Automated CVR Modification for Improving Perception of Stop Consonants

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

Abstract—**Increasing the intensity of the consonant segments relative to the nearby vowel segments, known as consonant-vowel ratio (CVR) modification, is reported to be effective in improving perception of stop consonants for listeners in noisy backgrounds and for hearing-impaired listeners. A technique for automated CVR modification using detection of acoustic landmarks corresponding to the stop release bursts, with high temporal accuracy, is investigated. Its effectiveness in improving perception of stop consonants in the presence of speech-spectrum shaped noise is evaluated by conducting listening tests on five normal-hearing subjects with VCV utterances involving six stop consonants and three vowels. The processing improved the recognition scores for stop consonants by nearly 7, 18, and 25% at SNR levels of 0, −6, and −12 dB, respectively.**

Keywords- Clear speech; acoustic landmarks; burst onset detection; CVR modification; speech intelligibility

I. INTRODUCTION

A talker in a difficult communication environment, such as a noisy background or while talking to a hearing impaired listener, usually alters the speaking style to make the speech more intelligible. This mode of speaking is known as “clear speech”. Compared to the conversational style speech, it is more intelligible for normal-hearing listeners under adverse listening conditions, hearing-impaired listeners, children with learning disabilities, and non-native listeners, [1]-[4]. The important acoustic properties identified to be contributing to the intelligibility advantage of clear speech include increased intensity of the consonant segments relative to the vowel segments, more targeted vowel formants, increased value and dynamic range of fundamental frequency, properly released stop bursts with increased intensity, expanded stop closure and formant transition durations, and increased 1-3 kHz energy in the long-term spectrum [1],[5],[6]. Out of these parameters, consonant intensity and duration are the two important parameters identified for speech intelligibility enhancement.

The ratio of RMS value of the consonant segment to that of the nearby vowel segment is known as the consonant-vowel ratio (CVR). Gordon-Salant [7] conducted listening tests on normal-hearing listeners using manually annotated consonant-vowel (CV) syllables involving 19 consonants paired with 3 vowels in 12-talker babble noise background. At 6 dB SNR, increasing CVR by 10 dB resulted in nearly 14% increase in recognition scores. Montgomery and Edge [8] used 100 consonant-vowel-consonant (CVC) syllables for tests on twenty subjects with sensorineural hearing impairment. Modification of the stimuli by equating the burst and vowel intensities resulted in nearly 10.5% increase in recognition scores. Kennedy *et al.* [9] investigated the CVR increase needed for maximizing consonant recognition using vowel-consonant (VC) syllables with 9 voiced and 7 unvoiced consonants in 3 vowel contexts (/*a*, *i*, *u*/) as the stimuli and subjects with sensorineural hearing loss. The optimum value of CVR increase ranged from 7.3 dB in /*a*/ context to 9.5 dB in /*i*/ context for voiced consonants, and from 9.5 dB in /*i*/ context to 12.5 dB in /*u*/ context for unvoiced consonants. Thomas [10] investigated the effect of CVR modification using synthesized VC and CV syllables consisting of six stop consonants in three vowel contexts. Testing with broadband noise background on normal-hearing subjects showed that increase in CVR by 12 dB improved recognition scores by 14% for CV syllables and 23% for VC syllables. The modification was more beneficial in VC context than in CV context, possibly due to the more effective suppression of forward masking.

The acoustic cues essential for speech identification are concentrated in certain regions in speech and these perceptually important regions are known as “landmarks” [11],[15]. Many researchers have used a perceptual-cue enhancement approach by modifying the speech characteristics around the landmarks. Hazan and Simpson [11] used 36 VCV syllables, comprising 12 consonants /*b*, *d*, *g*, *p*, *t*, *k*, *f*, *v*, *s*, *z*, *m*, *n*/ and three vowels /*a*, *i*, *u*/ as the test material, with manually annotated vowel onset and offset, burst, aspiration, frication, and nasal landmarks. Two sets of modified stimuli were prepared. In the first set, the modification involved increasing the intensity of closure release burst by 12 dB and that of frication and nasal regions by 6 dB. In the second set, the intensity of VC and CV transitions (four pitch cycles preceding the voicing offset and following the voicing onset) was also increased along with the earlier modification. Stimuli were combined with speech-shaped noise at SNR levels of 0 and –5 dB. Listening tests on 13 normal-hearing subjects showed 6 to 12% improvement in recognition scores, at SNR levels of 0 and –5 dB, respectively. Improvement gained from intensity modification of the transition segments was marginal.

Guelke [12] reported an automated technique for improving perception of stop consonants /*p*, *t*, *k*, *b*, *d*, *g*/ by detecting and amplifying stop release bursts by 9, 15, and 17 dB. Listening tests in the presence of masking noise using nonsense syllables showed the recognition scores increasing from 51% for the unprocessed speech to 90% for speech with 17 dB burst intensity enhancement. Colotte and Laprie [13] used a spectral variation function based on mel-cepstral analysis to locate the regions for enhancement, with 82% detection of manually located landmarks at a temporal accuracy of 20 ms. Bursts and fricative segments were amplified by 6 dB. Listening tests, involving identification of missing words in 50 sentences from TIMIT database on 13 normal-hearing subjects, showed an increase of 9% (from 72 to 81%) in the recognition scores. The enhancement was equally effective for stops and fricatives.

Ortega *et al.* [14] used a broad-class HMM classifier for identifying the burst, frication, nasal, vowel onset, and vowel offset landmarks with detection rates of 80, 75, 90, 78, and 76%, respectively. The corresponding insertion rates were 25, 9, 24, 16, and 17%. The test material comprised VCV utterances with 12 consonants /*p*, *t*, *k*, *b*, *d*, *g*, *m*, *n*, *s*, *f*, *v*, *z*/ and the vowels /*a*, *u*/. The enhancement involved amplification of fricatives and nasals by 9 dB, and stop bursts by 12 dB. The waveform discontinuities during intensity scaling were avoided using 5 ms raised half-cosine functions. Testing on 52 subjects using VCV syllables added with speech-shaped noise at −5 dB SNR resulted in recognition scores of 74, 76, and 72%, for the unmodified stimuli, stimuli with manual labeling and modification, and the automatic labeling and modification, respectively. The reduction in recognition score for the automated method was attributed to the limitations of the automated landmark detection method and the use of fixed amplification levels.

Thus the several earlier investigations show that CVR modification has the potential for improving consonant perception for normal hearing listeners in noisy backgrounds and for hearing-impaired listeners. A speech processing technique involving automated detection and enhancement of stop consonant landmarks is investigated. The burst onset, and the VC and CV transitions are detected and scaled in intensity, and the effect of modification is evaluated using VCV utterances.

II. Landmark Detection

The voicing offset and voicing onset landmarks are located using variation of peak energy parameter in the spectral band 0 – 400 Hz. We have used peak energy from the spectral band (1.2 to 5 kHz) and spectral centroid as the parameters for burst onset detection.

For speech sampled at 10 kHz, 512-point DFT is computed for 6 ms Hanning windowed frames, taken every one ms. The magnitude spectrum is smoothed by a 20-frame moving average. From the smoothed spectrum$|X(n,k)|$, the peak energy $E_g(n)$ in the 0 – 400 Hz band is computed as

$$E_g(n) = 10log_{10}(\max(|X(n,k)|^2, k_1 \leq k \leq k_2)) \qquad (1)$$

where n is the time index, k_1 and k_2 are the DFT indices corresponding to 0 and 400 Hz, respectively. The detection of voicing offset g– and voicing onset g+ is performed as reported in [15], using a rate of rise measure $r_E(n)$ computed with a time-step of 50 ms (K = 50) as

$$r_E(n) = E_g(n) - E_g(n-K) \qquad (2)$$

The crossing points below and above threshold values of −9 dB and +9 dB, respectively are taken as the voicing offset and voicing onset points.

Our earlier investigation [16] showed that the use of spectral moments, namely the centroid, standard deviation, skewness, and kurtosis related to the frequency of concentration of the spectral energy, spread of energy around this location, the symmetry of the spectrum, and its peakiness help in improving the detection of burst onset landmarks when combined with the band energy parameters. Further investigation has shown the peak energy in the spectral band 1.2 – 5 kHz (computed using Eq. 1 with corresponding indices k_1 and k_2) and the spectral centroid to be the major contributors.

During low energy segments such as stop closures, the spectral centroid may have random fluctuations, adversely affecting its usefulness. Addition of 100 Hz sinusoidal tone of −20 dB with respect to the maximum signal level was found to mask the random fluctuations during the closure without masking the bursts. The tone-added smoothed magnitude spectrum $|X_t(n,k)|$ for N-point DFT is normalized as

$$p(n,k) = |X_t(n,k)| / \textstyle\sum_{k=1}^{N/2} |X_t(n,k)| \qquad (3)$$

The centroid of the spectrum is computed as

$$F_c(n) = \textstyle\sum_{k=1}^{N/2} p(n,k)k \qquad (4)$$

A rate of change function based on Mahalanobis distance [17],[18] and with a short time-step is used as it is effective in taking care of the correlation and scale differences between the parameters [16]. The function is computed as

$$\mathrm{ROC}_{\mathrm{md}}(n) = (\mathbf{d}(n)\,\Sigma^{-1}\,\mathbf{d}(n)^{\mathrm{T}})^{0.5} \qquad (5)$$

$$\mathbf{d}(n) = \mathbf{y}(n) - \mathbf{y}(n{-}K) \qquad (6)$$

where $\mathbf{y}(n)$ is the parameter set of the frame n, $\boldsymbol{\Sigma}$ is the covariance matrix, and K is the time-step. The covariance matrix properly weights the parameters and $\mathrm{ROC}_{\mathrm{md}}(n)$ remains in a low range during steady-state segments and increases sharply during abrupt transitions. The covariance matrix is computed using parameters from frames having energy within 30 dB of the highest energy. An intervocalic burst onset (b) is located at the most prominent peak in the $\mathrm{ROC}_{\mathrm{md}}(n)$ between the voicing offset and voicing onset

TABLE I. LANDMARK DETECTION FOR VCV UTTERANCES

Detection rate (%)	Temporal accuracy (ms)			
	≤5	≤10	≤15	≤20
Burst onset (180)	96	97	99	99
Voicing offset (180)	36	64	73	83
Voicing onset (180)	76	93	98	98

points. The landmark detection was evaluated with a set of 180 utterances comprising VCV syllables with 6 stop consonants /*p*, *t*, *k*, *b*, *d*, *g*/ and three vowels /*a*, *i*, *u*/ from 5 male and 5 female speakers. Detection rates are listed in Table 1. Burst detection was nearly perfect at temporal accuracy of 20 ms and decreased to 96%, for temporal accuracy of 5 ms. Detection of voicing onset and offset was satisfactory at 20 ms but significantly decreased at 5 ms.

III. CVR MODIFICATION

Acoustic cues for the identification of stop consonants in nonsense VCV utterances are contained mainly in the VC and CV transitions and the stop release burst. In clear speech, the burst is more intense and clearly released. CVR modification involved increasing the energy of burst onset and the VC and CV transitions. With g_-, b, and g_+ representing the voicing offset, burst onset, and voicing onset points, respectively; the locations t_a, t_b, t_c, and t_d were selected as (g_- – 20 ms), g_-, (b – 10 ms) and (g_+ + 10 ms), respectively. The region selected for intensity enhancement extended from t_a to t_b in the VC transition and from t_c to t_d in the CV transition. The 20 ms margin in the VC transition ensures the enhancement to start 20 ms before the voicing offset. The 10 ms margins in the CV transition ensure that the enhancement starts 10 ms before the burst onset and extends into the voicing region by 10 ms. This also takes care of the possible temporal misalignment of the detected landmarks. The amplification was applied using a cosine tapered window with a smooth rise and fall times of 2.5 ms to avoid amplitude discontinuities. An example of the processing for CVR enhancement for a VCV utterance is given in Fig. 1. Informal listening tests conducted using VCV utterances with the three vowel contexts as test material showed the amplification by 9 dB to be most effective, in terms of improving the audibility of the consonant without introducing perceptible distortion.

IV. LISTENING TESTS

The test material consisted of 18 VCV syllables, with 6 stop consonants /*b*, *d*, *g*, *p*, *t*, *k*/ and three vowels /*a*, *i*, *u*/. The utterances were recorded from 5 speakers (2 male and 3 female) at a sampling frequency of 10 kHz. This test material helps in removing the semantic effects and bringing out the possible vowel context dependencies of the enhancement. The stimuli were normalized to have the same RMS value during the vowel segments. The stimuli were processed for 9 dB increase in CVR. Processing did not affect the vowel levels as it involved amplification of the transition and burst segments.

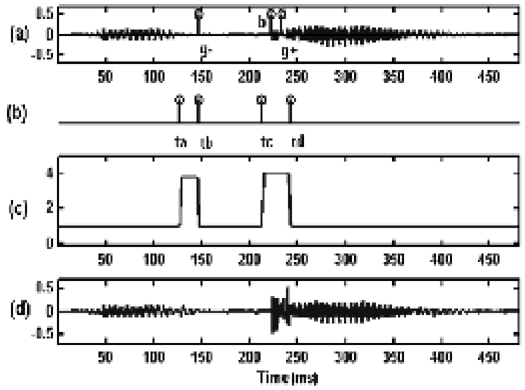

Figure 1. CVR modification of VCV utterance /aga/: (a) waveform with landmarks, (b) boundaries of windows selected for intensity modification, (c) scaling function for CVR enhancement, (d) modified waveform.

Speech-shaped noise, with long-term spectrum nearly flat from 100 to 1000 Hz and 12 dB/octave roll-off afterwards [11],[15] was used as the background noise. Noise was added to get the stimuli with SNR levels of ∞ (no noise), +12, +6, 0, –6, and –12 dB, with respect to the RMS value of the vowel segment. The noise extended for 1 s on either side of the stimulus. There were 12 listening conditions corresponding to two types of stimuli (unprocessed, processed) and six SNR levels. The stimuli were presented through Sennheiser PX80 headphones. The amplifier gain was set by the listener for the most comfortable listening level for the loudest set of sounds, corresponding to −12 dB SNR. The same setting was used for all the listening conditions. Thus the speech signal levels remained constant across the different SNR levels.

Consonant recognition tests were conducted on five normal-hearing subjects (3 male and 2 female, 18−40 years). Test for each listening condition consisted of six sessions, with a total of 360 presentations (5 speakers × 18 VCV utterances × 4 presentations). The order of listening conditions was randomized across the listeners to reduce the effects of practice and fatigue. A computerized test administration setup was used for conducting the listening test. After each presentation, the listener responded by clicking on one of the six response choices (corresponding to the six consonants) displayed on the computer screen. The position of the response choices displayed on the computer screen was randomized to avoid position bias in the responses. The response and the response time were recorded. Each test session involved 60 presentations, with the stimuli selected in a randomized order.

TABLE II. RECOGNITION SCORES FOR CVR MODIFICATION

Sub.	SNR ∞		12 dB		6 dB		0 dB		−6 dB		−12 dB	
	Unp.	Proc.	Unp.	Proc.	Unp.	Proc.	Unp.	Proc.	Unp.	Proc.	Unp.	Proc.
ARJ	97	96	96	94	93	94	79	94	57	78	35	56
KRI	97	95	89	89	87	86	81	86	75	89	50	79
ANU	86	87	84	84	83	85	80	86	67	84	37	63
AGM	97	98	97	98	97	97	91	96	72	95	46	68
RKS	87	87	84	84	80	80	76	82	67	84	37	63
Mean(%)	93	93	90	90	88	88	81	88	68	86	41	66
s.d.	5.8	5.2	6.3	6.2	7	6.9	5.7	5.9	6.8	6.4	6.6	8.5
Impr.(%)		-0.2		-0.2		0.4		7.4		18.4		24.8
p		0.37		0.35		0.24		0.01		2E-04		4E-05

V. RESULTS AND DISCUSSION

The recognition scores for the individual subjects along with the mean and standard deviation are given in Table 2 and a plot of the mean scores is shown in Fig. 2. The recognition scores of the unprocessed and the processed stimuli were almost the same as the SNR decreased to 6 dB. For SNR levels of 0, −6, and −12 dB, the recognition scores of the unprocessed stimuli were 81, 68, and 41%, respectively. The corresponding values for the CVR modified stimuli were 7, 18, and 25% higher.

Analysis of recognition scores from the stimulus-response confusion matrices in the three vowel contexts indicated CVR modification to be almost equally effective in the three vowel contexts. Consonant-wise analysis of the recognition scores for voiced and unvoiced stop consonants at different SNR levels are shown in Fig. 3 and 4. Compared to the labial (*b*, *p*) and velar stops (*g*, *k*), alveolar stops (*d*, *t*) were benefitted more by CVR modification at the lower SNR levels.

Information transmission analysis [19] was carried out on the stimulus-response confusion matrices to get the relative information transmitted in terms of overall, voicing, and place features. The results, listed in Table 3, show that the loss in information at the lower SNRs was mainly due to the place feature. Processing by automated CVR enhancement was effective in improving the transmission of the place feature. The improvement in transmission of the voicing feature by consonant intensity enhancement was lower compared to the corresponding improvement in transmission of the place feature.

The response time is an indicator of the perceptual load of the stimuli. As seen in Fig. 5, the mean response time increased as the SNR decreased to +12 dB and there was no significant change for further decrease in SNR. There were

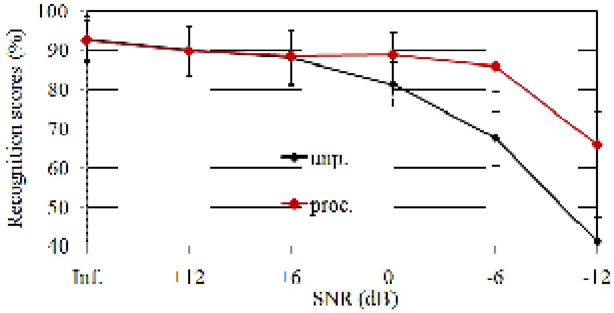

Figure 2. Recognition score (%) vs SNR (dB).

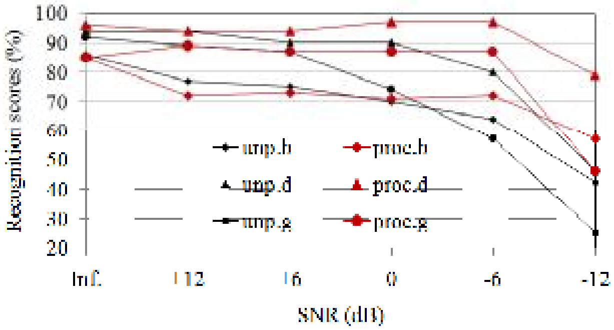

Figure 3. Recognition scores (%) for voiced stops vs SNR (dB).

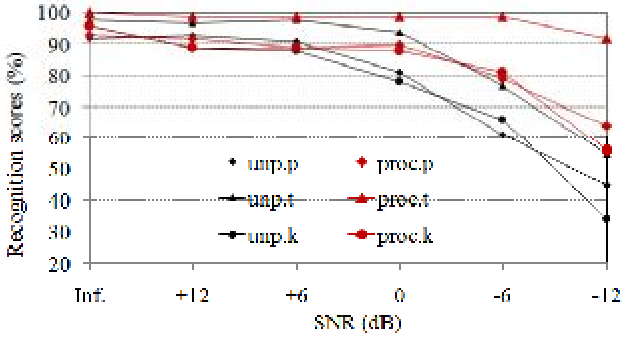

Figure 4. Recognition scores (%) for unvoiced stops vs SNR (dB).

TABLE III. RESULTS OF INFORMATION TRANSMISSION ANALYSIS. UNP.: UNPROCESSED STIMULI, PROC.: STIMULI PROCESSED WITH CVR MODIFICATION

SNR	Overall		Place		Voicing	
(dB)	Unp	Proc	Unp	Proc	Unp	Proc
Inf.	86	85	89	87	75	77
+12	79	79	84	82	67	69
+6	75	76	77	80	66	66
0	63	78	55	81	69	66
-6	45	72	29	70	61	69
-12	20	43	6	32	34	51

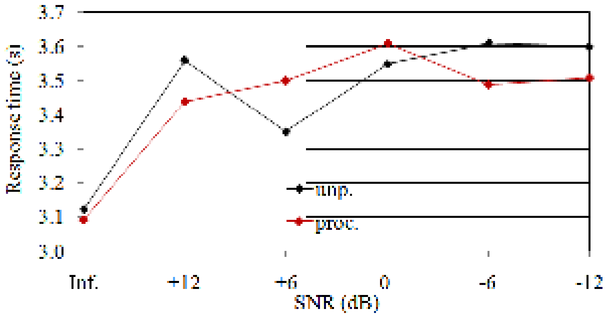

Figure 5. Response time (in s) *vs* SNR (dB)

no consistent differences in the values of the response time for the unprocessed and the processed stimuli indicating that the processing resulted in an increase in recognition scores without a change in the perceptual load.

VI. CONCLUSION

A method for CVR modification for improving perception of stop consonants has been presented. The method automatically locates the landmarks associated with stop consonants and amplifies the transition and burst segments, without introducing perceptible artifacts. The effect of processing was evaluated by conducting listening tests on five normal-hearing subjects, using VCV utterances from five speakers as the test material.

Intensity modification of the burst onset and transition segments improved stop consonant perception at lower SNR levels by 25%, and the improvement was contributed by better reception of the place feature of the stops. It did not introduce any change in the perceptual load. The improvements in recognition scores achieved by our technique for automated processing are similar to those reported in earlier studies using CVR modification with manually annotated landmarks. Thus an implementation of the technique with real-time processing has the potential of improving speech perception in noisy listening situations. Further listening tests are needed to establish the optimum value of CVR modification for different levels of background noise and to evaluate the processing for listeners with sensorineural hearing impairment.

REFERENCES

[1] M. A. Picheny, N. I. Durlach, 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, 96-103, 1985.

[2] K. L. Payton, R. M. Uchanski, and L. D. Braida, “Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing,” J. Acoust. Soc. Am., 95, 1581-1592, 1994.

[3] A. R. Bradlow, N. Kraus, and E. Hayes, “Speaking clearly for children with learning disabilities”, J. Speech Lang. Hear. Res. 46, 80-97, 2003.

[4] A. R. Bradlow and T. Bent, “The clear speech effect for non-native listeners,” J. Acoust. Soc. Am., 112, 272-284, 2002.

[5] F. R. Chen, “Acoustic characteristics and intelligibility of clear and conversational speech at the segmental level,” M.S. dissertation, Massachusetts Institute of Technology, 1980.

[6] M. A. Picheny, N. I. Durlach, and L. D. Braida, “Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech,” J. Speech Hear. Res., 298, 434-446, 1986.

[7] S. Gordon-Salant, “Recognition of time/intensity altered CVs by young and elderly subjects with normal-hearing,” J. Acoust. Soc. Am. 80, 1599-1607, 1986.

[8] A. A. Montgomery and R. A. Edge, “Evaluation of two speech enhancement techniques to improve intelligibility for hearing-impaired adults,” J. Speech Hear. Res., 31, 386-393, 1988.

[9] E. Kennedy, H. Levitt, A. C. Neuman, and M. Weiss, “Consonant-vowel intensity ratios for maximizing consonant recognition by hearing-impaired listeners,” J. Acoust. Soc. Am., 103, 1098-1114, 1997.

[10] T. G. Thomas, “Experimental evaluation of improvement in speech perception with consonantal intensity and duration modification,” Ph.D. dissertation, Dept. of Elect. Engg., IIT Bombay, 1996.

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

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

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

[14] M. Ortega, V. Hazan, and M. Huckwale, “Automatic cue enhancement of natural speech for improved intelligibility”, Speech, Hearing, and Language, 12, 42-56, 2000.

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

[16] A. R. Jayan, P.S. Rajath Bhat, and P. C. Pandey, “Detection of burst onset landmarks in speech using rate of change of spectral moments,” in Proc. 17th National Conference on Communications, Bangalore, India, 2011.

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

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

[19] G. E. Miller, P. E. Niceley, An analysis of perceptual confusions among some English consonants. J. Acoust. Soc. Am., 27, No.2 338-352, 1955.