5, pp. 2603–2618, Nov. 2009.

[7] E. Jacewicz, R. A. Fox, and J. Salmons, "Regional dialect variation in the vowel systems of typically developing children," *Journal of Speech, Language, and Hearing Research*, vol. 54, no. 2, pp. 448–470, Oct. 2011.

[8] N. Fakotakis, A. Tsopanoglou, and G. Kokkinakis, "A text-independent speaker recognition system based on vowel spotting," *Speech Communication*, vol. 12, no. 1, pp. 57–68, Mar. 1993.

[9] 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.

[10] 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.

[11] 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, Jul. 2013.

[12] S. Deb and S. Dandapat, "Emotion classification using segmentation of vowel-like and non-vowel-like regions," *IEEE Transactions on Affective Computing*, vol. 99, pp. 1–15, Jul. 2017.

[13] 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.

[14] 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.

[15] A. Kumar, S. Shahnawazuddin, and G. Pradhan, "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.

[16] 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.

[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. 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.

[19] 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.

[20] 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.

[21] A. K. Vuppala and K. S. Rao, "Vowel onset point detection for noisy speech using spectral energy at formant frequencies," *Int. J. Speech Technol.*, vol. 16, no. 2, pp. 229–235, Jun. 2013.

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

[23] 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.

[24] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effct of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, Jul. 1993.