INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Non-Local Estimation of Speech Signal for Vowel Onset Point Detection in Varied Environments

Avinash Kumar, S. Shahnawazuddin and Gayadhar Pradhan

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

Abstract

Vowel onset point (VOP) is an important information extensively employed in speech analysis and synthesis. Detecting the VOPs in a given speech sequence, independent of the text contexts and recording environments, is a challenging area of research. Performance of existing VOP detection methods have not yet been extensively studied in varied environmental conditions. In this paper, we have exploited the non-local means estimation to detect those regions in the speech sequence which are of high signal-to-noise ratio and exhibit periodicity. Mostly, those regions happen to be the vowel regions. This helps in overcoming the ill-effects of environmental degradations. Next, for each short-time frame of estimated speech sequence, we cumulatively sum the magnitude of the corresponding Fourier transform spectrum. The cumulative sum is then used as the feature to detect the VOPs. The experiments conducted on TIMIT database show that the proposed approach provides better results in terms of detection and spurious rate when compared to a few existing methods under clean and noisy test conditions.
Index Terms: non-local means estimation, vowel region, vowel onset point.

1. Introduction

Vowel onset points (VOPs) are the instants where vowel regions in a speech sequence start [1–3]. The vowels are near-periodic, high signal-to-noise ratio (SNR) and longer duration sound units [4]. Considering the difference in excitation source and vocal tract system characteristics in vowels, several signal processing [2, 3, 5–7] and statistical methods [8–12] have been proposed to detect VOPs/vowels in a speech sequence. The signal processing features mostly employed for these tasks include the difference in the energy of each of the peaks and their corresponding valleys in the amplitude spectrum [1], the zero-crossing rate, the energy and the pitch information [8], the wavelet scaling coefficients [13], the Hilbert envelop of the linear prediction (LP) residual [14], the spectral peaks, the modulation spectrum energies [2], the spectral energy around the glottal closure instants (GCIs) [3], the Mel-frequency cepstral coefficients (MFCCs) and the sub-band energies of the LP residual around the GCIs [12].

Detecting the VOPs in a given speech sequence, independent of the text contexts and recording environments, has remained a challenging area of research. In a continuous speech utterance, the change in the excitation source and vocal tract characteristics are not prominent at the semivowel to vowel transitions as well as for diphthongs [15]. In the case of noisy environments, most of these characteristics also deviate depending on type of noise and the SNR [6]. For instance, the VOP detection method reported in [2] depends on the spectral peaks and LP residual. Hence, this method may fail to detect the VOPs accurately due to the modification in the nature of LP residual as well as speech spectrum depending on the type of noise and the SNR. Similarly, the VOP detection method reported in [3], may fail in those cases where the noise is either impulsive or speech like. Hence, for a better detection of VOPs, the detection method is required to preserve the characteristics of the excitation source and the vocal tract system for the vowel sound units in a given speech sequence irrespective of the recording environments.

In the presented work, we have exploited the patch based non-local means estimation (NLM) [16–18] to detect the high SNR regions exhibiting periodicity over a longer duration. The NLM algorithm finds an estimate of each of the samples in the patch under consideration as a weighted sum of values at other similar sample points in a search-neighborhood of the given speech signal. Consequently, the estimation is not expected to critically depend on the recording environments. Furthermore, during the computation of weights in NLM, most of the low SNR and non periodic speech samples are deemphasized due to unavailability of similar samples in the neighborhood. As a result, the estimated speech contains only long duration high SNR speech regions. Mostly, those regions are the vowel regions. This helps in overcoming the ill-effects of environmental degradations. Next, for each short-time frame of denoised speech sequence, we cumulatively sum the magnitude of the corresponding discrete Fourier transform (DFT) spectrum. For this, the spectrum corresponding to the frequency range of 500-2500 Hz is considered. The magnitude spectrum in the stated frequency range has significantly higher values for the vowel sound units. Finally, the cumulative sum is smoothed and enhanced using a sigmoid function for better discrimination between vowel and non-vowel regions.

The rest of the paper is organized as follows: Section 2 discusses the NLM estimation of speech signal. The proposed method for detection of VOPs is explained in Section 3. The exiting VOP detection techniques are discussed in Section 4. The experimental results are presented in Section 5. Finally, the paper is concluded in Section 6.

2. NLM estimation of speech signal

Non-local means estimation is a very successful patched-based image denoising technique [16–18]. To the best of our knowledge, the NLM estimation has not been used for speech signal enhancement yet. In the case of NLM algorithm, an estimate of each of the samples in the signal is derived using other samples in a search-neighborhood. In other words, for a given sample j, the estimated signal $\hat{y}(j)$ is computed as the weighted sum of values at other sample points k. As illustrated in Figure 1 using a dummy signal, j and k corresponds to the centers of local patches, respectively, which lie within a search-neighborhood

http://dx.doi.org/10.21437/Interspeech.2017-624

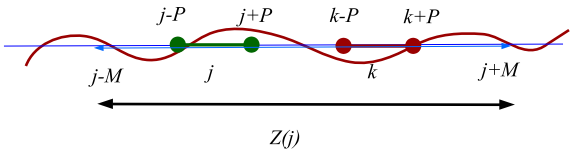

Figure 1: *Illustration of NLM algorithm using a dummy signal.*

of $Z(j)$. Each patch has L_δ samples with δ representing a patch of samples ranging from $-P : P$ such that $L_\delta = (2P+1)$. The estimated signal is computed as follows [18]:

$$\hat{y}(j) = \frac{1}{Y(j)} \sum_{k \epsilon Z(j)} w(j,k) y(k) \tag{1}$$

where, $Y(j) = \sum_k w(j,k)$ and $w(j,k)$ is the weight value given as:

$$w(j,k) = \exp\left(- \frac{\sum_{\Delta \epsilon \delta} (y(j+\Delta) - y(k+\Delta))^2}{2 L_\delta \kappa^2} \right) \tag{2}$$

Eq. (2) can also be written as:

$$w(j,k) = \exp\left(- \frac{\sum_{\tau \epsilon \delta} d^2(j,k)}{2 L_\delta \kappa^2} \right) \tag{3}$$

where, d represents the difference between sample points belonging to the patches centered at points j and k, respectively. The difference value is summed over δ and normalized in order to get the weight value. The bandwidth parameter κ controls the amount of smoothing to be applied. Several weighting techniques for NLM estimation have been proposed [17, 19]. Among those, square patches with centered point of reference is the most preferred one. In order to get a smoother output in the case of image denoising, a patch correction method is also employed as follows [17]:

$$w(j,j) = \max_{k \epsilon Z(j), k \neq j} w(j,k). \tag{4}$$

The above correction approach avoids $w(j,j) = 1$ which corresponds to weighing of similar patches. In this work, this over-smoothing step is avoided. From Eq. (2), it is evident that the NLM technique exploits the non-local information present in the signal. The patch difference is taken over whole neighborhood range $Z(j)$. It is to note that the weight value depends only on the similarity between the patches and not on the distance between them. Averaging over correlated patches provide better sample estimate.

The performance of NLM estimation depends on the critical parameters like bandwidth parameter κ, patch width L_δ (where $L_\delta = 2P + 1$) and neighborhood-width $N(m)$ (where $N(m) = 2M + 1$), where M is the half-neighborhood-width. The bandwidth parameter κ control the smoothness applied to the signal as mentioned earlier. For image denoising, the proper value of κ is selected to be 0.5σ, where σ is standard deviation of the noise [18]. Size of patch width, which is specified by patch-half-width (P) is often similar to the smallest feature size. Next, important parameter is width of neighborhood ($N(m)$) specified by half neighborhood-width M. Larger value of M will lead to enhanced performance due to better averaging with increased computational complexity.

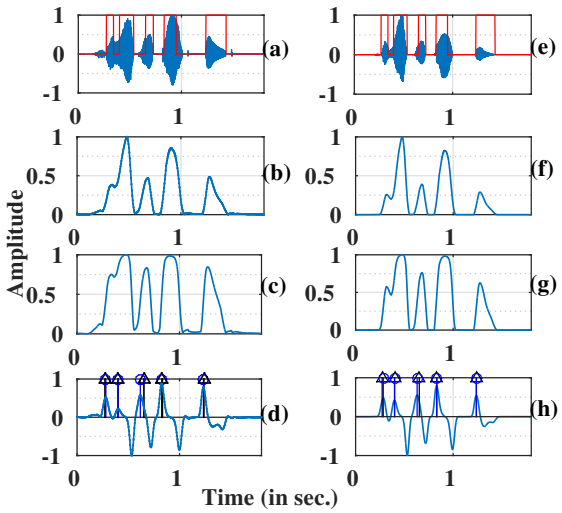

Figure 2: *(a) A segment of clean speech signal taken from TIMIT database with reference vowel regions. (b) Smoothed feature obtained by cumulatively summing the magnitude spectrum obtained from the short-term DFT. (c) Sigmoidal enhanced feature. (d) VOP evidence with hypothesized VOPs (circle symbol) and true VOPs (arrow symbol). (e) Speech signal estimated using NLM. (f)-(g) Corresponding smoothed feature, sigmoidal enhanced feature and VOP evidence, respectively.*

3. Proposed VOP detection method

In the proposed method, NLM is used first for preserving the high SNR regions in a speech sequence and suppressing the ill-effect of noise. Since speech is a non-stationary signal, in the present work, the patch-half-width and the half neighborhood-width are chosen to be 2 ms and 20 ms, respectively. This selection is made to maintain the stationarity property of speech signal within the analysis frame. The standard deviation of the noise is computed using ten frames in the entire signal having the least energy. For computing the energy, the given speech signal is processed in overlapping short-time frames of duration 20 ms with a frame-shift of 10 ms. The bandwidth parameter κ is fixed at 0.7σ. During the experiments performed on a development set, it was observed that the variation of κ from 0.5σ to 0.9σ does not greatly affects the performance. An example of a segment clean speech signal and corresponding signal degraded by 5 dB babble noise are shown in Figure 2(a) and Figure 3(a), respectively. The corresponding NLM estimated speech signals for clean noisy cases are shown in Figure 2(e) and Figure 3(e), respectively. It can be observed that the NLM estimated speech contains only high SNR periodic sound units.

After enhancing the given speech using NLM, the signal is processed through the following sequence of steps for obtaining the VOP evidence.

I. The enhanced speech signal is processed in short-time frames that are 20 ms in duration with 50% overlap for the computation of short-term DFT magnitude spectrum. Next, the cumulative sum of the magnitude spectrum corresponding to vowel regions (500-2500 Hz) is computed and smoothed over 50 ms duration with a frame-shift of 1 ms. The smoothed cumulative sum of the magnitude spectrum for the clean and noisy speech signals are shown in Figure 2(b) and Figure 3(b), respectively, while those

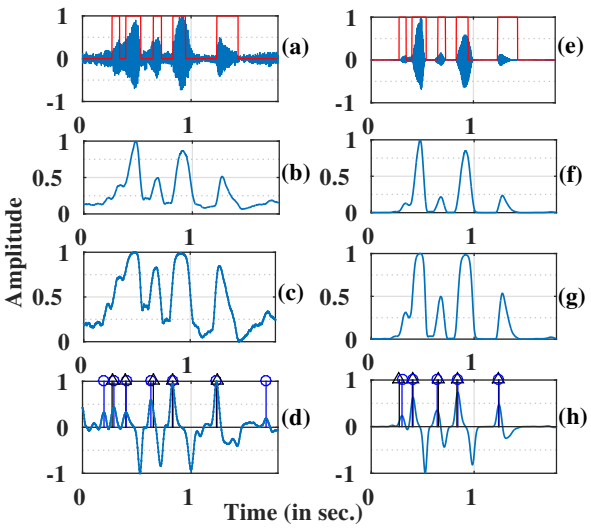

Figure 3: *(a) Speech signal (same as shown in Figure 2(a)) degraded by 5 dB babble noise with reference vowel regions. (b) Smoothed feature obtained by cumulatively summing the short-term DFT magnitude spectrum of the noisy speech signal. (c) Sigmoidal enhanced feature. (d) VOP evidence with hypothesized VOPs (circle symbol) and true VOPs (arrow symbol). (e) Speech signal estimated using NLM. (f)-(g) Corresponding smoothed feature, sigmoidal enhanced feature and VOP evidence, respectively.*

for the NLM enhanced cases are shown in Figure 2(f) and Figure 3(f), respectively.

II. The cumulative sum of the smoothed magnitude spectrum is then processed by sigmoidal function to enhance the relatively low SNR vowel regions as:

$$w_s(n) = \frac{1 - w_{sm}}{1 + \exp\{-\lambda(M(n) - T_h)\}} + w_{sm} \qquad (5)$$

where, $w_s(n)$ is sigmoidal function of smoothed magnitude spectrum $M(n)$, λ is the slope parameter set to 5, T_h is the threshold derived from the mean value of the signal $M(n)$ and w_{sm} is minimum value of sigmoidal function (set to zero in the present case). The output $w_s(n)$ is used as the feature for obtaining the VOP evidence. The smoothed cumulative sum of the magnitude spectrum after sigmoidal enhancement for clean and noisy cases shown in Figure 2(c) and Figure 3(c), respectively, while those for the NLM enhanced cases are shown in Figure 2(g) and Figure 3(g), respectively.

III. The VOP evidence from the feature is obtained by convolving it with a first order Gaussian differentiator (FOGD) window of length 100 ms and standard deviation being one sixth of the window [2]. Peak locations in the evidence are hypothesized as VOPs. The VOP evidences with hypothesized VOPs for the clean and noisy cases are shown in Figure 2(d) and Figure 3(d), respectively, while those for the NLM enhanced cases are shown in Figure 2(h) and Figure 3(h), respectively.

From the above discussions, it can be concluded that the ill-effect of noise is highly suppressed during NLM estimation. At the same time, the proposed feature is discriminative for vowel and non-vowel sound units.

4. Existing VOP detection methods

For comparing the performance of proposed VOP detection method, two state-of-the-art VOP detection approaches are considered in this study [2, 3]. As suggested in the first method [2], three features viz. the Hilbert envelope (HE) of the LP residual signal, the sum of the ten largest peaks in the DFT spectrum and the modulation spectrum energy of the input speech signal are computed. Next, the features are smoothed over 100 ms regions and then enhanced by computing the slope using the first order Gaussian difference. The evidences for each of those smoothed features are obtained by individually convolving with a first order Gaussian difference (FOGD) operator. Finally, a combined evidence is obtained by combining the three evidences sample by sample. In rest of the paper this approach is termed as COMB-EVI method.

In the second existing VOP detection approach [3], first the GCIs are determined by zero frequency filtering (ZFF) of the speech signal. Next, the DFT spectrum is computed for the speech samples present in 30% of glottal cycle starting from the GCI. The spectral energy is then computed within the frequency band of 500-2500 Hz. The smoothing and enhancement of spectral energy is done by following the similar procedure as suggested in [2]. The final VOP evidence is obtained by convolving the smoothed spectral energy with a FOGD operator. In rest of the paper this approach is termed as SE-GCI method.

5. Experimental results and discussion

In this section, we present the performance of proposed approach and compare it with two state-of-the-art VOP detection methods under clean and noisy conditions.

5.1. Experimental dataset

In this work, TIMIT database is used for performance evaluation. For better statistical validation of the experimental results, three dataset are prepared. Each dataset consists of 200 randomly selected utterances, equally divided between male and female speakers. Performance of the proposed and existing methods are evaluated on each of these datasets and the performance reported are averaged over those three datasets. A development set consisting of 400 utterances is used for optimizing the tunable parameters as discussed earlier. To simulate noisy test conditions, three different noises namely White, Factory and Babble noise from the NOISEX-92 database [20] are added to the speech files. The energy level of the noise is varied so that the SNR of the nosiy speech is either 0, 5 or 10 dB.

5.2. Metrics for performance evaluation

For the proposed and existing methods, the peaks in the respective VOP evidences are marked as VOPs. Using the manual markings given in the TIMIT database as the reference, the performances of the detected VOPs are measured using the following metrics:

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

In earlier reported works, the IR was evaluated only for deviation values ranging from 10-40 ms in steps of 10 ms [2]. For a better comparison, the IR in this work is measured by varying the deviation values from 5 ms to 40 ms in steps of 2 ms.

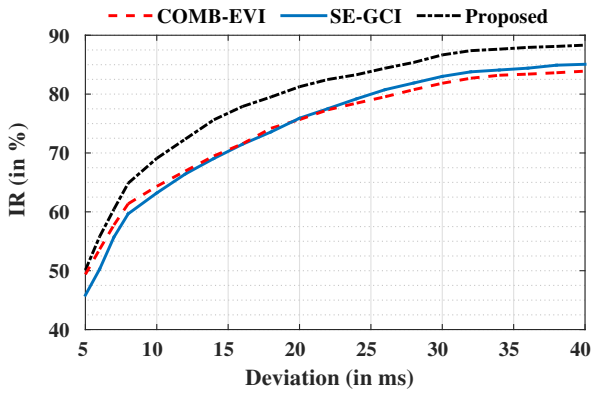

Figure 4: *The IR profiles for the VOP detections under clean testing conditions. The IR profiles are given for the COMB-EVI, SE-GCI and the proposed methods. The predefined deviation is varied from 5 ms to 40 ms in steps of 2 ms.*

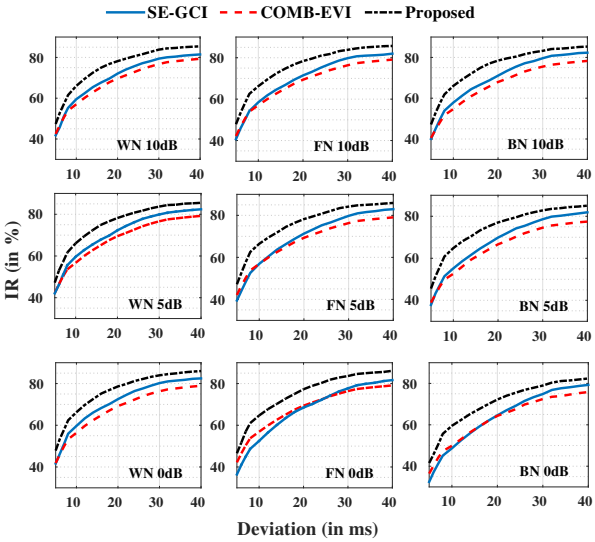

Figure 5: *The IR profiles for the VOP detections for noisy testing conditions. The IR profiles are shown for the COMB-EVI, SE-GCI and the proposed methods. For this study, three different noises (white (WN), factory (FN) and babble (BN)) are added to the given speech signal.*

5.3. Experimental results

The identification rate (IR) for the VOP detection with respect to the clean speech signal for the proposed, COMB-EVI [2] and SE-GCI [3] methods are summarized in Figure 4. The predefined deviation values considered in this study are varied form 5 ms to 40 ms in steps of 2 ms. It is evident from the shown profiles that, for all the considered deviation values, the proposed method provides significant improvement in IR compared to the existing techniques. For instance, the IR for the case when the deviation is 10 ms, is 69.09%, 64.33% and 63.20% for the proposed, the COMB-EVI and the SE-CGI, methods, respectively. Similarly, the IR for the 40 ms deviation case is 88.32%, 83.90% and 85.06%, respectively.

The IR profiles for the existing and proposed VOP detection methods under noisy testing conditions are summarized in Figure 5. It is evident from these results that, for all the considered deviation values and noises, the proposed method provides significantly improved IR compared to the existing techniques explored in this work. For example, for the 0dB SNR white noise case with deviation being 10 ms, the IR is 66.25%, 56.38% and 59.67% for the proposed, the COMB-EVI and the SE-CGI, methods, respectively. Similarly, for the 5dB SNR factory noise case with deviation being 20 ms, the IR is 78.21%, 69.29% and 71.29%, respectively. For the 10dB SNR babble noise case with the deviation being 30 ms, the IR is 83.23%, 75.48% and 79.60%, respectively. Furthermore, it can be observed that, for both clean and noisy speech signals, the IR improvements are much better when compared to the existing methods when the deviation is low. For most of the applications this is required in order to analyze the transition regions of the vowels.

Table 1: *Spurious rates for VOP detection using the COMB-EVI, SE-GCI and the proposed methods. The SRs are given with respect to the clean as well as noisy test conditions.*

Noise	SNR	COM-EVI	SE-GCI	Proposed
Clean speech		12.85	7.27	4.99
White	10	27.30	7.20	5.03
	5	27.70	7.34	5.23
	0	28.40	7.85	6.31
Factory	10	27.02	12.99	5.00
	5	27.44	29.94	10.30
	0	27.80	38.13	23.34
Babble	10	30.38	14.22	7.46
	5	31.84	29.12	16.32
	0	32.93	35.86	26.01

The spurious rates for the proposed and the existing VOP detection methods explored in this work are given in Table 1. It is to note that, the proposed approach provides the minimum SR for clean as well noisy cases for all the considered SNR levels. As discussed earlier, for the factory and the babble noise cases, the SE-GCI provides significant spurious detection. This may be due to impulsive and speech like behavior of the noise.

6. Summary and conclusion

The work presented in this paper deals with the detection of VOPs in varied environmental conditions. To do the same, a given signal is processed through the non local mean estimation to re-estimate the high SNR and periodic speech sound units and suppress the ill-effect of noise. Next, for each short-time frame of enhanced speech sequence, the cumulative sum the magnitude spectrum is computed for the range of frequencies lying between $500 - 2500$ Hz. The cumulative sum is then smoothed and processed through the sigmoidal function to enhance relatively low SNR vowel regions. The sigmoidal enhanced feature is used for the detection of vowel onset points. The proposed feature for detecting VOPs in a speech sequence is discriminative and robust towards environmental degradations. The experiments conducted on TIMIT database show that the proposed approach provides better results in terms of detection and spurious rates when compared to state-of-the-art methods under clean and noisy test conditions.

7. References

[1] D. J. Hermes, "Vowel onset detection," *Journal of the Acoustical Society of America*, vol. 87, no. 2, pp. 866–873, February 1990.

[2] S. R. M. Prasanna, B. V. S. Reddy, and P. Krishnamoorthy, "Vowel onset point detection using source, spectral peaks, and modulation spectrum energies," *IEEE Transactions on audio, speech, and language processing*, vol. 17, no. 4, pp. 556–565, May 2009.

[3] 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.

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

[5] K. Vuppala, K. S. Rao, and S. Chakrabarti, "Improved vowel onset point detection using epoch intervals," *AEU-International Journal of Electronics and Communications*, vol. 66, no. 8, pp. 697–700, August 2012.

[6] A. K. Vuppala and K. S. Rao, "Vowel onset point detection for noisy speech using spectral energy at formant frequencies," *International Journal of Speech Technology*, vol. 16, no. 2, pp. 229–235, 2013.

[7] B. Sarma, S. Prajwal, and S. M. Prasanna, "Improved vowel onset and offset points detection using bessel features," in *International Conference on Signal Processing and Communications*, July 2014, pp. 1–6.

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

[9] J. Y. S. R. K. Rao, C. C. Sekhar, and B. Yegnanarayana, "Neural network based approach for detection of vowel onset points," in *International Conference on Advances in Pattern Recognition and Digital Techniques*, vol. 1, December 1999, pp. 316–320.

[10] B. K. Khonglah, B. D. Sarma, and S. R. M. Prasanna, "Exploration of deep belief networks for vowel-like regions detection," in *Annual IEEE India Conference*, December 2014, pp. 1–5.

[11] B. Dev Sarma and S. R. M. Prasanna, "Analysis of spurious vowel-like regions (vlrs) detected by excitation source information," in *Annual IEEE India Conference*, December 2013, pp. 1–5.

[12] A. Kumar, S. Shahnawazuddin, and G. Pradhan, "Improvements in the detection of vowel onset and offset points in a speech sequence," *Circuits, Systems, and Signal Processing*, pp. 1–26, 2016.

[13] J. H. Wang and S. H. Chen, "A C/V segmentation algorithm for Mandarin speech using wavelet transforms," in *International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, March 1999, pp. 417–420.

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

[15] S. R. M. Prasanna and G. Pradhan, "Significance of vowel-like regions for speaker verification under degraded condition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2552–2565, May 2011.

[16] A. Buades, B. Coll, and J. M. Morel, "A non-local algorithm for image denoising," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR '05, vol. 2, 2005*, 2005, pp. 60–65.

[17] ——, "A review of image denoising algorithms, with a new one," *Multiscale Modeling & Simulation*, vol. 4, no. 2, pp. 490–530, 2005.

[18] V. Dimitri and K. Michel, "SURE-based non-local means," *IEEE Signal Processing Letters*, vol. 16, pp. 973–976, 2009.

[19] C. Deledalle, V. Duval, and J. Salmon, "Non-local methods with shape-adaptive patches (NLM-SAP)," *Journal of Mathematical Imaging and Vision*, vol. 43, pp. 103–120, 2012.

[20] 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 Communication*, vol. 12, no. 3, pp. 247–251, July 1993.