Detection of Burst Onset Landmarks in Speech Using Rate of Change of Spectral Moments

A. R. Jayan, P. S. Rajath Bhat, P. C. Pandey
Department of Electrical Engineering
Indian Institute of Technology Bombay
Powai Mumbai 400 076 India
Email: {arjayan,rajathbhat,pcpandey}@ee.iitb.ac.in

***Abstract*— Burst onset landmarks in the speech signal are transient segments with low energy and their accurate detection is important in applications involving landmark based speech modification, estimation of place of closure for speech training aids, and phoneme recognition. Rate of change measures of energy parameters from spectral bands with fixed boundaries are generally used for landmark detection. The differences in the parameter variation rates and ranges, correlations among them, and their dependencies on speakers, make them less suitable for precise time localization of burst onsets. A method for detection of burst onset landmarks is presented which uses rate of change of spectral moments, in addition to energy parameters of the short-time speech spectrum. Evaluation results indicate that this method can give high detection rates with improved temporal accuracy.**

Keywords-Speech landmarks; burst onset detection; spectral moments

I. INTRODUCTION

The perceptual cues essential for the identification of speech are non-uniformly distributed in the speech signal. Compared to the steady-state segments, transition segments are concentrated areas of acoustic cues. These acoustically salient regions are known as "landmarks" [1]. Speech intelligibility enhancement is possible by intensity, spectral, or duration modification of these landmarks so as to make them robust in the presence of noise [2]. The motivation for these intelligibility enhancement techniques is derived from "clear speech", a natural way adopted by a speaker to improve speech intelligibility when talking to a hearing impaired listener or in the presence of background noise. Clear speech is reported to be about 17% more intelligible than conversational speech [3]. This intelligibility advantage is mainly derived from the robustness of acoustic landmarks. Several intelligibility enhancement techniques based on acoustic properties of clear speech have been reported and many of these use detection and modification of speech landmarks for automated processing [4], [5], [6].

Burst onset landmark is associated with the abrupt release of frication energy after a closure interval, and it may be followed by aspiration and the transition of formants to the onset of the succeeding vowel. The durations of closure, burst, and voicing onset are dependent on the speaker, speaking style, type of the stop consonant, and the context in which it appears. The typical values of these parameters are in the range of 50-100 ms for closure, 5-10 ms for burst, 0-30 ms (voiced stops) and 30-150 ms (unvoiced stops) for voicing onset, respectively. An illustration of the landmarks associated with a typical stop consonant in a vowel-consonant-vowel (VCV) context is given in Fig. 1. Among the consonantal landmarks, burst onset landmark is most transient in nature. Accurate detection of burst onsets is important for improving performance of landmark based speech recognition schemes [7], [8]. For use in speech training aids, Pandey and Shah [9] proposed a method for estimating the vocal tract shape during the stop closures of vowel-consonant-vowel utterances by using bivariate surface modeling of the shapes obtained by LPC analysis during the transition segments. The detection of burst onset landmark can be helpful in locating the burst offset, aspiration, and consonant vowel (CV) transition boundaries. The properties of burst spectrum can be used to provide supplementary information regarding the place of articulation of the stop consonant. Detection of burst onset landmarks with good temporal accuracy is quite important in intelligibility enhancement applications [10], [11].

A rate-of-rise (ROR) measure based on the first difference of parameters extracted from different bands in the smoothened wideband spectrum (e.g. band energy, spectral peak, and spectral tilt) is generally used to locate landmarks [1], [12], [13]. Such methods have been reported to provide good detection rates (≈90%) at moderate temporal accuracies (20-30 ms). Overall detection rates for burst onset landmarks reported by Liu [1] are 41%, 68%, 85%, and 88%, for temporal accuracies of 5, 10, 20, and 30 ms respectively for sentences from TIMIT database. Salomon *et al*. [13] used temporal parameters like envelope, periodicity, and fine structure in addition to band energy parameters for landmark

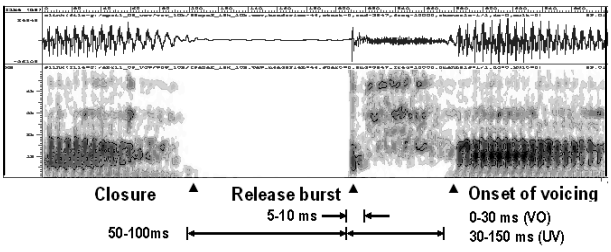

Figure 1. Stop consonant landmarks for VCV utterance /apa/: waveform and wideband spectrogram.

detection. For a temporal accuracy of 50 ms, 91% of stop closures and 96% of stop release bursts were detected, when evaluated using sentences from TIMIT database. Adaptive time steps (5 ms for stop bursts, 30 ms for frication, and 2 pitch periods in periodic regions) were used to improve temporal accuracy of detection.

Time step used in the computation of rate-of-rise (ROR) using first difference operation should be comparable with the transition duration of the landmark of interest. Landmarks with slow parameter variations can be reliably detected with a larger time step (20-30 ms), but for time-localizing abrupt landmarks, shorter time steps are needed. The use of a fixed time step for all the parameter tracks irrespective of the type of landmark has the drawback of spreading the actual center of transition when the transition duration is less than the time step used. When the time step is less than the transition duration, the peak amplitude in the ROR, representing the transition gets reduced [14]. Both these cases adversely affect the temporal accuracy and rates of landmark detection. Landmark detection schemes for speech recognition application typically use a two-pass strategy with a first pass with comparatively larger time step (26-50 ms) to locate the neighborhoods of landmarks of interest reliably, followed by a second pass with a lower time step (10 ms), to time-localize the landmarks [1], [14].

The energy variation associated with the burst onset may get distributed non-uniformly over the frequency bands, and the rate of variation may also be non-uniform depending on the band boundaries and type of the landmark. This makes the selection of a single fixed value for time step difficult when energy parameters from spectral bands with different bandwidths are used for landmark detection. A typical burst lasts for a time period of 5-10 ms after its onset, and temporal accuracies less than this order may adversely affect the processing based on the detected landmarks.

Performance of landmark detection methods based on spectral energy change degrades at regions of lower SNR levels, and in the presence of background noise. Sainath and Hazan [15] reported a sinusoidal model based scheme in which sharp changes in the signal harmonicity was used in addition to spectral energy variation for landmark detection. This method was able to locate landmarks more precisely, compared to the spectral energy based method.

Parameters characterizing the overall spectral shape are reported to be useful for improving the temporal accuracy of landmark detection. An approach based on parameters from the Gaussian mixture modeling (GMM) of short-time log magnitude spectrum [10], [11] was investigated to improve the accuracy of detection of stop landmarks. In this method, a rate of change (ROC) function defined on the parameters obtained by modeling of the log magnitude spectrum using 4 Gaussian components along with a voicing onset-offset detector and a spectral flatness measure was used for detection of stop landmarks. It was possible to detect over 90% of stop release bursts in VCV syllables and 73% of stop bursts in TIMIT sentences within 5 ms of manual landmarks. The iterative process of estimation of the Gaussian parameters is computation intensive and the method is not suited for a real-time implementation. A method based on spectral moments in addition to energy band parameters is investigated for locating burst onset landmarks, with the objective of reducing the computational complexity and for improving the temporal accuracy. Spectral moments are indicatives of spectral shape, and have been used as parameters supporting classification of Mandarin stops [7].

Combining different parameters to get a single parameter indicating the overall variation needs some form of normalization of the parameters, taking into account their ranges and correlations. Due to the speaker dependent variabilities in speech, defining a fixed weight for a parameter without over-sensitizing or desensitizing its variation is often difficult. We have investigated the effectiveness of a rate of change measure based on Mahalanobis distance for time localizing the burst onsets. Evaluation has been performed using VCV utterances and sentences from TIMIT database to quantify the effect of parameters and distance measures on the temporal accuracy of burst onset detection.

II. METHODOLOGY

We have investigated the use of peak energy from fixed frequency bands and the first four moments of the short-time speech spectrum as parameters for burst onset detection. A combined rate of change measure based on Mahalanobis distance, referred to as ROC-MD, is investigated and its performance is compared with that of sum of the individual ROCs. The use of a rate of change measure (ROC) for different values of time steps is investigated to examine the effect of time steps on temporal accuracy of burst onset detection.

A. Computation of energy band parameters

For speech sampled at 10 kHz, 512-point DFT is computed for 6 ms Hanning windowed frames, every 1 ms. The magnitude spectrum for each frame is smoothed by a 20-point moving average taken along the time index n. From the smoothed spectrum $X(n,k)$, peaks in three different frequency bands (1.2-2.0, 2.0-3.5, and 3.5-5.0 kHz) are selected as

$$E_b(n) = 10\log_{10}(\max|X(n,k)|^2) \quad (1)$$

where the band index b ranges from 1 to 3, and the frequency index k ranges from the DFT indices representing the lower and upper boundaries for each band.

B. Computation of spectral moments

Treating the normalized short-time speech spectrum as a probability density function, its first four moments are evaluated as a function of the position of the analysis window. The smoothed spectrum $X(n,k)$ is normalized as

$$p(n,k) = |X(n,k)| \Big/ \sum_{k=1}^{N/2} |X(n,k)| \quad (2)$$

where N is the number of points in the DFT computation.

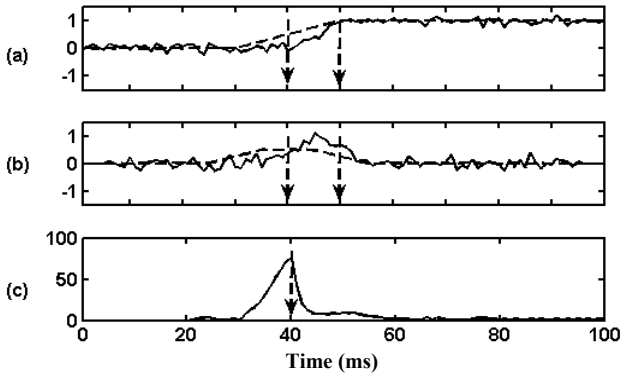

Figure 2. Simulated parameter tracks as a function of time (ms): (a) P1 (solid), P2 (dotted), (b) ROR tracks, and (c) Mahalanobis distance based ROC.

The centroid of the spectrum is computed as

$$F_c(n) = \sum_{k=1}^{N/2} p(n,k)\, f_k \tag{3}$$

where f_k is the frequency in Hz corresponding to the DFT bin with index k. The spectral moments (the centroid, standard deviation, skewness, and kurtosis) indicate the frequency of concentration of the spectral energy, spread of energy around this location, the symmetry of the spectrum, and its peakiness.

The second, third, and fourth moments, which are related to the variance $F_\sigma(n)$, skewness $F_s(n)$, and kurtosis $F_k(n)$, respectively, are computed as

$$F_m(n) = \left[\sum_{k=1}^{N/2} (f_k - F_c(n))^m\, p(n,k) \right]^{1/m} \tag{4}$$

where $m = 2$ for $F_\sigma(n)$, $m = 3$ for $F_s(n)$, and $m = 4$ for $F_k(n)$.

C. Computation of rate of change

Landmark detection involves measurement of rate of variation of a set of parameters extracted from the speech signal on a short-time basis, and locating regions with a significant variation characterizing the landmark. A rate of change measure based on first difference operation with a fixed time step is generally used to get the rate of variation of parameters. For band energy prameter $E_b(n)$, ROC measure is defined as

$$r_{Eb}(n) = E_b(n) - E_b(n-K) \tag{5}$$

where K is the time step. This measure indicates the difference in parameter value of the current frame, from a frame preceding it by K frames. An abrupt transition is indicated by a well defined peak in the ROC track, while the track has a very low value during steady-state segments.

The rate of change of different parameters are in different magnitude scales and they may be correlated to a certain extent. Mahalanobis distance [16], [17] can be used to take care of the correlations and scale differences of the parameters and to get a single rate of change indicating the overall variation of parameters. Mahalanobis distance of a point **y** defined by a set of parameters, from a cluster of previous parameters **x** is defined as

$$d = ((\mathbf{y}-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})^T)^{0.5} \tag{6}$$

where $\boldsymbol{\mu}$ is the mean and $\boldsymbol{\Sigma}$ is the covariance matrix of the cluster **x**.

Fig. 2(a) shows two simulated parameter tracks P1 (solid) and P2 (dotted) which may correspond to energy variations in two spectral bands at the onset of a burst landmark. The rate-of-rise tracks of P1 and P2, computed with a first difference with a 10 ms time step are shown in Fig. 2(b). For P1, the transition duration is equal to the time step. The ROC track has a peak at 45 ms, the center of transition of P1. In case of P2, the transition duration being more than the time step, the ROC track becomes flat-headed, loosing the time information. The amplitude of the ROC track for P2 has lower amplitude than that of P1, even though the two parameter variations are taking place with the same amplitude level. A combined ROC based on Mahalanobis distance (ROC-MD) of the two parameters with a time step of 10 ms is shown in Fig. 2(c). It has a well defined peak at 40 ms where the simultaneous variations in P1 and P2 take place, showing that the ROC based on Mahalanobis distance combines the individual variations in parameters and is able to give a single measure indicative of the overall variation.

D. Detection of burst onset landmarks

The detection of voicing offsets (*g*-) and voicing onsets (*g*+) are performed using the method reported in [1] using the peak energy variation in the frequency band from 0 to 400 Hz. The peak energy is computed as

$$E_g(n) = 10\log_{10}(\max|X(n,k)|^2) \tag{7}$$

where $k_1 \le k \le k_2$, k_1 and k_2 being the DFT indices corresponding to 0 and 400 Hz respectively. A rate of rise measure of $E_g(n)$ is computed with a time step of 50 ms (K = 50) as

$$r_{Eg}(n) = E_g(n) - E_g(n-K) \tag{8}$$

The crossing points $r_{Eg}(n)$ below and above threshold values of -12 dB and +12 dB respectively are taken as the voicing offset and voicing onset points. An intervocalic burst onset is located at the most prominent peak in the ROC, between the *g*- and *g*+ points.

Three different approaches were attempted for the computation of the covariance matrix, namely 1) dynamic computation from a cluster formed by parameter set of 20 frames preceding the current frame, 2) computation from the parameter set of the entire utterance, and 3) computation from the parameter set of the utterance excluding the silence and very low energy intervals.

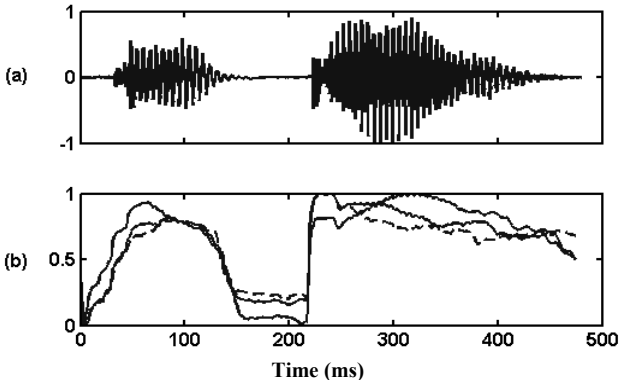

Figure 3. VCV utterance /aga/: (a) waveform, (b) E_{b1} (thick), E_{b2} (thin) and E_{b3} (dashed) tracks.

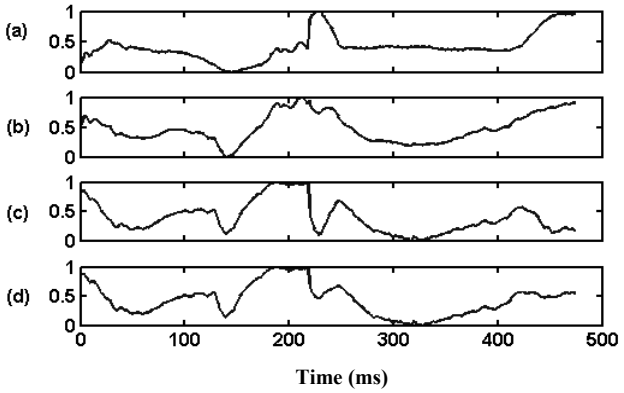

Figure 4. VCV utterance /aga/: (a) F_c , (b) F_σ , (c) F_s , and (d) F_k tracks.

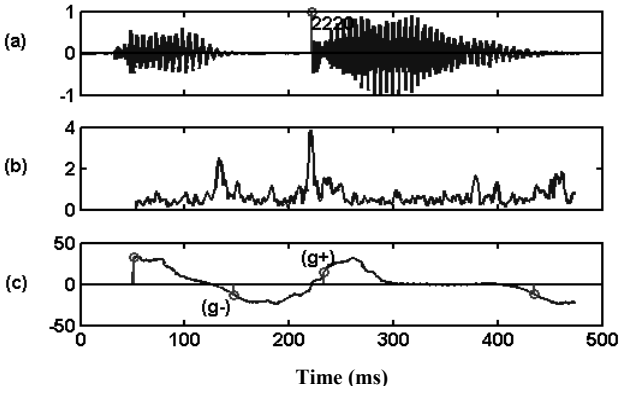

Figure 5. VCV utterance /aga/: (a) Waveform, (b) ROC-MD, (c) r_{Eg} track used for voicing offset and onset detection.

An examination of the ROC tracks resulting from these three approaches, and the individual values in the covariance matrix indicated method 3 to be most effective. The threshold used to demarcate the speech and silence was kept at 20 dB below the maximum signal level in the utterance. The absolute value of this threshold and the duration of parameters concatenated for getting the covariance matrix were individually examined and were found not very critical in getting a valid covariance matrix, provided the parameter set properly represented the long term variations in the high energy speech segments. The ROC-MD was computed as

$$\mathrm{ROC}_{\mathrm{md}}(n) = ((\mathbf{y}(n) - \mathbf{y}(n-K))\boldsymbol{\Sigma}^{-1}(\mathbf{y}(n) - \mathbf{y}(n-K))^T)^{0.5} \quad (9)$$

where $\boldsymbol{\Sigma}$ is the precomputed covariance matrix from the selected parameter set, $\mathbf{y}(n)$ is the parameter set of the current frame n, and K is the time step.

III. Evaluation and Results

A. Evaluation using VCV utterances

The method was evaluated using 180 VCV utterances from 10 speakers (5 male and 5 female) involving 6 stop consonants (/*p, t, k, b, d, g*/) recorded in three vowel contexts (/*a, i, u*/) from each speaker.

A total of seven parameters, namely three band energy (E_{b1}, E_{b2}, E_{b3}) and four spectral moments (F_c , F_σ , F_s , F_k) were used for burst onset landmark detection. Fig. 3 shows the energy tracks for a VCV utterance, on a normalized scale. The four corresponding spectral moments are shown in Fig. 4. Fig. 5 shows the VCV utterance, and ROC-MD computed from energy and spectral moment parameters, along with the r_{Eg} track. Four different cases were considered, based on the selection of distance measure and the parameter set used for burst onset detection.

Case 1: Sum of ROC of band energies: Individual ROCs were computed for 3 band energy parameters (E_b) with time steps of 3 ms and 6 ms. A combined ROC was obtained by normalizing absolute values of each ROC to the range of 0 to 1 and by summing them. The three normalized ROCs were assumed to contribute equally towards burst onset detection.

Case 2: ROC-MD of band energies: ROC-MD was computed using the 3 band energy parameters defining the parameter set $\mathbf{y}$ and the corresponding covariance matrix $\boldsymbol{\Sigma}$, for time steps of 3 and 6 ms.

Case 3: ROC-MD of spectral moments: Using the four spectral moments, ROC-MD was computed. To avoid computation overflow during silence and closure segments, addition of 1) single tone of 100 Hz, and 2) broad-band noise to the signal was investigated. Addition of tone had the drawback of offsetting the centroid to the tone frequency and thereby biasing the higher order moments. A broad-band noise with 40 dB SNR was added to the signal, which was found not to affect the ROCs associated with the frication noise at the burst onsets.

Case 4: ROC-MD of energy and spectral moments: ROC-MD was computed for all the utterances with the three energy parameters and the four spectral moments defining the parameter set $\mathbf{y}$ and the corresponding covariance matrix $\boldsymbol{\Sigma}$, for time steps of 3 and 6 ms.

The detection rates at the temporal accuracy levels of 3, 6, 10, 15, and 20 ms are listed in Table I for time steps of 3 and 6 ms, respectively, for the 4 different cases considered. The temporal accuracy of burst onset detection reduced as the time step was increased, in all the four cases. Compared to the use of sum of ROC's, ROC-MD improved the temporal accuracy of landmark detection. The spectral moments were less effective compared to the band energy

TABLE I. DETECTION RATES FOR VCV UTTERANCES

Case	Time step (ms)	Temporal accuracy (ms)				
		3	5	10	15	20
1	3	81	84	86	87	87
	6	48	79	85	86	87
2	3	86	91	93	96	97
	6	55	85	93	96	97
3	3	76	81	83	87	90
	6	52	66	72	73	76
4	3	90	95	96	98	99
	6	57	90	96	99	99

TABLE II. DETECTION RATES FOR TIMIT SENTENCES

Time step (ms)	Temporal accuracy (ms)				
	3	5	10	15	20
3	58	63	77	87	88

parameters. The combined use of energy and spectral moments with a Mahalanobis distance based ROC was most effective in localizing burst onset landmarks.

B. Evaluation using TIMIT sentences

The method was evaluated using 50 conversational style sentences from TIMIT database involving 10 sentences each from 5 speakers (2 male and 3 female). ROC-MD of spectral moments and energy parameters (time step = 3 ms) which gave the best burst detection rate in VCV utterances (Case 4) was used for evaluation. The covariance matrix was precomputed for each utterance excluding parameters during closure and silence intervals.

A single-pass processing was used for burst onset detection which located the most prominent peak in the ROC-MD for each voicing onset (*g*+) with a preceding voicing offset (*g*-) and a valid closure interval. A minimum duration of 10 ms with energy below 20 dB of the vowel energy level in the utterance was considered as a valid closure. For a total of 238 tokens, the detection rates at the temporal accuracy levels of 3, 6, 10, 15, and 20 ms are listed in Table II. The method also detected silence to vowel/semivowel onsets, frication onsets, glottal stops/clicks, with the insertion rates being 11%, 9%, and 4% respectively.

IV. CONCLUSION

The use of spectral moments as additional parameters for burst onset detection and the use of Mahalanobis distance based rate of change was investigated. Energy parameters are highly reliable and contribute more towards detection rate. Spectral moments are useful as additional parameters for improving detection rates of burst onset landmark, but need to be combined with energy parameters for reliable and accurate detection of landmarks. Rate of change obtained by Mahalanobis distance based first difference (ROC-MD) operation is more effective in combining parameters and deriving a single parameter indicative of the overall variation. It is less sensitive to the variations in time steps and it is effective for time-localizing the burst onsets. For both distance measures considered, short time steps performed better.

ACKNOWLEDGMENT

The research work is supported by the project “National Program on Perception Engineering”, sponsored by the Department of Information Technology, MCIT, Government of India.

REFERENCES

[1] S. A. Liu, “Landmark detection for distinctive feature based speech recognition,” *J. Acoust. Soc. Am*., vol. 100, pp. 3417-3430, 1996.

[2] V. Hazan and A. Simpson, “The effect of cue-enhancement on the intelligibility of nonsense word and sentence materials presented in noise,” *Speech Commun.*, vol. 24, pp. 211-226, 1998.

[3] M. A. Picheny, N. I. Dulrach, and L. D. Braida, “Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech,” *J. Speech Hear. Res*., vol. 28, pp. 96-103, 1985.

[4] R. W. Guelke, "Consonant burst enhancement: A possible means to improve intelligibility for the hard of hearing,” *J.Rehab. Res. Develop*., vol. 24, pp. 217-220, 1987.

[5] V. Colotte and Y. Laprie, “Automatic enhancement of speech intelligibility,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2000, Istanbul, Turkey, pp. 1057-1060.

[6] M. Ortega, V. Hazan, and M. Huckwale, “Automatic cue enhancement of natural speech for improved intelligibility”, *Speech, Hearing, and Language: Work in Progress*, vol. 12, pp. 42-56, 2000.

[7] C. Y. Lin and H. C. Wang, “Mandarin stops classification using spectral moments,” in *Proc. Int. Symp. Chinese Spoken Language Processing (ISCSLP)*, 2008, pp. 1-4.

[8] C. Y. Lin and H. C. Wang, “Using burst onset information to improve stop/affricate phone recognition”, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2010, Dallas, Texas, pp. 4862 - 4865.

[9] P. C. Pandey and M. S. Shah, “Estimation of place of articulation during stop closures of vowel-consonant-vowel utterances,” *IEEE Trans. Audio, Speech and Language Processing*, vol. 17, pp. 277-286, 2009.

[10] A. R. Jayan and P. C. Pandey, "Automated detection of speech landmarks using Gaussian mixture modeling,” in *Proc. Int. Symposium on Frontiers of Research on Speech and Music (FRSM)*, 2008, Kolkata, India, pp. 323-327.

[11] A. R. Jayan and P. C. Pandey, “Detection of stop landmarks using Gaussian mixture modeling of speech spectrum,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2009, Taipei, Taiwan, pp. 4681-4684.

[12] P. Niyogi and M. M. Sondhi, “Detecting stop consonants in continuous speech,” *J. Acoust. Soc. Am.*, vol. 111, pp. 1063-1076, 2002.

[13] A. Salomon, C. Y. Espy-Wilson, and O. Deshmukh, “Detection of speech landmarks: Use of temporal information,” *J. Acoust. Soc. Am.*, vol. 115, pp. 1296-1305, 2002.

[14] C. Park, “Consonant landmark detection for speech recognition,” Ph.D. dissertation, Massachusetts Institute of Technology, 2008.

[15] T. N. Sainath and T. J. Hazan, “A sinusoidal model approach to acoustic landmark detection and segmentation for robust segment-based speech recognition, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2006, Toulose, France, pp. 525-528.

[16] P. C. Mahalanobis, "On the generalized distance in statistics," *Proc. National Institute of Sciences of India*, vol. 2, pp. 49–55, 1936.

[17] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, Discrete-Time Processing of Speech Signals, New York: John Wiley, 2000.