Analysis of Stop Consonants in Indian Languages Using Excitation Source Information in Speech Signal

B. Yegnanarayana[1], *K. Sri Rama Murty*[2] *and S. Rajendran*[1]

[1]International Institute of Information Technology, Hyderabad-500032, India
[2]Dept. of Computer Science & Engineering, IIT Madras, Chennai-600036, India
yegna@iiit.ac.in, ksrmurty@gmail.com, su.rajendran@gmail.com

Abstract

In this paper we propose excitation-based features for extracting information about the manner of articulation for stop consonants. The excitation-based features are derived from very low frequency information in the signal and also from the normalized error computed from the linear prediction residual. The proposed zero-frequency filtered signal brings out the region of glottal activity during excitation. Likewise, the normalized error helps to distinguish regions of noise and pure voicing. These nonspectral methods of analysis of stop consonants seem to provide additional and some better features over the features derived from the traditional methods based on short-time spectrum analysis.

1. Introduction

Stop consonants of consonant-vowel (CV) type form an important subset of alphabet in most Indian languages. Table-1 lists the stop consonants with the vowel ending /a/ for Indian languages. The acoustic-phonetic description of the consonants in each of these characters is precise, and is expressed in terms of voiced (V), unvoiced (uV), aspirated (A) and unaspirated (uA) categories. They are also organized according to the place of articulation (velar, post-alveolar, dental and labial). One of the challenging tasks in speech analysis is to determine the acoustic correlates of the production of the CVs, which seems to be difficult to extract even from clearly articulated speech signals. An important acoustic correlate is the voicing onset, i.e., the onset of laryngeal vibration, and its location in relation to the burst release. The other acoustic correlates of interest are: (a) discrimination of voiced and unvoiced stop consonant, and (b) the discrimination of aspirated and unaspirated stop consonant. The presence of breathiness in the voiced aspirated stop makes the task of analysis of these utterances all the more interesting and challenging. "Breathiness is due to incomplete and non-simultaneous glottal closure during the closed phase of the phonation cycle" [1]. Breathy phonation is characterized by vocal cords that are fairly abducted (relative to modal and creaky voice), and have little longitudinal tension. This results in some turbulent airflow through the glottis, and the auditory impression of "voiced mixed in with breath" [2].

Accurate discrimination of voicing onset from the speech signal is not only important clinically, but also useful to determine the voice onset time (VOT). The VOT is defined as the difference between the time of burst release and the time of onset of voicing. The VOT forms an important clue to discriminate stop consonants between voiced and nonvoiced, and aspirated and unaspirated.

Since voicing onset is a characteristic of glottal activity, the onset time can be marked accurately using electroglottography (EGG). Sometimes even with EGG waveform it may be difficult to mark the voicing onset due to 'subcritical' vocal fold vibration in breathy voicing conditions [3]. Moreover, it is not possible always to obtain the EGG signal. Therefore it is necessary to derive the voicing onset information from the acoustic speech signal itself.

Table 1: Stop consonants (CVs) in Indian languages

	uVuA	uVA	VuA	VA
Velar	/ka/	/k^{h}a/	/ga/	/g^{h}a/
Post-alveolar	/ʈa/	/ʈ^{h}a/	/ɖa/	/ɖ^{h}a/
Dental	/ta/	/t^{h}a/	/da/	/d^{h}a/
Labial	/pa/	/p^{h}a/	/ba/	/b^{h}a/

Most commonly used methods for measuring the onset of voicing are based on the onset of periodicity in the acoustic waveform, possibly supplemented by spectrographic analysis, especially the onset of visible energy in the first and/or higher formants. It is also possible to measure the onset of voicing as the onset of energy visible in the voice bar, the region of the lowest frequency energy in a wide-band spectrogram corresponding to the fundamental frequency f_0, typically found below the first formant (F_1) [3]. The ideal acoustic measurement of voicing onset is one that is both accurate and relatively consistent. Comparative study of accuracy and variability of five acoustic (f_0, F_1, F_2, F_3 & periodicity in waveform) measures of voicing onset showed that measurements based on waveform or the voice bar seem to provide the best results. In all these cases the voicing onset determined by observing the EGG is taken as reference for accuracy. But, as mentioned earlier, there are situations in voiced aspirated stops, where even EGG may not provide a good reference, apart from the fact that EGG may not be available most of the time.

The main problem with these acoustic measures is that the desired information of the glottal activity is in a very low frequency region 0-100 Hz (within the f_0), where the amplitude of the acoustic signal is very low compared to the amplitude at other frequencies. In spectrographic analysis, the effects of block processing sometimes limit the visibility of formant features. The presence of noise and voicing in the aspirated (breathy) regions, and the low amplitude of the voice bar in voiced stops make the direct measurements from the waveform difficult. Thus analysis of stop sounds, especially the voiced aspirated stops, to extract information about the glottal activity remains a challenge.

Since the glottal activity is primarily the activity contributing to the excitation source of the vocal tract system, it is likely

that if the analysis is focused on the excitation component in the speech signal, it may provide new insights into the phonation characteristics present in the signal. It is also desirable to avoid spectral analysis, as it may invariably use block processing, resulting in blurring of details of voicing onset information. Thus in this paper we propose nonspectral processing of the speech signal to extract the low frequency information, as well as the excitation component information in the signal. Section 2 presents the basis for the proposed method of processing the speech signal. Section 3 examines the performance of the proposed method to distinguish voiced, unvoiced, aspirated and unaspirated features in stop sounds. In Section 4 the proposed analysis is demonstrated for all the CV units given in Table 1. The results of analysis are given in terms of the VOT, where the VOT is defined as the interval between the instant of release of burst and the onset of glottal activity (not necessarily the periodic vibration of vocal folds). The analysis is made for CV units with different vowel endings also. The detection of these phonation features in stop consonants in continuous speech for male and female voices is examined in Section 5. The results are compared with EGG signals corresponding to the continuous speech to demonstrate the effectiveness of the proposed method of nonspectral analysis. Finally, in Section 6 we discuss the potential of the proposed method for various applications such as deriving the instants of glottal closure in voiced speech, pitch period information and voice activity detection (VAD). Here VAD refers to the region of voiced speech in the signal, although normally VAD is used to distinguish speech and nonspeech regions. Our future effort will be focused on studying the effectiveness of these nonspectral methods for analysis of consonant clusters that occur in Indian languages.

2. Excitation-based nonspectral analysis of speech

The objective of this paper is to explore some nonspectral methods for analysis of stop sounds. The nonspectral analysis should highlight the information of the excitation source of the vocal tract system. The primary and most important mode of excitation is due to the activity at the glottis. In normal voiced excitation (called modal voicing) there will be vibrations of the vocal folds resulting in glottal opening, followed normally by an abrupt closure of the vocal folds, and then a closing phase of the glottis, before the glottis is opened again for the next cycle due to build up of pressure from the lungs. Other aspects of glottal activity includes vibration with large opening producing breathy voice, a complete opening for the production of voiceless sound, a partial closure of the vocal folds producing creaky voice, and finally a complete closure of the vocal folds such as for glottal stops. Fig.1 illustrates the continuum of phonation types as proposed by Gardon and Ladefoged [2]. The other excitation modes are due to frication at the glottis, due to turbulent noise at a narrow constriction and plosion due to release of pressure behind a closure. We focus on extraction of the excitation due to glottal activity, and try to derive the acoustic correlates of stop consonants from this excitation information.

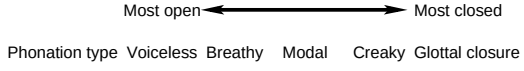


Figure 1: Phonation types [2]

The normal voiced excitation can be viewed as a sequence of approximate impulses due to rapid closure of the glottis in each cycle of vibration. The impulse-like excitation information can be extracted from the speech signal by reducing the effects of the resonances of the vocal tract system, and also the high frequency effects due to frication. The main assumption in the proposed method is that significant excitation of the vocal tract system takes place due to abruptness of the glottal closure rather than any other type of excitation. This will enable us to determine the regions where there is significant impulse-like excitation. From the impulse sequence one can determine whether the region has nearly periodic excitation as in voiced sounds, or somewhat aperiodic excitation as in parts of aspirated regions.

The information about the glottal activity appears mostly in the very low frequency region, i.e., less than the fundamental frequency. To extract this information, the speech signal is passed through an ideal resonator (in discrete-time case) at zero frequency. While this is equivalent to integration, we prefer to use the term resonator, as such an analysis may be useful for exploring features at other frequencies as well. The low frequency trend in the resulting output is removed. In the resulting signal all the high frequency information, including the resonances of the vocal tract and the frication effects, are reduced significantly. Only the contribution due to sequence of excitation impulses is reflected in the output. The excitation impulse information gets enhanced irrespective of the variation in the shape of the vocal tract system during production of different sounds. The excitation characteristics can be used to determine the regions of glottal activity and nonglottal activity in the speech signal.

The breathiness in voiced aspirated regions is due to noise and vibration at the glottis. In order to capture the effect of noise in this type of excitation, the excitation component of the speech signal is derived using linear prediction (LP) analysis. The linear prediction residual for a p^{th} order LP analysis is used as an approximation to the excitation component in the speech signal. The choice of the LP order p, the frame size and frame rate used for the LP analysis are not critical. In this work, a 10^{th} order LP analysis is performed on frames of 20 ms at a rate of 100 frames per second. The ratio of the energy of the LP residual and the speech signal for every block of the frame size and for every sample shift is computed. The resulting plot is called the normalized error as function of the sample index, and it is used to distinguish the excitation information due to noisy voiced segments and clean voiced segments. Note that the spectral information in LPCs is ignored here, by considering only the residual. However, block processing is used to derive the LP residual, but the effects of blocking are insignificant for the analysis of the acoustic correlates of the stop consonants under consideration.

The following steps are involved in processing the speech signal to derive the low frequency information and the breathy noise part of the glottal excitation.

(a) Difference the signal $s[n]$, i.e.,

$$x[n] = s[n] - s[n-1]. \tag{1}$$

(b) Pass the signal twice through an ideal resonator at zero frequency. That is

$$y_1[n] = -\sum_{k=1}^{2} a_k y_1[n-k] + x[n], \tag{2a}$$

and

$$y_2[n] = -\sum_{k=1}^{2} a_k y_2[n-k] + y_1[n], \qquad (2b)$$

where $a_1 = -2$ and $a_2 = 1$. This is equivalent to successive integration four times.

(c) Remove the trend in $y_2[n]$ by subtracting the average of $y_2[n]$ over 10 ms at each sample. The resulting signal $y[n]$ is called the zero-frequency filtered signal, or simply the *filtered signal*.

(d) The LP residual is computed as follows:

$$e[n] = x[n] - \sum_{k=1}^{p} a_k x[n-k], \qquad (3)$$

where a_ks are the LPCs, obtained by solving the autocorrelation normal equations,

$$\sum_{k=1}^{p} a_k R[m-k] = -R[m], \qquad m = 1, 2, \ldots, p, \quad (4)$$

where $R[m]$ is the autocorrelation sequence computed from the differenced signal $x[n]$.

(e) The normalized error for each sample shift is computed using

$$\eta[n] = \frac{\sum_{m=n-N/2}^{n+N/2} e^2[n+m]}{\sum_{m=n-N/2}^{n+N/2} x^2[n+m]}, \qquad (5)$$

where $N+1$ is the total number of samples in each frame.

The filtered signal $y[n]$ and the normalized error $\eta[n]$ are used to represent the excitation information derived from the speech signal. From the plots (see lower part of Fig. 2) for the utterance of the voiced aspirated syllable /g^{h}a/, one can notice that it is easy to determine the onset of glottal activity and the ending of the glottal activity (see Fig. 2k). In the initial voicing region the filtered output is relatively high as compared with the amplitude of the signal in that region. There is low value of normalized error in this region. At the release of the burst of the stop sound there is a significant increase in $\eta[n]$. The burst release cannot be seen either in the waveform or in the filtered output. During aspiration, the filtered output is large indicating significant glottal activity. It is difficult to distinguish the glottal activity during aspiration and the following modal voicing, as the filtered output appears nearly periodic in both regions. But in the normalized error, there is a significant raise in the value in the aspiration region compared to the values in the modal voicing region. The voicing onset is also easy to locate in the filtered output.

3. Analysis of manner of articulation for stop consonants

Isolated utterances of the CV units listed in Table 1 are used in this study. The utterances are produced by a male speaker. Each utterance is repeated 5 times. The speech signal is sampled at 8 kHz. The data is collected for all the five vowel endings for each of the consonants in CV units. All the data was collected in a laboratory environment using a close-speaking microphone. Thus the data can be considered as clearly articulated clean data.

Fig. 2 shows the waveform, filtered output and the normalized error plots for the following four velar stops: /ka/, /k^{h}a/, /ga/ and /g^{h}a/. The plots of the filtered outputs in each case clearly show the regions of glottal activity.

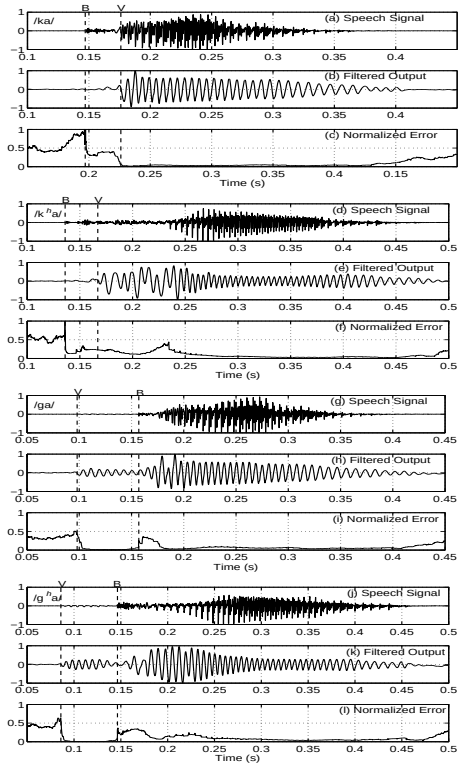


Figure 2: The speech signal, filtered output, and the normalized error for four different velar stop sound units

The following observations can be made to distinguish the four categories.

(a) Unvoiced unaspirated: There is sudden increase in the normalized error at the release of the burst. The normalized error is large in the short burst region relative to the modal voicing region.

(b) Unvoiced aspirated: There is sudden increase in the normalized error at the release of the burst. The large $\eta[n]$ is extended over the aspirated region due to the presence of breathy noise. The $\eta[n]$ is low in the modal voiding region. The filtered output is *somewhat less periodic* in the aspirated region.

(c) Voiced unaspirated: There is relatively large output in the filtered signal due to initial voicing compared to the relatively small amplitude in the waveform. There is an increase in the $\eta[n]$ during the short burst region.

(d) Voiced aspirated: There is large output in the filtered output during the initial voicing region, and then in the aspirated and model voicing regions. There is a dip in the filtered output at the burst release. But the $\eta[n]$ has an abrupt raise at the burst release, followed by large $\eta[n]$ in the aspirated region due to breathy noise. The filtered output is nearly periodic in the aspirated region as in the modal voicing region.

The burst release instant (marked as B) is determined as the instant where there is a large increase in $\eta[n]$. The starting instant of the glottal activity (marked as V) is derived from the

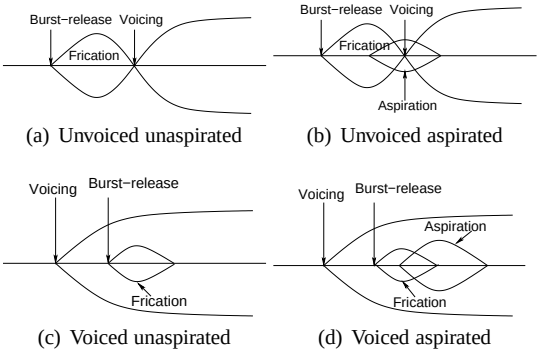
filtered output. In all the cases the burst release instant (B) and voicing onset (V) can easily be identified. The interval between these two instants is used as VOT in this study. We examine VOTs for different types of CV units in the next section.

4. Analysis of CV units - Estimation of VOT and burst durations

In this section the VOTs and burst durations are obtained for different categories of CV units for different vowel endings. All the VOTs are obtained manually from the filtered output and the normalized error as shown by the markers in Fig. 2. For unvoiced stops, the burst release (B) takes place before the onset (V) of the glottal activity. The interval between these two is the burst duration, and is also the VOT in this case. The VOT is generally larger for velar stops compared to the other three categories. The duration of aspiration is difficult to measure from the plot of the normalized error.

The VOT for voiced (unaspirated and aspirated) is measured from the instant of onset of glottal activity (V) to the instant of burst release (B). Since in the voiced case, the voicing due to the glottal activity starts first and continues through the vowel, and the burst release due to the articulatory gesture comes during the voicing, the beginning and end of the burst can be seen clearly in the region of glottal activity, especially for voiced unaspirated stops. In this case the burst duration is different from VOT. For voiced aspirated stops, the burst duration is sometimes difficult to identify as the breathiness due to the glottal activity and the burst duration may overlap. The normalized error may remain large throughout the aspiration region, even though there is burst release during that period.

All the above observations are valid for stop consonants with different vowel endings. Fig. 3 illustrates the relative placement of the important events in various stop consonants. Table 2 shows the average values of burst durations for different categories of CV units ending with the vowel /a/.



(a) Unvoiced unaspirated (b) Unvoiced aspirated

(c) Voiced unaspirated (d) Voiced aspirated

Figure 3: Schematic representation of the important events in the stop consonants

Table 2: The duration of burst in stop consonants (in ms)

Unvoiced	ka	k^ha	ʈa	$ʈ^ha$	ta	t^ha	pa	p^ha
Duration	32	36	16	20	23	18	12	12
Voiced	ga	g^ha	ɖa	$ɖ^ha$	da	d^ha	ba	b^ha
Duration	19	23	9	16	12	17	7	13

5. Phonation features in continuous speech

Finally, we have compared the filtered output for continuous speech utterances with the EGG waveform, to determine the accuracy of the extracted information of the acoustic correlates, especially the region of glottal activity in the utterance. We notice from Fig. 4 that the manual marking of glottal activity made from the filtered output match well with the information in the EGG waveform. In some cases, the filtered output shows the glottal activity better than even the EGG waveform as in the region around 1.7 s in Fig. 4.

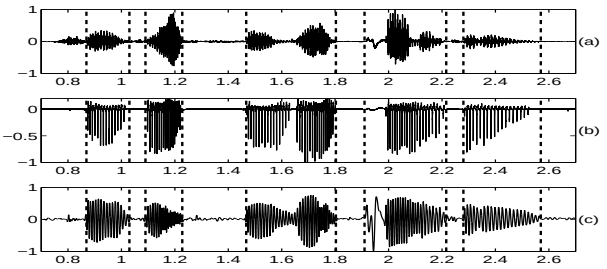


Figure 4: Continuous speech of a male voice for the utterance "She has left for a great party today" (http://archives.limsi.fr/VOQUAL/voicematerial.html).
(a) speech signal, (b) differenced EGG signal, and (c) filtered signal

6. Summary and conclusions

In this paper we attempted to make a case for nonspectral methods for analysis of stop consonants. The methods are intended to focus on excitation characteristics during the production of stop consonants. We have proposed a zero-frequency filtered signal to extract the region of glottal activity, and the normalized error from LP residual to determine the noise regions of excitation during burst release and during aspiration. The instants of burst release and glottal activity can easily be detected from these excitation features.

Additional excitation information such as strength of impulses during glottal action and the effects of coupling of various cavities might bring out better discrimination among these and other CV units. The results of the studies in this paper show that accurate acoustic-phonetic analysis of stop consonants may be possible by examining the features in the excitation source also, in addition to the commonly used spectral features such as harmonics and formants. Our preliminary studies show that the information in the filtered output together with other information can be used for the following applications: (a) estimation of instants of glottal closure, (b) estimation of voice periodicity and (c) voice activity detection.

7. References

[1] R. Wayland and A. Jongman, "Acoustic correlates of breathy and clear vowels: the case of Khmer," *Journal of Phonetics*, vol. 31, no. 2, pp. 181–201, 2003.

[2] M. Gardon and P. Ladefoged, "Phonation types: a cross-linguistic overview," *Journal of Phonetics*, vol. 29, no. 4, pp. 383–406, 2001.

[3] A. L. Francis, V. Ciocca, and J. M. C. Yu, "Accuracy and variability of acoustic measures of voicing onset," *Journal of the Acoustical Society of America*, vol. 113, no. 2, pp. 1025–1032, 2003.