# Bessel Features for Detection of Voice Onset Time using AM-FM Signal

Chetana Prakash
Speech and Vision Laboratory
International Institute
of Information Technology
Hyderabad, India
Email: chetana@research.iiit.ac.in

Dhananjaya N
Speech and Vision Laboratory
Indian Institute of Technology
Chennai, India
Email: dhanu@cse.iitm.ac.in

Suryakanth V. Gangashetty
Speech and Vision Laboratory
International Institute
of Information Technology
Hyderabad, India
Email: svg@iiit.ac.in

***Abstract*—Voice onset time is an important temporal feature which is often overlooked in speech perception, speech recognition as well as accent detection. The VOT in unvoiced stops varies with a number of factors, among which the most established one is the place of articulation. In this paper we propose an approach for the automatic detection of VOT. The proposed method uses Bessel expansion to emphasize the vowel and consonant regions of stop consonant vowel units (SCV) such as /ka/, /Ta/, /ta/ and /pa/. AM-FM signal has been emphasized after appropriate consideration of the range of Bessel coefficients, separately for the vowel and consonant regions of SCV units. The reconstructed signal from the Bessel expansion is a narrow-band AM-FM signal, therefore the amplitude envelope (AE) function for the emphasized signal can be estimated using discrete energy separation algorithm (DESA). For the detection of VOT, both the AE of vowel and consonat emphasized signal has been analyzed. Detection of VOT is analyzed for the continuous speech corpus consisting of recording television broadcast news bulletins.**

## I. Introduction

Voice onset time (VOT) is the duration between the release of closure of a stop consonant and the onset of voicing. Stop consonants are produced with a closure of the vocal tract at a specific place which is known as the place of articulation. During the closure, there is a build up sublaryngeal pressure. When the closure is released, there is a transient burst of air, then some frication due to turbulence at the place of articulation, and aspiration noise from turbulence caused by the glottis, being in an open or spread position [1].

VOT plays an important role in word segmentation, stress related phenomena, and dialectal and accented variations in speech patterns [2]. It can also be used for classification of accents. VOT combines the temporal and frequency structure over very short duration. This makes the VOT detection task difficult, but it is an important temporal feature. The sub-band frequency analysis is performed to detect VOT of unvoiced stops [3]. The amplitude modulation component (AMC) is used to detect vowel plus voiced onset region (VOR) in different frequency bands assuming the stop to vowel transitions has different amplitude envelopes for partitioned frequency ranges. In the above studies the standard band-pass filtering approaches are used for obtaining different frequency bands.

In this study we consider the important subset of basic units namely SCV. Stop consonant vowel (SCV) are the sound units produced by complete closure at some point along the vocal tract, build up pressure behind the closure, and releases the pressure by sudden opening. These units have two distinct regions in the production characteristics: the region just before the onset of the vowel (corresponds to consonant region) and steady vowel region. In this paper automatic detection of VOTs of SCV units in the continuous speech is presented. The technique presented is conceptually simple and can be implemented as a one step process, which makes the real time implementation feasible.

The paper is organized as follows: In Section II, we describe the Bessel expansion. The signal modeling based on AM-FM and its analysis using DESA is described in Section III. Section IV describes an approach for detection of VOT in stop consonant vowel (SCV) units. In Section V we describe the proposed approach for automatic detection of VOT in continuous speech.

## II. Bessel Expansion

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 [4]:

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

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

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

with $1 \leq p \leq P$, where *P* is the order of the Bessel expansion, and $J_1(\lambda_p)$ are the first-order Bessel functions. The order of the Bessel expansion *P* must be known a priori. It has been shown that there is one-to-one correspondence between the frequency content of the signal and the order $(p)$ at which the coefficient attains peak magnitude [5]. It is to be noted that the Bessel series coefficients $C_p$ are unique for the given signal.

TABLE I

*Bessel coefficient orders for emphasizing the vowel /a/ and different consonants /k/, /t/ and /p/ for sampling frequency of 8000 Hz and for the interval a = 20 ms*

| Region of speech signal | Required Bessel Coefficient order | Band limited frequency |
|---|---|---|
| /a/ | $P_1 = 12$ to $P_2 = 48$ | 300-1200 Hz |
| /k/ | $P_1 = 60$ to $P_2 = 100$ | 1500-2500 Hz |
| /t/ | $P_1 = 80$ to $P_2 = 120$ | 2000-3000 Hz |
| /p/ | $P_1 = 100$ to $P_2 = 140$ | 2500-3500 Hz |

The sinusoidal functions are periodic and ideal for representing general periodic functions. But it may not fully match the properties of other waveform. In case of nonstationary signals like speech, an aperiodic signal set may be more efficient for representation. The Bessel functions have regular zero-crossing and decaying amplitude that provides a better match to the behaviour of a speech waveforms [4]. The consonant and vowel regions of different stop consonant vowel (SCV) units can be obtained by taking the difference of Bessel expansion at two assigned order of the expansion as described below:

$$\begin{aligned} x_v(t) &= \sum_{p=1}^{P_2} C_p J_0\left(\frac{\lambda_p}{a}t\right) - \sum_{p=1}^{P_1-1} C_p J_0\left(\frac{\lambda_p}{a}t\right) \\ &= \sum_{p=P_1}^{P_2} C_p J_0\left(\frac{\lambda_p}{a}t\right) \end{aligned} \quad (3)$$

where $P_2 > P_1$. This is because the consonant and vowel region of SCV unit occupy different dominant frequency ranges, as such the onset regions of stops and vowel will be represented by the Bessel functions at different range of orders ($P_1$ to $P_2$). The selection of optimum window size $a$ is essential for effective analysis of consonant and vowel regions for the SCV unit. In (3) the signal $x_v(t)$ is a band-limited (narrow band) signal corresponding to the order range from $P_1$ to $P_2$ of the Bessel expansion. In order of low to high frequency components of the voice onset region (VOR), stops arranged as /k/, /t/ and /p/ [6]. The VOR of a given stop for adult speaker will fall in a certain frequency band as: **/k/ 1500-2500 Hz**, **/t/ 2000-3000 Hz**, **/p/ 2500-3500 Hz**. In addition, the vowel is assumed to have the most energy in the low frequency band (first formant, i.e., 300 to 1200 Hz). Table I shows the range of the Bessel coefficient orders required for emphasizing the vowel and consonant regions of the SCV units.

## III. AM-FM Model

A general monocomponent continuous time nonstationary signal has the form of a modulated signal defined as:

$$x(t) = A(t)\cos[\omega(t) + \phi(t)] = A(t)\cos[\Phi(t)] \quad (4)$$

Where $A(t)$ is the time-varying amplitude envelope (AE) of $x(t)$ with instantaneous frequency (IF) $\Omega(t)$ given by:

$$\Omega(t) = \frac{d\Phi(t)}{dt} = \omega + \frac{d\phi(t)}{dt} \quad (5)$$

Equation (4) has both amplitude modulation (AM) and frequency modulation (FM). These signals have been used in speech processing applications for modeling of speech resonances [7]. The discrete-time version of the monocomponent signal $x[n]$ is given by:

$$x[n] = A[n]\cos(\Phi[n]) \quad (6)$$

Both the instantaneous frequency and the amplitude envelope of the signal $x[n]$ can be derived from the Teager's nonlinear energy (NLE) operator. The NLE operator $\Psi(.)$ defined for the discrete signal $x[n]$ as [8]:

$$\Psi(x[n]) = x^2[n] - x[n-1]x[n+1] \quad (7)$$

It is applied to the AM-FM signal $x[n]$ and the difference signal $y[n] = x[n] - x[n-1]$. The amplitude envelope function $|A[n]|$ of the signal $x[n]$ is estimated by the discrete energy separation algorithm (DESA) as [7]:

$$|A[n]| \approx \sqrt{\frac{\Psi[x[n]]}{1 - \left[1 - \frac{\Psi[y[n]] + \Psi[y[n+1]]}{4\Psi[x[n]]}\right]^2}} \quad (8)$$

The amplitude envelope function of the AM-FM signal thus estimated exhibit ripples and therefore requires smoothing using a filter. The performance of the energy operator/DESA approach is vastly improved if the signal is first filtered through a bank of band pass filters, and at each instant analyzed (via $\Psi$ and DESA) using the dominant local channel response.

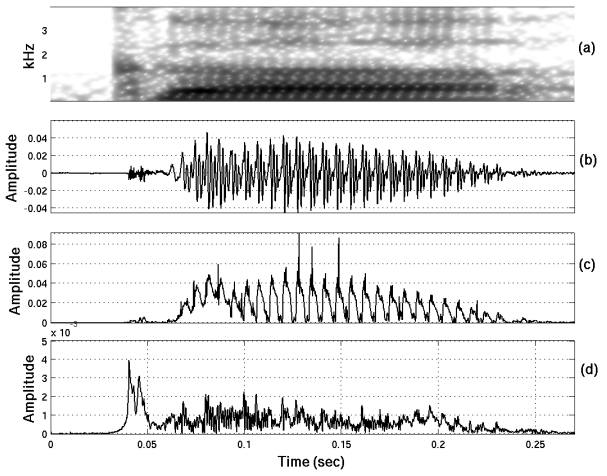


Fig. 1. *Plots of (a) Spectrogram, (b) Waveform, (c) AE estimation of the vowel (/a/) emphasized part, (d) AE estimation of the consonant (/k/) emphasized part for the speech utterence /ka/.*

## IV. Detection of VOT in SCV units

In order to detect VOT, the emphasized consonant and vowel regions of the SCV units are separated by using the Bessel expansion of appropriate range of orders by using (3). Since, the separated regions are narrow band signals, they can be modeled by using AM-FM model. The DESA technique is applied on the emphasized regions of the speech utterence in order to estimate the AE function for the detection of VOT. The beginning of the vowel region has been detected from

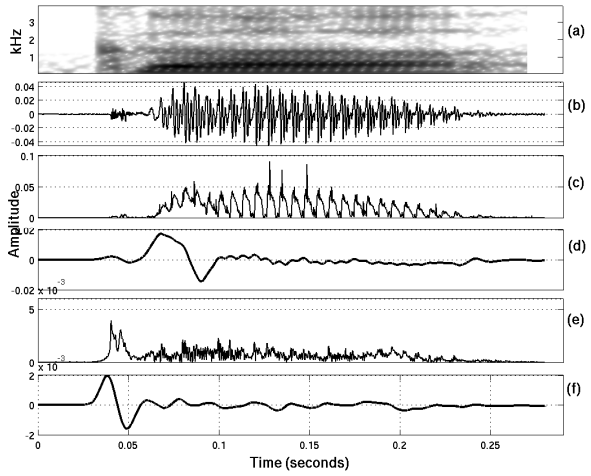


Fig. 2. *Plots of (a) Spectrogram, (b) Waveform, (c) AE estimation of the vowel (/a/) emphasized part, (d) Smoothed AE of the vowel, (e) AE estimation of the consonant (/k/) emphasized part for the speech utterence /ka/, (f) Smoothed AE of the consonant.*

the vowel emphasized part. From the beginning of the vowel region, by tracing back towards the beginning of consonant region in the consonant emphasized part, the beginning of the consonant regions has been detected. The VOT is obtained by taking the difference between beginning of the vowel region and beginning of the consonant region. The VOT is detected automatically by convolving the differenced Gaussian window of suitable length with the AE function of the vowel and consonant respectivelly. Beginning of the vowel has been detected by picking up the peak of the smoothed AE estimation of the vowel. From the beginning of the vowel, by tracing back towards the beginning of the consonant region in the smoothed AE of the consonant, the beginning of the consonant region has been detected by picking up the next peak. The VOT is the time difference between the beginning of the vowel and beginning of the consonant region.

For illustration we consider the SCV unit /ka/ whose waveform is shown in the Fig. 1(b). The spectrogram, amplitude envelope estimation for vowel and consonant emphasized regions of the speech utterence /ka/ are shown in the Figs. 1(a), 1(c) and 1(d) respectively. It is seen that the amplitude envelope corresponding to the vowel and consonant regions using proposed method are emphasized. This enable us to identify the beginning of the vowel $(t_v)$ and beginning of the consonant region $(t_c)$. Then VOT is given by $(t_{vot})$: $t_{vot}$ = $t_v$ - $t_c$. For the automatic detection of VOT, we considere the speech utterence /ka/ whose waveform is shown in Fig. 2(b). The spectrogram, amplitude envelope estimation for vowel and consonant emphasized regions of speech utterence /ka/ are shown in Figs. 2(a), 2 (c) and 2(e) respectively. Fig. 2(d) shows the smoothed AE function of the vowel, which is obtained by convolving differenced Gaussian window of length 20 ms with the amplitude envelope of the emphasized region of the vowel. Considering the peak of the smoothed AE function of the vowel as the beginning of the vowel region $(t_v)$. Fig. 2(f) shows smoothed AE function of the consonant, obtained by convolving differenced Gaussian window of length 20 ms with the amplitude envelope of the emphasized consonat region. From the beginning of the vowel, by tracing back towards the beginning of the consonat and picking up the next peak from the Fig 2(f) is considered as the beginning of the consonant $(t_c)$. Voice onset time $t_{vot}$ is given by $t_{vot}$ = $t_v$ - $t_c$. Table II shows the average duration of VOT for different categories of CV units ending with the vowel $/a/$. For unvoiced stops, the burst release takes place before the onset of the glottal activity. The VOT is generally larger for velar stops compared to the other three categories. The relatively smaller volume of the cavity behind the point of constriction in velar stops causes a geater pressure, which will take longer time to fall and allow an adequate transglottal pressure for the initiation of the vocal folds vibration. The extent of articulatory contact area in dental and velar stops is more, resulting in a slower release because of the Bernoulli effect pulling the articulators together. As the articulators come apart more slowly, there is a longer time before an appropriate transglottal pressure is produced. As a result, the durations of VOT for $/ka/$ and $/ta/$ are longer than those for $/Ta/$ and $/pa/$. The VOT durations for aspirated stop consonants are consistently longer than their unaspirated counterparts, as the aspiration region follows the closure release in case of aspirated stops.

TABLE II
*The average (across 50 speakers) duration of VOT for stop consonants (in ms)*

| **Unvoiced** | ka | $k^ha$ | Ta | $T^ha$ | ta | $t^ha$ | pa | $p^ha$ |
|---|---|---|---|---|---|---|---|---|
| **Duration (ms)** | 36 | 98 | 10 | 64 | 15 | 70 | 6 | 54 |

Detection of VOTs for SCV units in continuous speech requires the recognition of speech using suitable approach. Once the transcription of the speech utterence is obtained along with phone level boundaries, the regions of SCV units are identified. The VOT is then detected using the proposed approach described in this section. We develop HMM based system for recognition of continuous speech. We follow the approach for the development of HMM based speech recognition system as described in [9].

## V. PROPOSED APPROACH FOR DETECTION OF VOT IN CONTINUOUS SPEECH

The proposed approach for automatic detection of VOT for SCV units in continuous speech consists of the following steps:

**Step1:** The sequence of hypothesized phones along with phone segment boundaries are obtained for a given test utterance using HMM based speech recognition system. The hypothesized sequence of phones of test utterance is called test transcription.

**Step2:** Identify SCV units such as /ka/, /ta/ and /pa/ in the test transcription obtained in Step1.

The Step3 and 4 are repeated for all the SCV units of type /ka , /ta/ and /pa/ of test utterence.

**Step3:** Obtain the boundaries of identified SCV units in step2.

**Step4:** Type of SCV either (/ka/, /ta/ or /pa/) along with its segment boundaries is used to detect its VOT using the

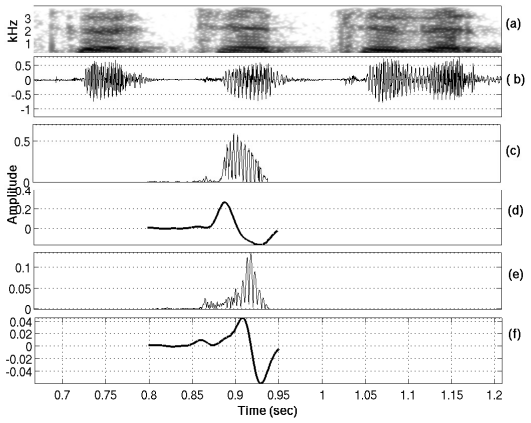


Fig. 3. *Plots of (a) Spectrogram, (b) Waveform of the speech signal, (c) AE estimation of the vowel (/a/) emphasized part, (d) Smoothed AE function of the vowel, (e) AE estimation of the consonant (/k/) emphasized part for the speech utterence /ka/ and (f) Smoothed AE function of the consonant for the Telugu language sentence /dvaipAkShi ka charcha la dvAra/*

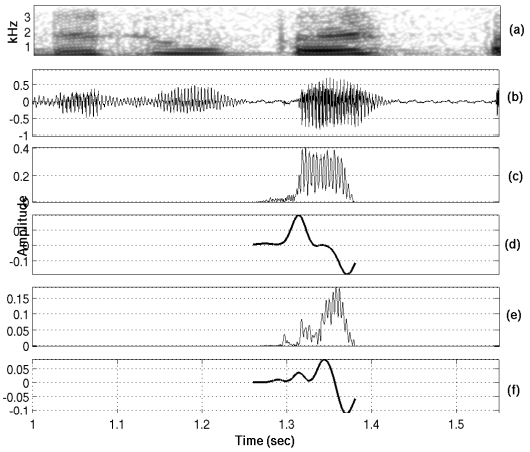


Fig. 4. *Plots of (a) Spectrogram, (b) Waveform of the speech signal, (c) AE estimation of the vowel (/a/) emphasized part, (d) Smoothed AE function of the vowel, (e) AE estimation of the consonant (/p/) emphasized part for the speech utterence /pa/ and (f) Smoothed AE function of the consonant for the Hindi language sentence /Telivijan nyuuj chenalo.n par paaba.ndii lagaadiihai/.*

approach described in Section IV For illustration, we consider a Telugu language continuous speech utterance /dvaipAkShi ka charcha la dvAra/, consisting of one SCV unit (/ka/) whose waveform is shown in Fig. 3(b). The hypothesized transcription of test utterence by HMM speech recognition system is /d v a i p A k Sh i k a ch a r ch a l a d v A r a/. The hypothesized boundaries of /ka/ is identified between 0.8 ms to 0.95 ms duration in the speech utterance. The segment of /ka/ is band limited for 300-1200 Hz using appropriate range of Bessel coefficients to emphasize vowel region /a/ of /ka/. Fig. 3(c) shows the amplitude envelope (AE) function of the vowel in segment /ka/. The amplitude modulation is performed using Teager energy operator (TEO) based approach. The AE function of the vowel is smoothed by convolving it with differenced Gaussian window of length 20 ms and is shown in Fig. 3(d). The location of peak $(t_v)$ in Fig. 3(d) corresponds to the begining vowel /a/ in SCV segment /ka/. Further, the segment of /ka/ is band limited for 1500-2500 Hz using appropriate range of Bessel coefficients to emphasize consonant region /k/ of SCV segment /ka/. Fig. 3(e) shows the AE function of the consonant in segment /ka/. The smoothed AE function of the consonant is shown in Fig. 3(f). The peak to the left of $(t_v)$ is hypothesized as the beginning of consonant $(t_c)$. The difference of two detected peak locations $(t_v) - (t_c)$ is hypothesized as the VOT of /ka/ in cotinuous speech test utterence. The value of $t_{vot}$ is found to be 26.3 ms. Similarly the VOT for /pa/ is detected for Hindi language test utterence as shown in the Fig. 4 respectively.

## VI. Summary and Conclusions

In this paper we have proposed an approach for detection of VOT based on Bessel expansion and amplitude modulation component of the Teager energy operator. The appropriate range of Bessel coefficients are used to construct the narrow band, band limited signal. The basis functions of Bessel transformation has damped sinusoids, hence Bessel transform is well suited for analysis of non stationary signals like speech. The DESA technique is used to estimate the amplitude modulation component of vowel and consonant regions of SCV units of speech. Since VOT detection approach requires accurate boundaries of SCV units in continuous speech. The HMM based speech recognition system has to be more accurate. When consonant region is articulated poorly then it is difficult to locate the peak corresponding to the beginning of consonant. Hence it is necessary to device an approach for the rejection of false VOT. The VOT information can be used for a wide range of accent classification. The combined approach based on Bessel expansion and AM-FM model is used for online implementation of VOT detection approach.

## References

[1] K. N. Stevens, *Acoustic Phonetics*, The MIT press, Cambridge, Massachusetts, USA, 1999.

[2] L. Lisker and A. S. Abramson, ''Some effects of context on voice onset time in English stops", *Language and Speech*, vol. 10, pp. 1-28, 1967.

[3] S. Das and J. H. L. Hansen, ''Detection of voice onset time (VOT) for unvoiced stops (/k/, /t/, /p/) using the Teager energy operator (TEO) for automatic detection of accented English", *Proc. 6th Nordic Signal Processing Symposium,* pp, 344-347, 2004.

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

[5] R. B. Pachori and P. Sircar, ''EEG signal analysis using FB expansion and second-order linear TVAR process", *Signal Processing*, vol. 88, no. 2, pp. 415-420, 2008.

[6] P. Ladefoged, *A* course in Phonetics, 3rd edition *Harcourt Brace College Publishers*, Fort Worth, 1993.

[7] P. Maragos, J. F. Kaiser, and T. F. Quatieri, ''Energy separation in signal modulation with applications to speech analysis", *IEEE Transactions on Signal Processing*, vol. 41, no. 10, pp. 3024-3051, Oct. 1993.

[8] J. F. Kaiser, ''On a simple algorithm to calculate the 'energy' of a signal", *IEEE ICASSP*, pp. 381-384, 1990.

[9] B. Yegnanarayana and Suryakanth V. Gangashetty, ''Machine learning for speech recognition- An illustration of Phonetic engine using hidden Markov models", *In Proc. Int. Conf. Frontiers of Interface Between Statistics and Science*, pp. 319-328, Jan. 2010.