Effect of CVR Modification on Automatic Recognition of Stop Consonants in Isolated VCV Utterances by an HMM based ASR system

Arjun P
M.Tech.,
Govt. Engineering College
Thrissur Kerala 680009 India
email: em.arjunp@gmail.com

A.R. Jayan
Associate Professor,
Rajiv Gandhi Institute of Technology
Kottayam 686501 Kerala India
email: arjayan71@rit.ac.in

***Abstract*— Consonant-Vowel-Ratio (CVR) modification is a technique involving selective amplification of consonant segments with respect to the nearby vowel segments. CVR modification is reported to be effective for improving recognition of consonants by normal hearing listeners in adverse listening conditions and also for listeners with hearing impairments. In this paper, we have investigated the effect of CVR modification on recognition of stop consonants in vowel-consonant-vowel (VCV) utterances by an automatic speech recogniser (ASR) based on hidden Markov model (HMM). The results indicate CVR modification to be effective in improving recognition of stop consonants by nearly 6 to 7 % in the presence of additive noise.**

Keywords—CVR modification, Stop consonants, ASR.

I. INTRODUCTION

The regions in the speech signal that contain important acoustic cues that are important for speech perception are called landmarks [1]. Intelligibility of speech may get adversely affected by the distortion of landmarks either by external noise or by the impairments in the hearing mechanism of the listener. Speech perception can be improved by making the acoustic cues more robust against subsequent degradations that could be introduced by the channel or by the hearing mechanism. The basic idea behind this approach of speech enhancement techniques is derived from "clear speech", the speech produced by a talker with an intention to increase intelligibility in a difficult communication environment such as talking to a hearing impaired listener/in presence of background noise. Speech produced by a talker in such situations with careful articulation is reported to be nearly 17% more intelligible than the normal (conversational style) speech of the same talker [2][3]. In clear speech, the landmarks are more robust against subsequent degradations. Clear speech is also reported to be more intelligible than conversation speech for children with learning disabilities and for non-native listeners [4][5].

Clear speech has distinct acoustic properties in the phoneme, syllable, and sentence levels, when compared with conversational speech. Reduced rate of articulation, more frequent and lengthy pauses, less sound deletions, well attained vowel formant targets, increase in the mean and dynamic range of fundamental frequency, increased intensity of consonant segments relative to nearby vowel segments, well defined stop release bursts etc., are some of the most noticeable acoustic properties of clear speech [6][7]. Increase in the consonant-vowel-ratio (CVR) manifested in the speech signal as increased intensity of consonant segments relative to the nearby vowel segments is reported to be a major contributor towards the increased intelligibility of clear speech. Even though, the reduction in articulation rate is the most noticeable parameter in clear speech, its contribution towards increased speech intelligibility is reported to be marginal [7][8][9].

Several speech processing techniques have been reported based on the modification CVR for improving speech intelligibility of conversational speech for normal hearing listeners in adverse listening conditions and also for listeners with hearing impairments [10][11][12[13][14]. Some of these methods make use of manual detection and modification of landmarks whereas some methods involve automated landmark detection and modification. Manual approaches are useful for investigating the effect of modification of acoustic cues on speech perception whereas the automated methods are useful in practical applications involving speech intelligibility enhancement.

A stop consonant is produced with a closure of the vocal tract followed by a sudden release. The signal energy is very low during closure and it is followed by a release burst caused by the sudden release of energy. Because of the weak and transient nature, perception of stop consonants by human listeners may get very much degraded in the presence of noise. Several investigations have indicated the importance of release burst on the perception of stop consonants [15][16]. CVR modification involving accurate detection of release bursts and subsequent modification of intensity of release burst by an appropriate gain function is reported to be effective for improving recognition of stop consonants by normal hearing listeners in noisy backgrounds [14].

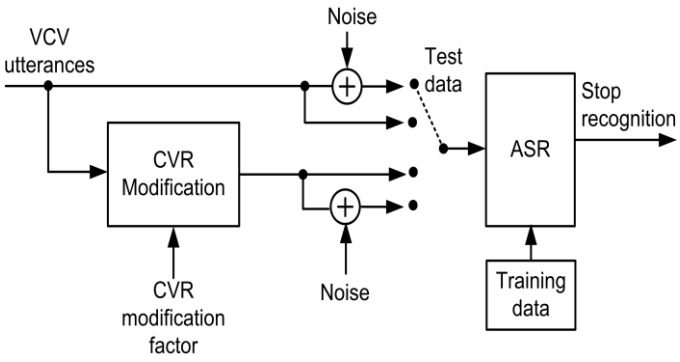

Fig. 1. The proposed method of ASR system.

CVR modification may improve the recognition of stop consonants in automatic speech recognition (ASR) system. The proposed method used in this investigation is illustrated in Fig.1.

We build an ASR system trained for recognising stop consonants in isolated vowel consonant vowel (VCV) utterances. The stop consonants in the original VCV utterances are enhanced in intensity by a CVR modification stage. CVR modification basically involves increasing the relative intensity of the release burst with respect to the nearby vowel by a certain factor. The recognition accuracy of stop consonants by the ASR system for both unprocessed VCV utterances (*unp*.) and the CVR modified VCV utterances (*cvr*.) are evaluated. Stop consonant recognition for different CVR modification factors are also analysed. For the optimum CVR modification factor, performances of the ASR stop consonant recognition in the presence of additive noise at different SNRs are also evaluated. This investigation is expected to give insights to the possible use of CVR modification for improving performance of ASR system for improving stop consonant recognition in the presence of additive noise.

This paper is organised as follows: The next section presents the development details of an ASR based on hidden Markov model (HMM) using HMM tool kit (htk). Section 3 presents the signal processing for CVR modification. In Section 4, the evaluation of the ASR using the test material at different levels of noise is presented and in Section 5 the results are discussed. Section 6 concludes the paper.

II. HMM BASED ASR SYSTEM

A. Block diagram of ASR system

This section describes the various steps involved in building an automatic speech recognition system for recognising the individual phonemes in VCV utterances. We have used htk running on Ubuntu operating system [17][18]. The block diagram representation of building the ASR system is shown in Fig. 2.

B. Audio recording

The recording was performed using a high quality recording device in a noise-free environment. We have used single channel 16 bit PCM format with a sampling frequency

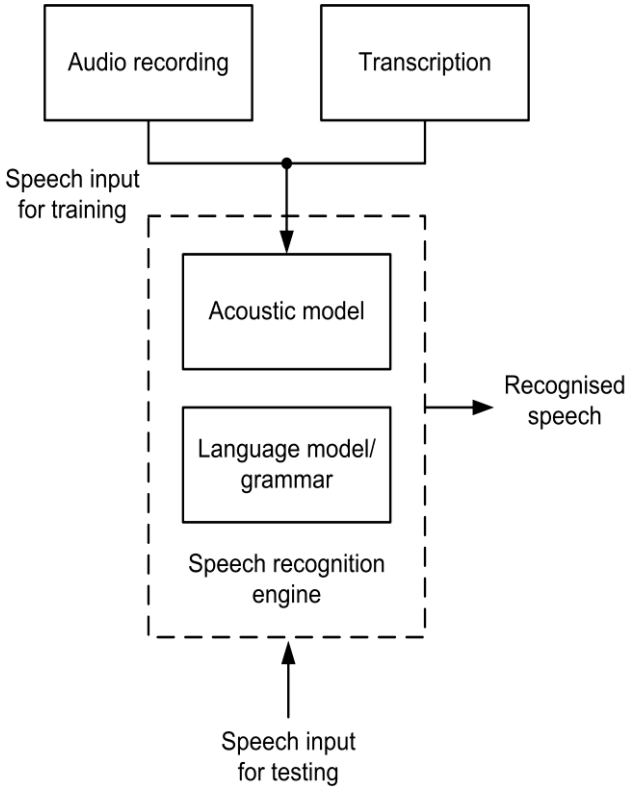

Fig. 2. Block diagram of the ASR system.

of 10 kHz. We have used six stop consonants (b,d,g,p,t,k) paired with three vowels (a,i,u) in VCV format. There are 54 VCV tokens per speaker (3 initial vowels × 6 stop consonants × 3 final vowels). Each VCV utterance was recorded twice, resulting in 108 VCV utterances per speaker. The database is created by recording VCV utterances from 11 speakers: 6 (3 male, 3 female) for the training data set and 5 (2 male, 3 female) for the test data set. The training data set consisted of 648 VCV utterances (108× 6) and the test data set consisted of 540 (108 ×5) VCV utterances. Out of the test data set, we have deleted 9 tokens based on audio and visual examinations. These tokens were observed to be articulated poorly and were not having proper release bursts/closures/vowel segments. The number of test tokens was limited to 531. The phoneme level details of the test VCV utterances are listed in Table 1 and Table 2.

C. Transcription

Each VCV utterance is labeled manually with the corresponding phone level details. The htk command HSLab was used to load wave file and to assign phone boundaries and labels. For each VCV utterance, we have used five distinct labels. The first and last segments are generally a silence segments labeled as sil. The second segment can be a vowel segment labeled as a, i, or u. The third segment is a closure part of the stop consonant and it is accordingly labeled as bcl, dcl, gcl, pcl, tcl, or kcl. The forth segment is the actual stop release

TABLE I. INITIAL AND FINAL VOWELS IN THE TEST DATA SET USED FOR CONSONANT RECOGNITION.

Vowel	Initial	Final
a	176	176
i	178	178
u	177	177
Total	531	531

TABLE II. CONSONANTS IN THE TEST DATA SET USED FOR CONSONANT RECOGNITION.

Consonant	No. of tokens	Consonant	No. of tokens
b	90	p	88
d	89	t	88
g	86	k	90

burst segment and it is labeled as b,d,g,p,t,k. The stop consonant segment is followed by the final vowel segment, labeled as a, i, or u. The labeling is performed for all the VCV utterances in the training set by careful audio and visual examination of each one of the 648 VCV utterances. Example for labeling a VCV utterance aka is given in Fig. 3.

D. Acoustic analysis

The acoustic signal is mapped to a parametric domain suited for speech recognition in the acoustic analysis stage. The HCopy tool of htk is used for this purpose. We have used Mel frequency cepstral coefficients (MFCC) as parameters for speech recognition. Each feature vector contains 39 values, namely: 12 MFCC values, 1 log energy value, 13 first order derivatives, and 13 second order derivatives. The settings used for acoustical analysis is listed in Table 3.

E. Acoustic model generation

HMM parameters need to be properly initialised for a fast and precise convergence of the training algorithm. We have used HInit command for initialising the HMMs. The HInit command uses a prototype file, labeled data, and MFCC data of the training data set. We have used a total of 16 HMMs (for 6 stops: b, d, g, p, t, k; 6 closures: bcl, dcl, gcl, pcl, tcl, kcl; 3 vowels: a, i, u, and 1 silence: sil). We have used an HMM topology with 4 active states and 2 non-emitting states (representing the initial and final states). After initialisation, the HRest command is applied iteratively. The HParse command is used to compile the grammar file and to generate a network file. The dictionary file lists the correspondence between the HMMs and the task grammar.

F. Speech recognition

Recognition of the stop consonant is performed using the HVite command. Each file in the test data set is converted to .mfcc files using HCopy command as described earlier. The HVite command returns a file containing the automatically

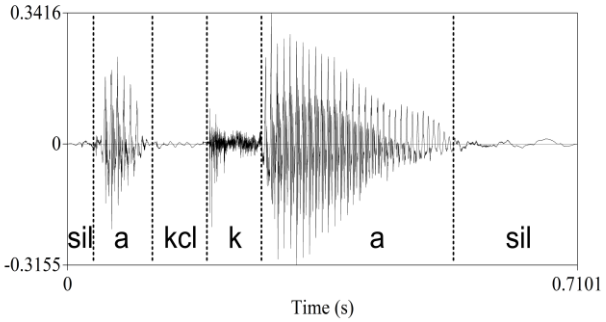

Fig. 3. Example for labeling of a VCV utterance /aka/.

TABLE III SETTINGS USED FOR ACOUSTICAL ANALYSIS.

Parameters	Specifications
Sampling rate	10000 Hz
File format	16 bit mono
Window length	25 ms
Window type	Hamming
Window overlap	15 ms
Pre-emphasis coeff.	0.97
Features	MFCC,ΔMFCC,ΔΔMFCC
Feature vector length	39
No. of filter banks	26
No. of MFCC coefficients	12

identified phone labels and boundaries. The evaluation using the training data set showed complete convergence of correct recognition of vowel and consonant segments in the VCV utterances.

III. SIGNAL PROCESSING FOR CVR MODIFICATION

CVR modification for stop consonants basically involves detection of the burst segment and modification of its intensity. The method described in [14] was used for CVR modification. The regions for CVR modification was selected by applying a thresholding logic to the rate of change of centroid frequency extracted from the magnitude spectrum of speech signal. The gain function applied on selected regions in the signal for a particular CVR modification factor was obtained by using peak energy and smoothed energy envelopes of the speech signal as described in [14]. An illustration of CVR modification of the speech signal for VCV utterance /aka/ is shown in Fig.4. It is clear that the release burst has been enhanced in intensity in the CVR modified utterance, compared to the unprocessed VCV utterance.

IV. EVALUATION OF ASR

The ASR was trained using the training data set. The test data set involved 531 VCV utterances recorded from 3 female (F18, F20, F22) and 2 male (M2, M4) speakers. The test data

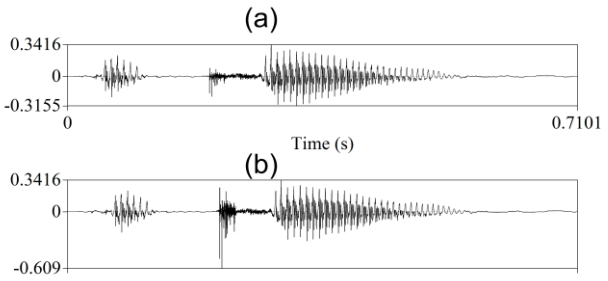

Fig. 4. CVR modification of a VCV utterance aka. (a) unprocessed VCV utterance, (b) CVR modified VCV utterance.

set was applied without any modification (*unp*.) and the consonant recognition score was obtained for individual speakers and also for individual stop consonants. The confusion matrix for stop consonant recognition was also analysed. The CVR modification was performed for modification factors of 3 dB, 6 dB, 9 dB, 12 dB, and 15 dB. Test data set with CVR modification factor *x* dB is denoted as *cvr_x*. For each of the CVR modified test data set, we analysed the stop consonant recognition scores for individual speakers and for individual stop consonants.

Two types of noise signals (spectrally shaped noise and white noise) were added to the test data set at 2 SNRs (20 dB and 10 dB). The VCV samples we have used for speech recognition experiments had an average duration less than 1 sec. We have used a spectrally shaped noise and white Gaussian noise (sampled at 10 kHz) mixed with the stimuli in our speech recognition experiments. The spectrally shaped noise was having a relatively flat spectral level in the 0 to 1 kHz band and its spectral envelope had a slope of -6dB for increase in frequency by every 1 kHz. The spectral envelope of white noise remained flat over the frequency range of 0 to 5 kHz. Both these noise maskers remained with same spectral properties over the entire duration of the VCV syllable. We have used an SNR level of 20 dB and 10 dB for evaluation. The noise signal was added to the VCV utterance and the level of noise for a required SNR was obtained relative to the maximum level of the vowel segment in the utterance. The VCV utterances were normalised to have same vowel level before addition of noise. It was also ensured that no clipping distortion took place during noise addition.

V. Results and Discussion

For the test data set, the initial vowel recognition was 100% for vowel /a/ and /i/ and 99% for vowel /u/. The overall initial vowel recognition score was 99.6%. For the final vowel, 100% recognition was obtained for /a/, 98.3% for /i/, and 98.8 % for /u/, resulting in an averaged vowel recognition score of 99%. For the stop consonants, the recognition scores were comparatively lower. The confusion matrix and the recognition scores (%) for the individual stop consonants are listed in Table 4. The voiced labial stop consonant /b/ had the minimum recognition score of 52.2%. It was mostly confused with the alveolar consonant /d/. The stop consonant /d/ was having a recognition score of 84.2% and it got mostly confused with velar stop /g/. The voiced velar stop /g/ had a recognition score of 77.9%. It got mostly confused with the unvoiced velar stop

TABLE IV. Confusion matrix for stop consonant recognition.

Stop	b	d	g	p	t	k
b(90)	47	26	8	7	0	2
d(89)	0	75	12	2	0	0
g(86)	1	5	67	1	0	12
p(88)	7	0	1	70	1	9
t(88)	0	1	0	1	52	34
k(90)	0	1	1	3	6	79

TABLE V. Speaker wise consonant recognition scores (%) without CVR modification.

Speaker (tokens)	**Rec. Score (%)**
F18(103)	74.8
F20(108)	76.8
F22(106)	67.9
M2(107)	77.5
M4(107)	70.1
Average	**73.4**

/k/. The recognition scores for unvoiced stop consonants /p/ and /k/ were 79.5% and 87.7%, respectively. Stop consonants /p/ got confused with /b/ and / k/. The recognition score for /t/ was 59.1% and it got mostly confused with /k/. The unvoiced velar stop /k/ had the maximum recognition score out of the 6 stop consonants.

Analysis was performed to see the recognition scores of individual speakers. The averaged recognition score for stop consonants was 73.4%. Out of the 5 speakers in the test data set, male speaker M2 had the highest consonant recognition score of 77.5% and the lowest score of 67.9% was observed for female speaker F22. The speaker-wise consonant recognition scores are listed in Table 5.

Effect of CVR modification by different scaling factors on consonant recognition was analysed. CVR modification factor was varied from 0 dB (unprocessed with no consonant modification) to 15 dB, in steps of 3 dB. It was observed that the CVR modification by 3 and 6 dB marginally improved the overall recognition scores. All other CVR modification factors (9 dB to 15 dB) resulted in reduction in the overall recognition scores. Table 6 lists the consonant recognition scores of 5 speakers for the different CVR modification factors. It may be concluded that the natural consonant-vowel ratio is giving best performance when the consonant recognition is performed in quiet condition and CVR modification by excessive factors adversely affect consonant recognition.

Table 7 lists the consonant-wise analysis of recognition scores with respect to CVR modification factors. It may be concluded that recognition of stop consonant /k/ is not much affected by the variation in CVR modification factors. Stop consonants /d/, /p/, and /t/ showed first increase in recognition

TABLE VI. CONSONANT RECOGNITION SCORES (%) FOR 5 SPEAKERS FOR DIFFERENT CVR MODIFICATION FACTORS.

Speaker. (tokens)	CVR Modification factor (dB)					
	0	**3**	**6**	**9**	**12**	**15**
F18(103)	74.8	70.8	70.8	69.9	70.8	70.8
F20(108)	76.8	77.8	79.6	77.7	75.0	72.2
F22(106)	67.9	66.9	64.1	64.1	62.2	60.4
M2(107)	77.5	82.2	81.3	80.3	76.6	74.8
M4(107)	70.1	74.6	72.8	71.9	72.8	71.9
Avg.	73.4	74.5	73.8	72.9	71.5	70.1

TABLE VII. CONSONANT RECOGNITION SCORES (%) FOR INDIVIDUAL STOP CONSONANTS FOR DIFFERENT CVR MODIFICATION FACTORS.

Consonant (tokens)	CVR Modification factor (dB)					
	0	**3**	**6**	**9**	**12**	**15**
b(90)	52.2	52.2	48.8	47.7	43.3	42.2
d(89)	84.2	86.5	84.2	82.0	79.7	78.6
g(86)	77.9	72.1	74.4	76.7	74.4	75.6
p(88)	79.5	82.9	79.5	77.3	78.4	75.0
t(88)	59.1	67.0	68.1	65.9	65.9	61.3
k(90)	87.7	87.6	87.8	87.7	87.7	87.8
Avg.	73.4	74.5	73.8	72.9	71.5	70.1

scores followed by reduction in recognition scores with further increase in CVR modification factors. Voiced stop consonants /b/ and /g/ displayed reduction in recognition scores with increase in CVR modification factor. CVR modification by factors of 6 to 9 dB is found to be most effective.

We have conducted two experiments to investigate the effect of additive noise on consonant recognition. Two types of noise (spectrally shaped noise and white noise) mixed with the test data set (unprocessed and CVR modified) at 2 SNRs (20 dB, 10 dB) was used for evaluation. The results are given in Fig. 5 and Fig. 6, respectively. It may be observed that the averaged consonant recognition scores at 20 dB SNR in the presence of spectrally shaped noise is maximum for CVR modification factor of 9 dB. For 10 dB SNR, the consonant recognition scores increase with increase in CVR modification factor.

For white Gaussian noise, it is observed that the recognition scores are lower than the corresponding scores for the spectrally shaped noise. This may be due to the increased masking effect of white Gaussian noise on consonant segments compared to spectrally shaped noise. At 10 dB SNR, it is observed that the averaged scores are insensitive to the CVR modification factor. This is due to the fact that the recognition scores for all stop consonants except /k/ reach 0% at 10 dB SNR, for white noise masker. All the stop consonants get recognized as /k/, resulting in 100% recognition score for /k/ and 0% score for all other consonants.

The speaker wise recognition scores for spectrally shaped noise and white noise for SNRs of 20 dB and 10 dB are listed in Table 8. The stop consonant wise recognition scores for spectrally shaped noise and white noise at 20 dB and 10 dB SNRs for CVR modification factor 9 dB are listed in Table 9. The recognition results indicate CVR modification to be

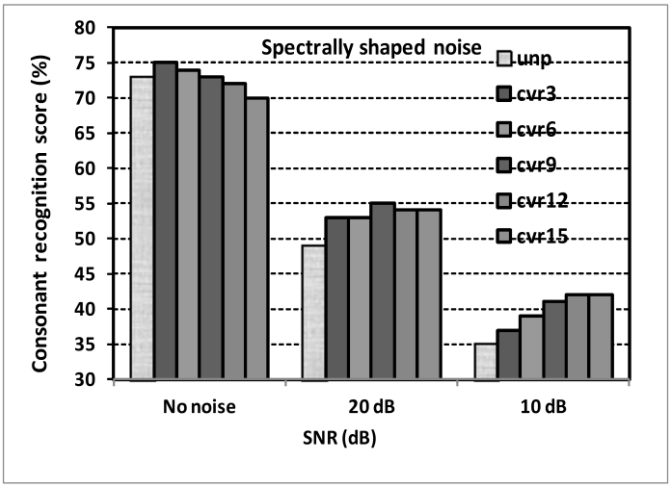

Fig. 5. Averaged recognition scores for different CVR modification factors for spectrally shaped noise.

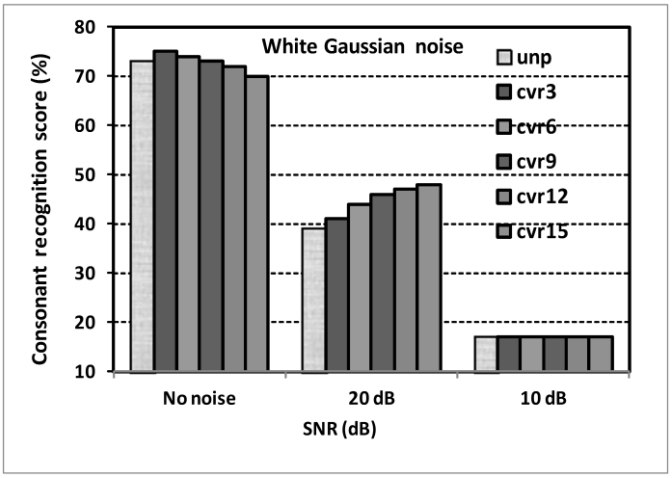

Fig. 6. Averaged recognition scores for different CVR modification factors for white Gaussian noise.

effective in increasing consonant recognition scores, for both spectrally shaped noise and white noise maskers.

VI. CONCLUSION

CVR modification may be used as an effective preprocessing technique for improving performance of ASR systems for stop consonant recognition. Nearly 6% improvement in stop consonant recognition score is obtained by a CVR modification factor of 9 dB at an SNR of 20 dB in presence of spectrally shaped noise (49 % to 55%). The improvement in consonant recognition score is nearly 6% (39 % to 45%) for white Gaussian noise under similar experimental conditions.

Both these results support the use of automatic CVR modification as an effective pre-processing step for improving performance of ASR systems for stop consonant recognition. Similar results are expected for fricative consonants and affricates. The possibility of extending the method for these

TABLE VIII. REC. SCORES FOR INDIVIDUAL SPEAKERS IN SPECTRALLY SHAPED NOISE AND IN (WHITE GAUSSIAN NOISE).

Speaker. (tokens)	SNR (dB)					
	No noise.		20		10	
	unp.	cvr9	unp.	cvr9	unp.	cvr9
F18(103)	75(75)	70(70)	54(46)	59(54)	40(17)	46(18)
F20(108)	77(77)	78(78)	52(37)	58(44)	37(17)	44(17)
F22(106)	68(68)	64(64)	40(30)	43(41)	35(17)	36(17)
M2(107)	76(76)	80(80)	50(31)	57(39)	33(17)	43(17)
M4(107)	70(70)	72(72)	50(50)	54(50)	31(17)	38(17)
Avg.	73(73)	73(73)	49(39)	55(46)	35(17)	41(17)

TABLE IX. REC. SCORES FOR STOP CONSONANTS IN SPECTRALLY SHAPED NOISE AND IN (WHITE GAUSSIAN NOISE) FOR CVR 9.

Speaker. (tokens)	SNR (dB)					
	No noise.		20		10	
	unp.	cvr9	unp.	cvr9	unp.	cvr9
b(90)	52(52)	48(48)	43(17)	44(17)	33(0)	32(0)
d(89)	84(84)	82(82)	75(58)	82(76)	54(0)	72(0)
g(86)	78(78)	77(77)	12(28)	17(30)	1(0)	1(0)
p(88)	80(80)	77(77)	77(32)	83(53)	63(0)	72(0)
t(88)	59(59)	66(66)	9(5)	16(3)	0(0)	2(0)
k(90)	88(88)	88(88)	79(93)	82(93)	60(100)	67(100)
Avg.	73(73)	73(73)	49(39)	55(46)	35(17)	41(17)

classes of consonants is to be investigated. The method also needs to be extended for improving consonant recognition in continuous speech material for making its effective use in practical applications.

REFERENCES

[1] Liu, S.A. (1996). Landmark detection for distinctive feature based speech recognition, J. Acoust. Soc. Am., 100(5), 3417-3430.

[2] Picheny, M.A., Durlach, N.I., & Braida, L.D. (1985). Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech, J. Speech Hear. Res. 28, 96-103.

[3] Payton, K. L., Uchanski, R. M., & Braida, L. D. (1994). Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing. Journal of Acoustical Society of America, 95, 1581–1592.

[4] Bradlow, A. R.,&Bent, T. (2002). The clear speech effect for non-native listeners. Journal of Acoustical Society of America, 112, 272–284.

[5] Bradlow, A. R., Kraus, N., & Hayes, E. (2003). Speaking clearly for children with learning disabilities. Journal of Speech, Language, and Hearing Research, 46, 80–97.

[6] Picheny, M. A., Durlach, N. I., &Braida, L. D. (1989). Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech. Journal of Speech and Hearing Research, 29, 434–446

[7] Picheny, M. A., Durlach, N. I., &Braida, L. D. (1989). Speaking clearly for the hard of hearing III: An attempt to determine the contribution of speaking rate to differences in intelligibility between clear and conversational speech. Journal of Speech and Hearing Research, 32, 600–603

[8] Vaughan, N. E., Furukawa, I., Balasingam, N., Mortz, M., & Fausti, S. A. (2002). Time expanded speech and speech recognition in older adults. Journal of Rehabilitation Research and Development, 39, 559–566

[9] Liu, S., & Zeng, F. G. (2006). Temporal properties in clear speech perception. Journal of Acoustical Society of America., 120, 424–432.

[10] Guelke, R.W. (1987). Consonant burst enhancement: A possible means to improve intelligibility for the hard of hearing, J. Rehab. Res. Develop. 24, 217-22.

[11] Hazan, V., and Simpson, A. (1998). The effect of cue-enhancement on the intelligibility of nonsense word and sentence materials presented in noise., Speech Commun. 24, 211-226.

[12] Skowronski, M.D., and Harris, J.G. (2006). Applied principles of clear and lombard speech for automated intelligibility enhancement in noisy environments, Speech Commun. 48, 549-558.

[13] Jayan, A.R., and Pandey, P.C. (2012). Automated CVR modification for improving perception of stop consonants, Proc. 18th Nat. Conf. Communications (NCC 2012), Kharagpur, India, 698-702.

[14] Jayan, A.R., Pandey, P.C. (2015), Automated modification of consonantvowel ratio of stops for improving speech intelligibility. Int. J. Speech Technol.18, 113-130.

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

[16] Kapoor, A., and Allen, J.B. (2012). Perceptual effects of plosive feature modification", J. Acoust. Soc. Am. 131, 478-491.

[17] Young, S. J., Evermann, G., Gales, M. J. F., et al. (2006) The HTK Book, version 3.4.

[18] Cambridge University Engineering Department (CUED), The Hidden Markov Model Toolkit (HTK)", [online] Available: http://htk.eng.cam.ac.uk/.

.