Improved Vowel Onset and Offset Points Detection Using Bessel Features

Biswajit Dev Sarma[1], Supreeth Prajwal S[2] and S. R. Mahadeva Prasanna[1]
[1]Department of EEE, IIT Guwahati, India
[2]Department of ECE, NITK Surathkal,India
Email:{s.biswajit, prasanna}@iitg.ernet.in[1],supreeth.neha@gmail.com[2]

***Abstract*—This work presents a method for improving accuracy of Vowel Onset Point (VOP) and Vowel End Point (VEP) detection in continuous speech. VOP and VEP are the instants at which the onset and offset of vowel takes place, respectively, during speech production. Speech signal is represented using Bessel functions with their damped sinusoid-like basis functions. Bessel expansion is used to emphasize the vowel regions by appropriate consideration of the range of Bessel coefficients. Bandpass filtered narrow-band signal is modeled as a monocomponent amplitude modulated-frequency modulated (AM-FM) signal. The amplitude envelope (AE) function of this vowel emphasized AM-FM signal gives strong evidence for the VOP and VEP. This evidence after adding with some of the existing evidences having source and system information, increases the detection rate as well as the accuracy of detection.**

Index Terms: VOP, VEP, Bessel Expansion, AM-FM signal model, AE function

I. Introduction

Vowel onset point (VOP) is the instant at which onset of the vowel takes place. Similarly, vowel end point (VEP) is the instant at which the offset of vowel takes place [1]. There are many changes that occur at VOP and VEP. Vowels are produced with the mouth wide open and consonants with a narrow or moderate constriction in the vocal tract. Due to this nature of speech production, a sudden change of energy is observed at the VOP and VEP. The characteristics of excitation source, vocal tract transfer function and modulation components etc., change around these instants [2]. There are many applications that may use the knowledge of VOP and VEP. Most of the syllable units in Indian Languages are consonant vowel (CV) and vowel consonant (VC) units and discriminatory information of CV and VC units are present in the region around the VOP and VEP [3]. [4] presents the significance of onset and offset of vowel like regions in speaker verification task. Apart from these, knowledge of VOP is useful in tasks like language identification, expressive speech processing etc [2]. There are many methods in the literature for detecting VOP and VEP. Compared to VOP, the task of detection of VEP is more recent. Some of the classical methods for VOP detection include use of energy, zero crossing rate, pitch information, resonances in the spectrum [1] and neural network models [3] etc. Some of the recent unsupervised methods for both VOP and VEP detection include the use of source information alone [4], and a combination of source, spectral peak and modulation spectrum information [2], [5].

In this work, we present a method for increasing the detection accuracy of these two techniques using Bessel expansion and amplitude modulated-frequency modulated (AM-FM) signal model. Bessel expansion and AM-FM model is used in literature for detection of glottal closure instants and voice onset time [6], [7]. Here, we demonstrate its use for VOP and VEP detection. Speech signal is approximated by a set of Bessel coefficients which emphasizes only low frequency components present in the vowel region. This bandpass filtered narrow-band signal can be modeled as a AM-FM signal. Such a narrow-band signal is observed to have sharp discontinuities at the onset and offset of vowel. The amplitude envelope (AE) of this signal can be obtained using discrete energy separation algorithm (DESA) [6]. The AE can be processed to enhance changes occurring at the onset and offset of vowel using a first order Gaussian differentiator and may be used as evidence for VOP and VEP. The conjecture is that the peaks in the evidence will be close to the actual VOP and VEP due to the sharp discontinuities in the AE. Apart from this, since the principle of extracting the evidence is different compared to existing methods reported in [2], [4], it may add well with them to further increase the combined evidence. This nature of evidence is therefore exploited to increase the accuracy of existing VOP and VEP detection.

The VOP and VEP detection methods using excitation source (ES) information and source, spectral peaks and modulation spectrum energy (SSM) information are implemented. In both the methods, VOP and VEP are detected by picking peaks in the evidences. Sometimes these peaks are much deviated from the VOP and VEP leading to reduction in the accuracy of detection. Such peaks are brought closer to the ground truth VOP and VEP by adding the evidence obtained from the AE function. This is done by exploiting the high resolution property of AE envelope function. The rest of the work is organized as follows: Section II describes analysis of VOP and VEP using Bessel Expansion and AM-FM model. Section III illustrates the VOP and VEP detection procedure. Section IV shows the performance evaluation of the method. Section V summarizes the paper.

978-1-4799-4665-5/14/$31.00 ©2014 IEEE

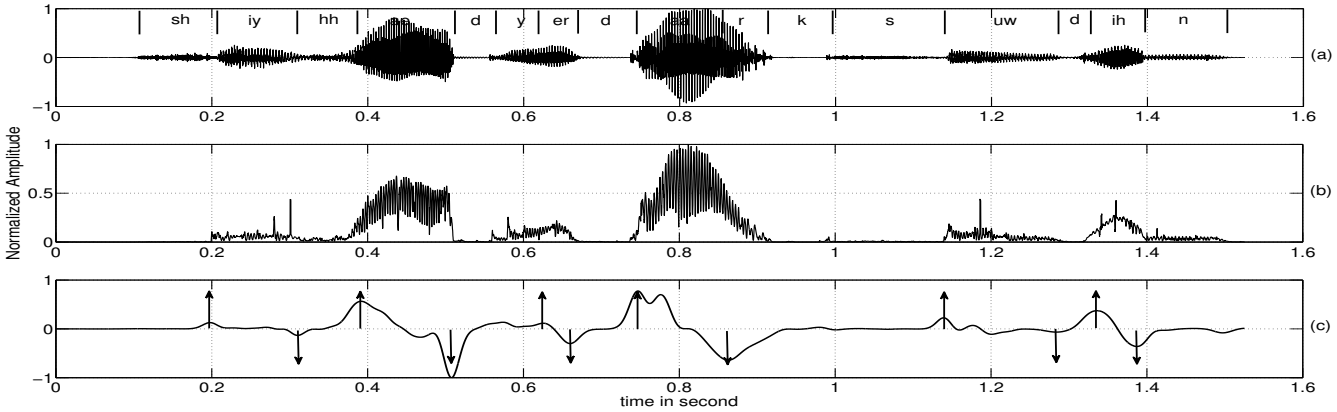

Fig. 1. *Illustration of the procedure for obtaining the VOP and VEP evidence using AE function. a) Speech signal with labels. b) Vowel enhanced AE function of the speech signal. c) Evidence obtained by convolving the vowel enhanced AE function with the first order gaussian differentiator. Arrows show the peaks close to VOPs and VEPs.*

II. Analysis of VOPs and VEPs using Bessel Expansion and AM-FM model

The sinusoidal functions are suitable for representing periodic signals. In case of non-stationary signals like speech, an aperiodic signal set is more efficient for representation. The Bessel functions have regular zero-crossing and decaying amplitude that makes the Bessel functions a good choice as basis functions for efficient representation of speech waveforms [6]. The series expansion of zeroth-order Bessel function of the first kind of a signal x(t) considered over some arbitrary interval (0, a) is expressed as [8]:

$$x(t) = \sum_{p=1}^{\infty} C_p J_0(\frac{\lambda_p}{a}t), \tag{1}$$

where, $J_0(\frac{\lambda_p}{a}t)$ are the zeroth-order Bessel functions and λ_p, $p = 1, 2, ..., \infty$ are the ascending order positive roots of $J_0(\lambda) = 0$. Bessel coefficients C_p are computed by using the orthogonality of zeroth-order Bessel functions $J_0(\frac{\lambda_p}{a}t)$ as:

$$C_p = \frac{2}{a^2[J_1(\lambda_p)]^2} \int_0^a tx(t)J_0(\frac{\lambda_p}{a}t)dt \tag{2}$$

with $1 \leq p \leq P$, where P is the order of Bessel expansion, and $J_1(\lambda_p)$ are the first-order Bessel functions. There is one-to-one correspondence between the frequency component (f_p) of the signal and Bessel coefficient index (p) at which the coefficient attains peak magnitude [6], given by

$$f_p = \frac{pf_s}{2D} \tag{3}$$

where, f_s is the sampling frequency and D is the number of samples in the analyzed signal. The speech signal can be modeled as a multicomponent AM-FM signal [9]. The signal components will be associated with various distinct nonoverlapping clusters of Bessel coefficients, if the AM-FM components of the speech signal are well separated in the frequency domain. Since vowels and consonants have different dominant frequency components, each class can be approximated by a different set of Bessel coefficients. In other words, the signal can be bandpass filtered to enhance only vowel regions by choosing appropriate Bessel coefficients. Bandpass filtering over a range of Bessel coefficients (C_{p1} to C_{p2}) can be computed as:

$$\hat{x}(t) = \sum_{p=p_1}^{p_2} C_p J_0(\frac{\lambda_p}{a}t). \tag{4}$$

where, $\hat{x}(t)$ is the bandpass filtered signal. The vowels have most of the energy in the low frequency band (300 to 1200 Hz) and accordingly Bessel coefficients from $C_{p1=12}$ to $C_{p2=48}$ are used for emphasizing vowel regions (applying Eqn. 3 for f_s=8000 Hz and D=160 samples). Now the bandlimited signal is considered as a monocomponent AM-FM signal and the discrete-time version of the vowel enhanced monocomponent AM-FM signal $\hat{x}[n]$ is given by:

$$\hat{x}[n] = A[n]cos(\phi[n]) \tag{5}$$

where, $A(n)$ is the time-varying amplitude envelope (AE) of $\hat{x}(n)$, with the time-varying phase $\phi[n]$. The amplitude envelope (AE) of the vowel enhanced signal can be obtained using discrete energy separation algorithm (DESA) [9].

$$|A[n]| \approx \sqrt{\frac{\psi[\hat{x}[n]]}{1 - [1 - \frac{\psi[\hat{y}[n]] + \psi[\hat{y}[n+1]]}{4\psi[\hat{x}[n]]}]^2}}, \tag{6}$$

where, $\hat{y}[n]$ is the differenced signal $\hat{y}[n] = \hat{x}[n] - \hat{x}[n-1]$. and $\psi(.)$ is the Teager's nonlinear energy operator given as

$$\psi[\hat{x}[n]] = \hat{x}^2[n] - \hat{x}[n-1]\hat{x}[n+1], \tag{7}$$

The amplitude envelope is approximately calculated by using DESA algorithm as shown in Eqn. 6. Moving average filtering of about 1 ms duration is carried out to smooth the AE function. Figure 1 illustrates the procedure for obtaining the VOP and VEP evidences. Figure 1(a) shows the speech signal for the utterance "She had your dark suit" taken from the TIMIT database. Figure 1(b) shows the vowel enhanced

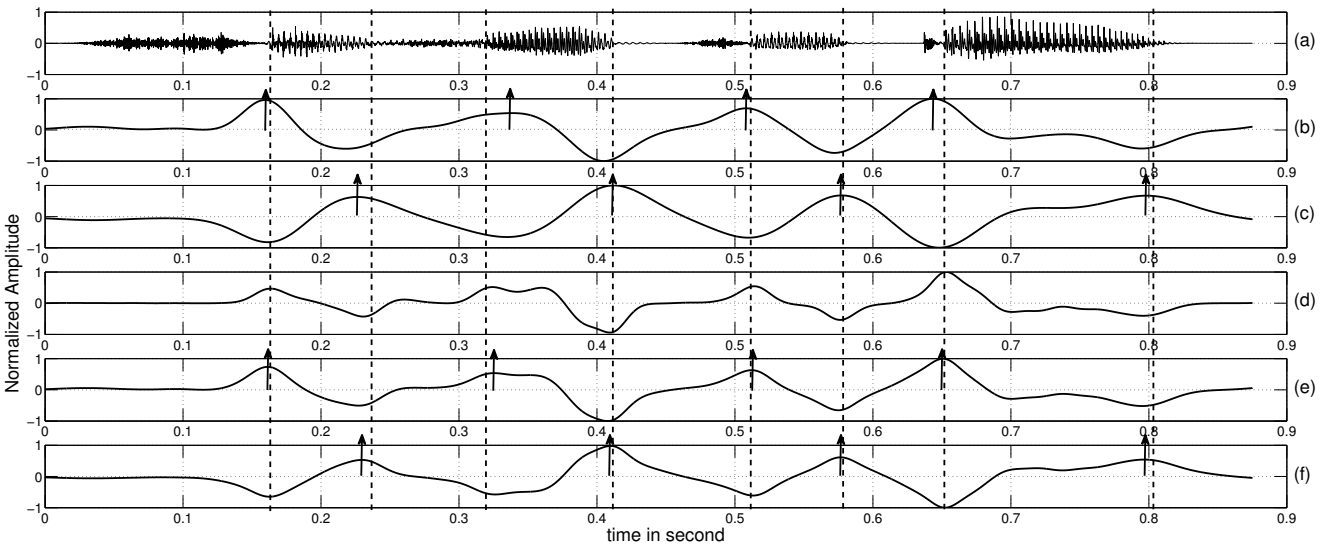

Fig. 2. *Illustration of the procedure for enhancing the ES method based VOP and VEP evidences using AE function. The dotted lines refer to the ground truth VOPs and VEPs. a) Speech signal for the utterance "she had your dark". b) VOP evidence obtained using ES method. Arrows refer to the detected VOPs. c) VEP evidence obtained using ES method. Arrows refer to the detected VEPs. d) VOP and VEP evidence obtained from AE function. e) VOP evidence obtained after adding AE evidence shown in (d) to the ES evidence shown in (b). Arrows refer to the detected VOPs. f) VEP evidence obtained after adding the inverted AE evidence shown in (d) to the ES evidence shown in (c). Arrows refer to the detected VEPs. Detected VOP and VEP are closer to the ground truth (dotted lines), after addition of the AE evidence.*

AE function of the speech signal. It can be seen from the figure that only the vowel region is emphasized and all other regions including fricatives and burst have been significantly attenuated. A close observation on the AE function of the vowel enhanced signal shows its potential in the VOP and VEP detection process. Figure 1(c) shows the evidence obtained by convolving the vowel enhanced AE function with the first order gaussian differentiator (FOGD) of size 100 ms and variance as 10% of window length. The convolved output is the evidence for VOPs and VEPs. The evidence gives a positive peak at the VOP, since the energy change is positive at VOP. Similarly, it gives a negative peak at the VEP because of negative nature of the energy change at VEP. The peaks near the VOPs and VEPs can be observed and are highlighted by arrows. These peaks are very close to the VOPs and VEPs and can help in the automatic detection process.

III. Detection of VOPs and VEPs

The evidence in the Figure 1 clearly shows the positive and negative peaks corresponding to the VOP and VEP, respectively. However, there are some peaks which are not at VOP or VEP. These peaks are because of the energy variation within the vowel region. Therefore, it is difficult to determine which peak is due to a VOP or VEP and which one is spurious. Increasing the variance of gaussian differentiator can eliminate some of the spurious peaks. But, this decreases the resolution of the hypothesized VOP or VEP. Therefore it may be difficult to detect the VOP and VEP using AE envelope, independently. However, this evidence can be used to enhance the performance of some of the existing VOP and VEP detection techniques. Some of the existing VOP and VEP detection methods are,

- by using excitation source information derived from Zero Frequency Filtered Signal (ZFFS) and Hilbert envelope (HE) of linear prediction (LP) residual of speech [4], and
- by using source, spectral peaks and modulation spectrum energies [2].

Evidences obtained from these techniques can be enhanced by adding the AE evidence and onset and offset can be detected more accurately. These two methods are briefly described in the following sub-sections:

A. VOP and VEP detection using excitation source (ES) information [4]

Method described in [4] was used for vowel like regions (vowel and semivowel) onset and offset points. Same method is used here for detection of VOP and VEP. Evidence from HE of LP residual of speech is derived as follows: The HE of LP residual of speech enhances information about GCIs. The smoothed excitation contour by taking maximum value of the HE of LP residual for every 5 ms block with one sample shift is convolved with a FOGD window of length 100 ms and a standard deviation of one sixth of window. The convolution result is the VOP evidence using HE. Evidence for VEP is obtained by doing the convolution operation from right to left instead of left to right as in the case of VOP.

Evidence from ZFFS is obtained as follows: The first order difference of the ZFFS can be treated as strength of excitation at the epochs. The second order difference of ZFFS contains change in the strength of excitation. This change is detected

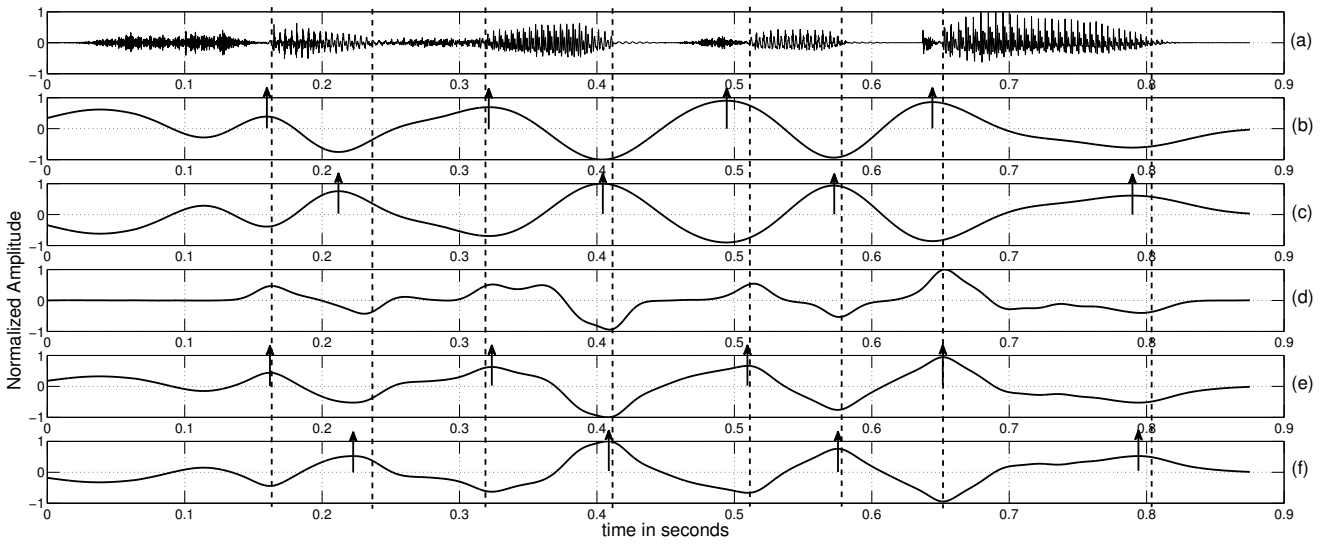

Fig. 3. *Illustration of the procedure for enhancing the SSM method based VOP and VEP evidences using AE function. The dotted lines refer to the ground truth VOPs and VEPs. a) Speech signal for the utterance "she had your dark". b) VOP evidence obtained using SSM method. Arrows refer to the detected VOPs. c) VEP evidence obtained using SSM method. Arrows refer to the detected VEPs. d) VOP and VEP evidence obtained from AE function. e) VOP evidence obtained after adding AE evidence shown in (d) to the SSM evidence shown in (b). Arrows refer to the detected VOPs. f) VEP evidence obtained after adding the inverted AE evidence shown in (d) to the SSM evidence shown in (c). Arrows refer to the detected VEPs. Detected VOPs and VEPs are closer to the ground truth (dotted line), after addition of the AE evidence.*

by convolving with a 100 ms long FOGD having a standard deviation of one sixth of window length. The convolved output is called the VOP evidence using ZFFS. VEP evidence is obtained by convolving from right to left.

The VOP or VEP evidence using the excitation source information is obtained by adding the two evidences and normalizing by the maximum value of sum. The locations of peaks between two successive positive to negative zero crossings of the combined evidence represent the hypothesized VOP or VEP. To reduce missing and spurious ones, an algorithm is used to force the detection of missing cases if other evidence is sufficiently strong, and reduce spurious detection of one event using knowledge of other event [4]. In this work, we will call this method as excitation source (ES) based method.

B. VOP and VEP detection using source, spectral peaks and modulation (SSM) spectrum energies [2], [5]

The Evidence from excitation source information is same as that of the HE of LP residual of speech described in the ES method. The evidence from spectral peaks energy is derived using the following sequence of steps: A 256 point discrete Fourier transform (DFT) is computed for 20 ms speech frame (with 10 ms shift), and ten largest peaks are selected from the first 128 points. The sum of these spectral peaks is plotted as a function of time. The change at the VOP and VEP available in the spectral peaks energy is further enhanced by computing its slope using FOD. These enhanced values are convolved with FOGD operator. The convolved output is the evidence using spectral peaks energy.

Slowly varying temporal envelope of speech signal can be represented by using modulation spectrum. VOP and VEP detection using modulation spectrum energy is obtained using the following sequence of steps: The temporal envelope of speech is dominated by low-frequency components. The VOP and VEP evidence due to modulation spectrum is derived by passing the speech signal through a set of critical bandpass filters, and summing the components corresponding to 4 to 16 Hz. The change at the VOP available in the modulation spectrum energy is further enhanced by computing its slope using FOD. These enhanced values are convolved with FOGD operator and the convolved output is the evidence using modulation spectrum energy.

All three evidences are combined to get final evidence for VOP and VEP. The positive and negative peaks in the combined evidence signal are marked as the VOP and VEP, respectively. In this work, we will call this method as source, spectral peaks and modulation spectrum (SSM) based method.

C. Improved VOP and VEP using Evidence from Bessel Functions

Evidences obtained using ES and SSM methods are enhanced in this work by adding the evidence obtained from the AE function of the vowel enhanced signal. Normally, the evidence from AE function will have a strong peak at the VOP and VEP compared to other speech region within the same vowel. Adding this evidence will enhance the ES and SSM evidence at the VOP and VEP. Even if the peaks at VOP or VEP are not strong enough, but almost comparable, in the combined evidence, the peaks will move towards the VOP or VEP. After adding the evidences same procedure is followed for the respective methods for obtaining the VOP or VEP.

Figure 2 and Figure 3 illustrate the enhancement procedure

for ES and SSM, respectively. Figure 2(a) shows the speech signal for the utterance "she had your dark". The dotted lines are ground truth VOPs and VEPs. Figure 2(b) and (c) show the VOP and VEP evidences of the speech signal shown in Figure 2(a), using ES. Figure 2 (e) and (f) show the VOP and VEP evidences after adding the evidence obtained from the AE function. The ES evidences in Figure 2(b) and (c) have some peaks which are much deviated from the ground truth VOPs/VEPs. In case of combined evidences in Figure 2(e) and (f), the peaks are comparatively closure to VOPs and VEPs. One such case for onset is the peak just right to 0.3 sec. In Figure 2(b), the peak is much deviated from the ground truth which comes closer in Figure 2(e) after combining the proposed evidence. Figure 3 shows similar plots using SSM method. In Figure 3 also, same trend can be observed. For example, the peak just left to 0.5 sec (in Figure 3(b)) is brought closer to the ground truth VOP (in Figure 3(e)) by adding the AE evidence. Similarly, the peak just left to 0.2 sec in Figure 3(c) is brought closer to the ground truth VEP as shown in Figure 3(f).

IV. Performance evaluation

100 sentences from 100 testing speakers of TIMIT database containing around 1000 VOPs and VEPs are used for testing. All VOPs and VEPs are manually marked to obtain the ground truth. The performance of VOP and VEP detection is measured using the following parameters:

- Detection rate (DR): Percentage of VOPs/VEPs that are detected within 40 ms of ground truth;
- Spurious rate (SR): Percentage of VOPs/VEPs that are detected beyond 40 ms of ground truth;
- Detection Accuracy: Percentage of VOPs/VEPs that are detected within 10 ms, 10 to 20 ms, 20 to 30 ms and 30 to 40 ms. This is shown by plotting histograms;

Table I shows the performance of VOP/VEP detection in terms of DR and SR. Performance of AE method is evaluated and it is found that SR is very high. This is due to the spurious peaks in the AE evidence as discussed in section III. Individual performances of ES and SSM are compared with corresponding combined performances (AE+ES and AE+SSM). Improvement is achieved in terms of both DR and SR in terms of increasing DR and reducing SR. Combining all three evidences increases DR, however, SR also increases significantly.

TABLE I
VOP/VEP Detection Performance

Method	VOP		VEP	
	DR (%)	SR (%)	DR (%)	SR (%)
AE	95.41	15.39	90.13	22.67
ES	94.06	8.17	92.14	10.09
ES+AE	95.33	6.63	92.95	8.82
SSM	93.56	9.03	87.21	14.76
SSM+AE	95.12	7.64	89.31	12.87
ES+SSM+AE	95.69	12.50	93.34	15.48

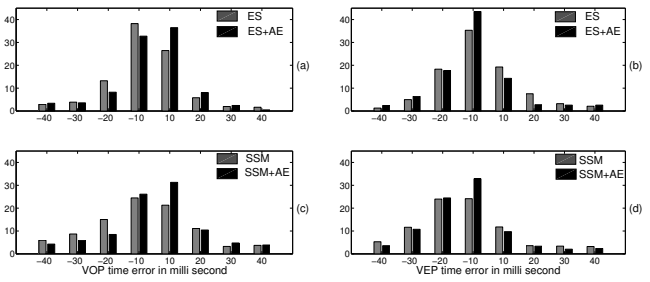

Fig. 4. *VOP/VEP detection accuracy in terms of percentage of VOP/VEP detected within 10, 20 30 and 40 ms time errors. a) VOP detection accuracy with ES and ES+AE evidence, b) VEP detection accuracy with ES and ES+AE evidence, c) VOP detection accuracy with SSM and SSM+AE evidence and d) VEP detection accuracy with SSM and SSM+AE evidence.*

Figure 4 shows the histograms showing the accuracy of VOP/VEP detection. Percentage of VOP/VEP detected within 10 ms, 20 ms, 30 ms and 40 ms region on both sides is plotted to illustrate the gain in accuracy. Figure 4(a) shows the histogram for VOP case before and after adding the AE evidence to the ES evidence. Figure 4(b) shows the histogram for the VEP case. In both the cases, it can be seen that, the percentage of VOPs/VEPs detected within 10 ms is significantly high in the combined evidence compared to the ES alone. Figure 4(c) and (d) shows similar histograms using SSM. Around 8% of improvement is achieved within 10 ms region after adding the AE evidence.

V. Summary and Conclusion

This work describes a method for increasing accuracy of VOP/VEP detection using Bessel expansion and AM-FM model. Speech signal is bandpass filtered to get a narrow-band signal having low frequency components. This is done by choosing appropriate Bessel coefficients to emphasize the vowel regions. Narrow-band signal is modeled as an AM-FM signal and its amplitude envelope is detected using DESA algorithm. It is shown that the evidence obtained from the amplitude envelope gives peaks very close to the VOPs and VEPs. This evidence when added to some of the recent existing evidences for detection of VOP/VEP, gives an improved result. Overall detection rate is increased and spurious rate is reduced for both VOP and VEP. Improvement is achieved in terms of accuracy. The percentage of VOP/VEP detected within 10 ms is increased significantly, compared to the existing methods.

Acknowledgment

This work is part of the ongoing project on development of Prosodically guided phonetic Engine for Assamese language funded by the Technology Development for Indian Languages (TDIL), Department of Electronics & Information Technology (DeitY), Govt. of India.

REFERENCES

[1] D. Herms, "Vowel onset detection," *Journal of Acoustic Society of America*, vol. 87, pp. 886–873, 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," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 4, pp. 556–565, May 2009.

[3] C. Sekhar, "Neural network models for recognition of stop consonant-vowel (SCV) segments in continuous speech," Ph.D. dissertation, Department of Computer Science and Engineering, IIT Madras, 1996.

[4] G. Pradhan and S. R. M. Prasanna, "Speaker verification by vowel and nonvowel like segmentation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 4, pp. 854–867, April 2013.

[5] J. Yadav and K. S. Rao, "Detection of vowel offset point from speech signal," *Signal Processing Letters, IEEE*, vol. 20, pp. 299–302, 2013.

[6] C. Prakash, D. N. Gowda, and S. V. Gangashetty, "Analysis of acoustic events in speech signals using Bessel Series expansion," *Circuits, Systems, and Signal Processing*, vol. 32, pp. 2915–2938, 2013.

[7] C. Prakash, N. Dhananjaya, and S. Gangashetty, "Bessel features for detection of voice onset time using AM-FM signal," in *Int. Conf. on Systems, Signal and Image Processing (IWSSIP-2011)*, 2011.

[8] J. Schroeder, "Signal processing via Fourier-Bessel series expansion," *Digital Signal Processing*, vol. 3, pp. 112–124, 1993.

[9] R. Pachori and P. Sircar, "Analysis of multicomponent AM-FM signals using FB-DESA method," *Digital Signal Processing*, vol. 20, pp. 42–62, 2010.