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

VOP Detection in Variable Speech Rate Condition

Ayush Agarwal, Jagabandhu Mishra, S. R. Mahadeva Prasanna

Department of Electrical Engineering
Indian Institute of Technology (IIT) Dharwad
Dharwad 580011, Karnataka, India
{160020008, jagabandhu.mishra.18, prasanna}@iitdh.ac.in

Abstract

The Vowel onset point (VOP) is the location where the onset of vowel takes place in a given speech segment. Many speech processing applications need the information of VOP to extract features from the speech signal. In such cases the overall performance largely depends on the exact detection of VOP location. There are many algorithms proposed in the literature for the automatic detection of VOPs. Most of these methods assume that the given speech signal is produced at normal speech rate. All the parameters for smoothing speech signal evidence as well as hypothesizing VOPs are set accordingly. However, these parameter settings may not work well for variable speech rate conditions. This work proposes a dynamic first order Gaussian differentiator (FOGD) window based approach to overcome this issue. The proposed approach is evaluated using a subset of TIMIT dataset with manually marked ground truth VOPs. The evaluated performance of VOP detection by using the proposed approach shows improvement when compared with the existing approach at higher and lower speech rate conditions.

Index Terms: vowel onset point (VOP), speech rate, excitation information, zero frequency filtered signal (ZFFS), first order Gaussian differentiator (FOGD) window.

1. Introduction

Vowel onset point (VOP) is a point at which the production of vowel initiates in the continuous speech production process. Vowels are produced by exciting the vocal tract system with the nearly periodic glottal vibrations. As per the speech production process, the consonants are always produced by preceding or following a vowel sound with it. The information of VOP can be used to find the start point of a vowel sound or the end point of a consonant sound. The vowels are the major energy carriers. The knowledge of VOP can be used to find the speech rate, as it gives a rough idea of the presence of the number of syllables. For the development of various speech processing based applications (like speech recognition, speaker recognition, spoken language recognition etc.), the VOP information can also be used as an anchor point in the feature extraction process. Hence motivated to develop methods for enhancing the detection of VOPs.

In literature there are many attempts to detect VOP from speech signals using different evidence present in the speech signal. These evidences are short term energy, zero crossing rate, spectral resonance, pitch [1], energy derivative [2], modulation spectrum [3] and excitation information like Hilbert envelope of LP residual signal and zero frequency filtered signal (ZFFS) [4] etc. This work uses the excitation source based evidence for detection of VOPs. From the speech production point of view, vowels are produced by exciting the vocal tract system by semi-periodic glottal vibrations and the shape of the vocal tract system changes to produce different types of vowel. Therefore excitation evidence provides stable performance to detect VOPs [4].

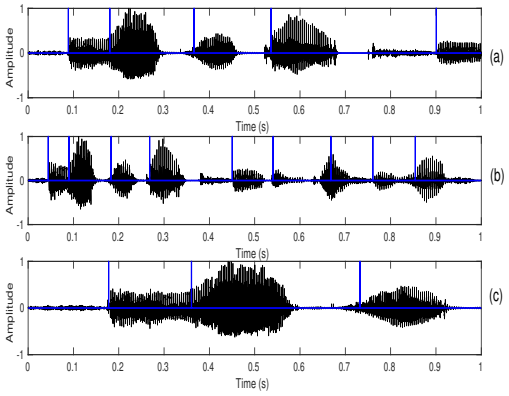

Figure 1: *A small speech segment of text "She had your dark suit in greasy wash water all year" from TIMIT dataset. Impulses are the ground truth VOPs. (a), (b), (c) are speech segments for normal, higher and lower speech rates respectively.*

The existing VOP detection methods are based on the assumption that the average speech rate of a speech segment is about four to five syllables (i.e four VOPs) per second [5]. But in reality the speech rate may be much higher or lower than the average speech rate. Figure 1, shows the number of VOPs present in the one second duration segment at normal, higher and lower speech rate, respectively. From the figure, it can be observed that, in an equal duration segment the number of VOPs increases with increase in the speech rate and vice versa. Therefore, the existing methods may result in miss detection and false detection of VOPs at higher speech rate and lower speech rate, respectively. These errors in VOP detection may cause performance degradation of VOP detection algorithm. This in turn will result in poor speech analysis and feature extraction for further processing. In this work, this issue is addressed by dynamically adapting the VOP detection algorithm parameters.

The chronological order of the paper is as follows: Section 2 gives a brief overview of the existing excitation based methods and a motivation for the proposed work. Section 3 describes the proposed method. Experimental setup, results and discussions are explained in Section 4. Finally, the conclusion and the future directions are presented in Section 5.

http://dx.doi.org/10.21437/Interspeech.2020-2326

2. Existing excitation based VOP detection and motivation

Vowel sounds are produced by exciting the time varying vocal tract system through an impulse like nearly periodic excitation, while the other sound units are produced by exciting through a random excitation or stop impulse. This motivated the community to analyse the excitation evidence to detect VOPs. The intuition behind using the excitation information is, there exists a sudden change in the short term energy contour of the excitation evidence while a speech production system starts producing a vowel. From linear prediction (LP) analysis of speech signal, LP residual signal is termed as the representation of excitation information, thus used in [4] to perform VOP detection. Similarly the positive zero crossing of the zero frequency filtered signal (ZFFS) of a speech segment also represents the excitation information [6], thus used in [4] to perform VOP detection. The brief overview of both the algorithms for detecting VOPs are presented in Sections 2.1 and 2.2, respectively.

2.1. VOP detection using LP residual signal

LP residual of the speech signal is a bipolar signal which conveys two information: (i) location of the glottal closure instants (GCIs) (ii) excitation source energy [3]. As mentioned above, the VOP is the sudden change in source energy. Hence, the contour of the Hilbert envelope (HE) of LP residual is used as it is the close representation of the excitation source energy. The procedure for obtaining the evidence contour is mentioned below [4]:

- Compute LP residual by processing the speech signal in framesize of 20 ms and frameshift of 10 ms with 10^{th} order LP analysis (8 kHz sampling frequency is taken).
- Compute HE of LP residual and determine the excitation source contour of the HE by collecting maximum value within a window of 5 ms and then shifting by 1 sample.
- The final evidence contour can be obtained by convolving the contour with a first order Gaussian differentiator (FOGD) window having window length of 100 ms and a standard deviation of one-sixth of the window length.

2.2. VOP detection using ZFFS

The zero frequency filter (ZFF) attenuates the information related to the vocal tract resonances and preserves the excitation information [6]. The procedure for obtaining the evidence using zero frequency filtered signal (ZFFS) is mentioned below [4]:

- Pass the speech signal through a zero frequency filter (ZFF) followed by a trend removal.
- Compute the second order difference of the ZFFS which represents the change in the strength of excitation at the glottal closure instants(GCIs) and then take its absolute value.
- ZFFS based excitation source contour is determined by collecting maximum value within a window of 5 ms and then shifting by 1 sample.
- The ZFFS based excitation source contour is convolved with the FOGD window to obtain the final evidence contour.

The final combined excitation based evidence contour is obtained by adding the evidence contour obtained from both the LP residual based method and ZFFS based method. To locate the VOPs, the peaks of the final combined evidence are picked. To remove spurious peaks, the peak picking is done such that the peak lies between the positive and negative zero crossing point and also crosses a threshold. An overview of the VOP detection procedure using excitation evidence is depicted in Figure 2.

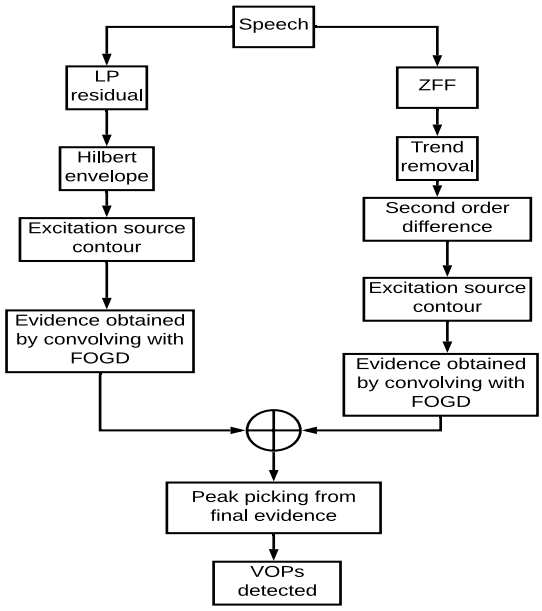

Figure 2: *Block diagram for VOP detection*

2.3. Motivation for the proposed work

The excitation evidence based methods uses FOGD window with a fixed window length of 100 ms and a fixed standard deviation of one-sixth of the window length for detecting the sudden changes present in the excitation source contour. The window length is chosen based on the assumption that the speech is produced at normal rate (average syllables of 4 to 5 per sec) [3]. It has been observed from the Section 1 and Figure 1, that this assumption is not valid when the speech rate changes. In the varying speech rate scenario, using the existing methods it has been found that, there exist many miss detection at higher speech rate and many false detection at slower speech rate. This argument can be clearly observed from the Figure 3. The miss and false detection happens due to the fixed window length, as from the intuition the window length of FOGD should be approximately same as the duration of the vowels. But due to the change in speech rate, the duration of vowel changes (i.e increases at lower speech rate and vice-versa).

If the fixed FOGD window length is used, then due to the change in vowel duration, this may smooth the peaks at higher speech rate and insert spurious peaks at lower speech rate. This argument can be clearly observed from the plots present in the Figure 3(b), (d) and (f). The plots in Figure 3(b), (d) and (f) shows the final combined excitation evidence contour, which are obtained by convolving with the FOGD contour with fixed window length at normal, higher and lower speech rate, respectively. Hence there is a need of changing the FOGD window length and standard deviation length in accordance with the speech rate parameter. In this work, this problem is addressed

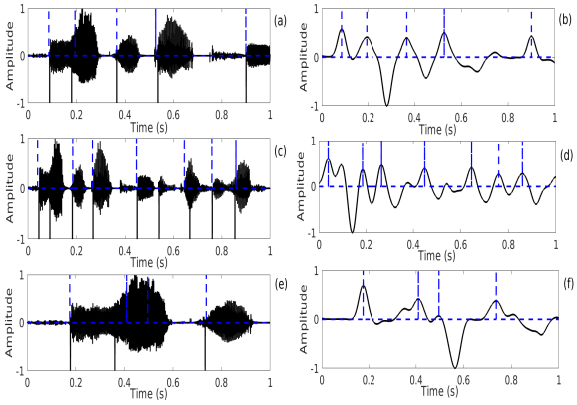

Figure 3: *(a), (c), (e) we have speech segments with speech rate modification factor =1, 2, 0.5 respectively. Impulses with dotted lines are the detected VOPs and negative impulses with solid lines are the ground truth.(b), (d), (f) are the evidence obtained by existing methods for (a), (c), (e) respectively.*

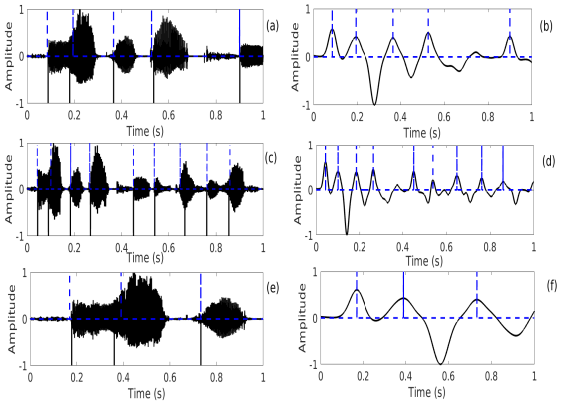

Figure 4: *(a), (c), (e) we have speech segments with speech rate modification factor =1, 2, 0.5 respectively. Impulses with dotted lines are the detected VOPs and negative impulses with solid lines are the ground truth. (b), (d), (f) are the evidence obtained by modified methods for (a), (c), (e) respectively.*

by using a dynamic FOGD window, in which the window length and the standard deviation adjusts itself according to the speech rate of the speech. The detailed description of the proposed method is given in the next section.

3. VOP detection under variable speech rate condition

In this work, for initial study the speech segments having higher and slower speech rate have been generated using the prosody duration modification algorithm given in [7], by changing the duration modification factor β. For $\beta > 1$, the duration of the speech increases, hence the speech rate decreases and for $\beta < 1$ the duration of the speech decreases, hence the speech rate increases.

3.1. Speech rate modification factor

Speech rate modification factor (γ) is the ratio between the modified speech rate and the actual speech rate of a speech segment as given in Equation 1. The Speech rate modification factor is inversely related with the duration modification factor β.

$$\gamma = \frac{MSR}{ASR} \tag{1}$$

where, MSR is modified speech rate representing the number of syllables per second of the modified speech segment [8]. ASR is the actual speech rate representing the number of syllables per second of the original speech segment [8].

3.2. Proposed method

In the proposed method the prime focus is to compute similar evidence irrespective of the speech rate of the speech. As mentioned in Subsection 2.1, the final evidence contour is obtained by convolving the excitation source contour with a FOGD window having fixed window length of 100 ms and the standard deviation of one-sixth of the window length irrespective of the speech rate. But as discussed in the Section 2.3, the window length should be vary dynamically in accordance with the speech rate parameter. In this work to detect VOPs under variable speech rate condition, a modified method of dynamic FOGD window is introduced. The window length of FOGD is

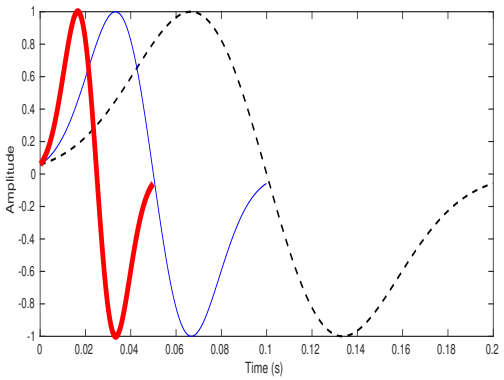

Figure 5: *Dotted line, thin solid line and thick solid line is waveform of FOGD window with γ = 0.5, 1, 2 respectively.*

given in Equation 2 and the standard deviation is given in Equation 3, where f_s is the sampling frequency and γ is the speech rate modification factor of speech signal.

$$w(n) = \frac{100 f_s}{\gamma} \tag{2}$$

$$\sigma(n) = \frac{100 f_s}{6 \times \gamma} \tag{3}$$

Dynamic FOGD window which adapts itself according to the speech rate modification factor is represented in Figure 5. As per Equation 2 and 3, for lower value of γ the window length and variance increases. This can be observed from the window function plotted in dotted line. Similarly, for higher value of γ the window length and variance decreases, can be observed from the thick solid line of Figure 5. The thin solid line window shows the reference window function at normal speech rate.

After adapting the dynamic update of the FOGD window length parameter, from the Figure 4(b), (d) and (f) it can be observed that the final evidence contour looks similar irrespective of the speech rate. This proposed method of dynamically adapting the window length solves the problem of miss and false detection of VOPs in variable speech rate condition.

Table 1: *Performance of VOP on manually marked TIMIT dataset. In the table DET VOPs, DR, IR refer to detected VOPs, deletion rate and insertion rate respectively in %. Dynamic refers to the dynamic FOGD window and standard deviation and Fixed refers to the fixed FOGD window and standard deviation .*

γ	DET VOPs(%)								DR(%)		IR(%)	
	$\pm10(ms)$		$\pm20(ms)$		$\pm30(ms)$		$\pm40(ms)$					
	Dynamic	**Fixed**	**Dynamic**	**Fixed**	**Dynamic**	**Fixed**	**Dynamic**	**Fixed**	**Dynamic**	**Fixed**	**Dynamic**	**Fixed**
0.25	63.29	28.27	82.71	53.01	89.05	70.54	93.06	77.39	6.94	22.61	14.73	175.81
0.5	63.61	51.24	82.71	53.01	89.49	70.54	93.51	86.56	6.49	13.44	13.12	48.92
1	68.51	68.51	85.11	85.11	90.23	90.23	96.14	96.14	3.86	3.86	10.35	10.36
1.5	66.36	70.14	83.41	86.09	89.75	91.88	93.14	93.38	6.86	6.62	14.73	3.15
2	63.21	63.96	81.17	81.01	88.47	84.48	93.14	85.71	6.86	14.29	13.74	0.28
2.5	61.32	59.94	81.41	74.95	89.12	77.15	92.63	77.67	7.37	22.33	16.03	0

4. Experimental setup, results and discussion

VOP detection task has been performed by using the 205 speech segment files of TIMIT dataset with manually marked ground truth of around 2500 VOPs as mentioned in [9]. The performance on the different speech rate with the temporal resolution of 10 ms, 20 ms, 30 ms and 40 ms has been evaluated. The following parameters are used to evaluate the performance.

- *Correctly detected VOPs (DET VOPs)* - The percentage of correctly detected VOP within a temporal resolution.
- *Deletion rate (DR)* - The percentage of VOPs that do not lie within a temporal resolution.
- *Insertion rate (IR)* - The percentage of VOPs that comes more than once within a temporal resolution.

The experiment is carried out in two setups (i) variable speech rate with fixed FOGD window length (ii) variable speech rate with dynamic FOGD window length. The speech with variable speech rate is generated using the suitable β factor in which the duration of the speech is changed keeping the pitch unchanged to preserve the naturalness of the speech [7]. Also the temporal resolution for performance evaluation has to be adjusted according to speech rate modification factor (γ), as the 20 ms temporal resolution for $\gamma = 2$ is equivalent 40 ms at normal speech rate (i.e $\gamma = 1$) and 80 ms for $\gamma = 0.5$. So, it is essential to adjust the temporal resolution according to γ to correctly evaluate the performance. Here, the performance window is adjusted by $temporal\ resolution/\gamma$.

Figure 6 shows the performance based on the percentage of correctly detected VOPs at different values of the speech rate modification factor (γ) under the temporal resolution of ±40 ms. The blue bars in the plot show the obtained performance using the modified method whereas the red bars show the obtained performance using the existing method. From Figure 6 we can observe that fixed FOGD window has a tolerance of γ ranging from 0.75-1.5, where it provides stable performance. Beyond this tolerance range i.e for fast or slow speech the existing excitation evidence based methods shows significant degradation in performance dropping as low as 77.67% at $\gamma = 2.5$ and 77.39% at $\gamma = 0.25$. On the other hand the modified method shows consistent performance under all γ values.

Table 1 shows the evaluated performance evaluation with all the performance parameters at the γ values of 0.25, 0.5, 1, 1.5, 2 and 2.5. From the table we can observe that for modified method, correctly detected VOPs under all temporal resolutions are consistent whereas the fixed window condition shows significant degradation at high and low values of γ. The insertion rate **(IR)** shows large deviation in the fixed FOGD case. For $\gamma = 0.25$ %IR is 175% because of the insertion of spurious

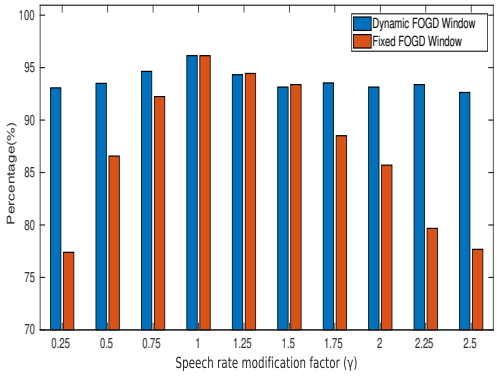

Figure 6: *Correctly detected percentage accuracy in case of dynamic FOGD window and fixed FOGD window for 40ms temporal resolution.*

peaks in the evidence of the excitation (refer Figure 3(f)). In this case, on an average the number of detected VOPs is 1.75 times more than the actual ground truth VOPs. When $\gamma = 2$ *or* 2.5 the %IR is approximately zero and the reason behind this is that the peaks of the evidence are far apart (refer Figure 3(d)). On the other hand, the %IR in case of dynamic window FOGD is consistent because the shape of evidence is similar for all γ values (refer Figure 4(b), (d), (f)). It can also be observed that for fixed FOGD as the γ decreases the %IR increases, but shows consistent %IR when evaluated using the proposed method.

5. Conclusion and Future Work

In this work, the shortcomings of existing excitation VOP detection method is identified and necessary modification is proposed to solve the problem. The existing methods were designed on the assumption that the speech is produced at the normal speech rate. But when the speech rate is high or low, existing methods show significant degradation in performance for VOP detection. After the proposed modification of the existing approach when tested, the obtained performance showed consistency, irrespective of the speech rate variation. In future the aim is first to compute the speech rate modification factor (γ) automatically from the speech signal. Secondly, as there is a possibility of speech rate variation in between a given speech segment. The aim is to dynamically capture the variation of speech rate in between the given speech segment, which may help to further enhance the performance of the proposed approach.

6. References

[1] J.-F. Wang, C.-H. Wu, S.-H. Chang, and J.-Y. Lee, "A hierarchical neural network model based on a C/V segmentation algorithm for isolated mandarin speech recognition," *Signal Processing, IEEE Transactions on*, vol. 39, pp. 2141 – 2146, 10 1991.

[2] C. C. Sekhar and B. Yegnanarayana, "Recognition of stop-consonant-vowel (SCV) segments in continuous speech using neural network models," *IETE Journal of Research*, vol. 42, no. 4-5, pp. 269–280, 1996. [Online]. Available: https://doi.org/10.1080/03772063.1996.11415933

[3] S. R. Mahadeva Prasanna, B. V. Sandeep 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, 2009.

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

[5] A. Cruttenden, *Gimson's Pronunciation of English*. Taylor & Francis, 2014. [Online]. Available: https://books.google.co.in/books?id=M2nMAgAAQBAJ

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

[7] K. S. Rao and B. Yegnanarayana, "Prosody modification using instants of significant excitation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 972–980, 2006.

[8] S. Narayanan and Dagen Wang, "Speech rate estimation via temporal correlation and selected sub-band correlation," in *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 1, 2005, pp. I/413–I/416 Vol. 1.

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